Multi-level Modeling and Inference of Transcription Regulation

Amos Tanay and Ron Shamir

School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel.

Tel: +972-3-6405394. Fax: +972-3-6409357.

Email: {amos,rshamir}@tau.ac.il.

Abstract

The understanding of transcription regulation is a major goal of today's biology. The challenge is to utilize diverse high-throughput data in order to infer mechanistic models of transcription control. We propose a new model which integrates transcription factor-gene affinity, protein abundance and gene expression profiles. The model provides detailed, yet computationally tractable description of the relations between transcription factors, their binding sites at genes promoters and the combinatorial regulation of transcription. At the core, our model manipulates dose-affinity-response functions that associate transcription factor concentrations and transcription factor-DNA affinities to determine the rate of transcription factor-DNA reactions. We study computational problems that arise in optimizing such models and develop polynomial algorithms for certain problems. We show how to assess missing values (notably protein abundance) and describe a novel framework to infer models from currently available data. On budding yeast carbohydrate metabolism data, our results demonstrate the sensitivity and specificity of the approach. They also suggest new active binding sites and a regulation model for the transcription program of the galactose system.

1 Introduction

The mechanisms of transcription control are of the most studied in biology. However, the way in which transcription factors induce and repress transcription is still poorly understood. 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. In this paper we introduce a new detailed, multi-level computational model of transcription regulation, and show how to use it to enhance our understanding of transcription programs.

A transcription program can be described by the following 3-level mechanistic model (compare Figure 6). First, certain transcription factors (TFs)[1] attain specific concentrations in specific post translational conformations. Second, as a result of sequence signals in the target gene's promoter and of the concentrations of certain TFs, the DNA in the proximity of that gene undergoes chromatin modifications and targeted TF binding. We say that the promoter binding sites become *active* in such cases. Third, the spatial organization of the promoter's binding sites and the physical properties of the active TFs induce a combinatorial regulation scheme which results in assembly of the transcription apparatus and thereby regulate the rate of transcription.

Several new experimental techniques provide data on different parts of this regulatory mechanism. Gene expression microarrays [21, 9] can measure genome-wide mRNA levels in many different conditions. High throughput DNA location analysis [19, 13, 14] measures genome-wide TF-gene affinities. Large scale proteomics [26] is rapidly developing, and will soon provide information on the level of the TF proteins themselves. It will then supply the last major missing type 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 is currently not sufficiently accurate. The most common method to analyze transcription programs today is to cluster the genes according to their expression profiles and search for enriched DNA motifs in the promoters of co-clustered genes [21, 25]. This basic method was extended in several directions: Pilpel et al. [18] searched for combinations of known motifs with significant co-expression in order to discover combinatorial transcription control patterns which are beyond the reach of clustering. Bussemaker et al. [2] used a linear model

[1] Abbreviations: DAR: dose-affinity response, LPWM: localized positional weight matrix, PWM: positional weight matrix, TF: transcription factor, TSS: transcription start site

to associate motifs and expression and applied a greedy algorithm for de-novo construction of such model from expression data. Segal et al. [20] presented a Bayesian graphical model integrating position weight matrices (PWMs) 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 levels to model both transcription levels and the abundance of the proteins (TFs) involved. In contrast, we model genes and TFs as separate entities. This improves the ability to model post-translational effects and to overcome poor correlation between mRNA and protein levels. Our model covers all three levels of transcription regulation: a) TF concentrations (or doses), b) TF-gene affinity and a resulting binding site activity, and c) the combinatorial regulation of genes. The model defines functions that determine binding site activity based on TF doses and TF-gene affinity (these are called dose-affinity-response functions, and are specific for each TF). It then uses the site activities as the inputs to combinatorial logic schemes that predict the rate of gene expression given TF concentrations and TF-gene affinities. A unique feature of our method is the global integration of TF doses and transcription rate via the dose-affinity-response functions: We use sequence and expression data from numerous of genes in order to infer 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 models from experimental data we must fill in unmeasured values (e.g., TF doses) and reconstruct the functions that define our model. We first study the case where all values are provided, including TF doses, TF-gene affinities and expression profiles. Our core algorithm then efficiently seeks a model with optimal transcription predictions given the inputs. In the more realistic scenario, 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 a method that can identify active DNA motifs and assess their relative activity across the given experimental conditions. Assuming each of the functionally important TFs in the system under study is characterized by a typical promoter motif, we use the site activity values as initial TF doses for the model optimization algorithm. We then apply an iterative algorithm which alternates between optimizing the model and the doses, and thus corrects biases in the initial estimate of TF doses.

