Modeling Transcription Programs: Inferring Binding Site Activity and Dose-Response Model Optimization

Amos Tanay * and Ron Shamir †

ABSTRACT

The modeling of transcription regulation programs is a major focus of today's biology. The challenge is to utilize diverse high-throughput data (gene expression, promoter binding site localization assays, protein expression) in order to infer the mechanistic models of transcription control. We propose a new model which integrates transcription factor-gene affinities, protein abundance and gene expression levels. Transcription factor binding site activity is represented by a dose-affinity-response function, and regulation is assumed to be a combinatorial function of the activities of the binding sites in the gene's promoter sites.

We develop algorithms that infer the model given complete data and give a fast polynomial time algorithm under reasonable assumptions. We also show how to assess initial values of missing data (notably protein abundance) using a novel framework for active motif detection, which may be of independent interest. We test the various components of the framework on gene expression data related to carbohydrate metabolism in yeast. The results demonstrate the high specificity and sensitivity of the approach and its advantages over extant motif activity detection methods. We are also able to predict new active motifs in the galactose pathway.

A key feature of our method is the global approach to transcription factor activity and to the relation between this activity and promoter signals. We use dozens of genes, with many different promoter signals and expression levels in order to draw conclusions on the function of a single transcription factor. This provides us the robustness necessary in order to overcome the considerable level of noise in the data.

*School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. E-mail: amos@post.tau.ac.il.

†School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. E-mail: rshamir@post.tau.ac.il.

RECOMB'03, April 10–13, 2003, Berlin, Germany.

1. CATEGORIES & SUBJECT DESCRIPTORS

J.3 Biology and genetics

2. GENERAL TERMS

Algorithms

3. INTRODUCTION

The mechanisms of transcription control are of the most studied in biology, yet, our understanding of the way in which transcription factors induce and repress transcription is very far from complete. In particular, biologists believe that important signals of the transcription program of each gene are "hardwired" into the DNA sequence around it, but we still cannot accurately predict this program from the DNA sequence.

A transcription program (TP) can be described by the following 3-level mechanistic model. First, certain transcription factors (TFs) attain specific concentrations in specific post translational conformations. Second, as a result of sequence signals in the gene's promoter and the concentrations of certain TFs, the DNA in the proximity of the target gene undergoes chromatin modifications and targeted TF binding. We say that the promoter binding sites becomes *active* in such cases. Third, the spatial organization of the promoter's binding sites and the physical properties of the TFs induce a combinatorial regulation scheme which results in conditional assembly of the transcription apparatus and regulate the rate of transcription.

Several new experimental techniques provide data on different components of this complex system. Gene expression microarrys [17, 7] generate large datasets on genomewide mRNA levels in many different conditions. High throughput DNA location analysis [15, 10] measures genomewide TF - gene affinities. Large scale proteomics may provide, in the near future, information on the level of the TF proteins themselves and will thus supply the last major missing part of the data from which the 3-layered model of transcription regulation can be built.

The computational challenge of building a consistent and predictive model for transcription is indeed considerable. DNA based prediction does not seem practical today. The most common method to analyze TPs today is to cluster the genes according to their expression profiles and search for enriched DNA motifs in the promoters of co-clustered genes [17, 21]. This basic method was extended in several directions: Pilpel et al. [14] searched for combinations of known motifs with significant co-expression in order to discover combinatorial transcription control schemes which are beyond the reach of one dimensional clustering. Bussemaker et al. [2] used a linear model to associate motifs and expression and applied a greedy algorithm for de-novo construction of such model from expression data. Segal et al. [16] presented a Bayesian graphical model integrating position specific score matrices (PSSMs) with expression profiles and suggested the use of DNA location experiments as noisy observations on TF-gene relations.

In this work we aim to study an extended transcription network inference problem. Previous studies used the mRNA expression levels to model both transcription levels and the abundance of the proteins (TFs) involved. In contrast, we model the two types of entities separately. This improves the ability to model post-translational effects and overcome poor mRNA-protein levels correlation. Our model covers all three levels of TPs: a) TF concentrations (or doses), b) TF-gene affinity and a resulting site activity, and c) the gene's combinatorial regulation scheme. The model defines functions that associate TF concentration and TF-gene affinity with binding site activity (these are called dose-affinity-response functions). It then uses them as the inputs to combinatorial logic schemes that predict the rate of gene expression given TF concentration and TF-gene affinities. The key feature of our method is the global approach to the integration of TF concentrations and transcription rate via the dose-affinity-response functions. We use dozens of genes, with many different promoter signals and expression levels, in order to draw conclusions on the function of a single TF. This provides us the robustness necessary in order to overcome the considerable level of noise in the data.

In order to build TPs from experimental data we must fill in missing details (e.g. TF doses) and reconstruct the functions that define our model. We first assume all details are provided, including TF doses, TF affinities and expression profiles. Our core algorithm then efficiently seeks a TP with optimal transcription predictions given the inputs. In the (practical) case where we lack experimental information on TF doses and/or TF-gene affinities, we must first generate initial assessments of these values. To achieve this goal, we have created an integrated active motif discovery framework that can identify active DNA motifs and assess their relative activity across the given experimental conditions (this framework can also be used as an independent gene expression analysis tool, improving, e.g., [2]). We can use the site activity values as initial TF doses for the TP optimization algorithm (Figure 1) then apply a two-phase iterative algorithm which alternates between optimizing the model and the TF doses.

