Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Inferring Regulatory Interactions between Transcriptional Factors and Genes by Propagating Known Regulatory Links

Qian Zhong*, Riccardo Boscolo*, Timothy S. Gardner†, Vwani P. Roychowdhury*

* Department of Electrical Engineering, University of California, Los Angeles, CA, USA

† Center for BioDynamics and Department of Biomedical Engineering, Boston University, Boston, MA, USA

Email: qianz@ee.ucla.edu, riccardo@ee.ucla.edu, tgardner@bu.edu, vwani@ee.ucla.edu

Abstract— **Determining transcriptional regulatory networks has been one of the most important goals in the field of functional genomics. Despite the recent advances in experimental techniques, complementary computational techniques have lagged behind. We introduce a novel computational methodology that uses DNA microarray data and known regulatory interactions to predict unknown regulatory interactions. Our method involves three steps: in the training stage, we utilize network component analysis (NCA) [1]–[3] to reconstruct the hidden activity profiles of transcriptional factors (TF); then we cluster TFs into functional modules according to the similarities of their reconstructed activity profiles; in the prediction stage, we infer additional TF-gene regulatory links by selecting TF profiles that best interpret genes expression profiles via a linear model. We applied the methodology to a gene expression dataset of bacterium Escherichia coli, whose partial TF-gene regulatory structure is obtained from RegulonDB [4]. Cross-validation results show that when the profiles of all TFs regulating a gene are reconstructed from NCA, we could identify 36% of the TF-gene interactions, and the prediction accuracy is 89%. And when the profiles of partial (50% or more) TFs regulating a gene can be reconstructed, we can identify 14% of the TF-gene interactions, and the accuracy rate is 69%. These represent some of the best known accuracy and coverage statistics reported in the literature so far.**

I. INTRODUCTION

Genes are regulated by transcription factors (TF), which assume an active conformation via post-transcriptional modification or ligand binding. TFs receive signals from sensor proteins and eventually positively or negatively regulate transcription in respondence to environmental changes. Determining transcriptional regulatory networks, i.e., a database of which TFs regulate which genes, has been one of the most important goals of the field of functional genomics.

Though experimental techniques, such as the genome wide location analysis of DNA binding regulators [5]–[8], identify TF-gene regulatory links, it's still important that computational tools could point a direction for the experiments, narrowing the experiment scale and reducing the cost. In general, computational tools utilize some *a priori* information about the TF-gene regulatory structure, build a model for it, and then predict additional unknown regulatory interactions. For example, certain computational studies focus on analyzing the sequence-specific TF binding sites (i.e., from *known* regulatory interactions) and discovering the motif patterns, and then perform a genome-wide scan to identify the target genes [9], [10] (i.e., the predictions are based on the stochastic modelling done in the first step). A major issue of these methods is how to control the amount of false positives in the prediction. Research was conducted regarding to better modelling the cis-regulatory elements [11], as well as more accurately computing p-values of putative binding sites [12]. Besides the sequence information, other researchers utilize the gene expression data, which are produced by high-throughput techniques such as DNA microarray [13], [14]. In [15], the authors extract the regulatory information using Bayesian statistics. And in [16], the authors design the experiments to measure the outputs of a targeted pathway in response to perturbations, and infer the causative input-output links based on a linearized model. In most cases, the environmental perturbations are so complex that multiple regulatory pathways are normally activated simultaneously. Therefore in order to discover system-level information regarding the underlying regulatory mechanisms, we need an appropriate approach to decompose the high-dimensional data.

