BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 4033–4038
doi:10.1093/bioinformatics/bti656

Systems biology

Inference of transcriptional regulatory network by two-stage constrained space factor analysis

Tianwei Yu and Ker-Chau Li*

Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA

Received on July 25, 2005; revised on August 24, 2005; accepted on August 30, 2005
Advance Access publication September 6, 2005

ABSTRACT

Motivation: Microarray gene expression and cross-linking chromatin immunoprecipitation data contain voluminous information that can help the identification of transcriptional regulatory networks at the full genome scale. Such high-throughput data are noisy however. In contrast, from the biomedical literature, we can find many evidenced transcription factor (TF)–target gene binding relationships that have been elucidated at the molecular level. But such sporadically generated knowledge only offers glimpses on limited patches of the network. How to incorporate this valuable knowledge resource to build more reliable network models remains a question.

Results: We present a modified factor analysis approach. Our algorithm starts with the evidenced TF–gene linkages. It iterates between the network configuration estimation step and the connection strength estimation step, using the high-throughput data, till convergence. We report two comprehensive regulatory networks obtained for *Saccharomyces cerevisiae*, one under the normal growth condition and the other under the environmental stress condition.

Contact: kcli@stat.ucla.edu

Supplementary information: http://kiefer.stat.ucla.edu/lap2/download/bti656_supplement.pdf

INTRODUCTION

Transcription network modeling is a major step towards deciphering the cellular regulation system. It involves two major tasks: (1) finding the target genes for each transcription factor (TF), and (2) correlating each TF's activity to its target transcripts as the condition varies. The first task specifies the network configuration. Several methods are available. The computational approach includes the inference of TF binding targets by drawing information from TF binding motifs (Qiu, 2003; Pritsker *et al*., 2004), and from gene-expression dynamics (Pournara and Wernisch, 2004; Qian *et al*., 2003; Rung *et al*., 2002; Segal *et al*., 2003; Zhu *et al*., 2002). A more direct approach is the genome-wide location analysis, or cross-linking chromatin immunoprecipitation (ChIP), which profiles each TF for its binding sites over the entire genome (Harbison *et al*., 2004; Lee *et al*., 2002). Combining ChIP data with microarray gene-expression data can give more interpretable network connectivity estimates (Bar-Joseph *et al*., 2003; Xu *et al*., 2004; Zhou *et al*., 2005). It also serves the purpose of elucidating the relationship between a TF's activity and the abundance of its target transcripts. Among many related works, of our special interest is the Network Component Analysis (NCA) model by Liao *et al*. (2003), which treats TF activities as latent variables. We shall incorporate this idea in developing our method.

Although ChIP and gene-expression data are invaluable for building the transcription network at the genome scale, they are both subject to high level noises. To minimize the noise interference in network construction, instead of taking a *de novo* approach which would require the simultaneous estimation of a tall magnitude of parameters, our idea is to use a set of highly reliable connections as the skeleton for network building. For yeast, more than 1000 evidenced TF–gene relationships exist in the literature and they have been organized into knowledgebase available from the internet (Lee *et al*., 2002; Wingender *et al*., 2001). This source of information provides an excellent starting point for network construction.

We present an algorithm that integrates ChIP data, microarray data and prior biological knowledge to obtain the transcription network. Our approach has several features. First, it utilizes the known TF–gene relationships. Second, it takes into account the combinatorial nature of transcription regulation. Third, it provides an estimate of TF activity, which can be used to further study the transcriptional regulation of the TFs themselves. Fourth, it takes into account the condition specificity in modeling the TF–target gene binding relationship.

METHODS

The two-stage constrained space factor analysis model

To relate TF–gene linkages with transcript abundance, we adopt the factor analysis model (Morrison, 1990), which takes the form of $\boldsymbol{X} = \boldsymbol{LY} + \boldsymbol{E}$. It is well-known that without any constraint on the loading matrix $\boldsymbol{L}$, the model is not identifiable. In practice, rotation on the loading matrix is taken to yield interpretable results. Quite often, this reduces the number of non-zero loadings (Morrison, 1990).

Suppose there are N genes, K TFs and J gene-expression conditions. We represent the configuration of a transcription regulatory network by a sparse connection matrix $\boldsymbol{C}_{N\times K} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K]$ between TFs and genes. Each column vector $\boldsymbol{c}_k$ is composed of 1 and 0s, indicating the binding (1) and non-binding (0) relationship of the k-th TF to each gene.