Our technical contributions include the following: We define a new TP model and the related optimization problem. The model combines monotone dose-affinity-response functions for each TF and gene-specific combinatorial regulation functions. We show that under reasonable assumptions, a restricted case of the model optimization problem can be solved polynomially and we use this special case as a subroutine in a global, hill climbing model optimization algorithm. We also show how to optimize TF doses given a fixed model and use the two algorithms (model optimization, doses optimization) in an alternating fashion to simultaneously derive the model and the TF doses. The success of the alternating algorithm depends on reasonable initial assessments for the collection of active TFs and their doses. To this end, we build a framework for detecting TF binding sites and to assess their activities. We slightly extend the notion of a PSSM to include information on site location distribution (denoted LPSSM - Located PSSM). We use a background model based on [20] to define a likelihood ratio based scoring scheme for assessing motif (LPSSM) activity. Our ideas generalize [2, 3] by replacing the mean of expression across a set of genes by a more sensitive statistical test using a descriptive background model. We use our scoring scheme in a motif optimization algorithm that is capable of tuning LPSSMs for optimal activity.

We tested our methods on gene expression data related to yeast carbohydrate metabolism. We first applied our active motif detection framework and were able to identify TFs that are known to be active in the system plus several new putative active motifs. We compared our scoring scheme to the simple mean-based scoring system from [2, 3, 9] and showed that our suggested methods have improved sensitivity and specificity. Finally, we used our algorithm to generate a TP model predicting the transcription of all significantly changing genes in a set of galactose related conditions. We show that by analyzing the structure of our optimized TP, we derive correct and highly specific results for known TFs. We also predict the activity of two previously uncharacterized binding sites involved in regulation under galactose related perturbations.

4. A MODEL FOR TRANSCRIPTION PROGRAMS

We first define our model for transcription programs. A set T of TFs controls the rate of transcription of a set V of genes. The control depends on the TF-gene affinities $A : T \times V \rightarrow R_+$ which express the strength of interaction between each TF and the promoter of each gene. The affinity is expected to be a function of the sequence in the gene's promoter and can sometimes be predicted using standard DNA motif models as we shall show below. A valid TP model should be able to predict the rate of transcription using only the concentrations of TFs. Our model is built upon two concepts (compare Figure 2). The first is a dose-affinity-response (DAR) function which assigns for each TF dose and affinity the strength of binding site activity : $DAR_t : R_+^2 \rightarrow R_+$. The function is specific to the TF t and independent of the regulated gene, though its value for gene g will depend on the affinity of t to g. The second concept is the gene's combinatorial regulation scheme (CRS) which computes the rate of transcription given the gene's binding sites activities: $CRS_g : R_+^{|T|} \rightarrow R_+$. Given TF doses $D_1 \ldots D_{|T|}$ we predict the expression of a gene $g \in V$ as $CRS_g(DAR_{t_1}(D_{t_1}, A(t_1, g)), DAR_{t_2}(D_{t_2}, A(t_2, g)), ..)$.

We make two assumptions on our models structure. First, the DAR functions are monotone increasing in each coordinate. This is a very natural restriction in the biochemical reactions we are addressing and provides additional coherence for our model. True active TFs would tend to behave monotonically, while artifacts would be unlikely to manifest such consistent pattern. Furthermore, currently available expression datasets exhibit this phenomenon in many cases (for examples see Figure 3). A second assumption is introduced for simplicity and relates to the values taken by our model variables. We assume that the site activities and expression levels attain discrete values, albeit of arbitrary

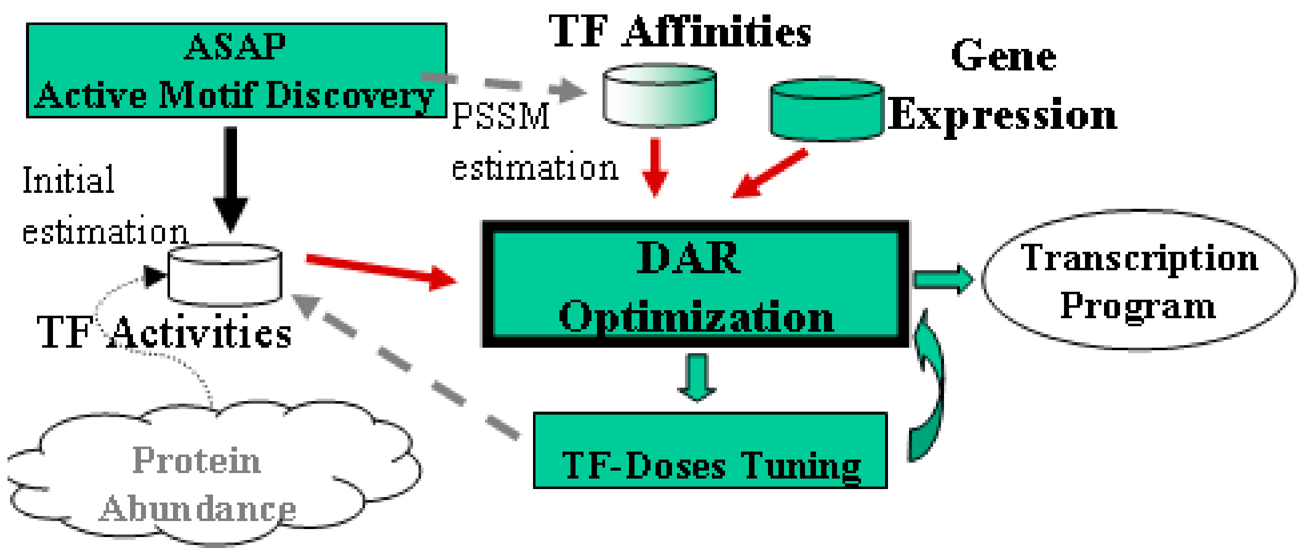


Figure 1: Transcription Program Inference. The core inference algorithm (DAR optimization) learns a transcription program given gene expression, TF affinities and TF doses. The transcription program model predicts the rate of gene expression in different conditions and formalizes both the logics and kinetics of transcription factor-genes interactions. Since not all of the required parameters are available from physical experiments, we apply an integrated framework to estimate and tune missing values. We use a novel algorithm for active DNA motif discovery (ASAP) to generate an initial approximation of the set of active TFs and their doses in each condition. We then alternate between model optimization and missing values tuning to optimize the predictions of the output transcription program model. Future proteome-wide experiments may enable direct assessment of active TF doses and improve the overall quality of the process.