Our framework for active binding site discovery can also be used as an independent gene expression analysis tool, improving, e.g., [2] with novel algorithms and model improvements. We extend the notion of a PWM to include information on the distribution of site locations relative to the TSS. This extended motif is called localized PWM (or LPWM). For gene expression data, we use probabilistic models similar to [24] to define a likelihood ratio based scoring scheme for assessing motif activity. Our score improves over [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 this scoring scheme in a EM-like motif optimization algorithm that is capable of tuning LPWMs for optimal activity.

We tested our methods on gene expression data related to yeast carbohydrate metabolism. Applying the active motif discovery framework, and using no prior knowledge on the TFs, we were able to identify TFs that are known to be active in the system, as well as several new putative active motifs. We compared our motif scoring scheme to the simple mean-based scores from [2, 3, 12] and showed that our scheme has improved sensitivity and specificity. Finally, we used our algorithm to generate a model which predicts the transcription of all significantly changing genes in a set of galactose-related conditions. We show that our optimized transcription program leads to correct and highly specific results for known TFs. We also predict the activity of three previously uncharacterized binding sites involved in regulation under galactose-related genetic perturbations.

A preliminary version containing some of the results of this paper has appeared in [23].

2 A model for transcription programs

We first define our model for transcriptional programs. The model is a generalization of simple genetic networks [15, 6]. In such networks, one studies a set of genes using a directed graph. The topology encodes the regulatory relations among genes, namely, the set S of genes with arcs into g are the *regulators* of g. g is also called the *regulatee* of S. For each regulatee one defines a function (logic) which determines the rate of transcription of the regulatee given the levels of its regulators. Basic network models thus make two critical assumptions on the regulatory system: a) the regulators states can be represented using the expression rate of the genes encoding them, and b) the relation between a potential regulator and a regulatee is a binary attribute (either there is an arc or there is none). The model we shall define and use below relaxes both assumptions in an attempt to build a mechanistic model which can follow more faithfully our knowledge on biological regulatory systems.

2.1 The topology

The transcription program model involves entities of three types. *Active TF*s represent transcription factors in their catalytically active from. *TF binding site*s represent loci at genes promoters where TFs bind. *mRNA*s represent the rate of transcription for genes in the system - the end product of the transcription program. We tie the three types of entities using a simple bipartite graph.

Definition 1 *A* **transcription program topology** *is a bipartite graph $M = (T, V, S)$ consisting of a set T of active TFs, a set V of genes and a set S of* **binding sites** *(or edges) connecting TFs and genes. We use the term* **regulators set** *of $g \in V$, or $N(g)$ to denote the set of edges adjacent to g, or the set of TFs adjacent to g, depending on the context.*

Note that we shall use V to denote both the genes and their mRNAs. For convenience, we shall use the term gene throughout for both. In each biological condition, we assume nodes and edges in M attain certain values. Active TFs are assigned **doses** (or concentrations). A binding site has a certain level of **activity** (or reaction) between the corresponding TF and gene's promoter. Genes have a **rate of transcription** (or of expression). A transcription program model is built over the topology by defining the way gene expression levels are determined once their binding sites activity levels are known, and by specifying how site levels are determined once their associated TFs levels are given.

2.2 Scoring a topology

We first focus on the dependency between sites and genes. We assume that both site activities and transcription rates attain values in a discrete alphabet $\mathcal{C} = \{1 \dots C\}$ and that their regulatory logic is a collection of unconstrained combinatorial functions. Specifically, the regulation of g is assumed to be via a **regulation function** $f_g : \mathcal{C}^{|N(g)|} \to \mathcal{C}$ which determines the transcription rate of g as a function of the activity levels of its regulator sites. The concrete functions will not play a major role in our paper, but their nature is implicitly used in the model. In reality, the levels of sites are hidden variables that will be predicted by our model from TF doses. We shall temporarily assume that sites levels, in addition to gene expression levels, are measured across a set of experimental conditions and given as input. We will show how to score the fit between these input data and the model topology. We will then turn to our main problem, which is the optimization of the regulator functions.

