Systems biology

Statistical methods to infer cooperative binding among transcription factors in *Saccharomyces cerevisiae*

Debayan Datta¹ and Hongyu Zhao²,³,*

¹Department of Biomedical Engineering, ²Department of Epidemiology and Public Health and ³Department of Genetics, Yale University, New Haven, CT 06520, USA

Received on August 5, 2007; revised on August 17, 2007; accepted on October 12, 2007

Advance Access publication November 7, 2007

ABSTRACT

Motivation: Transcription factors regulate transcription in prokarrovotes and eukaryotes by binding to specific DNA sequences in the regulatory regions of the genes. This regulation usually occurs in a coordinated manner involving multiple transcription factors. Genome-wide location data, also called ChIP-chip data, have enabled researchers to infer the binding sites for individual regulatory proteins. However, current methods to infer binding sites, such as simple thresholding based on p-values, are not optimal for a number of study objectives like combinatorial regulation, leading to potential loss of information. Hence, there is a need to develop more efficient statistical methods for analyzing such data.

Results: We propose to use log-linear models to study cooperative binding among transcription factors and have developed an Expectation-Maximization algorithm for statistical inferences. Our method is advantageous over simple thresholding methods both based on simulation and real data studies. We apply our method to infer the cooperative network of 204 regulators in Rich Medium and a subset of them in four different environmental conditions. Our results indicate that the cooperative network is condition specific; for a set of regulators, the network structure changes under different environmental conditions.

Availability: Our program is available at http://bioinformatics.med.yale.edu/TFCoooperativity

Contact: hongyu.zhao@yale.edu

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 INTRODUCTION

Regulation of gene expression is fundamental to all biological systems. With the advent of new experimental approaches and diverse data sources, it is now possible to study the mechanisms of regulation on a genomic scale. Transcription factors play a key role in regulating gene expression. Transcriptional regulation typically takes place by the binding of transcription factors to specific promoteor sequences of the gene. ChIP-chip experiments provide us with information about the binding targets of a particular protein (Buck *et al.*, 2004; Iyer *et al.*, 2001; Lieb *et al.*, 2001; Ren *et al.*, 2000).

Transcriptional regulation is combinatorial in nature, multiple factors work together to regulate the expression of a single gene or a set of genes. Therefore, it is crucial to study combinatorial regulation. Despite its importance, only a few research groups have developed algorithms to study transcription factor cooperativity. Pilpel *et al.* (2001) used the correlation between the co-occurrence of computationally derived motifs and expression data to screen for cooperatively binding transcription factors. Banerjee *et al.* (2003) integrated ChIP-chip and gene expression data, using the expression correlation score as a measure of cooperativity. Chang *et al.* (2006) used a stochastic system model to infer transcription factor cooperativity, while Balaji *et al.* (2006) used a network transformation procedure to determine the pairwise co-regulatory network from the transcriptional regulatory network.

One major limitation of these methods is that the binding targets of a transcription factor are determined based on an arbitrary p-value cutoff applied to the ChIP-chip data. For example, if the p-value denoting the significance of binding for a transcription factor to the probes associated with a gene is less than 0.001, the gene is inferred to be a binding target, and all subsequent analyses are based on such dichotomized and error prone data.

In this article, we propose to use the log-linear model to study ChIP-chip data. The statistical evidence for physical binding is explicitly modeled and the cooperativity among transcription factors is reflected as an interaction term in the model. Thus, our approach can more effectively use information in the data, leading to more accurate inference.

The rest of the article is organized as follows. In the Methods section, we describe the log-linear model and the Expectation-Maximization (EM) algorithm for statistical inference. The applications of our method and comparisons with simple thresholding are shown in the Results section. We construct the pairwise co-regulatory network for the transcription factors in *Saccharomyces cerevisiae*. Our results are consistent with many previously characterized interacting transcription factor pairs. We also compare the cooperative network structure under different environmental conditions. Our results show that the cooperative network structure changes significantly under different conditions. We conclude this article with the Discussions section.