To apply factor analysis model, we take $\boldsymbol{X}$ to be the microarray gene-expression profile matrix $\boldsymbol{G}_{N\times J}$, $\boldsymbol{Y}$ to be the TF activity profile matrix $\boldsymbol{T}_{K\times J}$, $\boldsymbol{L}$ to be the regulation strength matrix $\boldsymbol{B}_{N\times K} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K]$. We rewrite the model as

$$G_{N\times J} = B_{N\times K} T_{K\times J} + E_{N\times J}. \quad (1)$$

With constraints that $\boldsymbol{b}_k$ lies in the subspace confined by the projection matrix $\mathrm{diag}(\boldsymbol{c}_k)$

$$\mathrm{diag}(c_k) b_k = b_k, \quad k = 1, 2, \ldots, K, \quad (2)$$

*To whom correspondence should be addressed.

which means that for any (i, k) combination, $b_{i,k}$ can be non-zero only when $c_{i,k} = 1$.

In our analysis, the expression profiles are already in log ratios. If the network configuration matrix $\boldsymbol{C}$ were pre-specified, then model (1) would be reduced to the NCA model proposed by Liao *et al*. (2003), wherein conditions can be found with respect to parameter identification. But the more challenging task for us is how to estimate $\boldsymbol{C}$.

To guide the estimation of $\boldsymbol{C}$, we use the condition that elements in the configuration matrix $\boldsymbol{C}$ be bounded by the corresponding elements from two matrices $\boldsymbol{C}_{\mathrm{MIN}}$ and $\boldsymbol{C}_{\mathrm{MAX}}$:

$$c_{\mathrm{MIN}\ i,k} \le c_{i,k} \le c_{\mathrm{MAX}\ i,k}, \quad \forall i,k. \tag{3}$$

Note that elements of the connection matrices can only take values 0 and 1, thus for each (i, k) combination, one of the inequalities must be equality. The lower-bound $\boldsymbol{C}_{\mathrm{MIN}}$ represents the network configuration matrix for the higher-confidence set of TF–gene relationships. The upper-bound configuration matrix $\boldsymbol{C}_{\mathrm{MAX}}$ is composed of linkages from both the higher- and lower-confidence sets (see Data source section).

We start with $\boldsymbol{C} = \boldsymbol{C}_{\mathrm{MIN}}$. After stabilizing the initial estimates of $\boldsymbol{B}$ and $\boldsymbol{T}$ (see next section), we update the configuration by adding a new linkage that best agrees with current $\boldsymbol{B}$ and $\boldsymbol{T}$ estimates. We then update $\boldsymbol{B}$ and $\boldsymbol{T}$. This procedure is repeated many times till convergence.

The algorithm

We normalize each gene-expression profile to bring the mean to zero and standard deviation to one. We also normalize estimated TF activity profile in each iteration of our algorithm.

Step 1. Initial estimation of TF activity profiles T from higher-confidence set. Set $\boldsymbol{C} = \boldsymbol{C}_{\mathrm{MIN}}$. The initial estimate of the activity profile for a TF is constructed by the consensus of the expression profiles for those genes targeted by this TF, using the leading component of a weighted PCA (Morrison, 1990).

Step 2. Estimation of B and T. After the initial estimate of $\boldsymbol{T}$ is obtained, an alternating least-square procedure (Gifi, 1990) is applied to minimize the sum of square error $\|\boldsymbol{G} - \boldsymbol{BT}\|^2$.

(1) Estimating $\boldsymbol{B}$. Fix the $\boldsymbol{T}$ matrix. For each row vector $\boldsymbol{g}_i$ in matrix $\boldsymbol{G}$, find all k^*s such that $c_{i,k^*} = 1$. Regress $\boldsymbol{g}_i$ against the corresponding $\boldsymbol{t}_{k^*}$s. Replace the b_{i,k^*}s with the regression coefficients. Here we use ridge regression to deal with the stability issue arising from the collinearity between the regressor variables (Faraway, 2004).

(2) Estimating $\boldsymbol{T}$. Fix the $\boldsymbol{B}$ matrix, regress each column of matrix $\boldsymbol{G}$, $\boldsymbol{g}_j$ against $\boldsymbol{B}$. Replace the corresponding column of matrix $\boldsymbol{T}$, $\boldsymbol{t}_j$ with the estimated coefficients. The two steps are iterated until the sum of squared change of $\boldsymbol{T}$ is smaller than a cutoff value.

Step 3. Adding new TF–gene relations. The algorithm searches through all TF–gene pairs allowed by $\boldsymbol{C}_{\mathrm{MAX}} - \boldsymbol{C}$ to find a pair that best agrees with the current $\boldsymbol{B}$ and $\boldsymbol{T}$ estimates. Because all gene-expression profiles and TF activity profiles are normalized, this is done efficiently by finding the highest absolute covariance between the residual (unexplained part) of an expression profile and a TF activity profile.

