JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 4, 2008

Pp. 393–405
DOI: 10.1089/cmb.2008.0020

A Simple Model of the Modular Structure of Transcriptional Regulation in Yeast

VLADIMIR FILKOV[1] and NAMEETA SHAH[1]

ABSTRACT

Resolving the general organizational principles that govern the interactions during transcriptional gene regulation has great relevance for understanding disease progression, biofabrication, and biological systems in general. The available genome-level monitoring technologies and the best understood biological work on gene regulation are together providing us with unprecedented amounts of data and universal modeling frameworks in which to reason about regulatory systems on a computational level. Gene regulatory systems exhibit modularity in their regulatory sequences as well as in the corresponding gene expression. This modularity has a nontrivial, general combinatorial structure that can be studied and generalized to model classes of regulatory systems. Here, we study computationally the combinatorial nature of transcriptional regulation by assuming a one-to-one relationship between shared patterns in genome-wide gene-expression and cis-region modules. In our combinatorial framework, the DNA binding events are complementary to their expression counterparts, and together let us approximate the underlying regulation structure. Our model maps regulatory systems onto hierarchical structures which can be approximated by conflating existing large scale gene expression and ChIP-chip data. We have developed methods for building regulatory hierarchies and identifying the basic functional units, or modules, of transcriptional regulation. We validate our model using yeast data by showing agreement of our predictions with experimental data, and using the hierarchies to resolve a finer structure of co-regulation.

Key words: cis-regulatory modules, gene regulation modeling, regulation hierarchies.

1. INTRODUCTION

IN THIS POST-GENOMIC ERA, a critical challenge is to understand how genetic components interact to control (i.e., regulate) gene activity. Resolving the general principles behind such interactions would have wide-ranging implications. From a basic science perspective, it would allow us to better understand the nature of transcriptional regulation on a fundamental level. More practically, it could influence our interaction and manipulation of nature in different applications. Biofabrication is becoming possible as we understand better the biological systems and are able to engineer them. Knowing the rules in the

[1]Computer Science Department, University of California, Davis, California.

organization of regulatory systems will enable us to build better, task-specific genetic circuits. As many diseases have a genetic cause, knowing the pathways through which these diseases develop and progress can help us understand the diseases and, by modifying those pathways, even modify the disease itself. Understanding the regulatory control of genes that cause such diseases is key to our ability to do those things.

Computational data analysis methods are particularly important in genomics because large-scale technologies produce increasing amounts of data that necessitate the use of computers and sophisticated algorithms to provide meaningful information to users. The available genome-level monitoring technologies provide us with the means to observe complete biological systems. They also allow us to identify the elements of transcription and their interactions from underlying repeating patterns shared by regulatory DNA or expression data across the genome. On the other hand, the best understood biological work on gene regulation have revealed that gene regulatory regions consist of modules, each incorporating one or more binding sites. Such modules have been shown to have precise effects on the resulting gene expression and to cooperate through specific rules, or logic. Together, the unprecedented amounts of data and the biological principles of regulation modularity present us with universal modeling frameworks in which to reason about regulatory systems on a computational level.

Inspired by the best understood biological gene regulatory systems and the available large-scale genomic data, here we study computationally the modular nature of transcriptional regulation by modeling the structure of shared patterns in genome-wide gene expression and binding location data. In our combinatorial framework, DNA binding events are complementary to their expression counterparts, and together let us approximate the underlying regulation structure. Our model maps regulatory systems onto hierarchical structures which can be approximated by conflating existing large-scale genomic expression and ChIP-chip data. In this paper, we propose a simple combinatorial model of modularity in transcriptional gene regulation and evaluate its plausibility. Specifically, we

- Propose a simplified model of gene regulation based on a one-to-one correspondence between modules of binding sites in cis-regions and shared basic patterns in the gene expression.
- Present a graph theoretical structure, the Regulation Hierarchy, which captures our model and is an independent view of regulation from the reference points of cis-regulation and gene expression.
- Present the Expression Hierarchy and Transcription Factor Hierarchy, structures which approximate the Regulation Hierarchy, and methods to obtain them using biclusters of gene expression and ChIP-chip data.
- Present results that show significant agreement between the two approximations of a Regulation Hierarchy (expression and TF hierarchy), associating modularity in regulation to the granularity in gene expression.

Our method is novel in that it models the regulation modularity independently of the data types and utilizes the regulation modularity to enhance the results, while offering predictions that allow its validation from large-scale data. The model naturally encompasses the conflation of gene expression and TF-DNA binding data, of which public repositories are available. Its utility is wide-ranging, as it can be used to discover recurring cis-modules and their effects on regulation, as well as their co-occurrence with other modules. The model can also be extended to include functional roles of cis-modules, enabling language-theoretic treatments of the logic of cis-regulation.

This paper is organized as follows. In the next section, Section 2, we introduce the model and the regulation hierarchies following from it, including their properties. In Section 3, we make the case for using biclusters as expression patterns. Section 4 describes our methods used in building the hierarchies as well as metrics used to evaluate the results. We discuss the results in Section 5. Related work is given in Section 6, and we conclude in Section 7.

2. PROPOSED MODEL AND THE REGULATION HIERARCHY

We propose a simple (and very simplified) model of gene regulation based on a one-to-one correspondence between functional groups of binding sites in genes' cis-regions (CRs), which we call cis-regulatory modules (CRMs), and patterns in the gene expression matrix shared by the genes. The guiding principle is that CRMs, to a first approximation, are atomic and responsible for an atomic gene expression pattern.