*To whom correspondence should be addressed.
2 METHODS

2.1 Log-linear model

The log-linear model is a special case of generalized linear models (McCullagh et al., 1989). It is commonly used to analyze the association between two or more categorical variables, therefore, it can be thought as a statistical model to characterize multi-way contingency tables. There is no distinction between the dependent and independent variables and all the variables are treated as response variables whose mutual associations are explored.

The basic strategy in the log-linear model involves fitting models to the observed frequencies in the cross-tabulation of categoric variables. The models can then be represented by a set of expected frequencies that may or may not resemble the observed frequencies. In assessing how well the model fits the data, it depends on how well the frequencies expected under the model approximate the frequencies actually observed. The choice of an appropriate model is ultimately based on a formal comparison of goodness-of-fit statistics associated with hierarchically related models.

We describe the log-linear model for a 2 × 2 cross tabulation in the following as this is the focus of our analysis. Our goal is to ascertain whether a pair of transcription factors exhibit cooperative regulation, i.e., whether there is an association between their target genes. For a pair of transcription factors f1 and f2, their joint regulation pattern for a gene can be assigned to a cell in a 2 × 2 table, corresponding to whether this gene is regulated by neither of the factors (0,0), f1 only (1,0), f2 only (0,1) or both factors (1,1). In this case, summarizing over all the genes, factors.

For each transcription factor studied, the observed data consists of \( p \)-values denoting the statistical significance for binding of the factor to the probes associated with the genes. Thus, if there are \( M \) factors studied, and there are \( N \) genes, the observed data is an \( M \) by \( N \) matrix of \( p \)-values. The true binding matrix would be the \( M \) by \( N \) binary matrix of 0 and 1 with a value of 1 indicating binding and a value of 0 indicating no binding. The common procedure to determine whether a particular factor binds to a target gene is to apply a threshold to the associated binding \( p \)-value; genes with \( p \)-values less than the threshold are assumed to be the binding targets. Instead of simply applying a threshold to the ChIP-chip \( p \)-values to determine transcription factor binding targets, we use an EM-based approach for model inference.

We denote the observed binding pattern for a particular gene as \( b = (p_1, p_2) \), where each component is the \( p \)-value denoting the significance for the gene to be a binding target of the particular transcription factor. We define \( b \) to be the estimate of the true binding pattern for gene \( g \). Thus, for each gene \( g \), \( b \) can take on one of the four possible values \((0,0),(0,1),(1,0),(1,1)\). Under a specific log-linear model, with parameters \( \theta = (\mu, \lambda^1, \lambda^2, \lambda^1_{ij}) \), we can derive the probability of observing for observing each pattern \( P(b_{ij}) \), where \( b_{ij} \) denotes one of the four possible binding patterns for gene \( g \).

The probability of the observed binding data is then given by

\[
P(b) = \prod_{i=1}^{N} P(b_{ij})P(p_i|b_{ij}).
\]

Thus, for \( N \) genes, the probability of the observed data is

\[
P(p_1, p_2, \ldots, p_N) = \prod_{i=1}^{N} P(b_{ij})P(p_i|b_{ij}).
\]

We aim to estimate the parameter values that would maximize this probability. They are estimated using the EM algorithm, by treating the \( b_{ij} \) as the missing data.

E-Step: In the Expectation step, for every gene \( g \) with observed binding pattern \( p \), we estimate:

\[
P(b^{(m)}|p) = \frac{P(p|b^{(m)}) \times P(b^{(m)})}{\sum_{b} P(p|b^{(m)}) \times P(b^{(m)})}.
\]

where \( b^{(m)} \) is the estimate of the true binding state \( b \) at the \( m \)th step. Since \( b^{(m)} \) can have four possible values, at each step, we estimate four probabilities. The probability \( P(p|b^{(m)}) \) can be expanded as:

\[
P(p|b^{(m)}) = P(p_i|b_1^{(m)})P(b_2^{(m)}),
\]

\[
P(p_1, p_2|b^{(m)}) = P(p_i|b_1^{(m)})P(b_2^{(m)})P(p_i|b_2^{(m)})P(b_1^{(m)}),
\]

where \( b_1^{(m)} \) and \( b_2^{(m)} \) are the first and second components of the estimate \( b^{(m)} \) and take only binary values 0 or 1. The second equation in (4) results from the independence assumption for the data from two separate ChIP-chip experiments. We use local false discovery rates (Efron, 2004) to evaluate the two terms in the product in Equation (4) as described in Subsection 2.3.