In this paper, we propose a new method of modelling known regulatory interactions, and gene expression data via learning of linear models, as illustrated in Fig. 1. The purpose of the training part is to use the known links (representing regulatory interactions) and the final output (i.e., the gene expression data) to reconstruct the activities of a set of TF's that can be learned from the given data. Once the TF activity (TFA) patterns are determined, the regulatory interactions for the unknown genes can be determined via sparse regression, i.e., the best combination of TFAs, that best explains the expression profile of the given gene. The prediction results can be assessed by measuring both the false negatives and the false positives. We define the coverage rate as the number of true positives over the number of TF modules regulating this gene, and the accuracy rate as the number of true positives over the number of links we identified. Our results (see Section III), for example, are superior to what others have obtained previously in terms of both coverage rate and accuracy rate. In particular, authors of [17] combined microarray data

1-4244-0710-9/07/$20.00 ©2007 IEEE

with sequence information to infer the TF-gene regulatory network for Saccharomyces cerevisiae. They detected TF-encoding genes from the microarray data, and analyzed the relative over-abundance of cis-elements to identify TF-gene links. On average, they obtained a coverage rate at 15% and accuracy rate at 29% (see Table 2 in [17]). In addition, authors of [18] achieved a prediction accuracy rate as 63% by predicting target genes of a TF via support vector machine (SVM).

In order to build our linear regulatory model for estimating TFAs, we use the recently introduced Network Component Analysis (NCA). Traditional statistical techniques, e.g., principal component analysis (PCA) [19] and independent component analysis (ICA) [20], can successfully determine low-dimensional input signals of high-dimensional data set by imposing further constraint on the input. PCA requires the hidden input signals to be mutually orthogonal and ICA requires the signals to be statistically independent. Both constraints do not match the real biological regulatory system, therefore these methods are not suitable for deducing biologically significant information. Network component analysis (NCA) [1]–[3] resolves this issue by exploiting a certain type of priori knowledge about the connectivity pattern between the TF regulatory signals and the gene expression data, which can be obtained from experimental techniques as well as publicly available databases (e.g., RegulonDB [4], Ecocyc [21]).

With the reconstructed TF profiles from NCA, we infer the regulatory TF-gene interactions from a linear model. The inference procedure is centered around the regression problem of selecting a connectivity pattern that best fits the gene expression data. The methodology is specifically described in section II. In section III, we applied this methodology to a data set of bacterium Escherichia coli. 33 TFs profiles were reconstructed via NCA from three different subnetworks. Cross-validation results show that when all regulatory signals of a gene are reconstructed, our method could predict 36% of the TF-gene links, and the prediction accuracy is 89%; and in the case when only 50% or more of the regulatory signals can be reconstructed, we can still identify 14% of the TF-gene interactions at accuracy rate 69%.

II. Method

A complex biological control system can be approximated by a linear model including a set of L unknown input signals and a set of N measurable output signals as

$$e_n(t_m) = \sum_{l=1}^{L} a_{nl} p_l(t_m) + \gamma_n(t_m) \qquad (1)$$

$$n = 1, \cdots, N;\ m = 1, \cdots, M,$$

where $\gamma_n(t_n)$ is an error term representing both model bias and measurement noise, and a_{nl} can be considered as the control strength of the l_{th} input to the n_{th} output. Equation (1) can be visualized as a bi-partite network in Fig. 1. In the case of studying TF-gene regulation, this linear model (1) is derived from the Hill equation [1], where

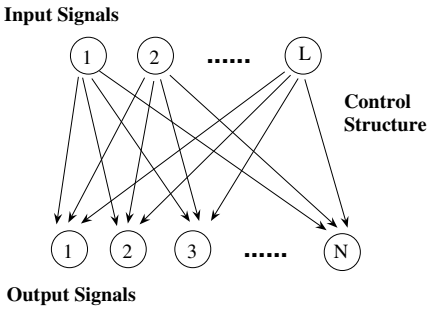

Fig. 1. A general biological control system is visualized as two layers: the hidden input signals and the measurable output signals. The links from the input to the output represents the control structure; the output is the gene-expression profiles of genes, and the input is the activity levels of the transcription factors. In the training phase, we use known regulatory interaction information (i.e., the links in the model) and gene-expression data (using the recently-introduced Network Component Analysis (NCA) procedure) to determine robust estimates of the transcription factors activities (TFAs) for a set of TFs. The estimated TFAs are then used in the prediction stage (via sparse regression) to estimate unknown regulatory interactions between the TFs and other genes (not included in the training set).

the output signals $\{e_1, \cdots, e_N\}$ can be considered as the logarithm of the expression data for N genes, and the input signals $\{p_1, \cdots, p_L\}$ represents the logarithm of the dynamic activities of L TFs.