Define matrix $\mathbf{D} = \boldsymbol{C}_{\mathrm{MAX}} - \boldsymbol{C}$. First we find the row-wise covariance matrix $\boldsymbol{V}$ between the residual expression matrix $\boldsymbol{R} = \boldsymbol{G} - \boldsymbol{BT}$ and the TF activity matrix $\boldsymbol{T}$, by $v_{i,k} = \mathrm{cov}(\boldsymbol{r}_i, \boldsymbol{t}_k)$. We then find the pair $\{i^*, k^*\} = \arg\max_{i,k}(|v_{i,k}| \times d_{i,k})$. We assign $c_{i^*,k^*} = 1$ and $b_{i^*,k^*} = \mathrm{cov}(\boldsymbol{r}_{i^*}, \boldsymbol{t}_{k^*})$. Then the estimates of $\boldsymbol{B}$ and $\boldsymbol{T}$ are stabilized as described earlier.

We iterate between Steps 2 and 3. In each iteration, we record the total reduction of residual sum of squares (RSS) $\|\boldsymbol{G} - \boldsymbol{BT}\|^2$. When the average reduction in RSS in the last 10 iterations is less than one-fifth of that of the initial 10, we consider most of the signals in the lower-confidence set have been picked up, and terminate the iteration.

Step 4. Fine-tuning of TF-gene relations. Once the convergence is reached, we use $\boldsymbol{T}$ as the final estimate of TF activity profiles. Based on this estimate, we make an additional effort to fine-tune the network configuration matrix $\boldsymbol{C}$, using regression variable selection techniques.

For each gene i, to determine its regulator TFs, we find k^*s such that $c_{\mathrm{MAX}\ i,k^*} = 1$, and consider the multiple linear regression model

$$g_{i,j} = \sum_k b_{i,k^*} t_{k^*,j} + e_{i,j}, \tag{4}$$

in which the b_{i,k^*}s are the coefficients to be determined. The Bayes Information Criterion (Faraway, 2004) is applied to find the best subset of regressors. We regress $\boldsymbol{t}_i$ against the best subset, then select the regressors for which the P-value is $<10^{-4}$. Set the corresponding positions of $\boldsymbol{C}$ to 1.

The data source

The higher-confidence set consists of known gene–TF relationships in the biomedical literature [see TRANSFAC (Wingender *et al*., 2001) and the website of Young's group (Lee *et al*., 2002)]. There are a total of 1089 TF–gene relationships, from which we shall construct the matrix $\boldsymbol{C}_{\mathrm{MIN}}$ used in Equation (3).

The lower-confidence set is based on the ChIP dataset (Harbison *et al*., 2004). We use all TF–gene pairs that were reported to have P-values <0.05. The use of this loose cutoff point is to lower the false-negative rate. We shall combine the lower- and higher-confidence sets and use matrix $\boldsymbol{C}_{\mathrm{MAX}}$ to represent the information.

Two large-scale microarray datasets are used in this study. The cell-cycle dataset (Spellman *et al*., 1998) is used for normal growth network estimation. The stress–response dataset (Gasch *et al*., 2000) is used for stress-specific network estimation.

Time-shifted activity–expression correlation

For the cell-cycle data, we further investigate the time-shifting behavior between a TF's expression profile and its activity profile. Denote the expression profile by $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$, the activity profile by $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)$ and time points by $\boldsymbol{t} = (t_1, t_2, \ldots, t_M)$. Let Δt be the amount of time-shifting in minutes, which takes an integer value between 0 and 20. We first estimate the correlation between $\boldsymbol{x}(\boldsymbol{t})$ and $\boldsymbol{y}(\boldsymbol{t} + \Delta t)$. We then find the delayed time Δt that maximizes the correlation in absolute value. We estimate $\boldsymbol{y}(\boldsymbol{t} + \Delta t)$ by fitting $\boldsymbol{y}$ with a cubic spline.

RESULTS

Regulatory network under rich-medium growth condition

Harbinson *et al*. profiled 203 TFs for their genome-wide DNA binding sites under rich medium growth condition (Harbison *et al*., 2004; Lee *et al*., 2002). We consider only those TFs that have evidenced binding targets. To avoid multiple counting of TF–gene relationships, if a group of TFs (e.g. HAP2/HAP3/HAP4/HAP5) always operate together as a functional unit according to the literature, we will count them as one TF. There are a total of 99 TFs used in our analysis. Their names and functions provided by *Saccharomyces* Genome Database (SGD) (Dwight *et al*., 2002) are given in Supporting Table 4.