M-Step: Once we have obtained \( P(b^{(m)}) \) for each gene, we cross-tabulate and form a two-way contingency table, with the 'count' for each of the four values \((0,0),(0,1),(1,0),(1,1)\) being the sum of the probabilities for that particular value, obtained from \( P(b^{(m)}) \) of each gene. We then use the log-linear model to ascertain whether there is any interaction between the two transcription factors. We define the interaction weight to be the negative logarithm of the \( p \)-value associated with the interaction term in the log-linear model. This analysis is repeated for every regulator pair to obtain the co-binding network.

\[
E_{ij} = \frac{\ln F_{ij}}{C_{21}},
\]

where \( i, j = 0,1 \) or both factors (1,1). In this case, summarizing over all the genes, factors.

### References


2.3 Local false discovery rate

In this subsection, we describe how we use local false discovery rates to estimate the terms in Equation (4). Local false discovery rate, introduced by Efron, 2004 in multiple hypothesis testing, can be used to quantify the plausibility of a particular hypothesis being true, given its specific test statistic. In our setting, for a particular regulator, the N binding p-values denoting the significance of binding to the N genes are first converted into the corresponding z-values using the transformation

\[ z_i = \Phi^{-1}(p_i), \quad i = 1, 2, \ldots, N \]

where \( \Phi \) is the standard normal cumulative distribution function and \( p_i \) is the binding p-value for the \( i \)th gene.

The \( N \) z-values are assumed to fall into two classes, ‘null’ or ‘non-null’ with prior probabilities \( \pi_0 \) and \( \pi_1 \), respectively. \( f_0(z) \) and \( f_1(z) \) are the prior densities for each class, \( z_i \) has density either \( f_0(z) \) or \( f_1(z) \) depending on its class. The local false discovery rate is defined as the a posteriori probability that a gene is not a binding target, i.e. it is a ‘null’ gene. To illustrate our calculation from Equation (4), let us for the sake of simplicity assume that a particular instance of \( b_{m,i} \) is \( b \), while the p-value denoting the significance for a gene \( g \) to be a binding target for a transcription factor to be \( p \) with corresponding z-value \( z \). For the case when \( b = 0 \), Equation (4) can be simplified by Bayes theorem:

\[
P(z|b=0) = \frac{P(z, b=0)}{P(b=0)} = \frac{P(b=0)\phi(z)}{P(b=0)}
\]

(5)

\( P(b=0|z) \) is then by definition the local false discovery rate for the gene \( g \), while \( P(b=0) \) is the ‘null’ prior probability. Equation (5) then reduces to:

\[
P(z|b=0) = \frac{P(b=0)\phi(z)}{P(b=0)} = \frac{\Phi(d(z))P(z)}{\pi_0} = \frac{\Phi(d(z))}{\pi_0}π_0 + \Phi(d(z))π_1 f_1(z)
\]

(6)

where \( \Phi(d(z)) \) is the local false discovery rate for the gene \( g \). Similarly, when \( b = 1 \),

\[
P(z|b=1) = \frac{P(b=1|z)\phi(z)}{P(b=1)} = \frac{(1-\Phi(d(z)))P(z)}{\pi_1} = \frac{(1-\Phi(d(z)))(π_0 + \Phi(d(z))π_1)}{\pi_1}
\]

(7)

The local false discovery rate, prior probabilities and the prior densities can be calculated by using the R package locfdr. We discuss this in Section 1 of Supplementary Material.

3 RESULTS

3.1 Comparison with simple thresholding

We compare the performance of our EM-based approach with simple thresholding in inferring cooperativity among transcription factors. In the simple thresholding approach, we apply a p-value cutoff to the ChIP-chip data, with p-values less than the predefined threshold signifying that a gene is a binding target for a transcription factor. For a pair of regulators, we cross tabulate the results in the form of a two-way contingency table and use the log-linear model to infer if there is any association between the regulator pairs. We compare the performance of these two methods for different p-value cutoffs on both simulated and real data.