Transcription regulators' targets typically vary in number depending on the structure of the pathways they involved into. Analogously, genes' regulatory sites are capable of binding to one or more regulator proteins. Because of the inherent system redundancy, often regulatory signals can be estimated even if partial information is available on such regulatory connectivity. Generally the problem of decomposing the output $e_1, \cdots, e_N$, into the combination of the regulatory inputs $p_1, \cdots, p_L$ through the coefficients of a regulatory matrix $[a_{nl}]$ results in an infinite number of solutions. Instead of imposing ad-hoc statistical constraints on the hidden regulatory signals, NCA [1]–[3] solves the problem by utilizing the a-priori knowledge on regulatory interactions, in order to constrain the solution set to biologically meaningful estimations.

We extended the traditional NCA framework to allow predictions of unknown regulatory interactions. Unlike traditional approaches, which are based on clustering possibly co-regulated genes based on their expression profiles, *our method relies on the reduced dimensionality of the estimated transcription factor activity profiles, in order to robustly predict possible additional targets of regulation*. The frameworks involves the following three separate steps:

I) NCA-based estimation of the activity profiles of a set of regulators of interest, based on known binding site connectivity data .
II) Clustering of the regulators activity profiles into groups that are functionally related .
III) Linear sparse regression of the expression profiles of genes with unknown regulators against the set of

clustered regulators' profiles .

Step I and II are based on NCA and standard hierarchical clustering, respectively. In step III, assuming that there are in total K clusters of TFs, our goal is to identify the most likely (in the Maximum Likelihood (ML) sense) candidate(s) in regulating the expression of a certain gene. The statistical significance of a set of predicted regulatory interactions is established by solving the following sparse regression problem:

$$\{j_1, j_2, \cdots, j_k\}_{opt} = arg\,min\|\vec{x}P_k - e_i\|^2, \qquad (2)$$

(where P_k is the reconstructed activities for the ${j_1}^{th}, {j_2}^{th}, \cdots, {j_k}^{th}$ cluster of TFs), and then by repeating the estimation multiple times while perturbing the system (typically with additive noise). We demonstrate (see section III for more details), that regulatory interactions that are consistently selected across multiple perturbations, reliably predict regulatory patterns.

III. RESULTS

In order to validate the accuracy of the proposed approach, we ran a standard cross-validation statistical test. The gene expression dataset of the bacterium Escherichia coli includes 448 microarrays under 190 conditions for 4345 genes, of which 445 microarrays for 189 conditions are the same dataset analyzed in [22], and the other 3 microarrays were assayed while treating E. coli MG1655 with serine hydroxamate at mid-log phase for 60 minutes. A subset of 941 genes and 126 transcriptional factors was selected since they have known regulatory interactions derived from the RegulonDB database. This set was further randomly subdivided into a training network including (627 genes with 123 known regulators) and a test network including the remaining 314 genes (regulated by 105 known TFs). The purpose of the experiment is to establish whether we are capable of estimating the regulatory interactions in the test network, once the activity profiles of the transcription factors common to both networks are estimated using NCA and used as predictors.

A. Estimating TF Profiles from NCA

In order to maximize the set of regulators whose activity can be estimated through NCA, we subdivided the training network into three subnetworks, with partially overlapping sets of genes and regulators (following the procedure described in [3]. By simultaneously running the estimation on all three sub-networks, we achieved a coverage of 341 genes and 33 unique TFs, whose activity profiles are shown in Figures 2 and 3. The figures show the results of 100 bootstraps, where in each bootstrap, Gaussian noise was added to the gene expression data in order to evaluate the overall convergency of the estimations. For each TF, all of its estimated profiles were pooled together and normalized. By doing a K-clustering (K is equal to 1 in this case) with squared Euclidean distance measure, we identified the center of the estimations and the distances from each estimation to the center.

