Journal of Bioinformatics and Computational Biology
Vol. 10, No. 5 (2012) 1250012 (20 pages)

DOI: 10.1142/S0219720012500126

Imperial College Press
www.icpress.co.uk

INFERRING THE REGULATORY INTERACTION MODELS OF TRANSCRIPTION FACTORS IN TRANSCRIPTIONAL REGULATORY NETWORKS

SHERINE AWAD[*,¶], NICHOLAS PANCHY[†,‖], SEE-KIONG NG[‡,**] and JIN CHEN[§,††]

[*]*Computer Science and Engineering Department*
Michigan State University, East Lansing, MI 48824, USA

[†]*Genetics Program, Michigan State University*
East Lansing, MI 48824, USA

[‡]*Institute for Infocomm Research, 1 Fusionopolis Way, Singapore*

[§]*MSU-DOE Plant Research Laboratory*
Computer Science and Engineering Department
Michigan State University, East Lansing, MI 48824, USA
[¶]*mahmoud4@msu.edu*
[‖]*panchyni@msu.edu*
[**]*skng@i2r.a-star.edu.sg*
[††]*jinchen@msu.edu*

Received 2 May 2012
Accepted 4 May 2012
Published

Living cells are realized by complex gene expression programs that are moderated by regulatory proteins called transcription factors (TFs). The TFs control the differential expression of target genes in the context of transcriptional regulatory networks (TRNs), either individually or in groups. Deciphering the mechanisms of how the TFs control the differential expression of a target gene in a TRN is challenging, especially when multiple TFs collaboratively participate in the transcriptional regulation. To unravel the roles of the TFs in the regulatory networks, we model the underlying regulatory interactions in terms of the TF−target interactions' directions (activation or repression) and their corresponding logical roles (necessary and/or sufficient). We design a set of constraints that relate gene expression patterns to regulatory interaction models, and develop TRIM (Transcriptional Regulatory Interaction Model Inference), a new hidden Markov model, to infer the models of TF−target interactions in large-scale TRNs of complex organisms. Besides, by training TRIM with wild-type time-series gene expression data, the activation timepoints of each regulatory module can be obtained. To demonstrate the advantages of TRIM, we applied it on yeast TRN to infer the TF−target interaction models for individual TFs as well as pairs of TFs in collaborative regulatory modules. By comparing with TF knockout and other gene expression data, we were able to show that the performance of TRIM is clearly higher than DREM (the best existing algorithm). In addition, on an individual Arabidopsis binding network, we showed that the target genes' expression correlations can be significantly improved by incorporating the TF−target regulatory interaction models inferred

[††] Corresponding author.

by TRIM into the expression data analysis, which may introduce new knowledge in transcriptional dynamics and bioactivation.

Keywords: Transcriptional regulation; hidden Markov model.

1. Introduction

The complex gene expression programs in living cells are moderated by regulatory proteins called transcription factors (TFs) that control the transcription of genes in the context of transcriptional regulatory networks (TRNs).[1] The TFs interact with their target genes in the TRNs to upregulate or downregulate gene expression. They can act independently or collaboratively with other TFs, leading to different *TF−target interaction models* that influence the regulation patterns of target genes in different ways.[2,3]

Various experimental methods have been developed to unravel the complex regulatory mechanisms behind biological processes. Recent developments in biotechnology (e.g. chromatin immunoprecipitation, yeast one-hybrid, and next-generation sequencing) have been used to indirectly or directly uncover TF binding relationships[4,5] to reconstruct draft regulatory circuits at a systems level.[2,3,6] To verify the TF−target gene relationships and to detect the TF functions *in vivo*, TF knockouts and/or overexpression experiments are usually carried out.[7]

However, single knockout or overexpression may not provide statistically significant evidence due to redundancy or confounding signals from indirect regulatory feedback.[8] For example, it has been shown that approximately 73% (about 4500) of the known genes of *S. cerevisiae* (yeast) are nonessential.[9] The results of the single knockout or overexpression experiments are therefore often nonconclusive, as it is highly likely that multiple nonessential genes can be involved. This has led to the development of automated experimental methods for double-knockouts to provide more statistically significant determination of the TF functions.[10] However, to systematically knockout (or overexpress) all possible combinations of the TFs in the whole genome is still challenging. Given an organism with k TFs, the total number of possible double-TF combinations is $k * (k - 1)/2$. For complex organisms, k can be easily in the range of thousands.

Instead of blindly trying out all possible TF pairs for double-knockout experiments, one solution is to select the TF pairs that are most likely to bring about the phenotypic change. To do so, we need to understand the interaction models employed by the TFs to influence the regulatory patterns of the target genes in the network. In other words, we need to uncover the models of TF−target regulatory interactions of the TF pairs and the target gene in terms of the TF−target interactions' directions (activation or repression) and their corresponding logical roles (necessary and/or sufficient). We design a set of constraints that relate gene expression patterns to regulatory interaction models, and propose an algorithm TRIM (Transcriptional Regulatory Interaction Model Inference) to systemically infer the regulatory interaction models between individual TFs, as well as any two

TFs, and their target genes, from wild-type time-series gene expression data. Our TRIM algorithm is based on a hidden Markov model (HMM). Experimental results on yeast data showed that TRIM outperformed the existing algorithms for inferring the regulatory interaction models of TFs and their target genes for individual TFs as well as pairs of TFs that are in collaborative regulatory modules. In addition, on an individual Arabidopsis binding network, we showed that the target genes' expression correlations can be significantly improved by incorporating the TF−target regulatory interaction models inferred by TRIM into the expression data analysis, which may introduce new knowledge in transcriptional dynamics and bioactivation.

2. Background

We define a regulatory module $R(TF, G, I)$ as a set of genes G regulated in concert by a group of one or more TFs that govern the target genes' behaviors via appropriate TF−target regulatory interactions in I.[11] There are two types of regulatory modules. In an *independent regulatory module*, the target genes are solely regulated by one TF. In a *collaborative regulatory module*, the target genes are regulated by multiple TFs.

A TF's regulatory interaction model can be defined in terms of two properties: the TF's functional role as an activator or a repressor, and its logical role as being necessary or sufficient. Table 1 shows such a regulatory interaction model as proposed by previous researchers.[3,11] Note that the categories in the TF's functional and logical roles can be combined. A TF−target interaction model in R can be Activator Necessary (AN), Activator Sufficient (AS), or Activator Necessary and Sufficient (ANS). Similarly for TFs that are repressors, they can be RN, RS, or RNS.[12]

Yeang and Jaakkola[3] attempted to characterize the combinatorial regulatory models of multiple TF−target interactions using a heuristic approach to measure how

Table 1. TF−target interactions can be modeled in terms of the TF's functional role as an activator (upregulates the target gene's expression) or a repressor (downregulates the target gene's expression), and the logical role of the TF as being necessary and/or sufficient. The two categories (functional and logical roles) can be combined.