3.2 Simulation

We consider an experiment involving two factors and 1000 genes. For each factor, we assume 200 genes are its targets. The test statistic z-values follow \( z_i \sim N(0,1) \) for the 800 non-target genes and \( z_i \sim N(2,1) \) for the 200 target genes. The p-values are obtained from the z-values using the transformation

\[ p_i = 1 - \Phi(z_i), \quad i = 1, 2, \ldots, 1000 \]

where \( \Phi \) is the standard normal cumulative distribution function. For each simulation, we create a two-way contingency table for the data. The degree of cooperativity is captured by the odds ratio of the target genes of the two factors. For example, when the odds ratio is 1, there is no cooperativity between the factors. With increasing odds ratio, there is progressively higher degree of cooperativity among the factors. We consider four different odds ratios (odds ratio = 1, 2, 4, 6). For each odds ratio, we performed 1000 simulations. For each simulated dataset, we obtained the statistical significance for the interaction term using the log-linear model for both methods. For the simple thresholding approach, we performed the log-linear analysis for three different cutoffs, \( p\text{-value} = 0.001, 0.005, 0.01 \). We selected different cutoffs for significance and plot the proportion of times the interaction p-values were less than the significance cutoffs versus the significance cutoffs (Fig. 1). Our plots indicate that the EM-based method consistently generates more significant interaction p-values for different odds ratios in the presence of cooperativity. Thus, our method is more powerful than the simple thresholding method. When the odds ratio is equal to 1, i.e. the null is true, we observe that our method does not lead to inflated positive results.

3.3 Datasets

We consider ChIP-chip data from Harbison et al. (2004). The binding profiles of 204 transcription factors for S. Cerevisiae were collected in Rich Medium. Eighty-four of these transcription factors were also profiled in at least one other experimental condition. Transcription factors were selected for profiling in a particular environment if they were essential for growth in that environment, or if there was other evidence suggesting their role in gene regulation in that environment. We inferred the co-binding networks of profiled regulators in five different conditions: Rich Medium, Amino Acid Starved, Moderately Hypoxic, Nutrient Deprived and Elevated Temperature.

3.4 Real networks

We compared the co-binding networks obtained by our EM-based method with the networks obtained by using the simple thresholding approach. For the simple thresholding approach, we used three different cutoffs p-value = 0.001, 0.005, 0.01 to determine the binding targets. In both the simple thresholding method and our EM-based approach, we considered a regulator pair to interact if the p-value for the interaction term in the log-linear model was less than 0.0001. For all the five environmental conditions studied, the EM-based method detected more interactions. For the simple thresholding method, as the cutoff threshold became larger, more and more
common interactions were detected by the two methods. The results are summarized in Supplementary Tables 1–4.

We varied the P-value threshold of the interaction term in the log-linear model and plotted the number of detected interactions versus the interaction weight cutoff (Supplementary Fig. 8). We note that for all the environmental conditions, the number of interacting regulator pairs starts falling as the interaction weight cutoff passes in between 3 and 4.

We also determined the cooperative binding network for 31 transcription factors in Rich Medium, which are known to be cell cycle regulators. We used gene expression data published by Spellman et al. (1998) to compare the correlation coefficients of the target genes obtained by our EM-based approach and simple thresholding. This dataset describes the mRNA levels of the genes in S. cerevisiae over 18 time points over two cell cycle periods. For each of the top 25 regulator pairs ranked based on their interaction weights, we ranked the genes based on their posterior probabilities to the binding targets of these two factors and selected the top 10 genes. We then calculated the average pairwise correlation coefficient for these genes. In the simple thresholding approach, we took the average of two binding ChIP-chip P-values for each gene, and selected the 10 genes with the lowest average for each of the 25 regulator pairs. We summarize the results in Supplementary Table 5. We observe that for most of the regulator pairs, the average pairwise correlation coefficient for the target genes determined by our EM-based approach is higher than those obtained from simple thresholding.