B. Clustering TF Profiles

Some regulatory profiles (Fur and OxyR are a good example), were observed to be highly correlated. TFs assuming similar dynamic activities might participate in common pathways and co-regulate several genes. Therefore, the reconstructed regulatory signals were further clustered into functional modules via a hierarchical clustering algorithm. The distance measure was chosen to be the correlation coefficient in order to better capture the correlation among the signals. As displayed in Fig. 3, 20 functional modules were identified, with 5 of them comprising multiple TFs.

C. Predicting TF-gene Regulatory Interactions

Given the clustered TF modules, our goal is to identify the regulatory interactions between TF modules and genes. The a-priori knowledge on the connectivity information for E. coli suggests that constraining the sparse regression to three or less TFs (see Fig. 4a) provides enough coverage while reducing the statistical variability. Therefore, following equation (2), for each gene we identify the 3 out of 20 TF modules that best describe the expression data.

The process is repeated over 100 bootstrap iterations, where, in each iteration the system is systematically perturbed. For every gene tested, the frequency with which each module is selected is recorded during the bootstraps. Fig. 4b shows a sample frequency plot for gene appA. Modules with higher selection frequency are the more likely candidates for a regulatory interaction. For example, in Fig. 4b, if the frequency threshold is set to be 0.8, only module 6 (TF AppY) is identified as a candidate regulator for this gene. Given the frequency statistics for each gene and TF module, we can establish the accuracy and coverage of the prediction by setting an arbitrary threshold and counting the number of false positives and false negatives.

In the test network, 70 of the genes are known to be regulated exclusively by the 33 TFs whose profiles were estimated from the training network, while another 30 genes are available whose regulators set is at least 50% covered in the training network. By selecting different frequency thresholds, we obtained a curve of the average coverage rate per gene and a curve of the average accuracy rate per gene (shown in Fig. 4c-d). In general, the coverage rate curve decreases and the accuracy rate curve increases as the chosen frequency threshold goes up.

When we selected a threshold of 0.98, we achieved an average coverage rate of 36% with an accuracy rate of 89% when all the regulatory signals of a gene are available from the NCA estimation. On the other hand, when only 50% or more of the regulatory activity profiles are available, the average coverage rate drops to 14% for an accuracy rate of 69%.

IV. DISCUSSION

As the amount of large-scale gene expression data obtained from high-throughput biological techniques, such as DNA microarray, increase rapidly, deciphering the complex