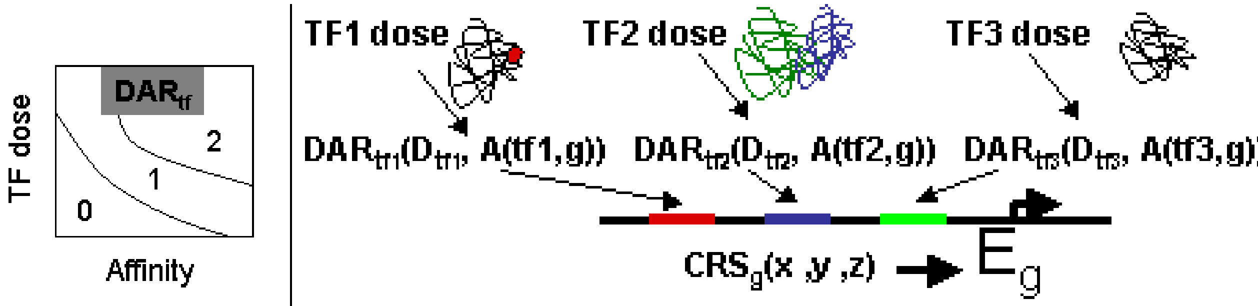
cardinality. The discretization assumption can be relaxed by using probability distribution over the discrete values in use (as in [18]). For simplicity, we omit such technical details and use concrete discretization. Affinities and doses are not discretized to a fixed set of values but we use only their relative order in all computations. Further restrictions may be made by constraining the CRS functions to a class of biologically reasonable logics (cf. [6]).

We now introduce the TP model optimization problem. We are given a set T of TFs, genes V, affinities A and a training set U consisting of expression profiles E_g^u and TF doses D_t^u for all $g \in V$ and $t \in T$ and for each condition $u \in U$. A *model scoring* (or fitness) *function* $SCORE(M)$ assigns a real value to every possible model and should provide indication to the quality of model predictions. We say that a scoring function is *decomposable over the genes* if we can write it down as: $SCORE(M) = \sum_{g \in V} SCORE_g(M)$ where $SCORE_g$ depends only on CRS_g and on the DARs with arguments restricted to affinities to g. A decomposable scoring function is *polynomial* if given a model for a gene g we can compute $SCORE_g$ in polynomial time. The goal of the TP model optimization problem is to find a model M optimizing $SCORE(M)$. There are many possible scoring functions we can use, including mutual information [13], consistency, rSpec [18], MDL/BDE [4] and more. In the experiments reported below we have used the mutual information p-value score (suggested first in [6]) which is derived by applying the chi square statistic to the mutual information between the site activity vectors and the gene's expression values. We shall assume throughout that the model scores are decomposable as is true for all the above mentioned scoring functions.

5. MODEL OPTIMIZATION

The TP model optimization problem can be decomposed into two interacting subproblems. In the CRS optimization problem, the DAR functions and the affinities are known and the goal is an optimal combinatorial regulation function. When using decomposable scoring schemes in any of the methods described above, CRS optimization is efficiently solved independently for each gene. We can thus assume the CRS optimization problem is done implicitly when calculating $SCORE_g(M)$ and focus on DAR optimization that uses $SCORE_g(M)$ queries.

We can use the monotonicity assumption on DARs to devise a polynomial algorithm for DAR optimization over a bounded number of TFs. We will show this for a special case in which site activities are binary (0/1) and we are fixing the DARs for all but one TF t. Assume we have such a base partial model M' and we search for an optimal M which realizes it via a specific DAR_t. We assume for convenience that the score function can be precomputed such that we can answer any query $SCORE_g(M)$ in $O(1)$ time. Let $g_1 \ldots g_n$ be the set of genes sorted by increasing affinities $a_1 < \ldots < a_n$ of t to them. Let $u_1 \ldots u_{|U|}$ be the set of conditions sorted by increasing doses of t, $d_1 < \ldots < d_{|U|}$. For simplicity, we assume that all values are distinct and add an artificial condition u^* with $d_* > d_{|U|}$ and an artificial gene g^* with $a^* > a_n$. Form a matrix with a column for each possible affinity value and a row for each possible dose value (see Figure 4). The key point here is that due to the monotonicity assumption, the activity levels are increas-



A B

Figure 2: Transcription program models. A: DAR functions map monotonically TF doses and affinity to binding site activity. B: Sites in the promoter region allow binding of TFs. Site activity is determined by the TF dose and its affinity to the gene, via the DAR function. Transcription rate is then determined by the CRS.

ing in each row and column, so we need to determine only a single threshold which divides the column (row) into two segments of 0 and 1 values (white and black in the figure). Furthermore, the decomposability of the score function ensures that for each gene g, $SCORE_g(M)$ is completely determined by the threshold for g's column (i.e., by values of the form $DAR_t(*, A(t,g))$. We can thus transform our dataset to a "down and right" directed grid graph over the matrix entries. The horizontal arc $((u_j, g_i),(u_j, g_{i+1}))$ corresponds to the partition of column g_i such that $DAR_t(x, a_i) = 0$ iff $x < d_j$. The cost of the arc is determined by applying $SCORE_g$ to the model in which this function $DAR_t(*, a_i)$ is used. Vertical arcs have 0 cost. All possible functions are represented as paths from (u^*, g_1) to (u_1, g^*) through this directed acyclic graph, and we can represent the global score of a model by the cost of such path. Furthermore, we can now find the DAR with optimal model score by searching for an optimal path in the graph.