Role	Concept	Description
Functional	Activator	Response of target gene is inline with the expression change of TF
	Repressor	Response of target gene is opposite to the expression change of TF
Direction of Logical	Necessary	Decreasing TF's expression level leads to responses opposite to its functional role
	Sufficient	Increasing TF's expression level leads to the responses consistent with its functional role
	Necessary and Sufficient	Increasing TF's expression level leads to responses consistent with its functional role, and decreasing the TF's expression level leads to the responses opposite to its functional role

well R fits the associated binding and gene expression data with a log-likelihood function. The regulatory module's likelihood is maximized with a greedy approach by incrementally adding genes to the module and monitoring the predictions of the TF−target interactions for optimality. However, this incremental approach does not study the functions of the TFs simultaneously because of the scalability issue introduced by the greedy search. Their method also uses a p-value−based approach to calculate the significance of the combinatorial property of a TF, determined by the gap of log likelihood scores between their model and a model built on the randomized gene expression data based on the entire timeframe. However, as stated in Ernst *et al.*,[2] a TF usually functions at specific "activation timepoints" instead of throughout the entire timeframe. This means that the identification of TF−target interaction modules should be focused on such activation timepoints rather than comparing with random gene expression data of the entire timeframe. In this work, we include a step to recognize the activation timepoints of the target genes in TRIM. Our experimental results will show that the target genes' expression correlation are indeed markedly improved by taking the actual activation timepoints into consideration.

Another related algorithm is DREM, which was proposed to derive dynamic regulatory networks that associate TFs with target genes and their activation timepoints.[2] To uncover transcriptional regulatory events leading to the observed temporal expression patterns and the underlying factors that control these events during a cell's response to stimuli, DREM integrates time-series gene expression data and protein−DNA binding data to build a global temporal map. The method mainly works by identifying bifurcation timepoints where the expression of a subset of genes diverges from the rest of the genes. The bifurcation points are then annotated, with the TFs regulating these transitions, which result in a unified temporal map. The method can therefore facilitate the determination of the time when TFs are exerting their influence, and assigns genes to paths in the map based on their expression profiles and the TFs that control them. Unlike the method by Yeang and Jaakkola,[3] DREM's ability to derive dynamic maps that associate TFs with the genes they regulate and their activation timepoints has indeed led to better insights for the regulatory module being studied. For example, one can identify master regulators that control the initial response and secondary regulators that are responsible for specific pathways. Numerous aspects of the observed response, including the condition-specific activity of factors and the activation of certain network motifs, can also be explained using DREM. However, unlike our method, DREM does not infer the logical roles of the TFs; for example, whether a specific master or secondary TF is necessary or sufficient for regulating a set of target genes. Such knowledge is essentially useful for understanding the complex regulatory mechanisms of many biological processes. We will show in this paper that by incorporating the knowledge of the regulatory interaction model, we can significantly improve the computation of gene expression correlation.

The kernel of our TRIM method is an HMM, which is a stochastic model that assumes the Markov property holds and all the states are unobserved (hidden).

A stochastic process is said to have the Markov property if the conditional probability distribution of its future states depends only upon the current state, as shown in Eq. (1):

$$p(x(t+1) \mid x(0), \ldots, x(t)) = p(x(t+1)|x(t)) \tag{1}$$

An HMM consists of a set of hidden states and each state has a probability distribution over the possible outputs.[13] In an HMM, a state k_i transits to another state k_j with a probability $P(k_j(t+1)|k_i(t))$. A state k can emit an output b with emission probability $e_k(b) = P(output = b|state = k)$.[13–15] In this work, each state refers to a possible TF−target interaction model, while the output emitted by a state will indicate whether a particular interaction model is true.[14]

By taking advantage of the HMM, our TRIM algorithm can consider the influence from multiple TFs *simultaneously*. Prior biological knowledge on the TFs, such as gene perturbation experiments, can also be incorporated in the setting of the initial emission probabilities (see Sec. 3.2), while the time-dependent regulatory relations can be effectively captured by preserving the time dependency within the HMM.

In this work, we assume that a TF−target interaction is consistent in the context of transcriptional control as long as the experimental conditions are unchanged. We also assume that the activity of a TF is proportional to its mRNA abundance over time. Although these assumptions may be violated in practice, existing algorithms for inferring TF−target interaction models at different levels of complexity[2,3,11,16] have all been developed with these assumptions. As such, we also make the same assumptions for TRIM in this current work, and we will leave the development of methods that consider the possible violations of such common assumptions for future work.

3. Methods

Given a large-scale TRN, we design our TRIM algorithm for inferring TF−target interaction models systematically from large-scale TRNs. A TRN can be represented as a directed graph, in which each node is a TF or a gene, and each edge represents a regulation relationship between a TF and a target gene. The framework of TRIM, as shown in Fig. 1, consists of two steps: the first step constructs the regulatory modules from the TRN; the second step infers the TF−target interaction models in each of the regulatory modules.

Figure 3 shows that the genes in yeast, one of the most well-studied eukaryotic organisms, are regulated mostly by individual TFs (69.6%) or pairs of TFs (18.2%). Therefore, in this paper, we focus on independent regulatory modules and collaborative regulatory modules with two TFs (we call these "2-TF collaborative modules"). Due to the model complexity, We will leave the development of the rules for inferring TF−target models for the TFs that are involved in more complicated regulatory modules with three or more TFs for future work.