We further ascertain whether the cooperative pairs detected by our EM algorithm are false positives. For the cell cycle regulator pairs detected by our method, we compared them with regulator pairs with known literature evidence (Section 2, Supplementary Material). We observe that the majority of the regulator pairs inferred by the EM algorithm have prior literature evidence. We also looked at Gene Ontology for the cooperative transcription factors. We randomly selected 250 cooperative regulator pairs determined by both the EM algorithm and simple thresholding, 250 cooperative regulator pairs determined by the EM algorithm and not simple thresholding, and 250 random regulator pairs. The number of regulator pairs sharing significant GO terms among them was 46, 48 and 22 for these three selected sets, respectively. Thus, the cooperative pairs detected by our EM algorithm appear to be real cooperative pairs.

### 3.5 Network properties

For each of the five environmental conditions, using the proposed EM algorithm, we obtained the pair-wise cooperative binding network for the transcription factors profiled in each condition. Figure 2 shows the network structure in Rich Medium. Supplementary Figures 1–3 show the network structure in other environmental conditions. Some simple network properties are summarized in Table 1. Apart from the Elevated Temperature condition where only six regulators where profiled, we observe a high percentage of association among transcription factor pairs in all the environmental conditions compared to Rich Medium. In particular, in the Nutrient Deprived condition and Amino Acid Starved condition, we observe a dense network structure with 81.32% and 71.12% of the total possible co-regulatory associations among transcription factor pairs, respectively. This is expected because...
Fig. 3. Histogram of the interaction weights among transcription factor pairs in four different Environmental conditions. We notice that for all the environmental conditions, while a small number of regulator pairs have large interaction weights, most regulator pairs have small interaction weights.

the regulators profiled in a particular condition were known to participate in gene regulation in that condition. The histogram of the interaction weights (Fig. 3) for each network reveals that while a small number of regulator pairs have large interaction weights, numerous transcription factor pairs have relatively small interaction weights.

For the transcription factor pairs having large interaction weights, they were often found to have physical interactions, or the transcription factors are closely related paralogous factors. For example, Swi4 and Swi6 are part of the SBF complex (Andrews et al., 1992, Nasmyth et al., 2004), and they have a large interaction weight between them (weight = 50.82).

Similarly, Ino2 and Ino4 form a heterodimeric complex (Ambroziak et al., 1994), and they also have a high degree of association (weight = 71.03), while Rtg1 and Rtg3 also form a complex (Rothermel et al., 1997) and have interaction weight = 49.08.

3.6 Network modules
To study the topology in the cooperative network structure, we consider network modules as a $k$-clique, defined as the largest complete subgraph of size $k$. All transcription factors comprising a $k$-clique have pair-wise co-regulatory association with each other. The possible biological interpretation for all transcription factors present in a module is that they participate in a common regulatory program.

It is interesting to note that not all transcription factor pairs present in a module have large interaction weights. Often, one or a few have a large interaction weight, while the rest have relatively low interaction weights. For example, in Rich Medium, we discovered a module consisting of the regulators Ash1, Cbf1, Ixr1 and Pdc2. The interaction weight between Ash1 and Cbf1 is 226.82, while the other 5 interaction weights are less than 55. However, this phenomenon is not that pronounced in environmental conditions other than Rich Medium. For example, in Nutrient Deprived condition, we discovered a module of size 9 consisting of Dal81, Dal82, Gat1, Hap2, Msn2, Msn4, Rtg3, Uga3 and Gln3. Out of the 36 pairwise interactions, 8 have interaction weights less than 25, while 9 have weights greater than 100.

3.7 Condition dependent binding
It has been previously reported (Harbison et al., 2004; Luscombe et al., 2004) that the binding patterns of transcription factors change with environmental conditions. Since all the regulators were profiled in Rich Medium, and a subset of them in different conditions, we compared the cooperativity network for the same set of regulators in two different environmental conditions, with one of the conditions being Rich Medium. Table 2 summarizes the results. We notice that the network structure is quite different under two different conditions. Compared to Rich Medium, we see a much larger connectivity in the network structure in other conditions.