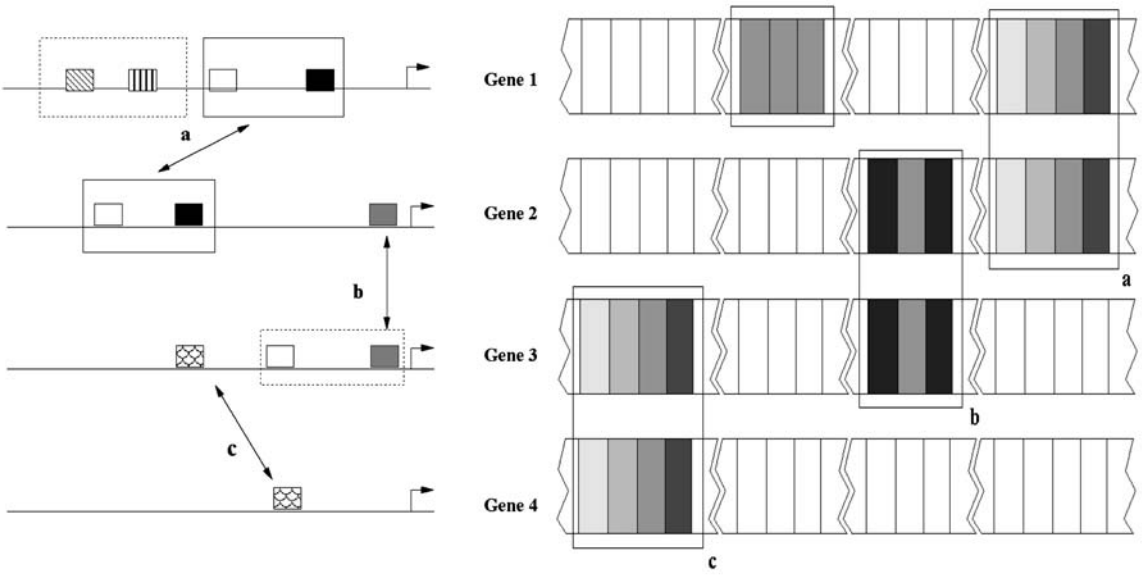

FIG. 1. Illustration of the simple model: the four cis-regions on the left are responsible for the expression patterns in the gene expression matrix on the right, where each box is an experiment (darker shades indicate higher gene expression). Cis-modules a, b, and c correspond to expression events a, b, and c.

They are the building blocks effecting gene expression, and their combinations completely determine any expression pattern during development. Figure 1 gives an illustration. On the right is a gene expression matrix corresponding to four genes, the horizontal stripes are the gene expression profiles, and each box on the y-axis corresponds to one microarray experiment. On the left are given the four genes' CRs, with binding sites organized in CRMs as indicated. The expression patterns a, b, and c on the right are consequences of the actions of the corresponding modules in the CRs.

We can formalize this notion of cis-modularity drawing from the discussion and illustration above with the following simple rules:

1. CRMs are the smallest sets of binding sites (or equivalently the TFs that bind there) which have a distinguishable function in the expression of a gene.
2. We call the smallest "significant patterns" of gene expression Fundamental Expression Patterns (FEPs), examples of which are the patterns a, b, and c on the right in Figure 1. We do not define FEPs completely here, but below we describe a practical approach to specify these patterns, for general gene expression matrices, using biclustering.
3. There is a one-to-one relationship between CRMs and FEPs.

Therefore, in our model of transcriptional regulation, a gene's expression is completely determined by the modules in its CR. Thus, two genes having the same CRMs will be expressed the same, while those that share CRMs will be co-regulated and those that share FEPs co-expressed. Differentially expressed genes will differ in at least one CRM in their CRs.

2.1. Regulation hierarchies

The third rule above naturally implies a partial order among genes based on the subset relationship between their sets of CRMs (or, equivalently FEPs), yielding a regulation structure that summarizes the co-regulation among genes.

The Regulation Hierarchy (RH) is meant to be an invariant view of regulation from both the sequence and gene expression, and a representation of both. RH is defined as a directed graph, $G_r = (V, E_r)$, where V is a set of nodes, or genes $\{g_1, g_2, \ldots, g_n\}$, and there is an edge between two nodes i and j, if the set of CRMs regulating gene i is a subset of the set of the CRMs regulating gene j, and the direction of

the edge is from the smaller toward the larger set of regulators. That is, if $\mathrm{Mod}(x)$ is the set of modules regulating node x, then for every pair of genes i and j, $(i, j) \in E_r$ if $\mathrm{Mod}(i) \subseteq \mathrm{Mod}(j)$. If g_i and g_j share CRMs but none dominates the other, then neither $(i, j) \in E_r$ nor $(j, i) \in E_r$. Equivalently, RH can be defined in terms of the shared FEPs in the gene expressions, hence its invariant nature. Although this structure set-theoretically is a partially ordered set, or poset (but not a lattice or even semi-lattice because the meet and join are not defined for all pairs of genes), we call it a hierarchy to capture its level structure and branching.

We define the following two additional hierarchy graphs, which, in contrast to RH, can be obtained from existing data. The first is the Transcription Factor Hierarchy (TFH), defined as the graph $G_{tf} = (V, E_{tf})$, where if $Tf(x)$ is the set of transcription factors that can bind to the CR of gene x then $(i, j) \in E_{tf}$ if $Tf(i) \subseteq Tf(j)$. In practice, we can construct the TFH from TF-DNA binding data by carefully evaluating the overlaps between TF regulators. The second hierarchy is the Expression Hierarchy (EH) in which nodes are FEPs, or significant sub-matrices in the expression matrix. The EH graph is defined as $G_e = (M, E_e)$, over a set of gene expression sub-matrices, $M = \{m_1(g_1, e_1), m_2(g_2, e_2), \ldots, m_n(g_n, e_n)\}$, where there is an edge between nodes $m_i(g_i, e_i)$ and $m_j(g_j, e_j)$ if $g_i \subseteq g_j$ and $e_j \subseteq e_i$. In practice, the nodes of the EH, respectively, TFH, will be sets of genes, such that within each set the genes will have the same FEPs, respectively the same TFs. The definitions above can readily be extended to incorporate gene sets at the nodes, instead of just single genes.