Assume we are given a topology M and a set E of experimental conditions, where each condition $u \in U$ has a vector of gene expression levels e^u. $e^u_g \in \mathcal{C}$ is the expression level of gene g in condition u and $E = \{e^u | u \in U\}$. Suppose we also have the set A of (measured or predicted) activity levels r^u_s of each site $s \in S$ under each condition $u \in U$. A **topology score** $\phi(M, E, A)$ will be a real valued function using M and E to evaluate a site activity matrix A. Put differently, the score should assess the dependencies among site levels and gene expression levels given the topology M. ϕ is closely related to commonly used gene network scoring schemes, but with altered roles for topology and regulators values: Most network schemes score the topology, while in our setting we score the regulator activity values given a fixed topology. We call a score **decomposable** if it can be expressed as a sum of separate contributions from individual genes, each depending only on the value of the gene and its regulators sites, i.e., $\phi(M, E, A) = \sum_{v \in V} \overline{\phi}(r_{N(v)}, e_v)$. Throughout this paper we assume that our score is decomposable.

Denote by n^v_r the number of conditions $u \in U$ in which v's regulator sites $N(v)$ attain the specific combination of activity values r (r is a vector of size $|N(g)|$). Denote by $n^{v,j}_r$ the number of conditions meeting the previous criterion which also have $e^u_v = j$. Also denote by n the number of conditions $|U|$ and by $n^{v,j}$ the number of times e^u_v equals j. We next show how several known network evaluation schemes can be used as topology scores.

- Consistency [22]. This simply sums over all genes the maximal possible number of correct model predictions:

$\sum_{v \in V} \sum_r \max_j n_r^{v,j}$.

- Mutual Information [11, 17]: $\sum_{v \in V} I(r_{N(v)}, e_v)$ where $I(r_{N(v)}, e_v) = -\sum_j \frac{n^{v,j}}{n} log \frac{n^{v,j}}{n} + \sum_r (\frac{n_r^v}{n} log(\frac{n_r^v}{n}) - \sum_j \frac{n_r^{v,j}}{n} log(\frac{n_r^{v,j}}{n}))$.

- Bayesian scores [6]: We can interpret our models probabilistically and attempt at inferring a joint distribution of the sites and genes. In this case, standard Bayesian scores (e.g., BDe) can be naturally applied.

- funcFit, or the mutual information p-value score [8]. The score equals the χ^2 value of the standard mutual information $I(r_{N(v)}, e_v)$. The χ^2 statistic is used with $(C-1)(R-1)$ degrees of freedom, where R equals the number of r vectors for which $n_r^g > 0$.

2.3 The Dose-Affinity-Response model

We now turn to describe the model that ties up TFs and sites. The model computes site activities given TF doses, using additional experimental information on the nature of relations between TFs and promoter binding sites. We denote by α_s the **affinity** of the site interaction represented by the edge s (recall that a binding site in our abstraction is a pair (TF,gene) so the affinity of a site actually refers to the affinity of the TF to the gene's promoter). Affinities are unbounded non negative integers and represent the strength of interaction (or biochemical reaction constant) between the DNA and the TF. Note that if several binding sites are present in the promoter of a certain gene, we will indicate this by increased affinity and not by several edges. We next define the function that computes site activities from TF doses and site affinities.

Definition 2 *A* **Dose-Affinity-Response (DAR) function** *is a non-decreasing function $\delta : Z^+ \times Z^+ \rightarrow \mathcal{C}$ computing for each TF dose and site affinity a site activity level. A* **DAR model** *is a triplet (M, α, Δ) where M is a transcription program topology, α is a set of site affinities α_s for each $s \in S$, and Δ is a set of DAR functions δ_t for each $t \in T$.*

Hence, if TF t has a DAR function δ_t, and $s = (t, g)$ has affinity α_s, then the activity of site s is determined by the dose d of t via $\delta_t(d, \alpha_s)$. Note that the affinity is a constant value per site, while the site's activity will vary (monotonically) with t's dose. The monotonicity of DAR functions means that for a fixed dose, response increases with affinity, and for a fixed affinity, response increases with dose. Suppose we are given a set of conditions U, where for each $u \in U$ we have a vector of gene expressions levels e^u and a vector of TF doses d^u. We can compute site activities r^u using the DAR functions by simple substitution:

$$r_s^u = \delta_{t_s}(d_{t_s}^u, \alpha_s). \tag{1}$$

Any topological score can be applied to the computed activities and the given gene expression to evaluate the fit of a DAR model to the experiments. We can now define our main model optimization problem w.r.t a topology score ϕ:

Definition 3 The DAR optimization problem. *We are given a transcription program topology M, site affinities α_s, and a set of experiments $U = (e_v^u, d_t^u)$. The goal is to find a set of DAR functions, giving rise to site activities $A = \{r_s^u | u \in U, s \in S\}$, such that the topology score $\phi(M, E, A)$ is optimized.*

A **transcription program model** is a DAR model along with concrete regulation functions for all regulatees. Given a DAR model and a set of experiments, it is easy to complete it to a transcription program model using majority voting: For the regulatee v and the combination r of activity levels of the sites $N(v)$, $f_v(r)$ is the most often observed expression rate of v when the inputs combination is r. This is the standard way by which generative gene network models are inferred once their topology is determined [17]. Note that the resulting functions will often be incomplete, as not all regulators combinations will be available. In what follow, we shall work directly with DAR models, using this natural completion to a full model only when analyzing our results.

2.4 Characteristics of the model

The DAR models we have introduced above serve several purposes and overcome some of the shortcomings of existing network models. Most importantly, we replace the "black box" used before to express the regulatory relations among genes with a detailed mechanistic model that incorporates new direct experimental data (e.g., TF-gene affinities) when available. The price we have to pay is the addition of hidden variables that increase the model's degrees of freedom and may lead to over-fitting. We control this effect by adding structure to our models, namely, constraints on the topology and restricted classes of DAR functions. A main structural constraint is the utilization of monotone DAR functions only. True active TFs tend to behave monotonically, while artifacts are unlikely to manifest such consistent pattern. We have observed this phenomenon in several cases, as shown in Figure 6. The integration of gene expression and TF affinities via the dose-affinity-response function thus models current biological understanding.

Most of the existing network inference algorithms [6, 22, 11, 17] employ discrete models. Such models are simpler computationally and enable easier statistical interpretation than in continuous (e.g., kinetic) models. High throughput experiments are inherently noisy, and thus they can rarely be translated to exact physical units. The experiments usually provide a normalized quantity which is only indicative of the actual physical quantity sought. For example, gene expression arrays results cannot be represented in absolute units of mRNA molecules per cell. Discrete models pre-process experimental information to transform a set of incomparable quantities into a small discrete set of semantically meaningful values in a small alphabet. The discretization process is known to be problematic in several respects: arbitrary thresholds must be set in advance and critical information may be lost. Our model is a novel compromise between discrete and continuous models: Gene expression is discretized as before, but TF doses and affinities are used as ranked integers. The assumption is that although we cannot interpret most experimental results in exact physical units, we can quite safely compare them. We thus assume the DAR functions take ranked integers as inputs and output a discrete binding site activity. Using this approach, we avoid the need to arbitrarily determine a TF affinity threshold that would simplistically divide the genes into two sets of "regulated" and "non regulated" ones. The representation of TF doses as ordered values enables their estimation using statistical tests for hypotheses like "a TF dose in condition X is larger than its dose in condition Y". A similar approach to gene expression may be exploited in future work.

Note that our transcription program model is acyclic. Indeed, unlike some gene networks, we do not aim to describe a time dependent dynamical behavior. Instead we focus on steady state analysis but aim to provide finer, mechanistic description of transcription control.

3 Model optimization

The monotonicity we impose on DAR functions is not only a biologically meaningful and powerful constraint, but also can be turned into an algorithmic advantage and enable efficient algorithms for the solution of the DAR optimization problem. We will deal with the case in which we optimize the DAR of one TF while fixing all others. We represent the dose-affinity plain as a matrix with a column for each affinity value and a row for each dose value (see Figure 6).

Our algorithm will represent single DAR optimization as a longest path problem in an appropriate grid graph built over that matrix. The idea, in the simple case where site activities are binary, is that since DARs are assumed to be monotone, we only need to determine one dose threshold for each affinity level (or vise versa) and ensure those thresholds are non-increasing with affinity. The threshold decision at each affinity level can be scored separately (since the topology score is decomposable), and the total score for a DAR function will equal the sum of contributions from each affinity level. We next define our algorithm formally.