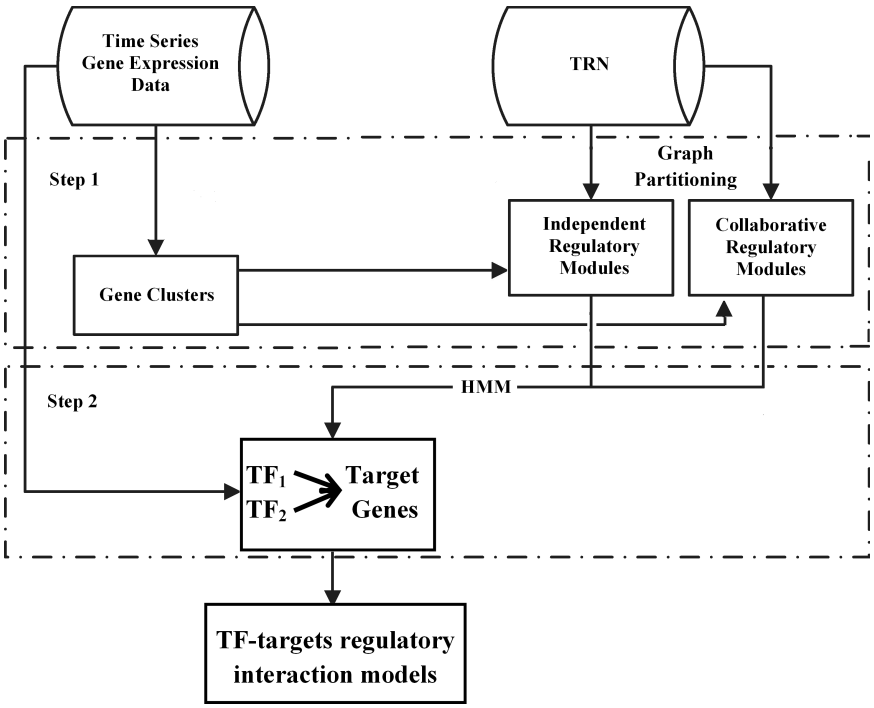

Fig. 1. TRIM framework for inferring TF−target regulatory interaction models. TRIM has two main steps: (1) a regulatory module construction step that includes both a clustering of co-expressed genes and a topological analysis to classify independent and collaborative regulatory modules from a given large-scale TRN, and (2) a HMM model to infer TF−target interaction models in the independent and collaborative regulatory modules identified in step 1.

3.1. *Constructing regulatory modules*

Given a large-scale TRN, a TF may regulate multiple target genes simultaneously but with different types of TF−target interactions. To construct regulatory modules, we extract the subnetworks using gene expression clustering followed by graph partitioning (Fig. 1, Step 1).

We first cluster all the target genes in a TRN based on their gene expression values with Cluster 3.0 (specifically, k-means), which uses Pearson correlation coefficient for gene similarity metric.[17] The clusters are then evaluated with Gene Ontology enrichment analysis using Bingo, and unenriched clusters are discarded.[18] Genes in the same cluster are considered to be co-expressed. And the co-expressed and co-regulated genes are usually weighed to be regulated by the same TF(s) with a similar interaction model. So for the target genes that are regulated by the same single TF (or the same TF pair), we partition them based on whether they are in the same cluster to construct independent regulatory modules (or collaborative regulatory modules). An illustrative example is shown in Fig. S1: TF_1 and TF_2 regulate

genes g_1 and g_2, and g_1 and g_2 belong to the same gene expression cluster, so this regulatory module contains TF_1, TF_2, g_1, and g_2.

3.2. *Designing the HMM model*

In the next step (Fig. 1, step 2), we design a new HMM model[15] to infer the regulatory interaction models for the TF−target interactions in every regulatory module detected above. For the regulatory modules with two TFs, we run the HMM model directly. For the regulatory modules with a single TF, we add a dummy TF with its expression value constantly zero. Some researchers have pointed out that designing an HMM model is a sort of art.[14] In the following text, we describe the details of how we design the structure, set the initial probabilities, and develop the updating method for emission probabilities and transition probabilities for our HMM for 2-TF collaborative regulatory modules.

Structure. To model 2-TF collaborative modules, our HMM consists of two TFs, TF_1 and TF_2, where each TF has four states (i.e., AS, AN, RS, and RN), as shown in Fig. 2. Each state emits two possible outputs, active or inactive. One can view a state as a representation of whether a particular regulatory interaction model for an *individual* TF−target interaction is valid (active) or invalid (inactive). In the training process, if one of the four states of TF_1 emits an active output, the HMM

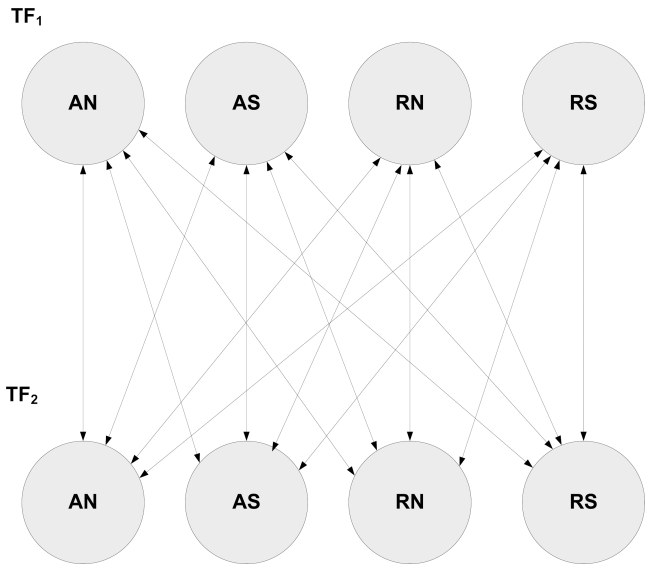

Fig. 2. A simplified representation of the HMM model for the collaborative regulatory module with 2-TF, TF_1, and TF_2, where each TF has four states (i.e. AS, AN, RS, and RN). Each state emits two possible outputs, active or inactive. One can view a state as representing whether a specific interaction model of an individual TF−target relation is valid or invalid. In the training process, if an active output is emitted from one of the four states of TF_1, the model will transit to a state belonging to TF_2 and *vice versa*.

will transit to a state belonging to TF_2 based on the constrains (to be introduced later), and *vice versa*.

Initial Probability. The initial emission probabilities of the states in the HMM are set equally (if we do not have any prior information), or they can be determined by prior knowledge obtained from TF perturbation experiments. In addition, we use uniform transition probabilities (1/32), if there is an arrow in Fig. 2.

Emission and Transition Probabilities. We train the HMM model with discretized gene expression data. We discretized the gene expression levels as upregulation or downregulation instead using absolute absence or presence, by comparing the expression changes between two consecutive timepoints to determine whether there was increase (+1)/decrease (−1).[19] For the regulatory module with multiple target genes, all the gene expression changes are used sequentially for HMM training.

In an mRNA transcriptional process, the expression change of a TF is usually earlier than the change of its targets. Therefore, we adopt the concept of time lagging in this HMM model training process.[20] At timepoint n, the input to the HMM is a set of gene expression changes of the TFs from timepoint $n - l$ to n, where l is a time lag that is defined by the user to capture the effects of the TFs at the earlier timepoints $(n - l, \ldots, n)$ on the target gene at timepoint n.

To update the emission probabilities properly, we design a set of constraints that relate the gene expression patterns of a TF and its target to the regulatory interaction model (see Table 2). The emission probability of a state (active/non-active) can then be updated with Eq. (2), where b and b' are the outputs of state k and $E_k(b')$ is the probability for emitting output b' from state k.[14] The final emission probability of each state reflects the likelihood of the state being active or inactive.

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \tag{2}$$

In an HMM, a path is a sequence of states that follows the Markov chain of hidden states, in which the probability of a state depends only on the probability of the previous state. In this way, we are able to consider the regulatory effects of the two TFs simultaneously, unlike previous works. The Viterbi algorithm is applied to effectively find the most probable path.[13,15] The final path contains two states with a state for each TF's TF−target interaction model. For example, a path "AS-RN" means that the first TF is activator sufficient and the second TF is repressor necessary for the same target genes.