2.2. *Properties and utility of the regulation hierarchies*

The EH and TFH are, in a way, an upper and lower bound (respectively) on the edges in the RH, because their sets of edges satisfy $E_e \subseteq E_r \subseteq E_{tf}$. Namely, $E_r \subseteq E_{tf}$, since CRMs are groups of TFs and there is no partial overlap between them. On the other hand, $E_e \subseteq E_r$ since FEPs correspond to CRMs, in the ideal case, but the data may not contain all possible FEPs (i.e,. all possible ways that a gene can be differentially expressed).

Figure 2 illustrates on an example the above inclusion relationship between RH, EH, and TFH. Each node represents a gene's regulatory region, shown with the list of TFs (squares) binding there, and its expression profile, comprising eight experiments. There are two additional expression experiments: the dashed rectangles to the right of the first eight, which are imaginary and indicate differential expression of which genes are capable but that is not detected in the data. Pairs of nodes are connected by edges if they share TFs or expression patterns, as specified by our model. Each gene expression pattern corresponds to a cis-module. Light gray edges are in the EH (and thus in all three hierarchies), and are obtained from evidence of shared FEPs in the gene expression data. Dark gray edges are in RH (and also in TFH), and their presence in the absence of an EH edge indicate there is modularity in the CRs unsupported by the expression data (potentially because more experiments are needed, as with the nodes with TF lists {A}, {B, C}, and {A, B, C} in the figure). Black edges belong only to the TFH. Their presence in the absence of the other two edge types indicates only shared TFs between nodes but not necessarily co-regulation, as with the nodes with TF lists {B, C}, {D}, and {B, C, D} in the figure.

The RH and TFH will not be equal in general because there can be TFs binding to a CR without having a functional effect on that gene's expression. Likewise, the EH and RH will not be equal in general because not all modules' functions are identifiable from existing data. With ideal (but not necessarily complete) data, these three hierarchies would be directed, and transitively closed graphs [i.e., if $(i, j) \in E$ and $(j, k) \in E$ then $(i, k) \in E$]. They would also be acyclic, except for the trivial cycles between two genes sharing exactly the same regulators.

From the RH one can readily answer if two genes are co-regulated by looking up if they have a common root node. Also, with the RH and the TFH one can explore TF modules, whereas from the RH and EH the basic expression signals, FEPs corresponding to CRMs can be found.

We call the lowest nodes roots, and the highest leaves (Fig. 2). From the definition of the hierarchies, and since the nodes in the EH and TFH will be sets of genes, the levels in which they occur in the hierarchy will be related to the co-expression and co-regulation among the genes in each node. Generally, the higher the node is in the hierarchy, the higher the level of co-expression and co-regulation among its genes. Genes in root nodes are co-regulated by a small set of TFs, each set representing a CRM. The number of root

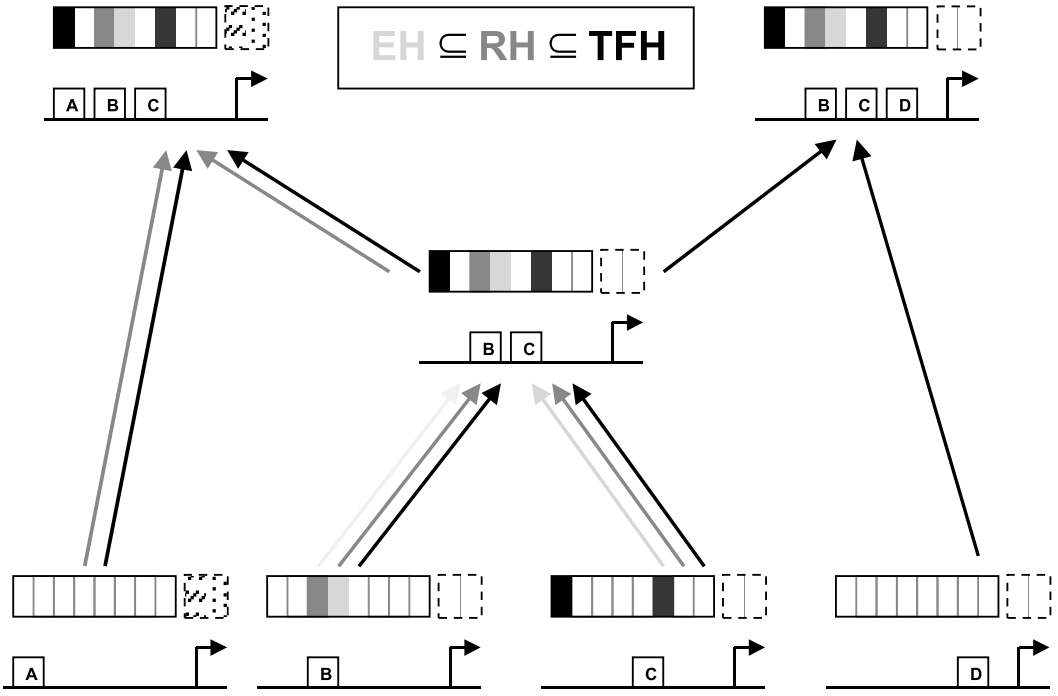

FIG. 2. Illustration of the regulation hierarchies and their relationship. EH, RH, and TFH are the expression, regulation, and TF hierarchies, respectively, and their edge sets satisfy the inclusion relationship shown. At each node are shown the cis-region with TF-specific binding sites, and an example corresponding gene expression profile under eight experiments. Two hypothetical gene expression patterns are shown dashed, to indicate that not all effects of TF binding are accounted for in the gene expression data. The light gray edges represent links in the EH (and RH and TFH), dark gray edges links in RH (and TFH), and the black edges links only in TFH. The lower nodes in the hierarchy have "simpler" signals and cis-regions, while the higher ones are more tightly co-regulated and co-expressed.

nodes for a given leaf node is an indicator of its regulation complexity. Edges in the hierarchy represent shared regulation.