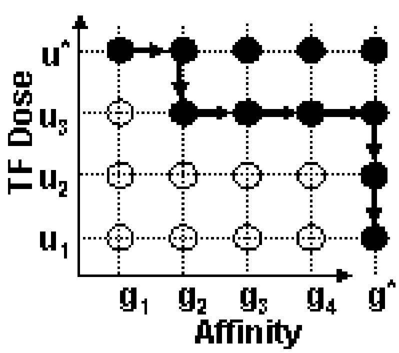


Figure 4: DAR function optimization. A shortest path in the above grid graph is equivalent to an optimal single DAR optimization, nodes on or above the path are assigned with 1 (black), nodes below are assigned with 0 (white).

We have thus shown:

Proposition 1. *The single DAR, binary site activities optimization problem admits a polynomial algorithm provided SCORE is decomposable and polynomial.*

We can extend the algorithm above to an arbitrary number k of site activity levels by an appropriate modification of the graph. The complexity is polynomial in $|V|$ and $|U|^k$, and hence less sensitive to large number of genes. The same can be done for optimizing simultaneously k DAR functions. In summary:

Proposition 2. *DAR optimization is polynomial for a fixed number of site activity levels when optimizing simultaneously any fixed number k of DAR functions, provided SCORE is decomposable and can be calculated in polynomial time.*

To illustrate our algorithm we demonstrate how to calculate arc costs in the case of consistency score. The same can be done with all other common model scoring functions. We assume again binary cardinality and single DAR optimization. Recall that the consistency score for a given set of regulators of a gene [18] simply counts the maximal number of correct predictions. (i.e., it implicitly finds the optimal function and counts how many times it is correct). We assume the entire model is given except for DAR_t and show how to calculate arc costs in column g. For each gene g the partial model induces a partition of the conditions into sets $\{U_i\}$ where in each set the level of activity of all TFs except t is constant. For each such set, we may have several gene expression readings for g. Optimal consistency is achieved when we can use DAR_t to partition each U_i into two subsets with constant g expression. It is now clear how to calculate all arc costs in $O(|U|)$ time. We initialize $U_i^0 = 0$ and $U_i^1 = U_i$ and set $p_i^j = \|\{u \in U_i^1, E_g^u = j\}\|$ and $n_i^j = \|\{u \in U_i^0, E_g^u = j\}\|$. We now iterate over the conditions ordered by the doses of t. Suppose the current condition u_k belongs to the set U_i and $E_{g_i}^{u_k} = \theta$. Then move u_k from U_i^1 to U_i^0 (to reflect a new threshold $D_t^{u_k}$), decrease p_i^θ and increase n_i^θ by 1. The consistency score of the model for dose threshold $D_t^{u_k}$ is $\sum_i(max(p_i^0, p_i^1) + max(n_i^0, n_i^1))$ and can be updated in constant time.

In practical settings, we globally optimize a model by repeatedly applying single DAR optimization to all TFs. The process is monotonically improving and can be terminated when no further improvement is possible. The optimized set of DARs immediately induces a set of CRS functions, since, as mentioned above, the scoring schemes in use implicitly select a CRS as part of the score calculation.

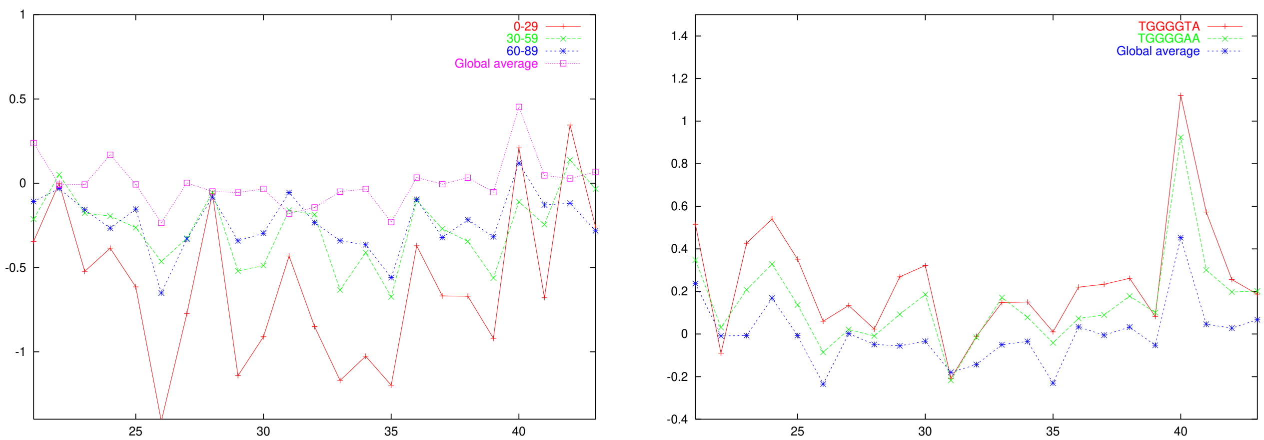


Figure 3: Effects of affinity on activity. X axis - different expression measurements from [8]. Y axis - mean expression. Left: gcn4 bounded genes show variable response as a function of binding strength. We sorted all yeast genes by gcn4 binding affinity from [11]. We collected the first, second and third groups of 30 genes and plotted the mean expression of each group over the different conditions. The magnitude of repression in gcn4 bounded genes depends on gcn4 affinity. The analysis, although ignoring important effects such as combinatorial regulation, strongly supports the hypothesis of monotonic relation between affinity and response in gcn4 regulated genes. Furthermore, we detect multiple levels of response, underlying the importance of using more than two response types. Right: mig1 bounded genes are undetectable in [11] but exhibit similar behavior as seen by consensus based testing of mean expression. We plotted the average expression of yeast genes with the mig1 consensus TGGGGTA and its minor perturbation TGGGGAA. Both sets manifest significant induction of expression, compared to the global mean, but the exact mig1 consensus is responding stronger.

6. INFERRING MOTIFS AND THEIR ACTIVITY

We now turn to the problem of inferring a set of active TFs and their doses. To achieve this goal in the absence of location data or direct dose measurements, we combine expression and sequence data. We shall identify a set of active binding sites and assess their activity at each condition. We can then use these active sites as the set of active TFs and the activity levels as the doses to perform TP optimization. By alternating between TP model optimization and dose optimization we can compensate for biases in the initial TF and dose predictions.