Note that the HMM is a probabilistic model which cannot assume combined events or none of the events to occur. Therefore, the model described above does not include a state for "Neither" or "Both Necessary and Sufficient (N+S)". To infer neither/N+S regulatory models, a post-processing step is required. We use the distribution of coefficient of variation (CV) of all the emission probabilities to determine whether a regulatory interaction model can be N+S or neither: if none of the probabilities are significant, the model outputs "neither"; if the probabilities of both

Table 2. Constrains for the inferring TF−target interaction models in a 2-TF collaborative regulatory module. They are based on the TF's expression direction and the response of the target gene.

TF_1	TF_2	Target-gene expression	TF−target interaction model
Up	Down	Up	TF_1 is Activator Sufficient or TF_2 is Repressor Necessary
Up	Down	Down	TF_1 is Repressor Sufficient or TF_2 is Activator Necessary
Up	Up	Up	At least one TF is Activator Sufficient
Down	Down	Up	At least one TF is Repressor Necessary
Up	Up	Down	At least one TF is Repressor Sufficient
Down	Down	Down	At least one TF is Activator Necessary
Up	—	Up	TF_1 is Activator Sufficient
Up	—	Down	TF_1 is Repressor Sufficient
Down	—	Up	TF_1 is Repressor Necessary
Down	—	Down	TF_1 is Activator Necessary

Table 3. An illustrative example of HMM training process.

Time	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
TF_1	0	0	1	1	1	1	0	1	1	1
TF_2	0	−1	0	−1	0	1	−1	0	−1	0
g	0	−1	−1	−1	1	−1	0	1	0	−1

Note: There are 10 observations of gene expression changes for a collaborative regulatory module with 2-TF: 0 means no significant gene expression change; 1 means significant upregulation and −1 means significant downregulation.

N and S states are significant, and if there is a significant difference between the probabilities of the two states, TRIM outputs the more significant state, otherwise our model outputs N+S.

We show an illustrative example of our HMM in Table 3. In this example, TF_1 and TF_2 regulate the same target gene g. For simplicity, no time lag is used in the example ($l = 0$). Given the gene expression changes of the two TFs and the target gene, we can infer the regulatory interaction models for both of the TF−target pairs using the HMM as follows. We first initialize the active emission probabilities of all the states equally to 0.5. At time t_0, as none of the expression changes is significant, nothing is done. At time t_1, the downregulation of both TF_2 and g triggers the active emission probability of state AN of TF_2 (Table 2, Row 10). The active emission probability of state (TF_2, AN) is updated by adding the new frequency and then being normalized i.e. $(0.5 + 1)/2 = 0.75$, while the inactive emission probability of (TF_2, AN) is 0.25. Meanwhile, the active emission probabilities of all other states are decreased to 0.25 and the inactive emission probabilities of them are increased to 0.75. See Table 4 for the updated emissions. At time t_2, the upregulation of TF_1 and the downregulation of g trigger the state RS of TF_1 (Table 2, Row 8). The active emission probability of (TF_1, RS) is updated to be 0.625 and the other active emission probabilities are adjusted accordingly. At time t_3, state RS of TF_1 and state AN of TF_2 are triggered given the two-TF constraint in Table 2, Row 2. The active emission probability of

Table 4. The emission probabilities of the first four observations of gene expression changes in the illustrative example shown in Table 3.

	AS	AN	RS	RN	AS	AN	RS	RN
			t_0				t_1	
TF_1	0.50	0.50	0.50	0.50	0.25	0.25	0.25	0.25
TF_2	0.50	0.50	0.50	0.50	0.25	0.75	0.25	0.25
			t_2				t_3	
TF_1	0.125	0.125	0.625	0.125	0.062	0.062	**0.813**	0.062
TF_2	0.125	0.375	0.125	0.125	0.062	**0.688**	0.062	0.062

Note: After processing these four observations, the model suggests that the most possible regulatory interaction model for (TF_1, g) is RS with probability 81% and for (TF_2, g) is AN with probability 68%.

(TF_2, AN) is updated to $(0.375 + 1)/2 = 0.688$ and the active emission probability of (TF_1, RS) to $(0.625 + 1)/2 = 0.813$. See Table 4 for the emission probabilities of the four observations. The training process continues until all the gene expression observations are processed. For this example, the RS's active emission probability (0.675) is the highest amongst all the active probabilities of TF_1 and the RS's active emission probability (0.203) is the highest amongst all the active probabilities of TF_2 in the end. The final output for the TF−target interaction model is subject to the distribution of all the active emission probabilities in all the regulatory modules. In this example, we conclude that the most possible regulatory interaction model for (TF_1, g) is repressor sufficient (RS) and for (TF_2, g) is Neither.

4. Experimental Results

We evaluate the performance of TRIM using the yeast regulatory network. For comparison, we applied DREM v3.0 on the same dataset. We did not compare TRIM with Yeang and Jaakkola[3] since their objective is to build a reliable TRN, and the method does not output the TF function, which is the focus of our work.

4.1. *Data preparation*

Using the same statistical approach in Reimand *et al.*,[6] we generated a large-scale yeast regulatory network with 2230 regulatory relations between 268 TFs and 1509 target genes by filtering the yeast ChIP-chip binding data[21] and the binding-cite predictions.[22,23] The in-degree distribution (number of TFs per gene) in Fig. 3 shows that 87.8% of the target genes are regulated by one or two TFs, indicating that studying independent and 2-TF collaborative regulatory models is sufficient to cover the majority of the yeast TRN.

To train and evaluate TRIM, four widely used time-series microarray datasets from yeast cell cycle studies were collected.[24] These datasets contain 73 timepoints in total. In these experiments, yeast cells were first synchronized to the same cell cycle stage, released from synchronization, and then the total RNA samples were taken at even intervals for a period of time (see details in Table S1).

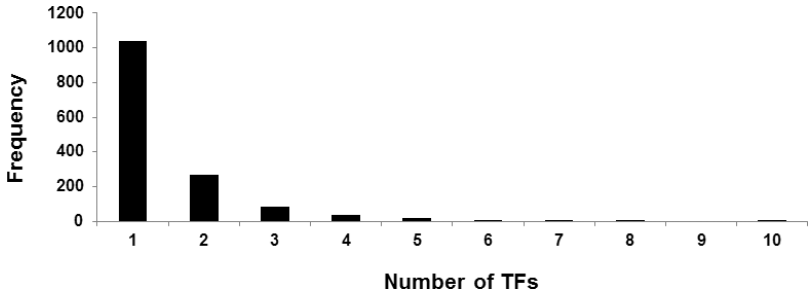

Fig. 3. In-degree distribution of the yeast TRN.