2.3. Cis-regulatory modules and the regulation hierarchy

The definition of RH gives us a way to identify potential CRMs, by (1) identifying the root nodes in the RH and looking up their TF sets, and (2) identifying the TF differences between consecutive levels in the hierarchy. The groups of TFs such identified form the set of CRMs for the hierarchy. All expression patterns of genes in the hierarchy can be decomposed to the expression patterns of the FEPs corresponding to those CRMs.

In Figure 3, RH nodes are shown with corresponding TFs to illustrate how CRMs are organized on the hierarchy. Nodes 1 and 2 are roots, and hence their TF sets, $\{TF_1, TF_2\}$ and $\{TF_3\}$, are CRMs. Since Node 3 has $\{TF_1, TF_2, TF_3, TF_4\}$ as a set of TFs, by taking the set differences along all edges in the hierarchy we get a total of three potential CRMs in this example: $\{TF_1, TF_2\}$, $\{TF_3\}$, and $\{TF_4\}$.

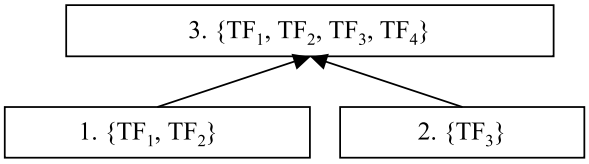

FIG. 3. The nodes in the regulation hierarchy organize the potential cis-regulatory modules (CRMs) in order of increasing complexity, from roots to leaves.

3. FEPs AS BICLUSTERS

To construct an EH from large-scale functional genomics data, we need a biologically meaningful way to identify FEPs from genome-wide microarray expression data. To that end, we define FEPs as biclusters. Such a definition allows us to use existing microarray data analysis methods and adapt them for this purpose. Although the definition of a bicluster varies for different algorithms, in essence all algorithms look for statistically significant submatrices in a given gene expression data matrix (Cheng and Church, 2000; Lazzeroni and Owen, 2002; Tanay et al., 2002). Finding the most significant biclusters is a computationally hard problem and all these algorithms use some heuristics to get close to a "near-optimal" solution.

The notion of a bicluster corresponds well to our FEPs, since a bicluster is an expression sub-pattern found in a significant number of genes. Empirically, variation in expression has been shown to correlate with CR complexity (Bilu and Barkai, 2005). Biclusters capture this expression variability, with each bicluster representing an expression pattern or event. Occurrence of a gene in multiple biclusters implies participation in multiple expression events. Genes in a bicluster may share a function, be co-regulated or be active in the same pathway. Exploring overlaps in biclusters can help differentiate among different biological concepts. To that end, a bicluster hierarchy, a structure to order biclusters representing an expression event and overlaps among biclusters, will be used to approximate RH.

4. METHODS

4.1. Cis-regions

A CR is approximated as a set of TFs, that may consist of multiple CRMs. Biclusters along with TF-DNA binding data are used to find CRs that are well represented in a given expression data set. CRs are obtained from sets of TFs that are enriched in a given hierarchy node (using the hypergeometric distribution). All subsets of an enriched TF set in a bicluster are considered and the set should be present in at least 10% of genes in that bicluster.

4.2. Building the expression hierarchy from biclusters

The EH is constructed from an initial set of biclusters obtained using the SAMBA algorithm (Tanay et al., 2002) on our gene expression data set. The biclusters that SAMBA produced are considered to be at level zero, or the root level of the hierarchy. New levels in the hierarchy are then constructed iteratively by creating nodes from overlapping pairs of biclusters from previous levels. New levels are generated until no significant overlaps among biclusters at the previous level are found. The specific steps are as follows.

Merge: When new nodes are created for all overlaps among biclusters, there is a likelihood of creating a large number of highly similar nodes which results in a "noisy" hierarchy. To handle this scenario, similar nodes are merged using average-link hierarchical clustering, using the following criteria for merging:

$$(|G_i \cap G_j|/|G_i \cup G_j| \geq 0.7 \;\&\; rEC < \forall EC)$$

$$\text{or} \quad (|G_i \cap G_j|/|G_i \cup G_j| > 0.5 \;\&\; |G_i \cup G_j| - |G_i \cap G_j| < 6),$$

where G_i is the set of genes for bicluster i, G_j is the set of genes for bicluster j, EC is the Co-Expression Index for gene set $G_i \cap G_j$ and rEC is the Co-Expression Index for a random gene set of size $|G_i \cap G_j|$ generated from $|G_i \cup G_j|$ (100 random sets are generated).