6.1 Active motif discovery

To measure the activity of a binding site motif, we first identify the set of genes that contain it in their promoters. We then evaluate the level of co-expression of that set in the data using a novel scoring method, which takes into account the individual expression distribution of each gene and condition. Motifs are defined in a more descriptive way than the commonly used PSSMs by taking into consideration their location distribution along the promoter. This method is used in a screening procedure that combines exhaustive search for k-mer seeds and their refinement to high-activity motifs. This procedure may be useful independently of the complete TP model inference algorithm.

We begin this section by defining the background model for evaluating sets of genes for co-expression (Section 6.1.1). We then define our extension to PSSM motifs (Section 6.1.2), and show how to assign an activity score and statistical significance to a motif (Section 6.1.3). We give a new algorithm to optimize motifs (Section 6.1.4) and finally show how to combine this algorithm into the motif discovery platform (Section 6.1.5).

6.1.1 A score for co-expressed gene sets

We model gene expression data by generalizing the SAMBA scheme [20]. In that scheme, a bipartite graph represents the entire data set. Vertices correspond to conditions and genes, and edges in the graph indicate expression level change. A subgraph corresponds to a subset of genes and subset of conditions (i.e. bicluster). Weights on the edges and non edges in the graph are used to transform the log likelihood ratio of a bicluster to a sum of weights over the corresponding bicluster edges.

We extend the original model in two ways. First, we consider probabilistic observations and thus probabilistic edges in the graph. Second, we relax the original binary discretization requirement in SAMBA to permit arbitrary cardinality discretization. A more detailed account on the extended model can be found in [19], and here we only provide the key ideas and necessary notation.

We are given a set of genes I, a set of conditions J and a real valued expression matrix $E = r_{ij}$. We introduce a set C of response types and translation functions $\phi_c : R \rightarrow [0, 1]$ for each $c \in C$. These functions assign to each expression value r the probability that the response c occurred given the observed expression r. Using ϕ_c, we transform E into a set of matrices q^c_{ij}, $c \in C$ by setting $q^c_{ij} = \phi_c(r_{ij})$. Note that response types need not be disjoint, so $\sum_c \phi_c(x)$ is not constrained to be 1. For example, response types may be weak to high, medium to high and high, and 2-fold expression change can be translated into 100% weak to high, 50% medium to high and 10% strong response.

We now construct the **expression response graph** as a union of bipartite graphs with a common gene side, $\cup_{c \in C} G_c$, $G_c = (U_c, V, E_c)$. Each G_c is created as in the binary discretization case, but we attach the value q^c_{ij} to an edge (i, j) and interpret it as the probability of the response type given the observed expression. To evaluate the significance of a bicluster, we compare its likelihood under a background model and under a co-expression model. The background model is constructed for each G_c as in [20], using degree preserving random graphs, and assigning the background probability p^c_{ij} to the edge between condition i at response type c and gene j. The co-expression model is a random subgraph with fixed edge probabilities p. The log likelihood ratio of a bicluster under the two models can be written as a sum of vertex pair weights. For an edge we have $\log \frac{p}{p^c_{ij}}$ and for a non edge we write $\log \frac{1-p}{1-p^c_{ij}}$. For the pair (i^c, j), one gets the weight:

$$w^c_{ij} = q^c_{ij} \log \frac{p}{p_{ij}} + (1 - q^c_{ij}) \log \frac{1-p}{1-p_{ij}} \tag{1}$$

The coefficients q^c_{ij} and $1 - q^c_{ij}$ reflect the relative certainty for the edge (i^c, j). We may regard w^c_{ij} as the expected likelihood ratio when edges in the graph are sampled with probabilities q^c_{ij} (not to be mixed with the background model probabilities p_{ij}).

6.1.2 A localized PSSM model

A position specific substitution matrix (PSSM) is a standard way for representing DNA motifs. A PSSM P is a vector of distributions over $ACGT$ denoted $P[0 \ldots l] : \{ACGT\} \rightarrow [0, 1]$. In practice, many binding sites motifs tend to concentrate in particular regions within the promoter. To model this phenomenon, we extend the standard PSSM definition by adding to it a distribution of its location: A Localized PSSM (LPSSM) is a PSSM with an additional location distribution p_l. The likelihood of an LPSSM match with a sequence s in location j is simply the product of profile probability and location probability: $Pr(P, s, j) = P_l(j) \prod_{0 \leq i \leq l} P(i, s[i + j])$.

The matching likelihood of a string s and an LPSSM P is $ML(P, s) = \max_j Pr(P, s, j)$. For each gene, we extract its promoter by taking a fixed-length sequence preceding the gene's start codon. Typical promoter lengths for yeast are 500-1000 bases. The *matching likelihood* of P with gene g, denoted $ML(P, g)$ is the $ML(P, s)$ where s is the gene's promoter sequence. Denote the set of genes with matching likelihood exceeding a threshold T by $J^P_T = \{g \in V \mid ML(P, g) \geq T\}$. PSSMs (and LPSSMs in particular) can be generalized to consider non directional hits (5' or 3') or to take into account multiple hits. We omit such details for simplicity but note that in some cases these generalizations are very important for successful modeling of TF-gene affinity.

6.1.3 Raw activity score and its approximated significance

We can use the weighted bipartite graph constructed above to score sets of genes for coregulation. This is done by collecting contributions from all conditions nodes with high total edge weight to the gene set. By using likelihood ratios w.r.t. a solid background model as our edge weights, we are able to obtain better sensitivity than the common models which use mean expression [3, 2].