We start with 891 evidenced relationships and 29 154 lower-confidence relationships. Using the cell-cycle microarray data by Spellman *et al*. (1998), we apply the algorithm as described in Methods section to reach a final network which has 3846 TF–gene connections. For each TF, we examine the biological processes that its target genes participate by GO Term Finder of SGD (Ashburner *et al*., 2000; Dwight *et al*., 2002). The list of genes

regulated by each TF can be found at http://www.stat.ucla.edu/~tyu/factor/. The over-represented terms are given in Supporting Table 5.

Some TFs are more specialized, whereas others act on a broader range of cellular processes. The GO slims define broad biological processes (Ashburner *et al*., 2000; Dwight *et al*., 2002). For each process, we identify TFs that regulate significant numbers of genes in it (Table 1). We find ABF1, FKH1/2 and INO2/4 to be the leading factors, each acting on 9 of the 33 processes. ABF1 and INO2/4 mostly act on metabolic and transport processes, whereas FKH1/2 mostly acts on cell cycle-related processes. Other widely influential factors include cell cycle-related SWI4, SWI6, and metabolism-related MSN2/4, HAP1, HAP2/3/4/5 and XBP1.

Figure 1 shows the regulatory relationship between TFs. An arrow points from a TF to another TF if the latter is the target gene of the former according to the TF–gene network we constructed. Consistent with their biological roles, we find the cell-cycle regulatory TFs SWI6, SWI4, FKH1/FKH2, ACE2/SWI5, MCM1 and STB1 (diamond nodes, Fig. 1) are linked together. Around the leading hub in the network, GAT1, we find a sub-network that involves nitrogen metabolism-related TFs GAT1, DAL80, DAL81, GZF3, GLN3, and stress-related TFs IXR1, XBP1, YAP1, RPN4 and HAP1 (square nodes, Fig. 1). Interestingly, the two regulators of GAT1 expression are cell-cycle TFs ACE2/SWI5 and FKH1/FKH2.

We investigate if the activity profile of a TF is correlated with its own gene-expression profile subject to a possible time delay. We consider the alpha-factor data, cdc-15 data and cdc-28 data separately. The elutriation-synchronized data are excluded from our analysis because the time interval (every 30 min) used in collecting the mRNA sample is too long. For each TF, we first compute the activity–expression correlation without time delay. Table 2 lists a total of 17 TFs which have correlation >0.4 in at least two of the three synchronization experiments. For each of the remaining TFs, we analyze the time-shifted activity–expression correlation as described in Methods section. We find 10 TFs showing delayed activity–expression correlation (see Table 3). As an example, the time-delay pattern for SWI4 is shown in Figure 2. Both the expression and the activity profiles exhibit cell-cycle pattern periodicity. The estimated time lag is $\sim$10 min (1/6 cycle).

As suggested by one referee, a related issue that can be addressed by using GO is about the functional homogeneity of transcription modules. Similar to the distance measure used by Ye and Godzik (2004) in studying protein domains, we compute the average length of the shortest GO-path between two genes linked to the same TF. The results are summarized in Supporting Figure 3. We only find a marginally significant (*P*-value 0.0324, one-sided signed rank sum test) decrease of distance when comparing with TF modules obtained by using CHIP data alone. Another measure based on GO-slims yields similar findings (see Supportive Information Text 1 for more discussion). Note that we have not paid attention to TF modules defined by a combination of TFs yet. Although ideally one would expect higher functional homogeneity for such better-defined modules, this is certainly a more complicated problem to address.

Table 1. TFs that regulate a significant fraction of genes (*P*-value <0.01) in each broad biological process as defined by GO slims