Proposition 1 *The optimization of a single DAR function with k site activity levels can be done in $O(vu^k(log(vu^k)+k)) + O(vu^k\Phi)$ time, where v is the number of genes, u is the number of conditions, and Φ is the time complexity of $\overline{\phi}$ computation. In particular, the problem is polynomial for fixed k.*

Proof: Assume first that $k = 2$. We have a partial DAR model (M, α, Δ') where Δ' defines all DAR functions except δ_t which is the target of our optimization. Let $g_1 \dots g_n$ be the genes regulated by t via sites $s_1 \dots s_n$ respectively. We assume the indices are sorted by increasing site affinities $a_1 = \alpha_{s_1} < \dots < a_n = \alpha_{s_n}$. Let $u_1 \dots u_{|U|}$ be the set of experiments sorted by increasing doses of t, $d_t^1 < \dots < d_t^{|U|}$. For simplicity, we assume that all values (a_i, d_i) are distinct. We add an artificial condition $u_{|U|+1}$ and an artificial gene g_{n+1}. We build a grid digraph on the pairs (u_j, g_i) and add horizontal arcs $((u_j, g_i), (u_j, g_{i+1}))$ and vertical arcs $((u_j, g_i), (u_{j-1}, g_i))$. The horizontal arc from (u_j, g_i) to (u_j, g_{i+1}) represents the determination of δ_t in the i'th affinity level $\delta(*, a_i)$. The function is set to 0 for doses strictly smaller than d_j and 1 otherwise. Given this dose threshold, we can determine the activity vector for the site s_i and may calculate $\overline{\phi}(r_{N(g_i)}, e_{g_i})$. This score will be used as the horizontal arc cost. We set the cost of vertical arcs to 0. It is now clear that directed paths in the matrix induce monotone DAR functions (and that different DARs induce distinct paths). Moreover, for each function, the score of the model with that DAR function equals the sum of arc costs along the path that induces it. We find the optimal DAR function by solving a longest path problem in the directed acyclic graph.

To handle $k > 2$ activity levels we shall use a similar construction in higher dimension. A monotone DAR function with k activity levels divides each column in the dose-affinity plain into k segments (some possibly empty). We uniquely represent this sub-division by a non decreasing sequence $h_1 \leq \dots \leq h_{k-1}$ where δ_t attains level i in the dose range $[h_i, h_{i+1})$ (we set $h_0 = u_1, h_k = u_{|U|+1}$). To fully define a monotone DAR function, we must determine for each affinity level a_j a sub-division $h_1^j \dots h_{k-1}^j$ such that $h_i^j \geq h_i^{j'}$ whenever $j < j'$. We thus define a digraph on pairs (h, a_i) where h is any sub-division and a_i an affinity level. We add arcs $(((h_1, \ldots, h_{k-1}), a_l), ((h'_1, \ldots, h'_{k-1}, a_l))$ whenever $h_i = h'_i$ for all i except one index j for which $h_j = h'_j + 1$, and we set their costs to 0. These arcs are analogous to the vertical arcs in the binary case. We also add arcs $((h, a_i), (h, a_{i+1}))$ and set their costs to $\overline{\phi}(r_{N(g_i)}, e_{g_i})$ using the DAR function $\delta(*, a_i)$ induced by the subdivision h to determine site activities $r_{N(g_i)}$. We can again optimize the DAR function using the longest path algorithm. For the implementation we may use a Fibonacci heap based algorithm [5] which requires $O(M + NlogN)$ time on a graph with N vertices and M arcs. In our case $N = O(vu^k), M = O(kvu^k)$. We also have to precompute all arc costs by applying $\overline{\phi}$ once for each node in $O(N\Phi)$. The total time complexity follows. ■

Note that in practice the number of activity levels is small (we used $k = 3$ in our analysis in Section 5), so the exponential dependence of the complexity on k is not a practical obstacle. However, optimizing several DAR functions simultaneously is NP-hard, as we now show.

