BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 17 2008, pages 1874–1880
doi:10.1093/bioinformatics/btn332

Gene expression

Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data

Huai Li and Ming Zhan*

Bioinformatics Unit, Branch of Research Resources, National Institute on Aging, NIH, Baltimore, MD 21224, USA

Received on March 12, 2008; revised on June 3, 2008; accepted on June 25, 2008

Advance Access publication June 27, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Unraveling the transcriptional regulatory program mediated by transcription factors (TFs) is a fundamental objective of computational biology, yet still remains a challenge.

Method: Here, we present a new methodology that integrates microarray and TF binding data for unraveling transcriptional regulatory networks. The algorithm is based on a two-stage constrained matrix decomposition model. The model takes into account the non-linear structure in gene expression data, particularly in the TF-target gene interactions and the combinatorial nature of gene regulation by TFs. The gene expression profile is modeled as a linear weighted combination of the activity profiles of a set of TFs. The TF activity profiles are deduced from the expression levels of TF target genes, instead directly from TFs themselves. The TF-target gene relationships are derived from ChIP-chip and other TF binding data. The proposed algorithm can not only identify transcriptional modules, but also reveal regulatory programs of which TFs control which target genes in which specific ways (either activating or inhibiting).

Results: In comparison with other methods, our algorithm identifies biologically more meaningful transcriptional modules relating to specific TFs. We applied the new algorithm on yeast cell cycle and stress response data. While known transcriptional regulations were confirmed, novel TF-gene interactions were predicted and provide new insights into the regulatory mechanisms of the cell.

Contact: zhanmi@mail.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genes are coordinately expressed under tight regulation by transcriptional factors to carry out complex and condition-specific biological functions in living cells. It is critical to develop computational approaches for deciphering transcriptional regulatory programs, in order to elucidate molecular mechanism of development or disease or identify biomarkers (Brunet *et al*., 2004; Hughes *et al*., 2000; Li and Zhan, 2006; Segal *et al*., 2004; Zhan, 2007). Microarray gene expression data have been extensively used for identifying transcriptional regulatory modules. Various computational methodologies have been introduced for those studies, including projection (e.g. principal component analysis, singular value decomposition, independent component analysis) (Alter *et al*., 2000; Lee and Batzoglou, 2003; Liebermeister, 2002), model-based approaches (e.g. network component analysis, probabilistic sparse matrix factorization) (Dueck *et al*., 2005; Liao *et al*., 2003) and conventional clustering (e.g. hierarchical clustering, self-organizing maps, K-means) (Eisen *et al*., 1998; Tamayo *et al*., 1999; Tavazoie *et al*., 1999). The projection methods decompose the original data into components that are constrained to be mutually either uncorrelated or statistically independent, and cluster genes into mutually non-exclusive modules based on their loading in the components. Since these methods do not cluster genes according to the pair-wise similarity, they can identify sets of coexpressed genes that are potentially co-regulated. Model-based approaches model microarray data as a linear mixture of latent variables that may correspond to specific biological sources. These methods usually incorporate prior knowledge on gene regulatory mechanisms as constraints for precisely estimating model's parameters. For example, the probabilistic sparse matrix factorization approach uses the 'sparse' constraint in the matrix decomposition to provide a combinatorial account of the gene expression in terms of a small set of factors. One challenge of such model-based approaches is the lack of sufficient data to estimate the parameters. Recent simulation studies suggest that transcriptional networks inferred from gene expression data alone can be considerably obscured by spurious interactions when the number of observations is small or the quality of the data is poor (Husmeier, 2003). Several approaches, including GRAM (Bar-Joseph *et al*., 2003), COGRIM (Chen *et al*., 2007) and ReMoDiscovery (Lemmens *et al*., 2006), have been developed to infer transcriptional regulatory networks by integrating gene expression data with transcription factor (TF) binding information. These approaches allow identification of more functionally coherent regulatory modules, in comparison with the analyses utilizing microarray data alone (Bernard and Hartemink, 2005; Joung *et al*., 2006; Kim *et al*., 2006; Yu and Li, 2005; Zhou *et al*., 2005).