Several transcription factor pairs exhibit a high level of interaction in one environmental condition, while no or negligible interaction in the other condition. For example, in the Nutrient Deprived condition, the transcription factors Rtg3 and Msn2 show a large degree of association (weight = 192.52), while in Rich Medium they do not show any association. Similarly, the regulators Gat1 and Gln3 show cooperativity (weight = 162.91), while in Rich Medium they exhibit a small amount of cooperativity (weight = 7.4). On the other hand, several regulators exhibit condition-independent cooperativity. For example, in Amino Acid Starved condition, Fhl1 and Rap1 have high association (weight = 126.52), while in Rich Medium their interaction weight equals 118.16.

For the set of transcription factors common to a particular environmental condition and Rich Medium, we consider the network modules in the two conditions. We note that the number of modules, as well as the size of the largest module is smaller for the regulators in Rich Medium (Table 2). Thus, compared to Rich Medium, the transcription factors profiled in the particular condition are more likely to take part in a combinatorial regulatory program.

<table>
<thead>
<tr>
<th>Table 1. Properties of co-binding networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental condition</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Elevated Temperature</td>
</tr>
<tr>
<td>Nutrient Deprived</td>
</tr>
<tr>
<td>Moderately Hyperoxic</td>
</tr>
<tr>
<td>Amino Acid Starved</td>
</tr>
<tr>
<td>Rich Medium</td>
</tr>
</tbody>
</table>

For each environmental condition, some simple properties of the connectivity structure of the co-binding network are tabulated.
Table 2. Comparison of co-binding networks

<table>
<thead>
<tr>
<th>Experimental condition</th>
<th>Number of transcription factors</th>
<th>Percent of total possible interactions</th>
<th>Maximum module size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevated Temperature</td>
<td>6</td>
<td>53.33</td>
<td>4</td>
</tr>
<tr>
<td>Rich Medium</td>
<td>6</td>
<td>66.67</td>
<td>3</td>
</tr>
<tr>
<td>Nutrient Deprived</td>
<td>14</td>
<td>81.32</td>
<td>9</td>
</tr>
<tr>
<td>Rich Medium</td>
<td>14</td>
<td>50.65</td>
<td>6</td>
</tr>
<tr>
<td>Moderately Hyperoxic</td>
<td>27</td>
<td>55.84</td>
<td>11</td>
</tr>
<tr>
<td>Rich Medium</td>
<td>27</td>
<td>39.89</td>
<td>9</td>
</tr>
<tr>
<td>Amino Acid Starved</td>
<td>34</td>
<td>71.12</td>
<td>17</td>
</tr>
<tr>
<td>Rich Medium</td>
<td>34</td>
<td>43.85</td>
<td>10</td>
</tr>
</tbody>
</table>

For the factors profiled in a particular condition, we compare their co-binding network properties in that particular condition and in Rich Medium.

3.8 Cell cycle transcription factors

Cell cycle is the process by which the cell reproduces by duplicating its contents and then dividing in two. It consists of four different phases—G1, G2, S and M phases. We obtained the co-binding network for 31 regulators in Rich Medium, which are known to regulate the cell cycle. We also compared this network with that obtained by simple thresholding.

The cooperative binding network for the cell cycle regulators (Supplementary Fig. 4) reveals a fair amount of association among the transcription factor pairs. Table 3 lists the top 25 regulator pairs with the highest interaction weights. For the 31 regulators studied, there are a total of 234 interactions. This equals to ~50.32% of the possible pairwise interactions. Our results also captured a large number of key interactions, which are known to occur during the different phases of the cell cycle.

The G1–S transition of the eukaryotic cell cycle is known to be mediated by protein complexes MBF and SBF. MBF is composed of Mbp1 and Swi6, while SBF is composed of Swi4 and Swi6 (Koch et al., 1993). Our analysis captured both these interactions with fairly large interaction weights. In addition, we found Mbp1 to be cooperative with Swi4.

An important regulator in the G2–M phase is the SFF factor, which is composed of Ndd1, Fkh1 and Fkh2 (Koranda et al., 2000; Kumar et al., 2000; Pic et al., 2000). We found strong pairwise association among these three factors. Fkh2 also forms a ternary complex in the presence of Mcm1 (Koranda et al., 2000; Kumar et al., 2000), while it has been experimentally found that Mcm1, Fkh2, Ndd1 form a complex to regulate the CLB2 gene and other genes (Kumar et al., 2000). Our results also show pairwise association between these factors.