Proposition 2 *DAR optimization is NP-hard when optimizing simultaneously an arbitrary number of TFs, if consistency is used as the topology score. This is true even if the set of regulators of each gene is of size 3.*

Proof: We will show that the DAR problem (in decision version) is NP complete using a reduction from 3SAT. We are given an instance I of 3SAT with variables $x_1, \ldots x_n$ and clauses $c_1, \ldots c_m$. Let $c_i = (z_{i1} \vee z_{i2} \vee z_{i3})$, and let the variables that correspond to these literals be x_{i1}, x_{i2} and x_{i3}, respectively. We construct a topology with a TF for each variable and a gene for each clause. For each i, we add sites between c_i and x_{ik} for $k = 1, 2, 3$. All sites have the same affinity. We now construct the set of experiments. We use three dose levels, denoted by 0, $\frac{1}{2}$, and 1 and binary expression levels 0 and 1. For each clause c_i we add eight experiments in which all doses are set to 0 except for the doses of x_{i1}, x_{i2} and x_{i3}. Those are set so that each of the eight possible 0/1 dose combinations is covered once. We have $8m$ such experiments in total, with the property that for each truth assignment of each clause, there exists at least one experiment in which the doses of the clause variables coincide with their truth assignment values. The gene expression rates at the $8m$ experiments are defined using the clauses, as follows: For each of the eight experiments of each c_i, the expression rate of each gene c_j is set to the truth value of clause c_j when substituting the dose values (0 as false, 1 as true) of x_{j1}, x_{j2} and x_{j3} in that experiment into c_j. Precisely, let u be an experiment with doses $d_1^u, \ldots, d_n^u$. For clause, set $e_i^u = 1$ iff $(d_{i1}^u \vee d_{i2}^u \vee d_{i3}^u) = TRUE$. Finally, add one experiment, denoted by

u', in which all doses equal $\frac{1}{2}$ and all expression rates equal 1. The total number of experiments is thus $8m+1$. The required bound for the total score is $(8m+1)n$. Note that this score requires perfect consistency. This concludes our reduction which is evidently polynomial. We claim that the instance i is satisfiable iff the DAR optimization problem has solution with consistency score $(8m+1)n$.

Assume first I is satisfiable with values $x_i = w_i$. Define a DAR solution in which $\delta_{x_i}(0) = 0, \delta_{x_i}(1) = 1$ and $\delta_{x_i}(\frac{1}{2}) = w_i$ for all i. Clearly all functions are monotone. We claim that the resulting consistency score is optimal. To see this we should prove that for each two experiments and each gene, identical site activities imply identical expression rates. This is true by construction for all pairs that do not include u', since their expression rates were determined by the clause in a consistent way. For a pair of experiments that includes u', we must ensure that for each gene, all experiments with site activities of input TFs that equals $\delta_{x_i}(\frac{1}{2})$ have expression level 1 (since in u' all expression rates were set to 1). The latter is true since $\delta_{x_i}(\frac{1}{2})$ were set to a satisfying truth assignment w_i.

To show the other direction, assume we have a solution for the DAR problem with $\phi = (8m+1)n$. In particular, this solution specifies values $\delta_{x_i}(\frac{1}{2})$ for all x_is. We shall use these values as our truth assignment. To see why this assignment satisfies all clauses, observe that the experiment u' gives rise for each gene to a specific site activity combination which is also present in some other experiment u'' (since we have covered, for each gene, all possible dose combinations). Since the expression value of all genes in u' is 1 and expression values in other experiments are determined by the clauses, consistency is possible only if $\delta_{x_i}(\frac{1}{2})$ satisfies all clauses. This conclude our proof. ∎

The proof above would apply to another topology score if the optimal DAR solution in the reduction under that score is the same as for the consistency score. This is indeed the case for the three other scores defined in the previous section, and so the hardness result holds for those scores.

In practical settings, a heuristic to globally optimize a model is to repeatedly apply the polynomial algorithm for single DAR optimization to one TF at a time. We start with some arbitrary set of DAR functions and repeatedly select one TF, reoptimize its DAR function and add it to the current set of functions. Since single DAR optimization is solved to optimality, the new function obtained must have a score that is equal or higher than that of the previous one. Hence, the whole process is monotonically improving, and convergence to a local optimum is guaranteed. We terminate the heuristic when the score converges. In the biological analysis reported below, our implementation converges within seconds.