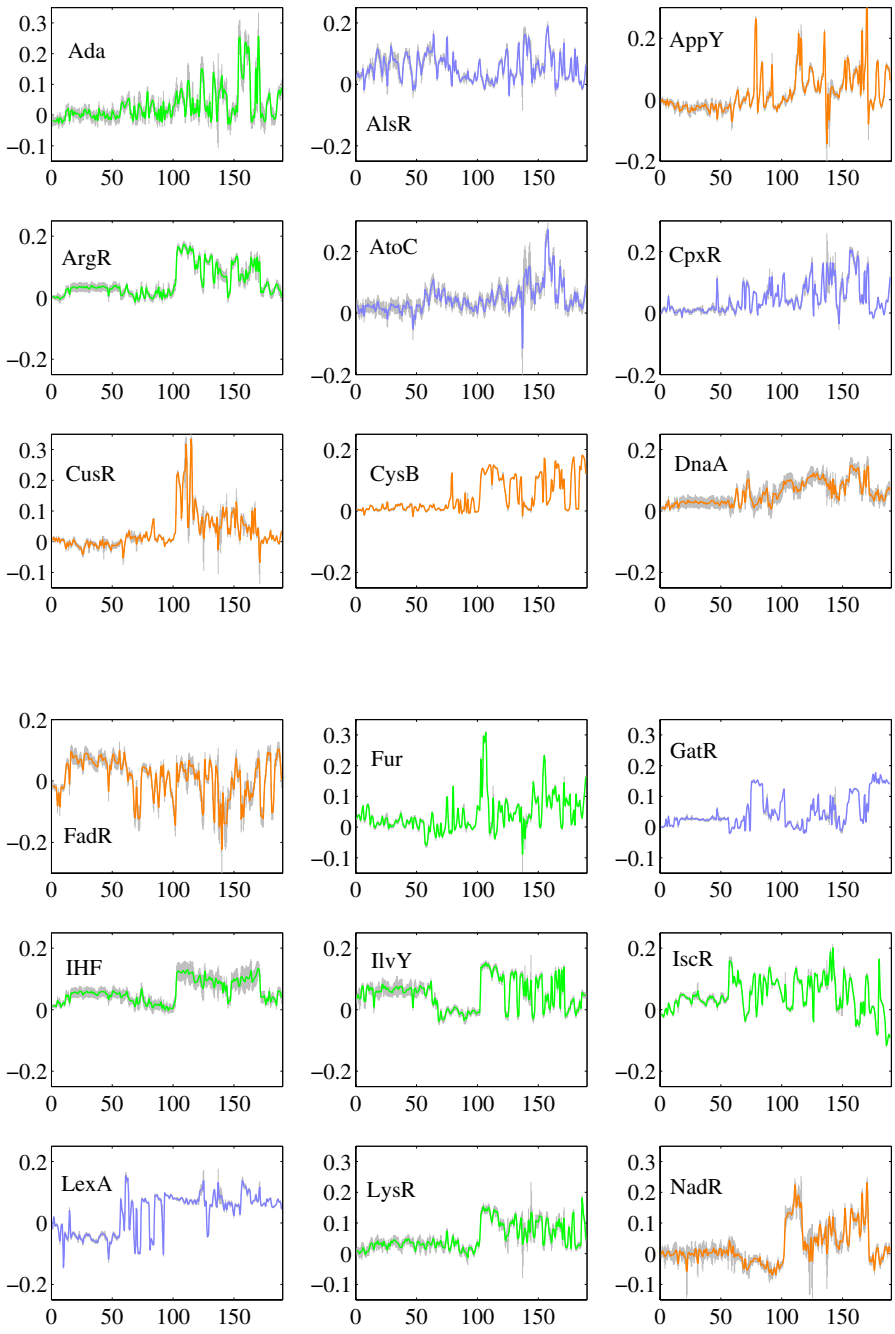

Fig. 2. Transcription factor activities of the bacterium E. coli, estimated from three regulatory subnetworks via NCA. Each color (orange, blue, or green) is associated with the number of subnetworks comprising this transcriptional factor (one, two or three respectively). For each TF, all estimations are normalized so that the norm is 1. The solid curve is the mean of the estimations, and the shaded area represents the standard deviations of the estimations.

physiological regulatory system with the aid of proper mathematical and statistical tools, becomes of particular interest. Unlike methods such as PCA and ICA, NCA doesn't make any assumption regarding the statistical properties of the regulatory signals. Rather, it provides simple means for incorporating the a-priori regulatory network structure and reconstructing hidden regulatory signals.