Biological process	Major TFs
Amino acid and derivative metabolism	GCN4, CBF1/MET4/MET31, LEU3, BAS1, PHO2
Carbohydrate metabolism	GCR1/GCR2, MSN2/MSN4
Cell budding	SWI6, SWI4
Cell cycle	FKH1/FKH2, SWI6, RPN4, ACE2/SWI5
Cell homeostasis	HAP1, MAC1, SKN7
Cell wall organization and biogenesis	INO2/INO4, SWI4, FKH1/FKH2, RLM1, TEC1
Cellular respiration	HAP2/HAP3/HAP4/HAP5, HAP1, RTG1/RTG3
Conjugation	STE12
Cytokinesis	FKH1/FKH2, ACE2/SWI5, XBP1
Cytoskeleton organization and biogenesis	ABF1, FKH1/FKH2, SWI4
DNA metabolism	SWI6, HIR3, SFL1, UME6
Electron transport	HAP1, INO2/INO4
Generation of precursor metabolites and energy	GCR1/GCR2, HAP2/HAP3/HAP4/HAP5, INO2/INO4, MSN2/MSN4
Lipid metabolism	HAP1, ABF1, INO2/INO4, SWI4
Meiosis	UME6, SFL1, FKH1/FKH2, RPN4, SUM1
Membrane organization and biogenesis	
Morphogenesis	SFL1
Nuclear organization and biogenesis	FKH1/FKH2, ABF1
Organelle organization and biogenesis	FKH1/FKH2, GCR1/GCR2, ADR1
Protein biosynthesis	RAP1, ABF1, PDR1, ROX1, CUP9, XBP1, GZF3, INO2/INO4
Protein catabolism	RPN4, REB1, GAT1, ABF1, XBP1, ADR1, SFL1
Protein modification	FKH1/FKH2, PHO2, INO2/INO4, SWI4, SWI6, REB1
Pseudohyphal growth	FKH1/FKH2
Response to stress	MSN2/MSN4, HSF1, INO2/INO4, GLN3, PHO2, XBP1, YAP1
Ribosome biogenesis and assembly	ABF1, HAP2/HAP3/HAP4/HAP5, STP1/STP2
RNA metabolism	ABF1, STP1/STP2, HAP2/HAP3/HAP4/HAP5, MSN2/MSN4
Signal transduction	STE12, SKN7
Sporulation	SUM1, UME6
Transcription	ABF1, PPR1
Transport	OAF1, DAL82, MIG1, RIM101, ADR1, INO2/INO4, XBP1
Vesicle-mediated transport	ABF1, GAT1, INO2/INO4, PPR1, REB1, XBP1
Vitamin metabolism	THI2, MAC1

Regulatory network under stress condition

TF binding to a subset of the regulatory sequences may be dependent on the environmental conditions of the cell. Harbison *et al*. (2004) analyzed the genome-wide binding properties of 84 TFs under multiple stress conditions. Combining this dataset with the stress–response microarray gene expression data (Gasch *et al*., 2000), we shall identify a network underlying the gene-expression regulation in stress conditions.

In the stress-specific ChIP dataset (Harbison *et al*., 2004), some TFs are profiled in multiple conditions. We include a TF–gene