In the M–G1 phase, consistent with previously published results, we found strong association between transcription factors Ace2 and Swi5.

In addition, we found several novel interactions that have not been previously reported. For example, significant interactions between Ndd1 and Rme1, Rme1 and Swi6, Hir3 and Ume1 were found using our method.

The co-binding network for the cell cycle regulators obtained by simple thresholding with a p-value cutoff of 0.001 contains 120 co-regulatory interactions, ~25.81% of the possible pairwise interactions. While simple thresholding captured most of the previously published interactions listed in Table 3, the interaction weights for some of these interactions are quite low. For example, the interaction strength between Fkh1 and Ndd1 is 17.3, between Fkh1 and Swi6 is 15.2, while that between Ace2 and Swi6 is 19.4. Some of the significant novel interactions detected by our method, like Ndd1 and Rme1, Hir3 and Ume1, also have low interaction strengths in the network obtained by simple thresholding (Interaction weights 24.7 and 20, respectively).

We also compared the performance of our results with those of Banerjee et al. (2003), Chang et al. (2006) and Tsai et al. (2005). The details are discussed in Section 2 of Supplementary Material. We observe that most of the literature confirmed interacting transcription factor pairs detected by Banerjee et al. (2003), Chang et al. (2006) and Tsai et al. (2005) were also detected by our EM-based method. Further, our method detected a larger number of interacting regulator pairs with literature evidence compared to their results.

4 DISCUSSION

We have developed an EM algorithm to use log-linear models to infer whether two transcription factors participate in co-operative gene regulation. Simulation studies show that our proposed approach outperforms the commonly used simple thresholding approach. By applying this method to all possible pairs of transcription factors in yeast, we have built the co-binding network under different environmental conditions. We notice that a few transcription factor pairs have large interaction weights, while the majority have relatively low weights. This perhaps reveals the centrality of these regulator pairs in the co-binding network. We also observe that regulator pairs that physically interact tend to have high interaction values.

Further, we have compared the structure and topology of the networks under different environmental conditions. We observe that the co-regulatory network changes under different conditions indicating differing transcriptional activity under different conditions. Some transcription factor pairs exhibit a large degree of interaction in one condition while showing negligible interaction in another condition. On the other hand, some regulator pairs show similar association in two different conditions. The reason for such a condition-dependent network structure is not easy to explain. It has been hypothesized (Balazsi et al., 2005) that transcription factors essentially act as sensors, adjusting their activities according to the way they perceive changing environmental conditions.

Our method readily identified previously confirmed cooperative transcription factors in the cell cycle. In addition, a few novel interactions are also detected. As ChIP-chip data tends to be noisy, integration of gene expression data to our analysis would be more comprehensive (Sun et al., 2006; Zhong et al., 2005).

Our method can be extended to determine higher order interactions among three or more transcription factors. However, the biological interpretation of such interactions is not trivial. For example, for three transcription factors, we might get significant pair-wise interactions but no
Table 3. Top 25 cooperative transcription factor pairs regulating cell cycle