We extended NCA's framework to inferring potential

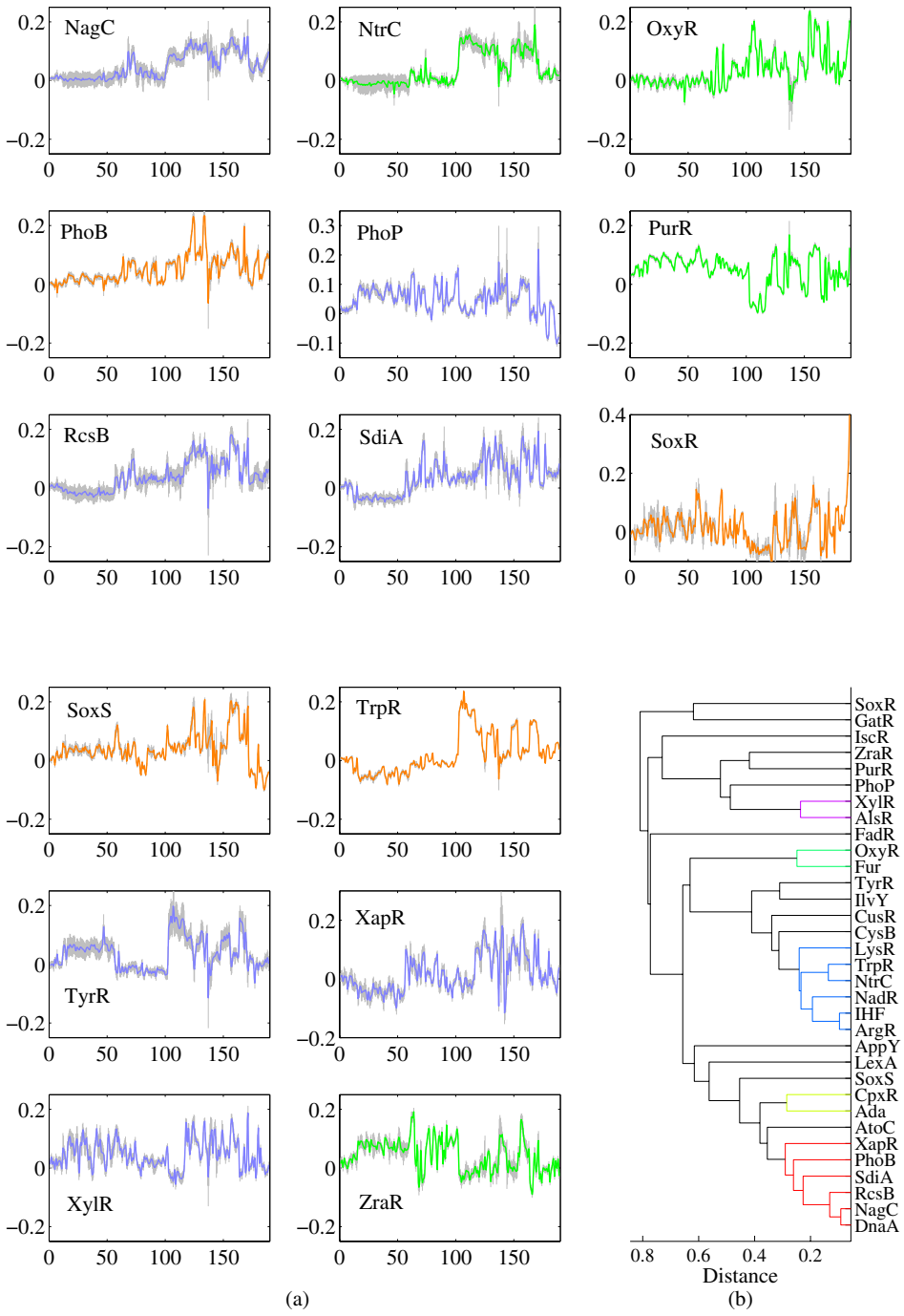

Fig. 3. a: Transcription factor activities of the bacterium E. coli, estimated from three regulatory subnetworks via NCA. Each color (orange, blue, or green) is associated with the number of subnetworks comprising this transcriptional factor (one, two or three respectively). For each TF, all estimations are normalized so that the norm is 1. The solid curve is the mean of the estimations, and the shaded area represents the standard deviations of the estimations. b: The clustering dendrogram for the 33 transcriptional factors. Transcriptional factors were clustered according to the correlation coefficient of their reconstructed activities. With the distance threshold chosen at 0.3, 20 modules are classified.

TF-gene links. By exploiting the reconstructed TF activity profiles, we identify candidate TF regulation targets by identifying those TF functional modules that best model a gene expression profile. The methodology was applied to a gene expression dataset of bacterium Escherichia coli, whose partial regulatory structure was obtained from RegulonDB.