Formally, we first define the raw activity score of a gene set $J' \in J$ in condition i as:

$$AS(J', i) = \max_{c \in C} \sum_{j \in J'} w^c_{ij} \tag{2}$$

We next assess the AS distribution for each condition and gene set size by sampling a large number of random gene sets. Given the empirical distribution we define the activity score approximated p-value (ASAP) score of a gene set J' and a condition i as:

$$ASAP(J', i) = -log(Pr(AS(V', i) \geq AS(J', i))) \tag{3}$$

where V' is a random gene subset of size $|J'|$. Using the empirical distribution is important since AS distributions (especially in smaller data sets and for low response conditions) are non normal.

We finally define the ASAP of an LPSSM P as the maximum ASAP score of the set of genes exceeding a matching likelihood threshold, taken over all possible thresholds:

$$ASAP(P) = \max_{0 \leq T \leq 1} \sum_{i \in I} ASAP(J^P_T, i) \tag{4}$$

6.1.4 LPSSM optimization algorithm

We shall next present an algorithm for activity optimization of an LPSSM. The algorithm is an EM-like heuristic with alternating phases of PSSM update and gene set optimization. Similar motif refinement procedures were previously used, e.g., in multiple sequence alignment [1] and motif finding. To the best of our knowledge, this is the first use of such procedure which combines expression and sequence data. We start from an initial LPSSM P_0 (possibly random). The first phase computes the $ASAP$ score of P_0 and determines the threshold T_0 for which $\sum_{i \in I} ASAP(J^{P_0}_{T_0}, i)$ is optimal. For each gene $j \in J^{P_0}_{T_0}$ we calculate its score contribution as $x^0_j = ASAP(J^{P_0}_{T_0}) - ASAP(J^{P_0}_{T_0} \setminus j)$. In the second phase we find for each gene $j \in J^{P_0}_{T_0}$ with positive x^0_j the maximal likelihood matching position of P_0 in the promoter s_j and use these positions as a gap-less alignment from which we extract the next iteration profile denoted as P_1. P_1 is formed by weighted counting in which gene j has a weight x^0_j, so higher activity genes have larger effect on the new PSSM. We use the ML location to recreate the position distribution (we add a Gaussian around each position, weighted by x^0_j). We continue iterating the two phases until $ASAP(P_k)$ does not improve.

6.1.5 A combined approach

We use the ASAP-EM algorithm as a subroutine in our platform for active motif discovery. We combine an exhaustive search for active k-mers and subsequent optimization of the highest scoring seeds. The combinatorial search examines all short DNA sequences with possible gaps, scans the entire set of promoters with the motif, extracts a set of genes and scores them using the ASAP scheme. We further constrain matches to location windows of different sizes and positions. We use a precomputed hash of short k-mers matches to speed up performance. The entire workflow is as follows:

1. Generate the expression weighted bipartite graph, sample random gene sets, assess AS distribution
2. Exhaustively screen all gapped k-mers in each window
3. Use ASAP-EM to optimize all k-mers with ASAP score above the random level
4. Cluster similar LPSSMs and output a concise set of motifs

6.2 Extracting affinity and activity

The outcome of the motif discovery process is a set of active LPSSMs which we can use as the set of active TFs in the TP model reconstruction algorithm. We can use the LPSSM match scores as the TF-gene affinities and the value $ASAP(J^{P_n}_{T_n}, j)$ as the TF dose in condition j. Using artificial computed values for concrete physical quantities such as binding affinity or protein abundance is rather arbitrary, but since we employ these values solely in comparisons we may still use these approximations as rough relative measures for our missing values. When direct measurements of affinities or protein abundances are available, it is of course preferable to incorporate them directly. For example, currently available DNA binding assays generate better estimates to TF affinities than consensus matching or even PSSM likelihoods (see Figure 5).

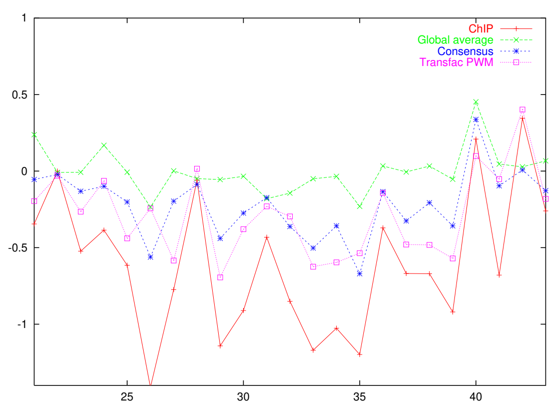


Figure 5: Different sources for TF affinities. We plot the mean expression of several groups of genes in galactose related conditions from [8]. X axis: different conditions. Y axis: Mean log expression ratio. The ChIP group consists of the 30 top bounded genes from a gcn4 location profile taken from [11]. The Transfac PSSM group consists of the 30 genes with top matching likelihood of their promoter to TRANSFAC PSSM M00038. The consensus group consists of all yeast genes with the exact motif TGACTCA in their promoter. Though all groups manifest significant co-expression compared to the global mean (implying that all sources of information may be use as affinities to some extent), the direct physical measurements outperform other methods.

6.3 Tuning TF activity approximations

The initial assessments of TF activities may be strongly biased by effects of combinatorial regulation. For example, whenever a TF is positively regulating a set of genes such that part of the set is also negatively regulated by another TF, we shall assign lower activity to the positive TF in cases where the negative TF is in action. We try to overcome such misleading initial values by tuning the TF activities given an optimized TP model. The TF activity optimization problem is defined by a model M, input expression profiles E^u_g and a score function f. The goal is to find TF doses D^u_t optimizing $f(M)$. This is an NP-hard problem (reduction from SAT, details omitted) but we can heuristically approach it by dealing with one TF and one condition at a time. The optimal value of D^u_t, fixing the rest of the parameters, is derived by maximizing the combination of score contributions from all the models genes: $argmax_{x \in \{D^{u'}_t, u' \neq u\}} \sum_g SCORE_g(M, D^u_t = x)$.