To decide whether a gene is significantly up or down regulated, we used a gene expression change cutoff at 0.35. In this experiment, we used three of them (CDC15, CDC28, and elu) as the training data and one (Alpha) as the testing data, since the Alpha dataset contains the most available gene expression values after applying the cutoff. A time lag $l = 2$ was used in this experiment.

For evaluating the inferred regulatory interaction models for the TF−target interactions of the independent regulatory modules, single TF knockout microarray data for yeast were collected.[9] A p-value cut-off at 0.05 was applied to determine whether a gene is significantly affected by a TF knockout.[9]

4.2. *Evaluation criteria*

To test whether our method can correctly predict the TF−targets interaction models, we adopt a similar approach as used in Segal *et al.*[11] to examine the distribution of the differentially expressed genes in the modules. With the regulatory interaction models learned from training data for the TF−target interactions in a 2-TF collaborative regulatory module, we can obtain the active timepoints on the testing data at which a TF functions. If the inferred regulatory interaction model is correct and the TF has the consistent function on the training and the testing data, then the gene expression correlations on the active timepoints should be significantly higher than on the whole timeframe of the testing data.

Mathematically, the set of activation timepoints T_x is defined as follows: let e_i be the gene expression change (1, −1 or 0) of a TF at timepoint t_i; for each TF and its TF−target interaction model x, timepoint t_i is in T_x if and only if $Indicator(t_i) = 1$ or $Indicator(t_{i+1}) = 1$:

$$Indicator(t_i) = \begin{cases} 1 & \text{if } x = \text{sufficient and } e_i = 1, \text{ or} \\ & x = \text{necessary and } e_i = -1, \text{ or} \\ & x = (\text{necessary and sufficient}) \text{ and } |e_i| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

To compute the gene expression correlation, we group together the target genes regulated by the same TF with the same interaction model. We then compute the

gene expression correlation on each group of the target genes over their activation timepoints, and normalize them. Mathematically, given a set of genes G and a set of timepoints T, if TF TF_k is predicted to be necessary for genes in G_n, sufficient for genes in G_s and both necessary and sufficient for genes in G_b, then the correlation score of all the genes regulated by TF_k is computed with Eq. (4).

$$COR(TF_k) = \frac{\sum_{g_i,g_j \in G_n} C(g_i, g_j, T_n) + \sum_{g_i,g_j \in G_n} C(g_i, g_j, T_s) + \sum_{g_i,g_j \in G_n} C(g_i, g_j, T_b)}{|G_n \cup G_s \cup G_b|} \tag{4}$$

where $G_x \subseteq G$ and $T_x \subseteq T$; $C(g_i, g_j, T_x)$ is the Pearson correlation score between g_i and g_j ($i \neq j$) at all the activation timepoints in T_x (x can be n, s, or b).

For comparison, we applied the same approach on DREM prediction results. Since DREM did not predict the effectiveness of TFs as necessary or sufficient, we grouped the target genes regulated by a TF based on the TF's dependence (activation or repression). The Pearson correlation of the same genes (grouped by TRIM prediction results) on all the timepoints were also computed.

4.3. *Evaluation of TRIM on yeast data*

In the yeast regulatory network, the 1509 target genes were grouped into 50 clusters with Cluster 3.0.[17] We then partitioned the yeast regulatory network into 1051 independent regulatory modules and 275 collaborative regulatory modules with two TFs. Finally, by combining the regulatory modules accordingly (see Sec. 3.1, the total number of the independent regulatory modules is reduced from 1051 to 640, while the number of 2-TF collaborative regulatory modules is reduced from 275 to 257. In the 640 independent regulatory modules, there are a total of 1051 TF−target pairs. TRIM was able to infer the TF−target interaction models for 833 pairs [see Table 5(a)]. On the same dataset, DREM was able to infer 873 TF−target relations with its p-value cut-off at 0.05.

We used the single TF knockout microarray data to directly verify our inferred TF−target interaction models for the independent regulatory modules. First, for the functional roles (activation or repression) of TFs, our TRIM successfully predicted 549 out of 815 models with a successful rate of 67.4%, which is clearly higher than the results of DREM (56.0%), as shown in Fig. 4. Second, for evaluating the inference performance of the logical roles of TFs, with knockout data, we can only look at the "necessary" logical role: a TF is necessary for its target genes if the expression values of the target genes are significantly changed when the TF is knocked out. In the single TF knockout data, a total of 815 independent regulatory modules were considered as necessary. Among them, TRIM predicted 682 pairs to be necessary or necessary and sufficient with a success rate of 83.6%. (We are unable to perform a similar comparison for DREM since DREM does not predict necessary TFs).

In addition, the gene expression correlation scores of the TF−target pairs that are predicted by both TRIM and DREM are shown in Fig. 5(a) (see Sec. 4.2 for the score

Table 5. Summary of TRIM's predictions for independent and two TFs regulatory modules on yeast and Arabidopsis.

	Activator	Repressor	Unknown
(A) Independent regulatory modules of yeast			
Necessary	37	35	3
Sufficient	68	50	17
Necessary & Sufficient	228	179	216
Neither		218	
(B) 2-TF collaborative regulatory modules of yeast			
Necessary	74	61	0
Sufficient	121	61	0
Necessary & Sufficient	6	5	0
Neither		212	
(C) 2-TF collaborative regulatory modules of of Arabidopsis			
Necessary	180	136	0
Sufficient	126	101	0
Necessary & Sufficient	6	0	0
Neither		87	

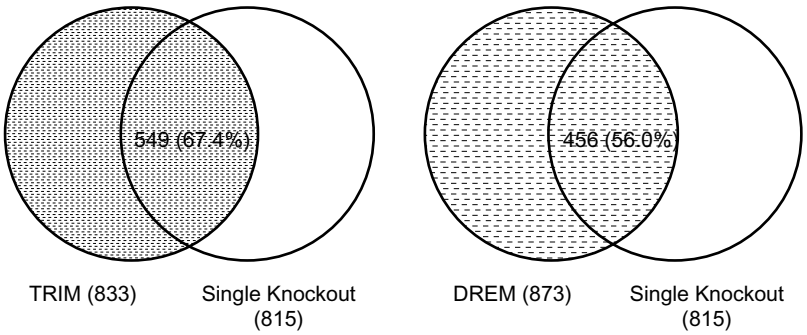

Fig. 4. In the independent regulatory modules, the overlap between the single TF knockout supported TF−target pairs (815) and TRIM prediction results (833) for the functional roles (activation or repression) is 549, greater than the overlap between knockout and the results of DREM (873), which is 456.

computation). The median correlation score for TRIM (0.52) is clearly higher than that of DREM (0.46) and than using all the timepoints (0.29). The detailed scores for individual TFs are in Fig. S2. To further evaluate the performance of TRIM on independent modules, we estimated the mRNA synthesis rate of each gene by applying decay adjustments[25] on the training and testing data:

$$\mu_g = e_g(\alpha + \lambda_g) \tag{5}$$

where μ_g is the synthesis rate of gene g; e_g is the gene expression of g; α is a consistent variable representing the growth rate of yeast; λ_g is the decay rate of each gene g,