We recently developed a two-stage matrix decomposition method that combine the characteristics of projection and model-based approaches for the discovery of transcriptional modules (Li *et al*., 2007). In the present study, we extend the two-stage decomposition method to incorporating TF binding data for unraveling TF-mediated regulatory programs. The new approach provides information of not only transcriptional modules, but also on which of the TFs control which target genes in which specific ways (either activating or inhibiting) in a regulatory program. Considering highly non-linear

*To whom correspondence should be addressed.

interactions between TFs and their target genes, we first adopt a non-linear independent component analysis (NICA) method to reduce the non-linear distortion in the data and decompose the data into independent latent components. Next, we develop a constrained probabilistic sparse matrix factorization (cPSMF) approach that models the expression of each gene across the independent latent components as a linear weighted combination of activity profiles of a small number of TFs. The model takes into account of the combinatorial and sparse nature of gene regulation by TFs. By incorporating TF-target gene relationships derived from ChIP-chip data into the probabilistic sparse matrix factorization, the cPSMF approach infers the network structure in a more accurate and robust manner. Finally, we fine-tune the transcriptional network by selecting target genes whose promoter regions contain a sequence that matches with the binding site of the corresponding TF. In comparison with other methods, our algorithm shows better performance in identifying functionally coherent transcriptional modules relating to specific TFs. We demonstrate the usefulness of the new method in a case study on yeast cell cycle and stress response data. While known transcriptional regulatory interactions were confirmed, novel TF-gene links were also predicted, providing new insights into the regulatory network of the cell.

2 METHODS

2.1 The two-stage constrained matrix decomposition model

In this proposed model, TF-mediated transcriptional regulatory programs are inferred based on integrated results from microarray, ChIP-chip and TF binding motif data. Suppose there is a microarray gene expression data matrix $\mathbf{X}\in\Re^{N\times M}$ with N genes and M samples and a configuration matrix $\mathbf{C}\in\{0,1\}^{N\times L}$obtained from ChIP–chip data with L TFs, where element C_{il} = 1 represents gene i is regulated by TF l. We first take the NICA step to de-nonlinearize microarray data into independent latent components. The model can be written as (Jutten and Karhunen, 2004; Lappalainen and Honkela, 2000; Li *et al*., 2007):

$$\mathbf{X}=\mathbf{F}(\bar{\mathbf{S}}\mathbf{A})+\mathbf{N} \tag{1}$$

where $\bar{\mathbf{S}}\in\Re^{N\times M'}$ denotes the independent latent source matrix and $\mathbf{A}\in\Re^{M'\times M}$ is the mixing matrix. M' is the number of latent sources. $\mathbf{N}$ is a white Gaussian noise matrix. F(·) is a non-linear mixing function, which is modeled using a multilayer perceptron (MLP) network with one non-linear hidden layer (Haykin, 1999). $\bar{\mathbf{S}}$ can be obtained from Equation (1) using the variational Bayesian learning (Lappalainen and Honkela, 2000) and the FastICA algorithm (Hyvarinen and Oja, 2000).

Next, we take the cPSMF stage to model the expression profiles of genes across the independent latent components as linear weighted combinations of L TF activity profiles

$$\bar{\mathbf{S}}=\mathbf{YZ}+\mathbf{N} \tag{2}$$

where $\mathbf{Y}\in\Re^{N\times L}$ is the weighting matrix, and $\mathbf{Z}\in\Re^{L\times M'}$ is the matrix that contains activity profiles of L TFs across the independent latent components. For each TF, we obtain an activity profile from the centroid of the expression profiles of the target genes across the independent latent components. The target genes of each TF are chosen from the configuration matrix $\mathbf{C}$, constructed from the ChIP-chip data. The $\mathbf{Y}$ matrix is inferred from Equation (2) by variational Bayesian learning (Dueck *et al*., 2005; Jordan *et al*., 1999) with constraints that

$$\mathbf{Y}=\mathbf{C}\cdot\mathbf{Y} \tag{3}$$

where · denotes an element-by-element product of $\mathbf{C}$ and $\mathbf{Y}$, which means that the element Y_{il} of $\mathbf{Y}$ can be non-zero only when C_{il} = 1. $\mathbf{C}$ is pre-specified from ChIP-chip data and has a sparse property since any target gene can only be regulated by a small number of TF's biologically. A detail description of the NICA and cPSMF approaches can be found in the Supplementary Material.