Move-up: If the gene set of a node at a current level is a subset of some node at the same level then that node is moved up to a higher level. Some fuzziness is allowed for the subset relationship, in order to move a node higher up, based on

$$(|G_i \cap G_j| \geq 0.7 \times |G_i| \;\&\; |G_i| - |G_i \cap G_j| < 6) \quad \text{or} \quad (|G_i \cap G_j| \geq 0.9 \times |G_i|).$$

Create Nodes: New nodes are created from "significantly overlapping" gene sets from all pairs from the current level. The following condition is used to determine the significance of an overlap:

$$((|G_i \cap G_j| > 20) \quad \text{or} \quad (|G_i \cap G_j| > 4 \;\&\; (|G_i \cap G_j|/|G_i \cup G_j| \geq 0.3$$

$$\text{or} \quad (|G_i \cap G_j| \geq 0.4 \times |G_i| \quad \text{or} \quad |G_i \cap G_j| \geq 0.4 \times |G_j|)))$$

or

$$(|G_i \cap G_j| > 9 \;\&\; (|G_i \cap G_j|/|G_i \cup G_j| \geq 0.2 \quad \text{or} \quad |G_i \cap G_j| \geq 0.25 \times |G_i|$$

$$\text{or} \quad |G_i \cap G_j| \geq 0.25 \times |G_j|)) \;\&\; (EC > \forall rEC).$$

In theory, the number of newly created nodes can be exponential but for real data this number is bound by the small degree of overlap among original biclusters. Choices for parameters are ad hoc and driven by specific data sets, but are conservative.

Create Edges: An edge (i, j) is created from node i to node j if: $|G_i \cap G_j| \geq 0.8 \times |G_i|$.

Eliminate Shortcuts: Transitive reduction is then applied to get the bicluster hierarchy.

A random EH was generated that has the same number of nodes and edges as the empirical one. In it, the edges remain the same but the gene sets in the nodes are different, as follows. For each node in the empirical EH a node of the same size is created by uniformly at random selecting genes from the union set of genes of the parent nodes. (Recall that in EH genes in a node are the intersection set of genes of its parent nodes.) As a result, root nodes are the same for both a random and a real bicluster hierarchy.

4.3. Metrics

The Co-Expression Index is a measure of the level of co-expression in a gene expression matrix. The Co-Expression Index for a given matrix with g genes and e experiments is defined as the average standardized expression value (z-score) for all genes over the experiments in the matrix:

$$\sum_{j=0}^{e} \left| \sum_{i=0}^{g} \frac{e_{ij} - \overline{e}_j}{\sigma_j} \right| \Bigg/ (e \cdot g),$$

where e_{ij} is the expression value for gene i and experiment j, $\overline{e}_j$ is the average of expression values for experiment j over all genes, and σ_j is the standard deviation of expression values for experiment j over all genes.

The Co-Regulation Index is a measure of the level of co-regulation in a gene expression matrix. A higher Co-Regulation Index means a "tighter" level of co-regulation. For a given matrix with g genes and e experiments, the Co-Regulation Index is defined as that fraction of TF's regulating either of two genes that regulate them both, summed over all gene pairs from g:

$$\sum_{i=0}^{g} \sum_{j=1+1}^{g} \frac{\mathrm{TF}_i \cap \mathrm{TF}_j}{\mathrm{TF}_i \cup \mathrm{TF}_j} \Bigg/ \binom{g}{2},$$

where TF_i is the set of TFs that bind to gene i and TF_j is the set of TFs that bind to gene j.

5. RESULTS AND DISCUSSION

This simple model of transcriptional regulation is meant to be an invariant view of regulation from either the CRs or the gene expression patterns. To validate this property and demonstrate the soundness of our model, we construct the two enveloping hierarchies, EH and TFH, in order to estimate the RH for yeast. Then we evaluate their overlap, as estimators of RH, as well as the properties we expect them to satisfy, from Section 2.2.

We used the following publicly available large-scale genome-wide yeast data sets to construct the hierarchies:

- Expression data were downloaded either from the Stanford Microarray Database (Gollub et al., 2003) or from the authors' publication supporting websites, following a comprehensive list of expression data sources provided at *http://cs.tau.ac.il/~rshamir/simba,* encompassing about 1000 experimental conditions and 6200 genes.
- TF-DNA binding (ChIP-chip) data was obtained from the supporting website for Harbison et al. (2004). Orfs bound by TFs with binding p-value of <0.001 and no conservation criteria were used. The dataset consists of binding data for a total of 352 TFs (some TFs were tested under different environmental conditions) with about 6200 genes.

To construct the EH, we started with a total of 507 biclusters obtained from the expression data matrix using the SAMBA algorithm (Tanay and Sharan, 2002) and default settings. New nodes were constructed from overlapping biclusters and similar nodes were merged, as described in Methods. The final EH has 1379 nodes, including the original biclusters, connected with 7227 edges, with 217 leaf and 217 root nodes. The root nodes are characterized by overlapping gene sets, and small nonoverlapping experiment sets, while the leaf nodes are nearly nonoverlapping in their gene sets with highly overlapping experiment sets, as expected.

To make sure we had comparable nodes between the hierarchies, we constructed only that portion of the TFH whose nodes overlap with the EH. Thus, the nodes in TFH contain the same gene sets as the nodes in EH, and are represented by the CRs obtained from the gene sets in those nodes (see Methods). An edge is established between TFH nodes for every subset relationship among sets of TFs representing the nodes' CRs. Only those CRs that are well-represented in EH (see Methods) are used instead of using all CRs which may be obtained solely from TF-DNA data. We found a total of 2150 edges in EH and 761 edges in TFH between nodes which had nonempty CRs.