7. RESULTS

We tested the different components of our model inference framework by applying it to data from carbohydrate metabolism-related experiments. 61 relevant expression profiles were selected from [8, 5, 7, 12]. The combined expression matrix was translated into 9 different response type matrices. To evaluate Activity Score distributions, we set the model edge probability p to 0.6, and randomized 10000 gene sets of each size from 1 to 2000 genes. We used 500 bases upstream the start codon as the promoters.

7.1 Comparing ASAP to mean expression methods

To compare the performance of our ASAP scheme to the mean expression methods we have used the 32 yeast PSSMs from TRANSFAC [22] version 5.1. The mean-based score of a set of genes $G \in V$ was defined (following [9]) as follows: each condition in the expression matrix was normalized to mean 0 and standard deviation 1. Let E be the normalized matrix, let s^G_c be the mean expression on the condition c of genes in G. Let s^G be the mean value of S^G_c over all conditions $c \in C$.

The collection of conditions with significant means is then extracted as $S = \{c \in U | |s^G_c - s^G| \geq \theta\sigma\}$ where $\sigma = \frac{1}{\sqrt{|G|}}$ and θ is a parameter. The mean score is finally defined as:

$$MeanScore(G) = \sum_{c \in S} (|s^G_c - s^G|) \tag{5}$$

For each PSSM we ordered all yeast genes by the likelihood of their promoter match to the PSSM. We then computed the score for each set of genes G exceeding a threshold θ on the likelihood for every possible θ, and chose the best score. To determine significance of the scores we used the maximum score obtained over 5000 instances in which the genes' promoters were randomly shuffled. The procedure was repeated for ASAP and for the mean score. As can be seen in Figure 6, both methods identified the GAL4 and RAP1 sites as active but only the ASAP scheme succeeded in identifying the activity of additional transcription factors which are known to be functionally associated with carbohydrate metabolism (MIG1, ADR1 and more).

7.2 Effect of motif localization on activity

We tested the utility of incorporating motif location distribution into the PSSM by re-analyzing TRANSFAC's known PSSMs on the same dataset using different positioning windows. We tested the activity of each PSSM restricted to

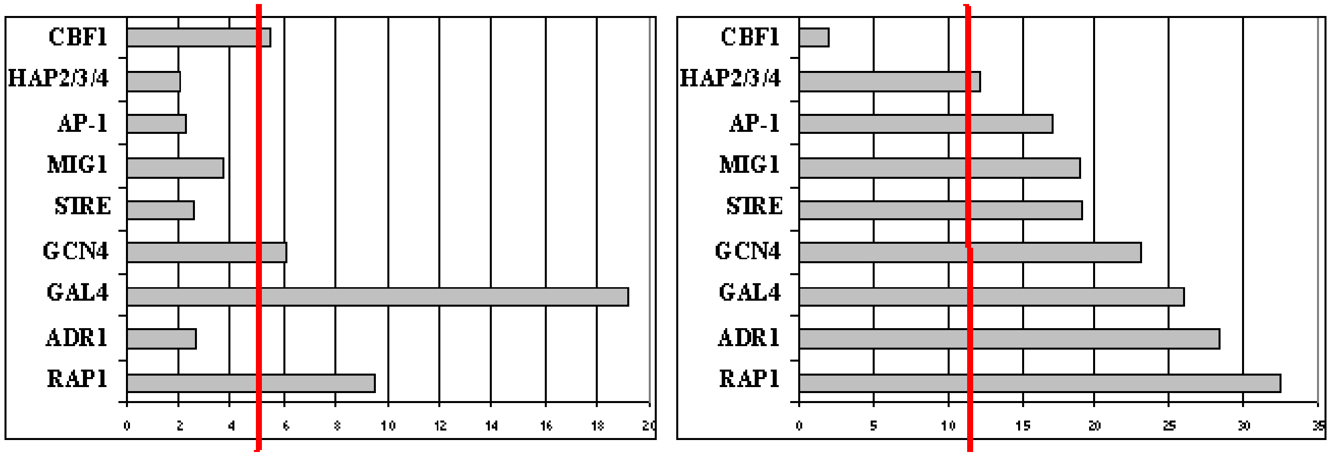


Figure 6: Comparing the ASAP scheme to mean-based scoring. All yeast PSSMs from TRANSFAC were scored against a combined carbohydrate data set. Left: Mean-based scoring. X axis: Mean score. Only RAP1, GAL4 and possibly GCN4 are detectable above the noise level (vertical line). Right: ASAP scoring. X axis: ASAP score. Additional relevant active motifs (ADR1, MIG1, STRE, AP-1) are detected.

windows of 100bp starting at -500 with steps of 50bp and ending at 0. The resulting histograms (Figure 7) show that some of the motifs (but not all) tend to be active when localized to specific positions. In several cases the ASAP of the restricted motif is better than that of the unrestricted one. Using LPSSM can improve the specificity of TF-gene association and refine our understanding of the relation between DNA and TFs.

7.3 Transcription program in the galactose system

To infer a TP for the galactose system we first applied our active sites discovery algorithm to the dataset of [8]. This is a subset of 23 conditions out of the 61 analyzed above. We first repeated the screening with 32 TRANSFAC motifs as in Section 7.1. On this dataset, only GAL4, GCN4, and MIG1 were detectable as significantly active, in accordance with biological literature on the galactose pathway. We then screened all 6-mers with one gap of size 0-12 bases and optimized all hits above the noise level. The active motifs are listed in Table 1. We were able to accurately rediscover de-novo the known GAL4, GCN4 and ADR1/MIG1 motifs. We also discovered 3 additional motifs (CANCCCC, TT(9N)CCCC, CCG(5N)CCG) which achieve scores above or equal to ADR1/MIG1. Note that this test assumes absolutely no knowledge on possible transcription factor motifs and scans a space of about 240,000 possibilities, so the high specificity is encouraging.