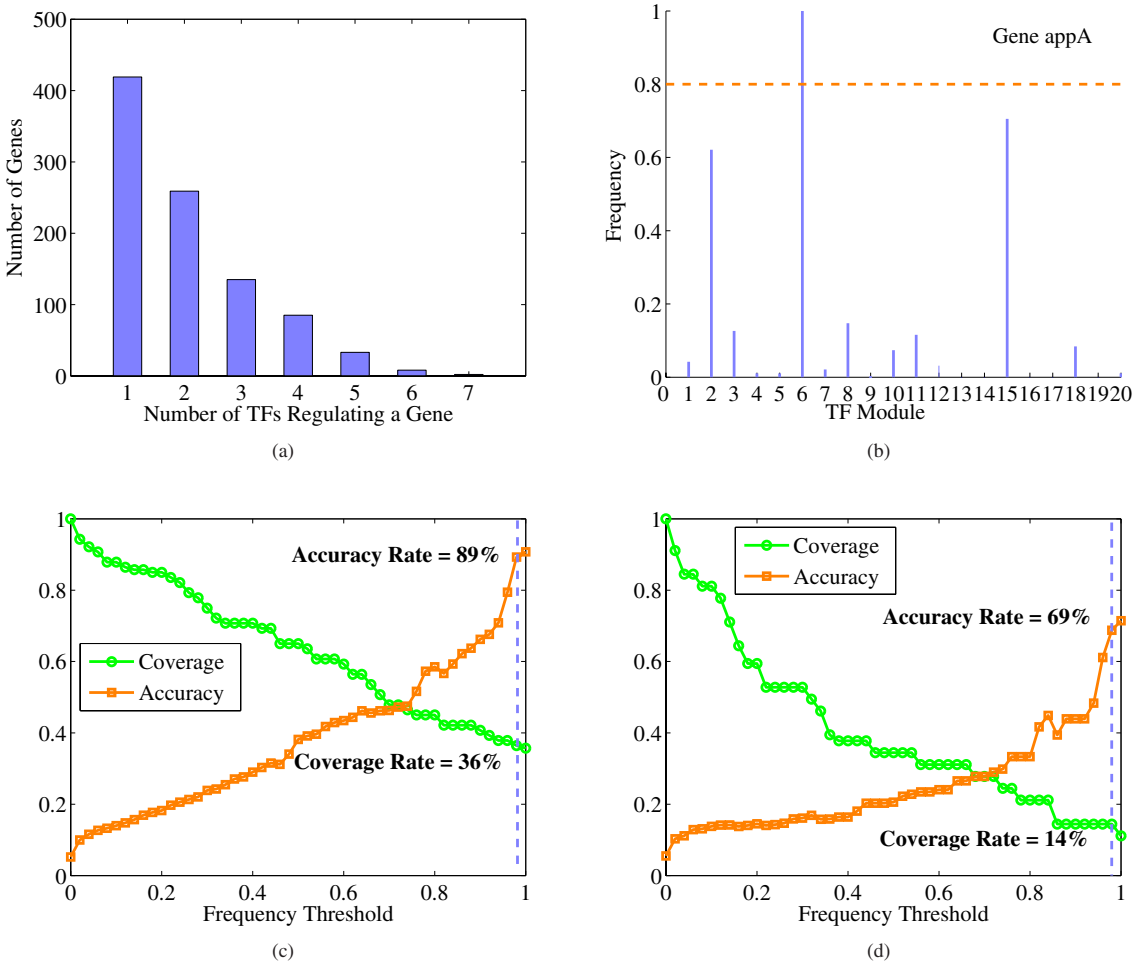

Fig. 4. a: Histogram of the genes that are regulated by different number of transcriptional factors in bacterium Escherichia coli. The regulatory information is obtained from RegulonDB, which includes 941 genes and 126 transcriptional factors. b: The frequency of each module being selected for testing gene appA. When the frequency threshold is set to be 0.80, only module 6 (TF AppY) is identified to regulate this gene. c: The coverage rate curve and accuracy rate curve for 70 testing genes that are only regulated by the 33 TFs. The coverage rate and accuracy rate are labelled in the figure when the frequency threshold is set to be 0.98. d: The coverage rate curve and accuracy rate curve for 30 testing genes that are partially regulated by the 33 TFs. The coverage rate and accuracy rate are labelled in the figure when the frequency threshold is set to be 0.98.