4 Estimating hidden variables

The algorithms described in the previous section rely on the knowledge of the active TF doses - the concentration of TFs in their catalytically active conformations. Since detailed measurements on protein abundance and conformation are not available yet, we must revert to estimation of TF doses. We do so by combining expression and sequence data. A common working assumption (see, e.g., [2]) is that TFs have specific DNA binding motifs, and that we can estimate the active TF doses by analyzing the expression of genes with such motifs in their promoters. We show below how to use this assumption to generate initial estimation of the TF doses. We also show how to rectify initial bias in dose estimates, by reevaluating the doses with respect to a given transcription program model, and, in later stages of the procedure, by alternating between model optimization and dose optimization.

In Section 4.1 we introduce our extension to motif models, and then show in Section 4.2 how to assign an activity value and statistical significance to such motifs. Section 4.3 describes a new algorithm to screen and optimize active motifs . Finally, Section 4.4 discusses the dose optimization problem, when a model is given.

4.1 A location-dependent motif model

A position weight matrix (PWM) is a standard way for representing DNA motifs (cf. [4]). A PWM P is a vector of distributions over $ACGT$. We denote by $P(i,t)$ the probability of observing character t in position i of the motif. In practice, many binding sites motifs tend to concentrate in particular regions within the promoter. To model this phenomenon, we introduce a distribution of locations for the motif: A *localized PWM* (LPWM) is a PWM with an additional location distribution p_l, where $p_l(j)$ is the probability of having the motif starting at position j in the target promoter. The likelihood of an LPWM 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 LPWM 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 $ML(P,s)$ where s is the gene's promoter sequence. Denote the set of genes with matching likelihood exceeding a threshold T by $J_T^P = \{g \in V \ | ML(P,g) \geq T\}$. In practice, one often has to generalize PWMs (and LPWMs in particular) to consider non-directional hits (5' or 3') and to take into account multiple hits. We omit such details for simplicity.

4.2 Estimating motif activity

The standard way of assessing the activity of a putative motif given a normalized expression profile is to calculate the Z score of the mean expression taken over all genes with the motif in their promoter [2, 3]. Here, we improve this strategy by a) replacing the normalization procedure by calculation of log likelihood ratios using the random graph models introduced in SAMBA [24] and b) assessing the p-value of the log-likelihood ratio in order to improve the statistical significance of obtained motif activities.

For the sake of our discussion here one can view the SAMBA approach as a method for replacing the raw value in the genes- by-conditions expression matrix by new weights. The method uses a positive model and a null model. The former assigns probabilities to sets of co-expressed genes (as manifest by a bicluster). The null model, which corresponds to uncorrelated, random sets of genes, takes into account the prevalence of each gene's (and condition's) expression. Thus, the presence of a commonly-expressed gene in a set would be more probable under the null model than that of a rarely expressed gene. The SAMBA weights are defined so that the sum of weights of a set of genes S in a given condition, denoted $W(S)$, is $-log(p_+(S)/p_0(S))$, where $p_+(S)$ (resp., $p_0(S)$) is the probability of observing S under the positive (resp., null) model. See [24] for details. Since the null model takes into account the data properties in a detailed way, one gets a more sensitive evaluation than using mean expression.

To further improve our sensitivity, we use re-sampling of random gene sets to assess the empirical distribution of $W(S)$ for each condition and gene set S of size i. That distribution is not normal, and we compute its mean $\mu(i)$ and maximal deviation $\beta(i)$. We define the activity score of a gene set J and a condition j as $AS(J, j) = |W(J) - \mu(|J|)|/\beta(|J|)$. Finally, given an expression matrix on a set of conditions I, and a motif P, the **activity score** for the motif P is:

$$AS(P) = \max_{0 \leq T \leq 1} \sum_{i \in I} AS(J_T^P, i) \tag{2}$$

Given the threshold $Th(P)$ that maximized $AS(P)$, we can compute the activity of P on each of the conditions $i \in I$ as $AS(P, i) = AS(J^P_{Th(P)}, i)$.