5.1. Co-expression and co-regulation between nodes in EH

We measured the levels of co-expression and co-regulation between genes in the nodes of the EH. The results are shown in Figure 4, where we see that a significant fraction of leaf nodes have higher Co-Expression and Co-Regulation Indices than root nodes. The Co-Expression and Co-Regulation Indices were also compared between connected nodes. For 7076 edges out of 7227 edges (97.9%), the child node had a higher Co-Expression Index than the parent node. For random nodes in 2664 edges (36.7%) the child node had a higher Co-Expression Index. For 5280 edges out of 7227 edges (73.1%), the child node had a higher Co-Regulation Index than the parent node. For random nodes in 2763 edges (38.2%) the child node had a higher Co-Regulation Index. For 5180 edges (71.7%), the child node had both Co-Expression and Co-Regulation Indices higher than the parent node. For random nodes in 1120 edges (15.5%), the child node had a higher Co-Expression and Co-Regulation Indices. These results show that the EH has the desirable property of regulatory complexity increasing from root to leaves.

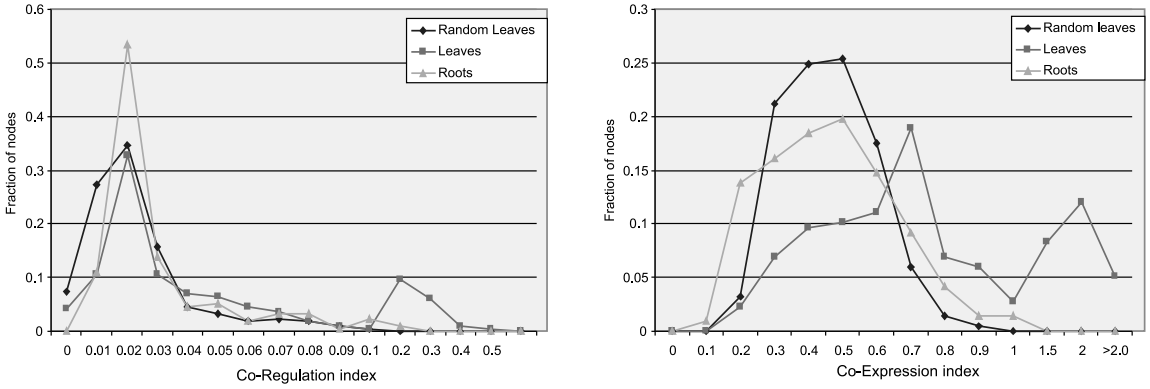

FIG. 4. The co-regulation (left) and co-expression (right) are overall higher in the leaves than in the roots of the hierarchy, compared to random subsets of elements (i.e., leaves) from the root sets.

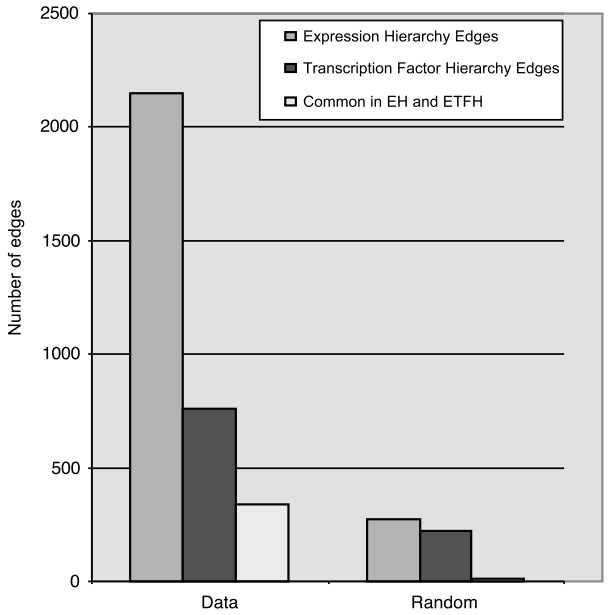

FIG. 5. The agreement between expression and TF hierarchies, in terms of number of shared edges and as compared to a random assignment of edges, is surprisingly strong in spite of the noisiness of the data and simplicity of the model.

5.2. Agreement between the expression and TF hierarchies

Here, we sought to evaluate the agreement between EH and TFH. Out of the comparable 2150 edges in EH and 761 edges in TFH, 341 edges were common to both. The EH and TFH obtained from a randomly created bicluster hierarchy (see Methods) had 276 and 222 edges, respectively, with only nine edges in common. The results are shown in Figure 5. The significant overlap in the hierarchies implies that our combinatorial model captures a good portion of modular relationships, despite its simplicity, and the noisiness of the data.

5.3. GO annotations and levels in the hierarchy

The GO hierarchy was used to compare the generality of processes vis-à-vis the levels of their genes in EH. Levels in the GO hierarchy are considered bottom-up. The most specific GO terms are at level zero. For all root node-leaf node pairs, their GO annotations were compared. 31.5% of root node annotations are found in their descendant leaf nodes out of which 43.2% have a better p-value in the child node. For random data, 7.7% of root node annotations are found in their descendant leaf nodes out of which only 3.9% have a better p-value in child node. The average level in the GO hierarchy for annotations with better p-value in descendant nodes is 0.36 compared to 0.41 for the annotations that did not have a better p-value in descendant nodes. Table 1 shows a comparison of enriched GO annotations in a root versus its descendant leaf nodes. Most of the root node annotations are more significantly enriched in one of the descendant leaf nodes. Also, leaf nodes have a number of enriched GO annotations that were not enriched in the root node. Root nodes at the lowest levels in the hierarchy represent more general processes, and nodes at higher levels represent more specific processes.

6. RELATED WORK

Much work has been done in the area of finding regulatory elements computationally from DNA sequence data (Bailey and Elkan, 1995; Sinha and Tompa, 2003; Thijs et al., 2001; Bussemaker et al., 2000). With the availability of genome-wide gene expression data, simple models for transcriptional regulation by single TFs have been used to find regulatory elements (Bussemaker et al., 2001; Chiang et al., 2001).