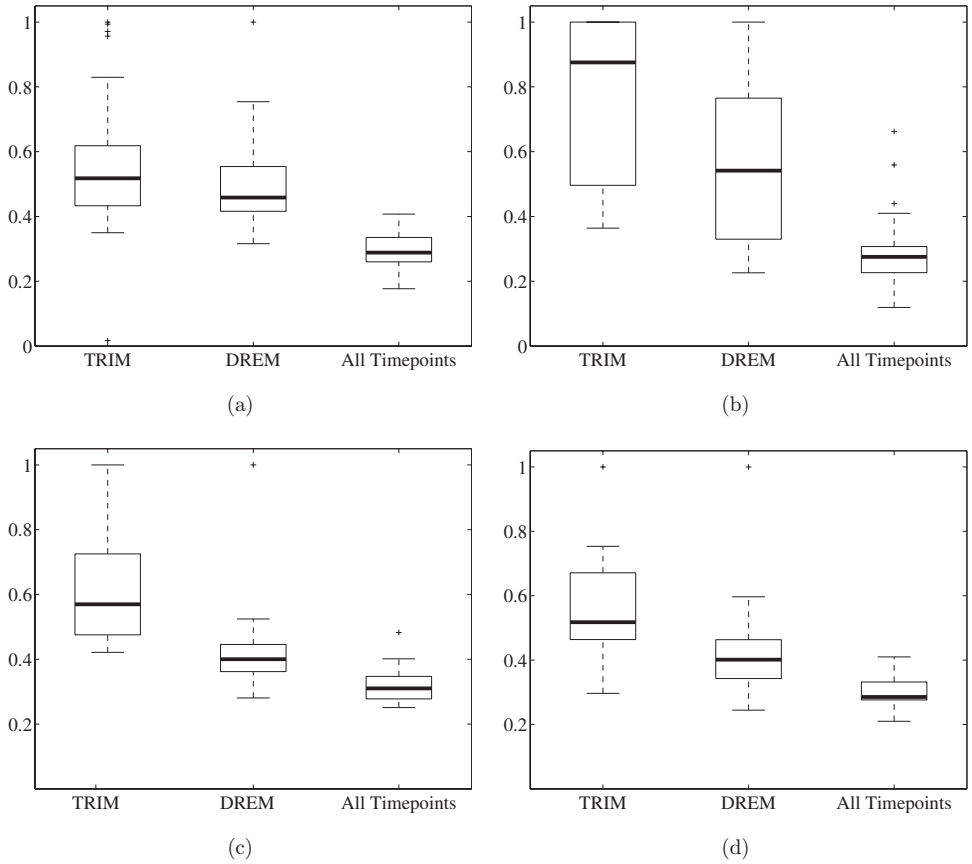

Fig. 5. Distribution of gene expression correlation of (a) the independent regulatory modules, (b) 2-TF collaborative regulatory modules, (c) and the independent regulatory modules after decay adjustment, (d) the 2-TF collaborative regulatory modules after decay adjustment.

which could be estimated with the mRNA half-life (L_g) for each gene with equation $\lambda_g = ln(2)/L_g$, where the mRNA half-life data was obtained from Ref. 25. For the growth rate of yeast (α), we used the same value as used in Ref. 25, i.e., $ln(2)/cell\_cycle\_length$ and the cell cycle length is set to 150 min. After applying the decay adjustments, the median of the expression correlation scores of TRIM further increased from 0.52 to 0.57, while the scores of DREM and all-timepoints are almost unchanged [0.40 and 0.31, respectively, see Fig. 5(c)], indicating that by focusing on the mRNA synthesis rates, TRIM model is able to predict the TF−target interaction models more precisely.

In the 540 2TF-target pairs in the 257 2-TF collaborative regulatory modules, TRIM was able to infer the interaction models for 328 pairs. Table 5B shows the summary of the number of interaction models predicted by TRIM for 2-TF modules. On the same dataset, DREM was able to infer 440 TF−target pairs. Figure 5(b)

shows the expression correlation scores of the 2TF-target pairs that are predicted by both TRIM and DREM. The median correlation score for TRIM (0.87) is significantly higher than that of DREM (0.54) and than using all the timepoints (0.27) (the detailed correlation values are shown in Fig. S3). We also applied similar decay estimation approach on the 2-TF collaborative modules. The result in Fig. 5(d) indicates the median correlation score after decay adjustment for TRIM is still higher than that of DREM. The overall performance, however, was not improved as significantly as for the single TF modules, probably because the decay model should be applied on the target genes only. This problem will be addressed in our future work.

In summary, the experimental results on the yeast data showed that TRIM can successfully predict TF−target interaction models for both independent and 2-TF collaborative modules.

5. Application of TRIM on Arabidopsis Data

In order to maintain a stable intracellular environment, living cells utilize complex and specialized transcriptional regulatory systems to react against a variety of external perturbations, such as temperature change, drought, UV, etc. Many of the adaptive mechanisms contributing to cellular homeostasis operate through TRNs to regulate the expression of anti-stress genes.[26] The key step to understand the TRN behavior is therefore to explore the roles of relevant TFs by using time-series gene expression data under different stress conditions. In this study, we applied TRIM to infer the roles (i.e. interaction models) of TFs in 2-TF regulatory modules of Arabidopsis TRN under eight different abiotic stress conditions. The objective of this study is to infer the TF−target interaction models in Arabidopsis TRN with all the available abiotic stress data, so that by looking at the gene expression patterns, we are able to tell whether a TF is a general abiotic stress TF, a specific abiotic stress TF, or a TF that does not function under abiotic stress. In addition, since our TRIM (like all the other current algorithms) assumes that TF−target interaction models will remain consistent under different abiotic stress conditions, we also need to verify whether this assumption holds for the various biological functions of the target genes.

We obtained Arabidopsis regulatory data from AtRegNet, which contains 11,355 direct interactions between TFs and target genes.[27] In AtRegNet, a direct interaction means either a TF binds directly to the target gene (detected by electromobility shift assay, yeast one-hybrid, or ChIP), or a TF directly regulates the target gene based on use of transgenic plants expressing an inducible TF-GR fusion protein. We partitioned the Arabidopsis genes in AtRegNet into 50 clusters with Cluster 3.0. A total of eight abiotic stress data sets (Cold, Drought, Genotoxic, Heat, Osmotic, Salt, UVB, and Wounding) were collected from AtGenExpress to train and test our TRIM's performance in Arabidopsis.[28] The datasets were filtered for significant variability in mRNA expression using Bonferroni corrected p-values at 0.05. For each experimental run, three of the datasets were reserved to evaluate the output of the model, while the remaining five were used for model training. In the experiment, six runs were

conducted each with a different combination of training and testing data (see detail in Table S2). A time lag $l = 2$ was used in this experiment.