The proposed algorithm is summarized as follows:

Input

- Microarray data matrix $\mathbf{X}$ = $[X_{ij}]$, where the element X_{ij} represents the expression level of gene i associated with the j-th sample, $i=1,\ldots,N$, j = 1,…, M
- Configuration matrix $\mathbf{C}$ = $[C_{il}]$, where C_{il} = 1 represents gene i is regulated by TF l, obtained from ChIP-chip data, i = 1,…, N, l = 1,…, L, L is the number of TFs.
- Maximum number of effective TFs, K. K should be less or equal to L.

Output

- The topology of transcriptional regulatory networks.

Algorithm

- De-nonlinearize $\mathbf{X}$ using the NICA approach.
  - Find the non-linear principal components matrix $\mathbf{S}=\bar{\mathbf{S}}\mathbf{A}$ from Equation (1) by variational Bayesian learning.
  - Decompose $\mathbf{S}=\bar{\mathbf{S}}\mathbf{A}$ by the linear FastICA algorithm to obtain linear independent components in $\bar{\mathbf{S}}$.
- Decompose $\bar{\mathbf{S}}$ from Equation (2) using the cPSMF method.
  - Construct $\mathbf{Z}$ as follows: For l = 1,…, L cluster the expression profiles of target genes of TF l into two groups (activated or inhibited pattern) using K-means approach. Then choose the centroid profile with the larger variance as the activity profile of TF l.
  - Initialize the element Y_{il} in $\mathbf{Y}$ to an arbitrary value if $C_{il}=1$, else set the value of Y_{il} to zero.
  - Infer $\mathbf{Y}$ from Equation (2) by factorized variational inference with constraint $\mathbf{Y}$ = $\mathbf{C}\cdot\mathbf{Y}$.
- Reconstruct transcriptional networks regulated by given TFs based on $\mathbf{Y}$.
  - For l = 1,…, L and i = 1,…, N, if $Y_{il}>\alpha$ then gene i is positively regulated or activated by TF l. α is a predetermined weight cutoff.
  - For l = 1,…, L and i = 1,…, N, if $Y_{il}<-\alpha$ then gene i is negatively regulated or inhibited by TF l.
- Fine-tune gene transcriptional networks by selecting target genes whose promoter regions contain a sequence that matches with the binding site of the corresponding TF.
  - For$l$ = 1,…, L and $\forall i$:$Y_{il}\neq 0$, compute the core similarity score O_{il} between the position-specific weight matrix of the TF l binding motif and the promoter sequence of the target gene i, using the MATCH software searching against the TRANSFAC database (http://www.biobase-international.com). Select the candidate target gene i regulated by the TF l if $O_{il}\geq\beta$; β is a predetermined cutoff value.

2.2 Biological assessment and visualization of inferred regulatory networks

To assess biological relevance of inferred regulatory network, we examined whether the identified TF-regulated transcriptional modules accumulate in certain Gene Ontology (GO) categories by conducting two different analyses: over-representation analysis and gene set enrichment analysis. The over-representation analysis was used to detect if a GO term is enriched in a transcriptional network. In specific, the hypergeometric probability p that a GO term is significantly enriched in a network is calculated as:

$$p=1-\sum_{i=0}^{k-1}\frac{\binom{A}{i}\binom{G-A}{n-i}}{\binom{G}{n}} \tag{4}$$

where k is the number of genes in the group, G is the total number of genes, n is the number of genes in the module with a given GO term and A is the total number of genes with a given GO term.

The gene set enrichment analysis is based on the non-parametric test. When considering an arbitrary GO category, it evaluates if the genes in the identified transcriptional modules that belong to the category are uniformly distributed or accumulated in the list sorted by some specific criteria (Backes *et al*., 2007). Since both the over-representation analysis and the gene set enrichment analysis were applied to many GO categories, we further conducted the false discovery rate (FDR) correction, which provides strong control to have less false negatives at the cost of a few more false positives.

We took advantage of Cytoscape's versatile visualization environment (Shannon *et al*., 2003) to produce graphic representation of the resulting regulatory networks.

2.3 Experimental datasets

Two publicly available yeast microarray datasets were used for algorithm evaluation and case studies. The first was a cell cycle dataset, which was determined under the normal growth condition (Spellman *et al*., 1998). The second was a stress response dataset, determined under different experimental conditions such as temperature shocks, amino acid starvation, nitrogen source deletion and progression into stationary phase (Gasch *et al*., 2000). The normalization of the datasets was conducted by zero transformation (Gollub *et al*., 2006), and the missing data were filled using the KNNimpute approach (Troyanskaya *et al*., 2001).