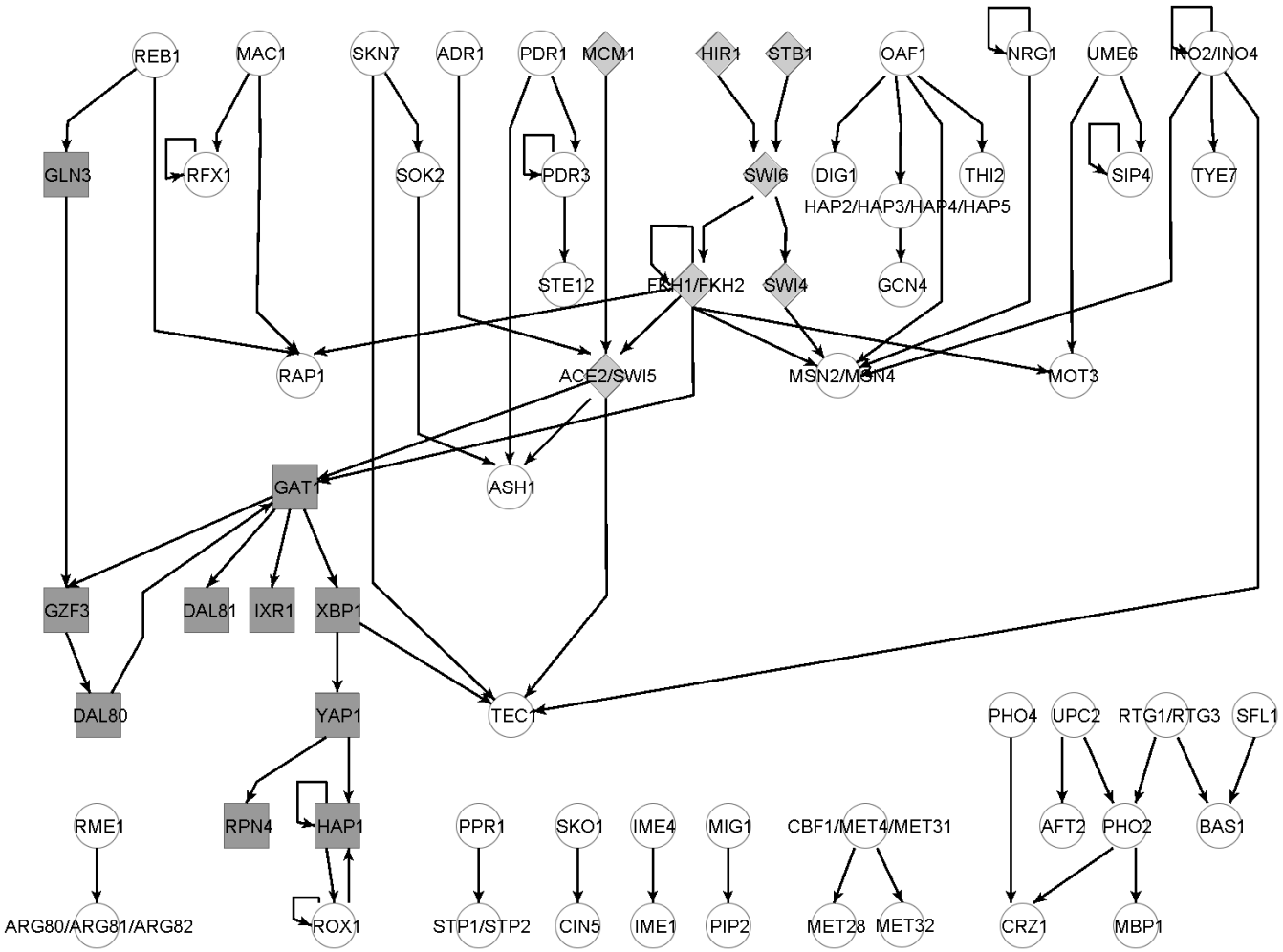

Fig. 1. The regulatory relationship between TFs under normal growth condition. An arrow points from a TF to another TF if the latter is the target gene of the former according to the TF–gene network we constructed.

linkage in the lower-confidence set as long as it is observed in one of the conditions. Starting with 579 higher-confidence TF–gene relationships and 29 316 lower-confidence relationships, we apply our algorithm and obtain a network of 8183 TF–gene connections, which involve 49 TFs. The list of genes regulated by each TF can be found at http://www.stat.ucla.edu/~tyu/factor/. For each TF, we find biological processes that are over-represented by its targeted genes (see Supporting Table 6).

We further use the regulatory network to study the 868 environmental stress–response (ESR) genes reported by Gasch *et al*. (2000). Among the 585 genes repressed in ESR, 98% have connections in our network, compared with 78% for non-ESR genes. At the significance level of 10^{-5}, seven TFs are identified as major regulators of these genes. The most prominent among them is RAP1, which regulates 93 genes. The others are ARG80/ARG81/ARG82 (regulating 56 genes), RCS1 (52 genes), CBF1/MET4/MET31 (46 genes), HSF1 (45 genes), RTG1/RTG3 (43 genes) and GAT1 (42 genes). Among the 283 upregulated genes in ESR, 97% have connections in our network. At the significance level of 10^{-5}, six TFs are identified as major regulators of these genes. They are MSN2/MSN4 (79 genes), PHO2 (32 genes), HAP2/HAP3/HAP4/HAP5 (26 genes), AFT2 (26 genes), ROX1 (26 genes) and RPH1 (26 genes).

DISCUSSION

We have presented a method to infer the transcriptional regulatory network at the full genome scale, by integrating information from microarray gene-expression data, genome-wide location (ChIP) data and the evidenced TF–target gene relationships in the biomedical literature. Our method is based on a constrained space factor analysis model, which treats TF activity as hidden variables.

In the analysis of transcriptional regulation, one central theme is how to describe TF activity. By finding co-regulated gene modules, some authors implicitly infer building blocks of the network without modeling the TF activity (Eisen *et al*., 1998; D'haeseleer *et al*., 2000; Ihmels *et al*., 2002, 2004; Kwon *et al*., 2003; Toh and Horimoto, 2002). Studying co-expression dynamics with other gene-expression levels as indicators of cellular state changes also by-passes the TF activity modeling issue (Li, 2002; Li *et al*., 2004).