Cross-validation results show that when all TF signals regulating a gene are available, 36% of the TF-gene interactions can be identified, with a 89% prediction accuracy. However, when only a minimum of 50% of the regulators are covered, 14% of the TF-gene interactions are correctly identified, with a 69% prediction accuracy.

The performance of our methodology depends on the availability of the reconstructed TF profiles, which primarily depends on the availability of the priori TF-gene regulatory information utilized by NCA. As experiments reveal more about the underlying network, our framework is expected to be more frequently applied in the situation where all TF profiles regulating a gene are available from NCA estimation.

ACKNOWLEDGMENT

This work was supported in part by the NSF grant ITR 0326605.

REFERENCES

[1] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 15 522–15 527, 2003.

[2] K. C. Kao, Y.-L. Yang, R. Boscolo, C. Sabatti, V. P. Roychowdhury, and J. C. Liao, "Transcriptome-based determination of multiple transcription regulator activities in escherichia coli by using network component analysis," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 641–646, 2004.

[3] R. Boscolo, C. Sabatti, J. C. Liao, and V. P. Roychowdhury, "A generalized framework for network component analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, pp. 289–301, 2005.

[4] H. Salgado, S. Gama-Castro, A. Martínez-Antonio, E. Díaz-Peredo, F. Sánchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jiménez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martínez, and J. Collado-Vides, "Regulondb (version 4.0): transcriptional regulation, operon organization and growth conditions in escherichia coli k-12." *Nucleic Acids Research*, vol. 32, no. Database-Issue, pp. 303–306, 2004.

[5] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J.

Wilson, S. P. Bell, and R. A. Young, "Genome-wide location and function of dna binding proteins," *Science*, vol. 290, pp. 2306–2309, 2000.

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

[7] J. Zeitlinger, I. Simon, C. T. Harbison, N. M. Hannett, T. L. Volkert, G. R. Fink, and R. A. Young, "Program-specific distribution of a transcription factor dependent on partner transcription factor and mapk signaling," *Cell*, vol. 113, pp. 395–404, 2003.

[8] e. a. C. T. Harbison, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99–104, 2004.

[9] G. D. Stormo and G. W. Hartzell, "Identifying protein-binding sites from unaligned dna fragments," *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 1183–1187, 1989.

[10] C. E. Lawrence and A. A. Reilly, "An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, pp. 41–51, 1990.

[11] Q. Zhou and J. S. Liu, "Modelling within-motif dependence for transcription factor binding site predictions," *Bioinformatics*, vol. 20, pp. 909–916, 2004.

[12] Y. Barash, G. Elidan, T. Kaplan, and N. Friedman, "Cis: compound importance sampling method for protein-dna binding site p-value estimation," *Bioinformatics.*, vol. 21, pp. 596–600, 2005.

[13] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, pp. 467–470, 1995.

[14] A. Campbell and L. Heyer, *Discovering Genomics, Proteomics, and Bioinformatics*. Benjamin/Cummings, 2002.

[15] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, pp. 166–176, 2003.

[16] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, pp. 102–105, 2003.

[17] P. M. Haverty and a. Z. W. U. Hansen, "Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification," *Nucleic Acids Research*, vol. 32, pp. 179–188, 2004.

[18] J. Qian, J. Lin, N. M. Luscombe, H. Yu, and M. Gerstein, "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data," *Bioinformatics*, vol. 19(15), pp. 1917–1926, 2003.

[19] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series," *Pac. Symp. Biocomput.*, vol. 5, pp. 455–466, 2000.

[20] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51–60, 2002.

[21] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp, "Ecocyc: a comprehensive database resource for escherichia coli," *Nucleic Acids Research*, vol. 33, pp. D334–D337, 2005.

[22] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale computational mapping and experimental validation of escherichia coli transcriptional regulatory interactions from a compendium of expression profiles," *PLoS Biology*, 2006.