The yeast ChIP-chip dataset used for the studies was obtained from a previous study (Harbison *et al*., 2004). The dataset contains 203 TFs, with all profiled in a rich medium and 84 profiled under multiple stress conditions. We used the TF-gene pairs that have P-values <0.05 to construct the configuration matrix C in our analysis.

The known TF-gene interaction data used in the studies were obtained from the yeast transcriptional regulatory network previously published (Herrgard *et al*., 2003; Milo *et al*., 2002). The yeast promoter sequences and TF binding motifs data were obtained from Harbison *et al*. (2004).

3 RESULTS AND DISCUSSION

3.1 Feature of the algorithm

The motivation of this work was to provide a mathematical framework for identifying condition-specific transcriptional regulatory networks by integrated analysis of microarray, ChIP-chip and TF binding motif data. An important feature of our method is the utilization of two-stage data decomposition (NICA + cPSMF). The NICA transformation captures the non-linear structure in the data and represents the data with independent latent components. Inspired by the fact that gene expression is regulated by a small set of TFs that act combinatorially, the cPSMF models the expression profile of each gene represented by the independent latent components as a linear combination of activity profiles of a small number of TFs. A configuration matrix (C matrix) is incorporated into the modeling as constraint for precisely estimating the influence of TFs. The configuration matrix, derived from ChIP-chip data, has a sparse property. That is, there are only a few non-zero elements in the matrix, as target genes can only be regulated by a small number of TF's. Through learning the parameters of the cPSMF model, the algorithm can infer activating and inhibitory regulatory relationships between TFs and their gene targets. The strength and direction of transcriptional regulation that a TF applies on its target genes are reflected by the weight of the TF presented in the Y matrix.

Most methods for inferring TF-regulated transcriptional modules are based on the assumption that there exists a correlation on the mRNA expression level between TFs and their target genes (Kim *et al*., 2006; Zhu *et al*., 2002). This assumption is however not always true, since the activation or inhibition of a target gene by a TF can be influenced by not only the mRNA expression of the TF and their targets, but also by post-transcriptional modification of the genes, as well as the concentration, post-translational modification and cellular localization of their protein products. Because of these, our algorithm uses the expression patterns of TF target genes, instead of TFs themselves, to deduce the activity profiles of TFs.

When applying our algorithm, we set the number of independent latent components equal to the number of experimental conditions for simplicity. For more accurate non-linear mapping, we set the number of hidden neurons in the MLP network as twice as the number of independent latent components. We also set K, the maximum number of effective TFs bound on target genes, equal to two in our algorithm. The choice of the parameters α (weight cutoff) and β (PWM matching score cutoff) is important for the structure of the inferred network. Since in general, ground-truth data are hardly available for condition-specific situation, we take a conserved approach in setting up the parameters in our case studies. We set weight cutoff $\alpha=0.05$ and PWM matching score cutoff $\beta=0.94$. Similar conserved parameters are also adopted in other similar studies (Kim *et al*., 2006).

3.2 Comparison with other methods

To evaluate our algorithm, we compared its performance with those by other similar methods, including GRAM (Bar-Joseph *et al*., 2003), COGRIM (Chen *et al*., 2007) and ReMoDiscovery (Lemmens *et al*., 2006). The latter three methods can predict transcriptional modules that are coregulated by TFs through integrated analysis of microarray, ChIP-chip and TF motif data. GRAM is based on an iterative search, in which genes with common TF binding sites on the promoters are first identified using ChIP-chip data and the clustered gene sets are further refined by shared expression profiles. COGRIM is derived from a Bayesian hierarchical model, while ReMoDiscovery is a non-iterative approach. While our method and COGRIM can infer activating or inhibiting relationships between TFs and their target genes, GRAM and ReMoDiscovery can not predict such relationships.