<table>
<thead>
<tr>
<th>TF1</th>
<th>TF2</th>
<th>Interaction weight</th>
<th>Correlation Coefficient</th>
<th>Literature Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hir3</td>
<td>Yox1</td>
<td>142.51</td>
<td>0.5831</td>
<td>Tsai et al. (2005)</td>
</tr>
<tr>
<td>Ndd1</td>
<td>Swi6</td>
<td>131.63</td>
<td>0.6544</td>
<td>Chang et al. (2000)</td>
</tr>
<tr>
<td>Mbp1</td>
<td>Swi6</td>
<td>126.31</td>
<td>0.3133</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Koch et al. (1993); Manke et al. (2003)</td>
</tr>
<tr>
<td>Ndd1</td>
<td>Rme1</td>
<td>111.21</td>
<td>0.3445</td>
<td></td>
</tr>
<tr>
<td>Fkh1</td>
<td>Fkh2</td>
<td>88.27</td>
<td>0.0541</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Kumar et al. (2000); Manke et al. (2003)</td>
</tr>
<tr>
<td>Skn7</td>
<td>Swi6</td>
<td>85.61</td>
<td>0.0106</td>
<td>Chang et al. (2006)</td>
</tr>
<tr>
<td>Ndd1</td>
<td>Skn7</td>
<td>70.12</td>
<td>0.4016</td>
<td>Tsai et al. (2005)</td>
</tr>
<tr>
<td>Rme1</td>
<td>Swi6</td>
<td>65.52</td>
<td>0.4443</td>
<td></td>
</tr>
<tr>
<td>Fkh2</td>
<td>Ndd1</td>
<td>62.16</td>
<td>0.0257</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Manke et al. (2003); Tsai et al. (2005)</td>
</tr>
<tr>
<td>Hir3</td>
<td>Ume1</td>
<td>61.46</td>
<td>0.3318</td>
<td></td>
</tr>
<tr>
<td>Fkh1</td>
<td>Swi6</td>
<td>60.09</td>
<td>0.8733</td>
<td>Tsai et al. (2005)</td>
</tr>
<tr>
<td>Ace2</td>
<td>Mbp1</td>
<td>54.10</td>
<td>0.2925</td>
<td>Tsai et al. (2005)</td>
</tr>
<tr>
<td>Ace2</td>
<td>Swi6</td>
<td>53.70</td>
<td>0.0365</td>
<td>Tsai et al. (2005)</td>
</tr>
<tr>
<td>Ace2</td>
<td>Ndd1</td>
<td>52.78</td>
<td>0.0272</td>
<td></td>
</tr>
<tr>
<td>Mbp1</td>
<td>Swi4</td>
<td>51.51</td>
<td>0.7738</td>
<td>Chang et al. (2006); Koch et al. (1993); Manke et al. (2003)</td>
</tr>
<tr>
<td>Swi4</td>
<td>Swi6</td>
<td>50.82</td>
<td>0.0546</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Koch et al. (1993); Kumar et al. (2000)</td>
</tr>
<tr>
<td>Fkh2</td>
<td>Mbp1</td>
<td>49.87</td>
<td>0.7292</td>
<td>Chang et al. (2006); Manke et al. (2003); Tsai et al. (2005)</td>
</tr>
<tr>
<td>Ash1</td>
<td>Hir3</td>
<td>48.93</td>
<td>0.1653</td>
<td></td>
</tr>
<tr>
<td>Ash1</td>
<td>Yhp1</td>
<td>48.11</td>
<td>0.1148</td>
<td></td>
</tr>
<tr>
<td>Mbp1</td>
<td>Rme1</td>
<td>47.97</td>
<td>0.3893</td>
<td></td>
</tr>
<tr>
<td>Fkh1</td>
<td>Ndd1</td>
<td>47.53</td>
<td>0.7866</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Tsai et al. (2005)</td>
</tr>
<tr>
<td>Mbp1</td>
<td>Ndd1</td>
<td>46.69</td>
<td>0.2948</td>
<td>Chang et al. (2006); Manke et al. (2003)</td>
</tr>
<tr>
<td>Ume1</td>
<td>Yox1</td>
<td>45.03</td>
<td>0.5832</td>
<td></td>
</tr>
<tr>
<td>Met31</td>
<td>Rme1</td>
<td>43.33</td>
<td>0.0785</td>
<td></td>
</tr>
<tr>
<td>Ace2</td>
<td>Swi5</td>
<td>42.61</td>
<td>0.3313</td>
<td>Banerjee et al. (2003); Chang et al. (2006); Manke et al. (2003)</td>
</tr>
</tbody>
</table>

From the cooperative network obtained for the cell cycle regulators, the top 25 interacting pairs based on their interaction weight. The correlation coefficient of the expression profiles for the regulator pairs as well as literature evidence for the interactions is also tabulated.

three-way interactions. In such a case, it is challenging to interpret the difference between these two types of interactions.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive comments. This work was supported in part by NIH grant GM 59507 and NSF grant DMS 0714817.

Conflict of Interest: none declared.

REFERENCES