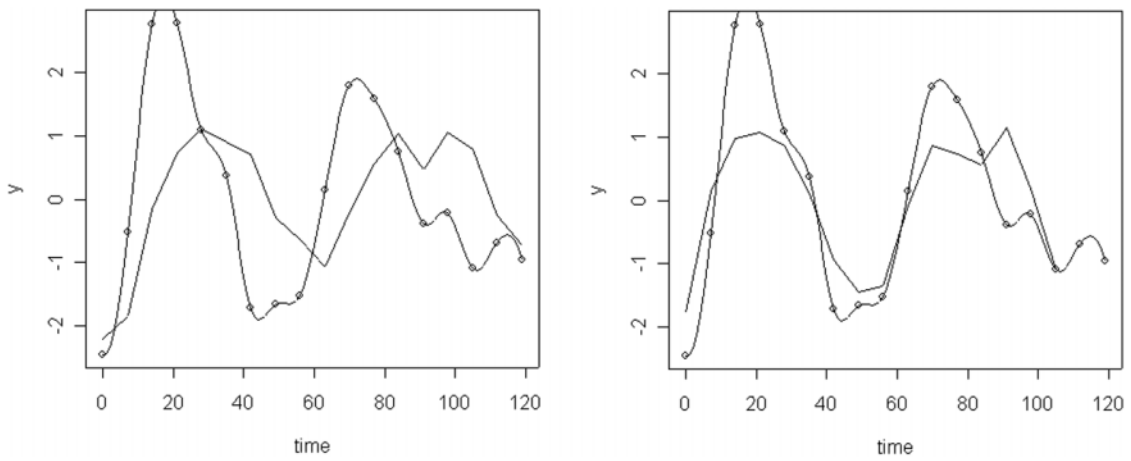

Fig. 2. Time-shifting relationship between the expression profile and activity profile of SWI4; alpha-factor synchronized cell cycle, 10 min shift. Left panel: the expression profile (curve) and activity profile (broken line); right panel: the expression profile (curve) with shifted activity profile (broken line).

Table 2. TFs that exhibit correlated expression and activity in at least two of the three synchronized cell cycle experiments

Factor	Gene	Correlation Alpha	cdc-15	cdc-28
GZF3	*GZF3*	−0.12	0.55	0.54
IME1	*IME1*	0.49	0.71	0.44
ASH1	*ASH1*	0.43	0.72	0.57
HAP2/3/4/5	*HAP2*	0.61	0.44	0.26
IME1	*IME1*	0.49	0.71	0.44
MAL13	*MAL13*	0.46	0.65	0.58
FKH1/2	*FKH1*	−0.66	−0.65	−0.81
NRG1	*NRG1*	−0.25	0.66	0.52
PDR1	*PDR1*	0.16	0.61	0.57
PHO2	*PHO2*	−0.41	−0.65	−0.14
RCS1	*RCS1*	0.51	−0.21	0.43
RFX1	*RFX1*	−0.34	−0.65	−0.96
RME1	*RME1*	0.34	0.87	0.73
ROX1	*ROX1*	−0.51	−0.45	−0.37
RPN4	*RPN4*	−0.47	−0.72	−0.09
THI2	*THI2*	0.59	0.88	0.66
YAP1	*YAP1*	0.39	0.82	0.56

Table 3. TFs that have activity lagging behind expression in at least two of the three synchronized cell-cycle experiments

Factor	Gene	Alpha-factor Shift (min)	Alpha-factor Correlation	cdc-15 Shift (min)	cdc-15 Correlation	cdc-28 Shift (min)	cdc-28 Correlation after shifting
CHA4	*CHA4*	4	0.66	19	0.40	17	0.49
HAA1	*HAA1*	8	0.79	14	0.32	20	0.66
HAP2/3/4/5	*HAP4*	15	−0.93	10	0.32	10	−0.74
HIR2	*HIR2*	7	−0.20	20	0.68	7	0.53
MET28	*MET28*	5	0.48	16	0.73	12	0.43
MIG2	*MIG2*	8	0.68	4	0.69	18	−0.35
PDR3	*PDR3*	18	−0.36	20	0.60	20	0.65
PHO4	*PHO4*	17	0.92	1	0.45	10	−0.34
SWI4	*SWI4*	10	−0.87	20	−0.29	17	−0.74
SWI6	*SWI6*	15	0.59	0	−0.32	20	0.86

Some authors tried to connect the TFs' activity directly with their gene-expression levels (Qian *et al*., 2003; Segal *et al*., 2003; Zhu *et al*., 2002). Bayesian learning by perturbed expression aims at directly finding the network structure, without the need to estimate TF activities (Pe'er *et al*., 2001; Pournara and Wernisch, 2004; Rung *et al*., 2002).

As in Liao *et al*. (2003), our method treats TF activities as hidden variables which help both the network configuration specification and the TF-binding strength modeling simultaneously. The estimation of a TF's activity is independent of the information about its own transcription profile, which allows further analysis of TF behavior as we demonstrated in the Results section.