We identified TF-mediated transcriptional modules using our method as well as the three other methods, based on the same set of data of microarray (from both cell cycle and stress response), ChIP-chip and TF binding motifs derived from the yeast (see Section 2.3). We chose 19 TFs that are involved in the cell cycle and stress response, and identified their target genes. We then examined the functional relevance of the target gene clusters based on the GO using over-representation analysis and gene set enrichment analysis. In the gene set enrichment analysis, the input set was sorted by the variance of the expression profile of the target genes. Table 1 shows

Table 1. Comparison with other methods based on GO functional enrichment

Gene set[a]	Over-representation analysis				Gene set enrichment analysis			
	Our algorithm	GRAM	COGRIM	ReMoDis.	Our algorithm	GRAM	COGRIM	ReMoDis.
ABF1	5.93	6.12	5.74	4.55	4.56	4.60	5.05	3.76
ACE2	4.43	1.60	3.59	4.47	2.06	5.23	3.43	1.53
FKH1	4.62	4.79	1.76	4.05	4.51	1.91	2.99	2.04
FKH2	6.10	1.28	5.82	6.09	2.86	3.60	4.02	2.41
GCN4	7.23	6.64	7.40	7.61	5.71	2.09	1.48	4.72
LEU3	7.74	7.29	6.44	5.20	2.70	1.32	1.05	1.83
MBP1	6.08	6.15	4.93	6.17	4.17	4.21	4.91	3.40
MCM1	6.13	6.87	5.97	6.74	2.71	1.19	2.93	1.40
NDD1	1.64	1.82	1.66	4.49	2.67	2.40	3.09	3.47
RAP1	8.39	7.37	8.92	6.6	6.49	0.85	2.17	2.04
REB1	5.52	5.03	5.53	5.10	4.13	5.09	4.92	3.46
STB1	4.72	4.43	3.91	6.51	2.11	0.50	2.89	5.09
SWI4	7.06	5.61	5.76	6.42	5.22	4.25	4.73	1.56
SWI5	4.41	2.36	5.7	4.78	6.01	5.79	5.54	0.94
SWI6	5.66	4.58	4.79	4.90	2.81	4.08	4.23	4.23
HSF1	6.53	4.42	4.4	4.38	3.47	1.64	2.92	3.52
MSN4	6.31	5.51	5.67	5.72	2.77	0.86	4.58	1.93
SKN7	5.86	6.75	4.55	2.91	1.29	2.88	1.27	0.85
YAP1	6.08	5.69	6.64	5.92	3.10	2.62	2.81	1.90
Averaged over TFs	5.81	4.96	5.22	5.40	3.65	2.90	3.42	2.64

The comparison is conducted based on two enrichment analysis methods: over-representation analysis and gene set enrichment analysis. The target gene sets of 19 TFs relating to cell cycle and stress response are evaluated. The GO functional enrichment levels of target genes identified by each method are shown. The enrichment level is calculated by transforming the enrichment P values after FDR correction to the negative log values and averaged over all functional modules for corrected $P<0.05$. If no functional modules are found for corrected $P<0.05$, the smallest value of corrected P is taken for calculating the enrichment level.
[a]each gene set is named by the TF that regulates the genes.

the statistical enrichment of functional GO terms in the target gene clusters based on these two analysis approaches. The enrichment level was calculated by transforming the enrichment P-values after FDR correction to the negative log values and averaged over all functional modules for corrected $P < 0.05$. If no functional modules are found for corrected $P < 0.05$, the smallest value of corrected P is taken for calculating the enrichment level. As illustrated, our algorithm out-performs the other methods on the functional enrichment in the target gene clusters. The averaged enrichment level over all the 19 clusters in over-representation analysis was the highest by our algorithm (5.81), followed by ReMoDiscovery (5.40), COGRIM (5.22) and GRAM (4.96). The averaged enrichment level over all the 19 clusters in gene set enrichment analysis was also the highest by our algorithm (3.65), followed by COGRIM (3.42), GRAM (2.90) and ReMoDiscovery (2.64). This implies that our algorithm can identify more functionally coherent gene clusters relating to specific TFs. In terms of the number of target genes identified, the average gene number per cluster was the highest by our algorithm (91), followed by COGRIM (85), ReMoDiscovery (74) and GRAM (32). Interestingly, our algorithm identified more target genes that are annotated with known functions, indicating that our approach provides more functional information about transcriptional regulatory program. The better performance of our algorithm in comparison with others indicates that the mathematical framework we choose for modeling the transcriptional regulatory program is appropriable.