We have used the six significantly active motifs discovered above to build a TP model explaining the gene expression of all the variable genes in the data set. We selected 140 genes with significant expression change in the data set and applied our model reconstruction algorithm using as initial values for TF doses and affinities the results of the above procedure. For model optimization, site activity and gene expression were assumed to have three possible values. The optimized DARs are shown in Figure 8. We tested the score contribution of each of the six putative motifs by comparing for each ones the global model score with and without the motif. Results show (see Table 1) that the known motifs are scored higher than the putative ones, and that at least one of the putative motifs does not contribute to the model score. We conclude that two of the putative motifs (U1,U3) may be promising targets for further experimentation.

8. ACKNOWLEDGMENTS

We thank Irit Gat-Viks and Roded Sharan for helpful discussions. The research was supported in part by a pilot grant from the McDonnell Foundation and by Israeli Science Foundation (grant no. 309/02).

9. REFERENCES

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[2] HJ. Bussemaker, H. Li, and ED. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–71, 2001.

[3] DY. Chiang, PO. Brown, and MB. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, 17(Suppl 1):S49–55, 2001.

[4] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *JCB*, 7(3-4):601–20, 2000.

[5] A. P. Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241–57, 2000.

[6] I. Gat-Viks and R. Shamir. Canalyzing functions and scoring functions in genetic networks. Submitted for publication, 2002.

[7] TR. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–26, 2000.

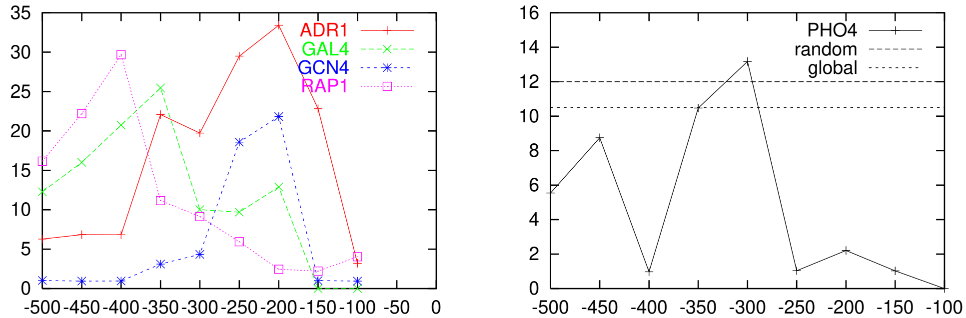


Figure 7: Effect of motif position on activity. Plots of activity as a function of the range window. X axis: offset of the range window from promoter TSS, Y axis: ASAP score. The left figure shows the behavior of several known motifs: RAP1 is strongly biased to -400 bases from the TSS, GAL4 to -300 bases, GCN4 to -200 bases. ADR1, on the other hand, is active in the range -350 to -100. The right figure shows the global and localized score for PHO4. The peak of the local score at -300 bases exceeds the noise level, while the global score is below it.

LPSSM Consensus	ASAP Score	Remark	Gain
CGG(11N)CCG	30.8	GAL4 consensus	5.7636
TGACTCAWT	22.57	GCN4 consensus	4.2718
TGGGGTA	22.03	ADR1/MIG1 consensus	4.7752
CANCCCC	26.92	Unknown, denoted U1	3.7714
CCG(5N)CCG	26.6	Unknown, denoted U2	-0.8498
TT(9N)CCCC	22.03	Unknown, denoted U3	2.8597

Table 1: De-novo identification of active motifs in the galactose system. For each detected active site we list both the initial ASAP score and the global TP model score gain derived by comparing models score with and without each TF. Negative scores are possible due to rounding errors or sub-optimality of the solution. Note that TFs that had high activity scores (e.g., U2) may be completely redundant and that this was discovered due to the additional structure imposed on the model.

[8] T. Ideker, J.A. Thorsson, V. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 291:929–34, 2001.

[9] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–7, 2002.

[10] VR. Iyer, CE. Horak, CS. Scafe, D. Botstein, M. Snyder, and PO. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–8, 2001.

[11] T.I. Lee et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[12] M.C. Lopez and H.V. Baker. Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns. *J Bacteriol*, 182(17):4970–8, 2002.

[13] D. Peer, A. Regev, and A. Tanay. A fast and robust method to infer and characterize an active regulator set for molecular pathways. *Bioinformatics*, 18(Suppl 1):S258–67, 2002.

[14] Y. Pilpel, P. Sudarsanam, and GM. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29(2):153–9, 2001.

[15] B. Ren et al. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–9, 2000.

[16] E. Segal, Y. Barash, N. Friedman, and D. Koller. From promoter sequence to expression: A probabilistic framework. In *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology (RECOMB 2002)*, 2002.

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

[18] A. Tanay and R. Shamir. Computational expansion of genetic networks. *Bioinformatics*, 17(Suppl 1):S270–8, 2001.

[19] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Integrated analysis of diverse genomic data. Submitted for publication, 2003.

[20] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(Suppl 1):S136–44, 2002.

[21] S. Tavazoie, JD. Hughes, MJ. Campbell, RJ. Cho, and GM. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

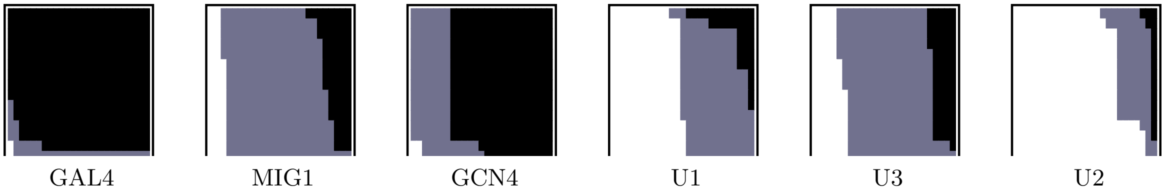


Figure 8: TP model DAR's for the galactose TFs. X axis: ranked gene affinity. Y axis: ranked TF dose. The colors represent the response (or derived TF activity). Darker color indicates stronger response.

[22] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, 29:281–3, 2001.