Note that in combining data from different abiotic stress experiments, we have assumed that the TF−target interaction models will remain consistent under different abiotic stress conditions. This is also a common assumption by previous researchers.[28,29] In this study, we compare the results of the TRIM model across multiple combinations of training and testing data. If our hypothesis is true, the interaction model and subsequent correlation scores given by our model should be consistent across all groups and the results would not be sensitive to how we grouped the data. Otherwise, it means that the TF−target interaction model may vary under different conditions. In this study, we found that the hypothesis is true for selected classes of TFs e.g. developmental TFs.

On average, our TRIM identified 318 2-TF collaborative regulatory modules involving 32 TFs (the prediction results is shown in Table 5C). Ten of these TFs had sufficient data to be analyzed by gene expression correlation. Detailed results are shown in Table S3. In all, we found that the developmental TFs were the most consistent in terms of prediction and scoring, though their correlation scores were not always the highest on average. The only exception to this rule, APETALA2, is involved in seed development and thus, we would expect it to be more differentially expressed, since the production of seeds involves a reproductive decision which is likely to be stress sensitive. Other TFs, such as specific stress response genes and those associated with photosynthetic processes, are less consistent.

We would also expect general stress response factors to show very consistent predictions and score highly with the given data, yet they are absent from our dataset. However, we missed these TFs because they are involved in more complicated regulatory modules with three or more TFs and extending TRIM to k-TF modules, as our future work, would capture their behavior.

In addition to grouping these TFs by function, we analyzed three individual TFs (ATGL1, WRKY53, and RD26) in depth. We show their Pearson correlation distributions in Fig. 6. The TF with the highest average correlation score, ATGL1, a developmental gene involved in trichome pattering, is an example of a case where our hypothesis held, as $\sim 80\%$ of the interactions that could be predicted were either AN or RN across all six experiments. The correlation scores were the highest when the variable modules were predicted as either AN or RN. Alternatively, the TFs with the two lowest average scores, WRKY53 and RD26, biotic- and drought-stress-associated regulators, showed very inconsistent prediction patterns. The highest Pearson correlation scores of RD26 occur when drought was included in the evaluation set. This is confirmed by the fact that there is little difference in the correlation whether we apply TRIM or use all timepoints to calculate the score. Hence, RD26 is an example of condition-sensitive TF which violates our assumption. WRKY53, on the other hand, has low and similar correlation scores whether we use activation timepoints predicted by TRIM or all the timepoints, which is true across all six runs. Therefore, the issue here is most likely not sensitivity to particular grouping of data,

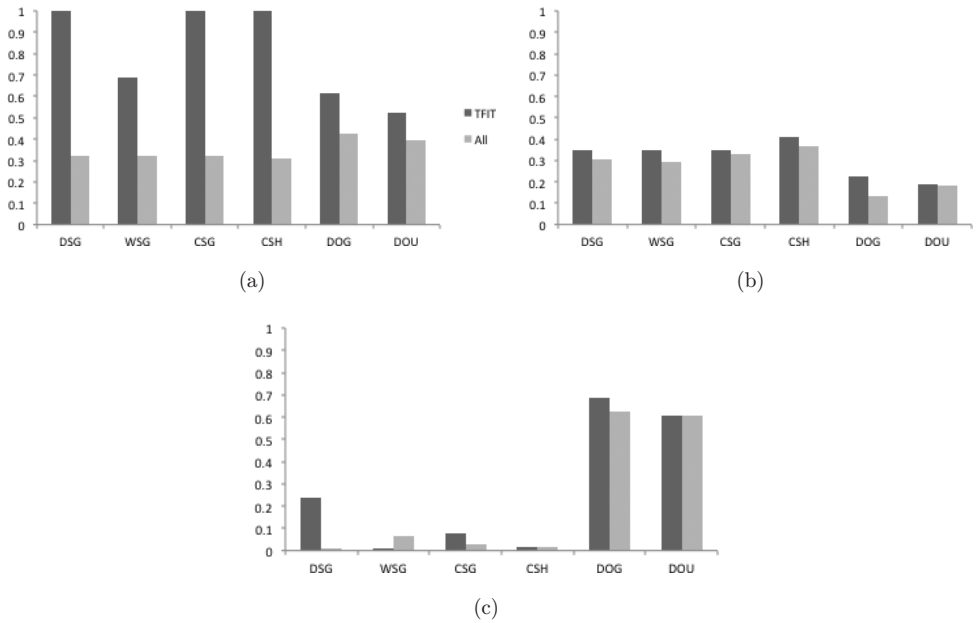

Fig. 6. Pearson correlation of target genes of (a) ATGL1, (b) WRKY53 and (c) RD26 using different combinations of training and testing data with TRIM.

but rather that the data is uninformative for this particular case. This conclusion is not unexpected given WRKY53 responds primarily to biotic rather than abiotic stress.

6. Conclusions

Revealing the mechanisms of the transcriptional regulatory programs in TRNs is essential for understanding the complex control by which genes are expressed in living cells. In this work, we model the interactions between the TFs and the target genes in terms of both the TF−target interaction's function (activation or repression) and its corresponding logical role (necessary and/or sufficient). Based on the characterizations proposed by Yeang and Jaakkola,[3] we define the combinatorial regulatory interaction models for possibly multiple TF−target interactions in TRNs.

We used DNA−protein binding and gene expression data to construct regulatory modules for inferring the transcriptional regulatory interaction models for the TFs and their corresponding target genes. Our TRIM algorithm is based on an HMM and a set of constraints that relate gene expression patterns to regulatory interaction models. In this work, we have shown how to apply TRIM to infer the transcriptional regulatory interaction models for TFs in collaborative regulatory modules involving two TFs. It can thus be used to help predict the phenotype of TF double-knockouts to reduce the number of double knock-out or overexpression experiments needed.

Our TRIM has the following advantages. First, TRIM is able to use only the wild-type gene expression data, which are available in many organisms. Second, prior knowledge of TFs can be accommodated in our model as follows: if the gene perturbation experiments for certain TFs are available, we can initialize the emission probabilities in the HMM model using the prior information. Third, rather than simply counting the hits that matches the constraints, our TRIM is able to capture the temporal regulation relations by preserving the time dependency with HMM. It helps capture the real regulatory behavior of the TFs on their target genes.

Our experimental results showed that TRIM is able to achieve a high true positives rate for inferring the necessary TF−target regulatory interactions for the independent regulatory modules. For the 2-TF collaborative regulatory modules, TRIM also resulted in consistently higher expression correlations of the co-regulated genes. Its application on Arabidopsis TRN reveals the functions of 10 key TFs in abiotic stress response pathways, showing that the inferred knowledge of TRIM can provide useful biological insights. For future work, we plan to extend the current 2-TF model into k-TF model for inferring TF−target models for the TFs that are involved in more complicated regulatory modules with three or more TFs.