Nevertheless, challenges still remain for identifying TF-regulated transcriptional programs. There are limited ChIP-chip data available for such analysis. The identification of TF binding sites on the promoter sequences is often associated with high false positive or negative errors. Moreover, ChIP-chip data and gene expression profile data are often generated under different experimental conditions. It is not clear how this difference effects the identification of condition specific regulatory programs by integrated analysis of these data.

3.3 Case study: regulatory networks of the yeast

3.3.1 Cell cycle We applied our algorithm to infer the transcriptional regulatory program of the cell cycle in the yeast under the rich medium growth condition. The analysis was based on the microarray data and TF binding information determined by ChIP-chip and promoter sequence analysis (see Section 2.3). From the total 203 TFs, we selected 32 TFs for the analysis according to their ranked activity profiles, which were sorted by the variance of the activity profiles of TFs. These TFs should be top cell cycle regulated TFs. Cheng and Li recently proposed a two-step method to identify the cell cycle regulated TFs by integrating microarray data with ChIP-chip data (Cheng and Li, 2008). We compared these 32 TFs with the putative cell cycle TFs identified by Cheng and Li's method, as well as the TFs identified by Tsai *et al.* (2005). Among them, 15 are known cell cycle related factors from previous experiments and 8 are also suggested as putative cell cycle TFs by these studies (Cheng and Li, 2008; Tsai *et al.*, 2005). Figure 1A shows the inferred regulatory network. The network contained 2017 TF-target interactions. Among these