Different from Liao's NCA model, however, we did not consider the network configuration as being given. Nevertheless, we consider both network configuration and connection strength estimation as integrative components of a general factor analysis model. We fit the model by iterating between the step of network configuration search and the step of parameter estimation. Several factors necessitate this adaptive model fitting approach. First, high-throughput data contain high levels of biological and measurement noise. Second, we have only incomplete knowledge about the network configuration. Third, there are probably other hidden variables, e.g. unknown TFs that are not included in the model. They may have confounding effects with the variables under study. The use of prior knowledgebase of TF–target gene relationship and our stepwise expansion of the network connection make our approach more immune to these confounding variables. Conceptually, our evolving model approach is analogous to model building by the neural network approach. In neural network modeling, the number of parameters that have to be estimated from the data is overwhelming. Yet with proper training, the network can converge to a useful local optimal solution. Likewise, although the full size factor analysis model (1) has multiple solutions, we aim at converging to a local optimum by adaptive learning. The available TF-target knowledge

serves us well in providing a reasonable starting point. As the repertoires of data and knowledge grow richer and richer in the future, we can expect our approach to become even more powerful.

We report two regulatory networks under different growth conditions for *Saccharomyces cerevisiae*. The network under the normal growth condition is estimated from cell-cycle microarray data and normal growth ChIP data. Based on the TF activity identified from the cell-cycle time-series, new time-shifting relationships are found between the activity and expression of some TFs. The stress–response network is estimated by using stress–response ChIP data and gene-expression data, pooling many stress conditions together. This network explains the expression of 98% of the ESR genes identified by Gasch *et al*. (2000), and correctly identifies several leading regulators. Somewhat expected, a comparison between the two networks shows that most TFs are regulating different sets of genes. An interesting exception is RAP1 (repressor activator protein). RAP1 regulates 45 genes for the network under normal growth condition, whereas it regulates 211 genes under stress condition. Among these two sets, 27 genes are shared (P-value $\sim 10^{-27}$). Furthermore, we find that 26 of the 27 shared genes are associated with protein biosynthesis, a process that is repressed under stress conditions. This is consistent with RAP1's role in ESR regulation (Gasch *et al*., 2000; Li *et al*., 1999).

In this report, all gene-expression profiles are standardized before the network estimation starts. We did not filter out genes with less-varying expressions in their original scale. As suggested by a referee, proper pre-screening should help reduce the instability in estimating our model parameters associated with such non-informative genes. During the revision of this paper, we examined the standard deviations in the original expression profiles and compared those for genes in our final network with those for the remaining genes. We find significantly higher expressional variation for the genes in the network [(P-value: 1.4×10^{-77} for cell-cycle data and 4.4×10^{-129} for stress–response data); see Supporting Figure 9]. These findings suggest that our results are not overwhelmed by the non-varying expression profiles.

ACKNOWLEDGEMENTS

We thank Dr Chiara Sabatti and Dr James Liao for helpful discussions. This work is supported by NSF grants DMS-0201005, DMS-0104038 and DMS-0406091.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Bar-Joseph,Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol*., **21**, 1337–1342.

D'haeseleer,P. *et al*. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Dwight,S.S. *et al*. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*., **30**, 69–72.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Faraway,J.J. (2004) *Linear Models by* R. Chapman & Hall/CRC, Boca Raton, FL.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gifi,A. (1990) *Nonlinear Multivariate Analysis*. John Wiley & Sons, New York.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Ihmels,J. *et al*. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet*., **31**, 370–377.

Ihmels,J. *et al*. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Kwon,A.T. *et al*. (2003) Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, **19**, 905–912.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Li,B. *et al*. (1999) Transcriptional elements involved in the repression of ribosomal protein synthesis. *Mol. Cell. Biol*., **19**, 5393–5404.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

Li,K.C. *et al*. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15566.

Liao,J.C. *et al*. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Morrison,D.F. (1990) *Multivariate Statistical Methods*. McGraw-Hill Publishing Company, New York.

Pe'er,D. *et al*. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17** (Suppl. 1), S215–S224.

Pournara,I. and Wernisch,L. (2004) Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, **20**, 2934–2942.

Pritsker,M. *et al*. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res*., **14**, 99–108.

Qian,J. *et al*. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.

Qiu,P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun*., **309**, 495–501.

Rung,J. *et al*. (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics*, **18** (Suppl. 2), S202–S210.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Toh,H. and Horimoto,K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**, 287–297.

Wingender,E. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*., **29**, 281–283.

Xu,X. *et al*. (2004) Learning module networks from genome-wide location and expression data. *FEBS Lett*., **578**, 297–304.

Ye,Y. and Godzik,A. (2004) Comparative analysis of protein domain organization. *Genome Res*., **14**, 343–353.

Zhou,X.J. *et al*. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol*., **23**, 238–243.

Zhu,Z. *et al*. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol*., **318**, 71–81.