TABLE 1. GO ANNOTATION COMPARISON BETWEEN LEAF AND ROOT NODES SHOWS MUCH HIGHER (OR REFINED) FUNCTIONAL ENRICHMENT IN THE LEAVES FOR THE SAME FUNCTION PRESENT IN THE NODES

GO annotation	*Root node p-value*	*Lowest leaf node p-value*
Cellular macromolecule metabolism	1.91E-20	Not present
Cellular metabolism	4.96E-15	Not present
Macromolecule biosynthesis	3.82E-40	Not present
Main pathways of carbohydrate metabolism	3.09E-10	Not present
Organic acid metabolism	4.92E-09	Not present
Aerobic respiration	1.33E-08	1.29E-20
Biosynthesis	4.71E-43	0
Cellular biosynthesis	1.65E-43	0
Electron transport	2.54E-10	6.62E-23
Generation of precursor metabolite and energy	7.16E-11	3.77E-21
Heme-copper terminal oxidase activity	1.86E-07	5.01E-13
Hexose catabolism	1.08E-08	1.51E-13
Hydrogen ion transporter activity	3.45E-12	1.91E-19
Oxidative phosphorylation	6.48E-09	1.59E-25
Oxidoreductase activity	1.25E-11	1.80E-12
Structural constituent of ribosome	0	0
Amino acid biosynthesis	Not present	4.50E-17
Amino acid metabolism	Not present	1.08E-20
Carrier activity	Not present	3.40E-15
Gluconeogenesis	Not present	4.31E-12
Ion transporter activity	Not present	6.54E-17
Methionine metabolism	Not present	3.99E-08
Protein biosynthesis	Not present	0
Regulation of protein biosynthesis	Not present	2.11E-07
Ribosomal subunit assembly	Not present	1.95E-12
Translation	Not present	1.56E-07
Transport	Not present	5.04E-10

The logic by which cis-elements interact to effect/modulate transcription is not known but some patterns are becoming apparent that indicate such logic is very likely to exist (Istrail and Davidson, 2005). Such patterns include, for example, logical operations based on motif occupancy (Yuh et al., 2001; Buchler et al., 2003), inhibition (Kulkarni and Arnosti, 2005), amplification, the modularity of cis-elements, their geography on the DNA with respect to other modules, modality of multiple modules. Knowledge of combinatorial regulation directed researchers towards developing algorithms to search for groups of binding sites that act in concert to effect expression (Sharan et al., 2003). Recent studies have been able to successfully predict gene expression data from sequence data by taking into consideration few of the complex rules utilizing logic operations and constraints like strength, orientation, relative position, and multiplicity of binding sites that govern transcriptional regulation (Beer and Tavazoie, 2004; Nguyen and D'Haeseleer, 2006).

Although gene's expression is determined by the combinatorics of TF interactions with cis-elements, it is not trivial to establish a correspondence between co-expression and co-regulation. Methods for grouping genes by similarity of expression profiles across multiple experiments have been partially successful in identifying functionally related genes (Eisen et al., 1998). But since co-expression does not imply co-regulation in general such methods have been limited to the identification of gross functional features and categories. Differentiating between co-expressed and co-regulated genes is important in particular for gene network inference. Pilpel et al. (2001) proposed and later improved (Lapidot and Pilpel, 2003) methods to identify clusters of genes which are co-regulated and co-expressed at the same time. They achieved this by scoring co-expression for genes which share overrepresented elements in the upstream regions. Such studies provide mostly empirical results but not fundamental understanding or combinatorial models

which relate co-regulation with co-expression, and suffer from false positives from the DNA motif search. A few studies recently have focused on identifying modules of genes by considering variety of available data: gene expression, sequence, and TF-DNA location (Pilpel et al., 2001; Bar-Joseph et al., 2003; Stuart et al., 2003). The working definition for a module in them varies between a group of strongly co-expressed genes in a subset of experiments (Bergmann et al., 2003) to a group of genes co regulated by the same factors and sharing a function (Segal et al., 2003). In both extremes, though, the definition of a module is imprecise and mostly empirical.

The small number of different patterns evident in time-course gene expression data, especially the cycling genes set by Spellman et al. (1998), has motivated several studies into evaluating the possibility of decomposing the expression signals into a combination of a few basic signals. In particular, the study by Holter et al. (2000) identified a small number of characteristic modes in microarray time-series data, as discovered by Singular Value Decomposition. Such studies although informative about the range of the transcriptional signals under specific conditions, and arguably successful in correlating functional gene categories with specific modes of regulation, do not address the issue of co-regulation, nor model the causes for it.

Clustering of genes has been effectively used for analysis of gene expression data, although such clustering techniques are limited as they depend on the global similarity of genes. Often, groups of genes are similarly expressed only under certain experimental conditions while their expression pattern is uncorrelated under other conditions. To overcome this limitation various algorithms have been developed that search for biclusters, i.e., a group of genes that have similar expression pattern under a subset of experimental conditions (Cheng and Church, 2000; Getz et al., 2000; Tanay et al., 2002; Lazzeroni and Owen, 2002; Bergmann et al., 2003). Genes in a bicluster may share a function, be co-regulated, or be active in the same pathway. They are proven to be more effective than standard clustering methods in case of gene expression data from multiple studies (Ihmels et al., 2004; Tanay et al., 2005a) as they better capture the biology of transcriptional regulation. Ihmels et al. (2002) and Tanay et al. (2004) used biclustering to reveal the hierarchical modular organization in the yeast transcriptional network. Bussemaker et al. (2001) and Beer and Tavazoie (2004) have shown the combinatorial effects of individual Transcription Factor Binding Sites on gene expression.