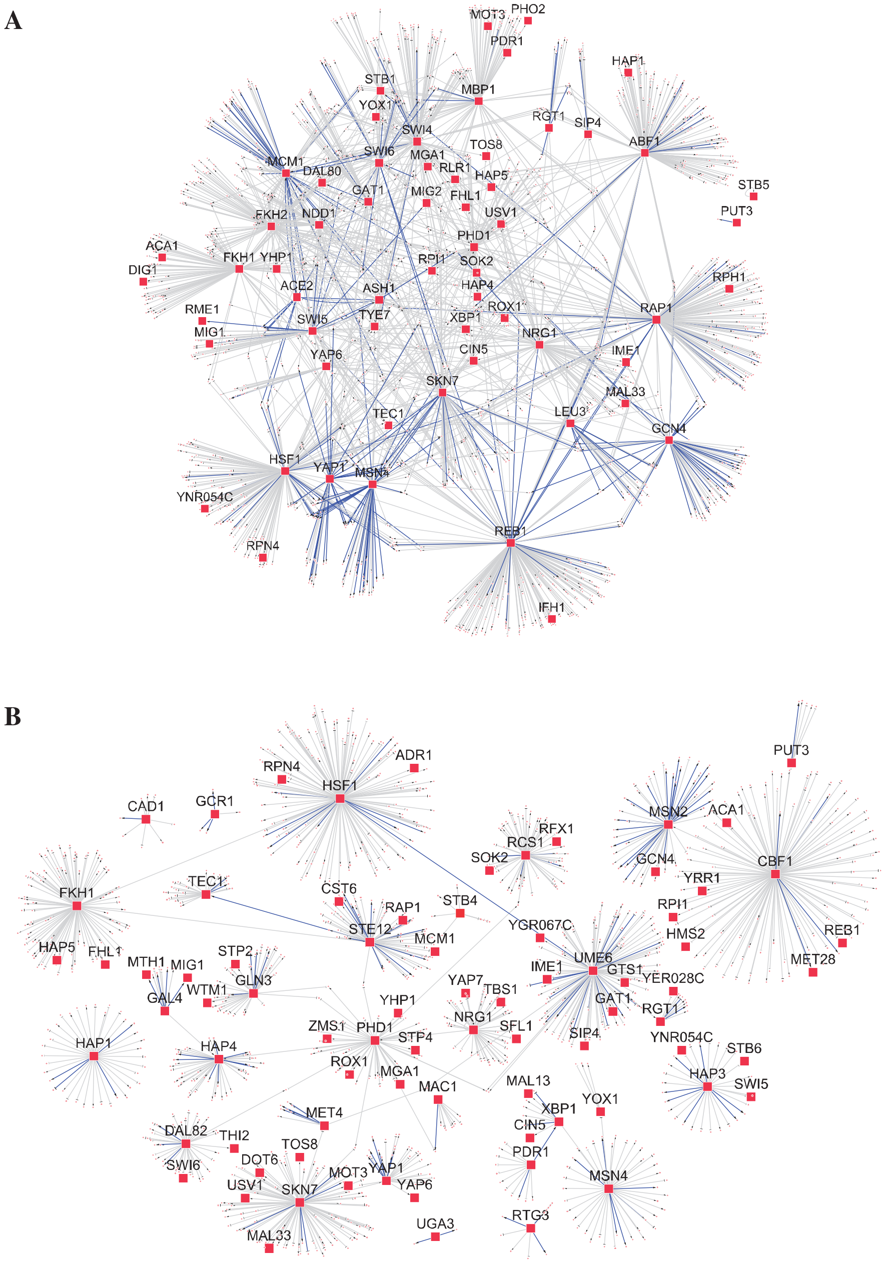

Fig. 1. Visualization of the transcriptional regulatory networks of the yeast inferred using our algorithm. TFs are represented by red squares and their target genes by small circles. Blue and gray lines indicate known and predicted regulatory interactions, respectively. (**A**) cell cycle, (**B**) stress response.

interactions, 160 are the known regulatory links (blue solid lines) from literature (Herrgard *et al*., 2003; Milo *et al*., 2002), while the others are new predictions. We found that the known cell cycle regulatory TFs FKH1, FKH2, MBP1, MCM1, NDD1, REB1, SKN7, SWI4, SWI5 and SWI6 were the predominant hubs in the network. The hub genes also included ABF1, GAT1, HSF1, MSN4, NRG1, PHD1, RAP1 and YAP1. Among them, ABF1, MSN4, NRG1, PHD1 and RAP1 are putative cell cycle TFs that previous studies have also suggested. For example, NRG1 was identified as a cell cycle regulated TF by (Cheng and Li, 2008). There were 76 target genes that were regulated by NRG1 in our inferred network, and these genes showed significant biological relevance (GO enrichment level 5.66). Similarly, PHD1, also previously predicted as an cell cycle regulated TF (Tsai *et al*., 2005), regulated 70 target genes, which were biologically significant (GO enrichment level 5.10). GAT1 was a major hub gene on the yeast transcriptional network and was regulated by cell cycle related TFs ACE2/SWI5 and FKH1/FKH2 in our network. These results are consistent to previous observations (Yu and Li, 2005). Besides showing TF-target interactions, our regulatory network further identifies activating and inhibitory relationships between TF and target genes. For example, SWI5 activated the expression of ASH1, while ASH1 repressed SWI5 (Fig. 1A). These findings are once again consistent to previous observations (Herrgard *et al*., 2003; Milo *et al*., 2002). The consistency of our findings from the inferred network with experimental data and predictions by others support the validity of our methods in inferring the regulatory networks.

3.3.2 Stress response The inference of this regulatory network using our algorithm was based on the stress response microarray data under heat shock from 25°C to 37°C, along with 203 TFs and their genomic binding sites determined by ChIP and promoter sequence analysis (see Section 2.3). Firstly, we chose top 32 active TFs according to their ranked activity profiles. These 32 TFs include all stress response related factors that are experimentally confirmed. Figure 1B shows the inferred network. Among the 1403 detected TF-target interactions in the network, 153 are the known regulatory links (blue solid lines) previously confirmed (Herrgard *et al*., 2003; Milo *et al*., 2002), while the others are novel predictions. Among the hub genes of the network, HSF1 was the most predominant, regulating 155 target genes. The other hub genes included CBF1 (115 targets), UME6 (115 targets), SKN7 (113 target genes), FKH1 (95 targets), STE12 (67 targets), MSN2 (48 targets), MSN4 (44 targets), PHD1 (43 targets), NRG1 (44 targets), YAP1 (29 targets) and CAD1 (8 targets). Six of the hub genes are particularly related to the stress response (HSF1, SKN7, MSN2, MSN4, YAP1 and CAD1) (Gasch *et al*., 2000). CBF1 and UME6 were also reportedly involving in stress response by previous studies (Sweet *et al*., 1997; Yu and Li, 2005). Interestingly, HSF1, FKH1, NRG1 and PHD1 were hub TFs on both stress response and cell cycle networks. The heat shock related TF HSF1, in particular, regulated 170 and 155 genes in the two networks, respectively, in which 136 target genes were common between the two networks. Reportedly, HSF1 is activated in G1 of the cell cycle under non-stress conditions and may play a role in the G1 regulation that does not involve the transcription of heat shock genes in the yeast (Bruce *et al*., 1999).