References

1. Qiu P, Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, *Biochem Biophys Res Commun* **309**: 495−501, 2003.
2. Ernst J, Vainas O, Harbison C, Simon I, Bar-Joseph Z, Reconstructing dynamic regulatory maps, *Mol Sys Biol* **3**(74): 1−13, 2007.
3. Yeang H, Jaakkola T, Modeling the combinatorial functions of multiple transcription factors, *J Comput Biol* **13**: 463−480, 2006.
4. Deplancke B, Mukhopadhyay A, Ao W, Elewa M, Grove A, Martinez J, Sequerra R, Doucette-Stamm L, Reece-Hoyes S, Hope A, Tissenbaum A, Mango E, Walhout M, A gene-centered *C. elegans* protein−DNA interaction network, *Cell* **125**: 1193−1205, 2006.
5. Ren B, Robert F, Wyrick J, Aparicio O, Jennings E, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert T, Wilson J, Genome-wide location and function of DNA binding proteins, *Science* **290**: 2306−2309, 2000.
6. Reimand J, Vaquerizas J, Todd A, Vilo J, Luscombe N, Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets, *Nucleic Acid Res* **38**: 4768−4777, 2010.
7. Hoth S, Morgante M, Sanchez JP, Hanafey MK, Tingey SV, Chua NH, Genome-wide gene expression profiling in arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant, *J Cell Sci* **115**: 4891−4900, 2006.
8. Honkela A, Girardot C, Gustafson E, Liu Y, Furlongb E, Lawrence N, Rattray M, Model-based method for transcription factor target identification with limited data, *Proc Nat Acad Sci* **107**: 7793−7798, 2010.
9. Killion P, Hu Z, Iyer V, Genetic reconstruction of a functional transcriptional regulatory network, *Nature Genetics* **39**: 683−687, 2007.
10. Tong A, Boone C, Synthetic genetic array analysis in *Saccharomyces cerevisiae*, *Meth Mol Biol* **313**: 171−191, 2006.

11. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N, Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data, *Nature Genetics* **34**: 166−167, 2003.
12. Babur O, Demir E, Gonen M, Sander C, Dogrusoz U, Discovering modulators of gene expression, *Nucleic Acid Res* **38**: 5648−5656, 2010.
13. Duda R, Hart P, Stork D, Pattern classification, 2001.
14. Durbin R, Eddy S, Krogh A, Mitchison G, Biological sequence analysis: Probabilistic models of proteins nucleic acids, 1998.
15. Mukherjee S, Mitra S, Hidden markov models, grammars, and biology: A tutorial, *J Bioinform Comput Biol* **3**: 491−526, 2005.
16. Bar-Joseph Z, Gerber G, Lee T, Rinaldi N, Yoo J, Robert F, Gordon B, Fraenkel E, Jaakkola T, Young R, Gifford K, Computational discovery of gene modules and regulatory networks, *Nature Biotechnol* **21**: 1337−1342, 2003.
17. Eisen M, Spellman P, Brown P, Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci* **95**: 14863−14868, 1998.
18. Maere S, Heymans K, Kuiper M, Bingo: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, *Bioinformatics* **21**: 3448−3449, 2005.
19. Ong I, Glasner J, Page D, Modelling regulatory pathways in *e. coli* from time-series expression profiles, *J Bioinform* **18**: S241−S248, 2002.
20. Schmitt WA, Raab RM, Stephanopoulos G, Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data, *Genome Res* **14**(8): 1654−1663, 2004.
21. Lee T, Rinaldi N, Robert F, Odom D, Bar-Joseph Z, Gerber G, Hannett N, Harbison C, Thompson C, Simon I, Zeitlinger J, Jennings E, Murray H, Gordon B, Ren B, Wyrick J, Tagne J, Volkert T, Fraenkel E, Gifford D, Young R, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science* **298**: 799−804, 2002.
22. Erb I, Nimwegen E, Statistical features of yeast's transcriptional regulatory code, *Proc Int Conf Comput Sys Biol* **1**: 111−118, 2006.
23. MacIsaac K, Wang T, Gordon B, Gifford D, Stormo G, Fraenkel E, An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*, *BMC Bioinform* **7**: 113, 2006.
24. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B, Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell* **9**: 3273−3297, 1998.
25. Miller C, Schwalb B, Maier K, Schulz D, Dumcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dolken L, Martin DE, Tresch A, Cramer P, Dynamic transcriptome analysis measures rates of mrna synthesis and decay in yeast, *Mol Sys Biol* (458): 1−13, 2010.
26. Zhang Q, Andersen ME, Dose response relationship in anti-stress gene regulatory networks, *PLoS Comput Biol* **3**: e24, 2007.
27. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E, Agris and atregnet: A platform to link cis-regulatory elements and transcription factors into regulatory networks, *Plant Physiol* **140**: 818−829, 2006.
28. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K, The atgenexpress global stress expression data set: Protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses, *Plant J* **50**: 347−363, 2007.
29. Kreps JA, Wu Y, Chang HS, Zhu T, Wang X, Harper JF, Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress, *Plant Physiol* **130**: 2129−2141, 2002.

Sherine Awad is a Ph.D. candidate in Computer Science and Engineering Department at Michigan State University. Her main area of research concerns inferring gene regulatory network using machine learning and data mining techniques. She received her B.S. in Computer Science from Suez Canal University in 2002 and M.S. in Information Systems from Helwan University in 2006.

Nicholas Panchy is a graduate student in Genetics Program at Michigan State University. He is primarily interested in the application of computational modeling to discerning gene function and expression. He received his B.S. in Biology and B.S. in Mathematics from University of North Carolina at Chapel Hill in 2011.

See-Kiong Ng is currently the Program Director and Advisor to the Data Mining Department of the Institute for Infocomm Research at the Agency of Science, Technology and Research (A*STAR) of Singapore. The department houses more than 50 international researchers and engineers, focusing on research and development in machine learning and data mining. See-Kiong obtained both his B.S. and Ph. D. in Computer Science from Carnegie Mellon University, with an M.S. from University of Pennsylvania. His primary research is in data mining and machine learning, with applications in text mining, bioinformatics, privacy-preserving data mining, and social network mining.

Jin Chen is an Assistant Professor in MSU-DOE Plant Research Laboratory, Computer Science and Engineering Department at Michigan State University. He is interested in constructing the plant bioenergy network with machine-learning methods to better understand the energy conversion systems. He obtained his B.E. in Computer Science from Southeast University in 1997 and Ph.D. degree in Computer Science from National University of Singapore in 2007.