7. CONCLUSION

In this paper, we presented a simple combinatorial model of causal modularity in transcriptional gene regulation. We presented the theoretical RH, and showed that it captures the complexities of genes' CRs. We introduced the concept of an FEP as a general identifiable functional event of a gene set. FEPs are used to build an EH, a lower envelope to the RH, from gene expression data. We identified FEPs with biclusters and used them to build the EH, a structure that can be of independent interest for integrative gene expression data analysis (Tanay et al., 2005b). A TF hierarchy, as an upper envelope to the RH, can be constructed from TF-DNA interaction data. Our results showed significant overlap between the two empirical hierarchies, in spite of the noisiness in the data, thus providing evidence for the proposed model of regulation.

Our choice of methods at each step was guided by usability and may not be the most appropriate. SAMBA was used with default settings to generate biclusters from expression data for expression events. Other algorithms such as Independent Component Analysis (Lee and Batzoglou, 2003) or Iterative Signature Algorithm (Bergmann et al., 2003) for the same purpose can be explored and generalized further.

The construction of EH using all overlaps among biclusters is a computationally expensive task. We used simple heuristics to limit overlaps to a small number; more general, topology-based optimization methods would likely yield better results. CRs and CRMs are approximated as sets of TFs using TF-DNA data. This is a very simplified view of a CRM. Information from sequence data, i.e., information about how these TFs bind, binding site locations, distance among sites, number of binding sites, ordering of binding sites, and inhibitory effects of TFs (Nguyen and D'Haeseleer, 2006), should be considered. But because of its simplicity this model of regulation is very flexible and can be extended by incorporating, for example, inhibitory effects into it, extending it to other organisms, and developing visualization tools for exploring hierarchies effectively.

Modularity, used as a model design paradigm here, helps us to scale the phenomenon of transcriptional regulation so that we can think of it not in terms of biochemistry but in terms of abstract processes and ideas, and in terms of its expressive language. The underlying meaning is that there are semantic building blocks that transcriptional regulation reuses to make genes active and to make networks connected. If, perhaps, there are a finite number of such semantic blocks, then there might be a language of transcription and gene regulation that is very much like the programming languages that we know, written in the genetic codes of animals. The regulation hierarchies may help us identify such modules. We hope that this work can serve as a stepping stone towards more complex combinatorial models which can help identify the elements of the language of transcriptional regulation.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Bailey, T.L., and Elkan, C. 1995. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 3, 21–29.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., et al. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotech.* 21, 1337–1342.

Beer, M., and Tavazoie, S. 2004. Predicting gene expression from sequence. *Cell* 117, 185–198.

Bergmann, S., Ihmels, J., and Barkai, N. 2003. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67, 031902.

Bilu, Y., and Barkai, N. 2005. The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol.* 6, R103.

Buchler, N.E., Gerland, U., and Hwa, T. 2003. On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA* 100, 5136–5141.

Bussemaker, H.J., Li, H., and Siggia, E.D. 2000. From the cover: building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* 97, 10096–10100.

Bussemaker, H.J., Li, H., and Siggia, E.D. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–174.

Cheng, Y., and Church, G.M. 2000. Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8, 93–103.

Chiang, D.Y., Brown, P.O., and Eisen, M.B. 2001. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics* 17, S49–S55.

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

Getz, G., Levine, E., and Domany, E. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA* 97, 12079–12084.

Gollub, J., Ball, C.A., Binkley, G., et al. 2003. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* 31, 94–96.

Harbison, C.T., Gordon, D.B., Lee, T.I., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.

Holter, N.S., Mitra, M., Maritan, A., et al. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* 97, 8409–8414.

Ihmels, J., Bergmann, S., and Barkai, N. 2004. Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993–2003.

Ihmels, J., Friedlander, G., Bergmann, S., et al. 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31, 370–377.

Istrail, S., and Davidson, E.H. 2005. Logic functions of the genomic cis-regulatory code. *Proc. Natl. Acad. Sci. USA* 102, 4954–4959.

Kulkarni, M.M., and Arnosti, D.N. 2005. Cis-regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. *Mol. Cell Biol.* 25, 3411–3420.

Lapidot, M., and Pilpel, Y. 2003. Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. *Nucleic Acids Res.* 31, 3824–3828.

Lazzeroni, L., and Owen, A. 2002. Plaid models for gene expression data. *Stat. Sin.* 12, 61–86.

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

Nguyen, D.H., and D'Haeseleer, P. 2006. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.* 2, 12.

Pilpel, Y., Sudarsanam, P., and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159.

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

Sinha, S., and Tompa, M. 2003. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 31, 3586–3588.

Spellman, P.T., Sherlock, G., Zhang, M.Q., et al. 1998. Comprehensive identification of cell cycle–regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297.

Stuart, J.M., Segal, E., Koller, D., et al. 2003. A gene co-expression network for global discovery of conserved genetics modules. *Science* 302, 249–255.

Tanay, A., Sharan, R., Kupiec, M., et al. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA* 101, 2981–2986.

Tanay, A., Sharan, R., and Shamir, R. 2002. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18, S136–S144.

Tanay, A., Sharan, R., and Shamir, R. 2005a. *Biclustering Algorithms: A Survey. Handbook of Computational Molecular Biology*. Chapman & Hall/CRC, New York.

Tanay, A., Steinfeld, I., Kupiec, M., et al. 2005b. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol. Syst. Biol.* 1, E1–E10.

Thijs, G., Marchal, K., Lescot, M., et al. 2001. A Gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes. *J. Comp. Biol.* 9, 447–464.

Yuh, C.H., Bolouri, H., and Davidson, E.H. 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development* 128, 617–629.

Address reprint requests to:
Dr. Vladimir Filkov
Computer Science Department
UC Davis
One Shields Ave.
Davis, CA 95616

E-mail: filkov@cs.ucdavis.edu