4 SUMMARY

In this study, we present a novel methodology for unraveling transcriptional regulatory networks by integrated analysis of microarray, ChIP-chip data and TF motif information. The method is based on a two-stage constrained matrix decomposition model. The new method offers several advantages over previously published algorithms: (1) it takes into account the non-linear structure existed in the data, particularly in the TF-target gene interactions; (2) the model considers the combinatorial nature of gene regulation by TFs; (3) it predicts not only TF-target interactions, but furthermore the activating or inhibitory relationships; (4) the model does not assume the correlation between TFs and their target genes on the mRNA expression. We demonstrated the usefulness of the new method on the discovery of condition-specific regulatory networks in the yeast. While known transcriptional regulations were confirmed, novel TF-target interactions were predicted and provide new insights into the regulatory mechanisms of the cell.

ACKNOWLEDGEMENTS

This study is supported by the Intramural Research Program, National Institute on Aging, NIH.

Conflict of Interest: none declared.

REFERENCES

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Backes,C. *et al*. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–192.

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

Bernard,A. and Hartemink,A.J. (2005) Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac. Symp. Biocomput.*, 459–470.

Bruce,J.L. *et al*. (1999) Activation of heat shock transcription factor 1 to a DNA binding form during the G(1)phase of the cell cycle. *Cell Stress Chaperones*, **4**, 36–45.

Brunet,J.P. *et al*. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Chen,G. *et al*. (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Cheng,C. and Li,L.M. (2008) Systematic identification of cell cycle regulated transcription factors from microarray time series data. *BMC Genomics*, **9**, 116.

Dueck,D. *et al*. (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, **21** (Suppl. 1), i144–i151.

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

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

Gollub,J. *et al*. (2006) The Stanford Microarray Database: a user's guide. *Methods Mol. Biol.*, **338**, 191–208.

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

Haykin,S. (1999) *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ.

Herrgard,M.J. *et al*. (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res.*, **13**, 2423–2434.

Hughes,J.D. *et al*. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Hyvarinen,A. and Oja,E. (2000) Independent component analysis: algorithms and applications. *Neural Netw.*, **13**, 411–430.

Jordan,M.I. *et al*. (1999) An introduction to variational methods for graphical models. In Jordan,M.I. (ed.) *Learning in Graphical Models*. MIT Press, Cambridge.

Joung,J.G. *et al*. (2006) Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation. *Bioinformatics*, **22**, 2005–2011.

Jutten,C. and Karhunen,J. (2004) Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *Int. J. Neural Syst.*, **14**, 267–292.

Kim,H. *et al*. (2006) Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 165.

Lappalainen,H. and Honkela,A. (2000) Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Girolami,M. (ed.) *Advances in Independent Component Analysis*. Springer-Verlag, Berlin, pp. 93–121.

Lee,S.I. and Batzoglou,S. (2003) Application of independent component analysis to microarrays. *Genome Biol.*, **4**, R76.

Lemmens,K. *et al*. (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.*, **7**, R37.

Li,H. *et al*. (2007) The discovery of transcriptional modules by a two-stage matrix decomposition approach. *Bioinformatics*, **23**, 473–479.

Li,H. and Zhan,M. (2006) Systematic intervention of transcription for identifying network response to disease and cellular phenotypes. *Bioinformatics*, **22**, 96–102.

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

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.

Milo,R. *et al*. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Segal,E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

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.

Sweet,D.H. *et al*. (1997) Role of UME6 in transcriptional regulation of a DNA repair gene in Saccharomyces cerevisiae. *Mol. Cell Biol.*, **17**, 6223–6235.

Tamayo,P. *et al*. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tsai,H.K. *et al*. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl Acad. Sci. USA*, **102**, 13532–13537.

Yu,T. and Li,K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, **21**, 4033–4038.

Zhan,M. (2007) Deciphering modular and dynamic behaviors of transcriptional networks. *Genomic Med.*, **1**, 19–28.

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.