4.3 Screening and optimization of motifs

We shall next present an algorithm for activity optimization of an LPWM. The algorithm is an EM-like heuristic with alternating phases of PWM 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 iterative procedure to train PWMs by combined expression and sequence data. We start from an initial LPWM P_0 (possibly random). The first phase computes the activity score of P_0 and determines the threshold T_0 for which $\sum_{i \in I} AS(J_{T_0}^{P_0}, i)$ is optimal. For each gene $j \in J_{T_0}^{P_0}$ we calculate its score contribution as $x_j^0 = AS(J_{T_0}^{P_0}) - AS(J_{T_0}^{P_0} \setminus j)$. In the second phase we find for each gene $j \in J_{T_0}^{P_0}$ with positive x_j^0 a position of P_0 in the promoter s_j that maximizes the matching likelihood, and use these positions as a gap-less alignment from which we extract the profiles P_1 for the next iteration. P_1 is formed by weighted counting in which gene j has a weight x_j^0, so higher activity genes have larger effect on the new PWM. We continue iterating the two phases until $AS(P_k)$ does not improve. The algorithm is described in Figure 6.

We use the AS-EM algorithm as a subroutine in our platform for discovery of active motifs. 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 one possible gap, scans the entire set of promoters with the motif, extracts a set of genes and scores them using the AS scheme. We use a precomputed hash of short k-mers matches to speed up the procedure. The entire workflow is as follows:

1. Generate the SAMBA weights; sample random gene sets; assess likelihood distributions; sample random motif scores and determine score significance thresholds (for single and multiple testing)
2. Exhaustively screen all k-mers with a single gap of size $\leq l$
3. Use AS-EM to optimize each gapped k-mer with AS value above the random level
4. Cluster similar LPWMs and output a concise set of motifs

4.4 Refining transcription factor dose estimations

The initial estimation of TF doses using the expression of regulated genes may be strongly biased by effects of combinatorial regulation. For example, whenever TF A is positively regulating a set of genes and part of the set is also negatively regulated by TF B, we shall assign lower activity to A in cases where B is in action. We try to overcome such misleading initial values by tuning the TF activities given a TP model. The **TF dose optimization problem** is defined given a DAR model (M, α, Δ), input expression profiles E and a score ϕ. The goal is to find

TF doses d_t^u optimizing ϕ (recall that given the DARs δ_t and the doses d_t^u we can calculate site activities r and apply the topological score ϕ). One can show that this problem is NP-hard by a simple reduction from SAT. We can heuristically approach it using hill climbing, optimizing the dose of one TF in one condition at each step. The optimal value of d_t^u, fixing the rest of the parameters, is derived by maximizing the combination of score contributions from all the model's genes: $argmax_{x \in \{d_t^{u'}, u' \neq u\}} \sum_{v \in V} \overline{\phi}(r_{N(v)}, e_v)$, where the site activities r are defined as usual 1.

By alternating between model optimization and dose optimization, we finally converge to a locally optimal solution. Our empirical studies show that the rough initial estimation of TF doses using motif activity scores generates a good starting point for this iterative algorithm.

5 Results

We tested the different components of our model inference framework by applying it to experimental data related to carbohydrate metabolism. 61 relevant expression profiles were selected from [10, 7, 9, 16]. We applied SAMBA to transform gene expression to log likelihood weights (using probability $P_c = 0.6$ for the positive model, see [24]). To assess p-values of weights we randomized 10000 gene sets of each size from 1 to 2000 genes and calculated their total weight. We used 500 bases upstream from the start codon as the promoters.

5.1 Comparing activity score and mean expression

To compare the performance of our activity score to the mean expression methods we used the 32 yeast PWMs from Transfac [27] version 5.1. Following [12], the mean-based score of a set of genes $G \subset V$ was computed as follows: Each condition in the expression matrix was normalized to mean 0 and standard deviation 1. Let s_c^G be the mean expression of genes in G on condition c. Let s^G be the mean value of s_c^G over all conditions $c \in C$. The collection of conditions with significant mean is then extracted as $S = \{c \in U || s_c^G - s^G| \geq \theta\sigma\}$ where $\sigma = \frac{1}{\sqrt{|G|}}$ and θ is a parameter. The mean score is finally defined as:

$$MeanScore(G) = \sum_{c \in S} (|s_c^G - s^G|) \qquad (3)$$

For each PWM we ordered all yeast genes by the matching likelihood of their promoter to the PWM. We selected an optimal threshold for maximal activity score (Formula (2)). To determine significance, we used the maximum score obtained over 5000 instances in which the genes' promoters were randomly shuffled. The same process (selecting an optimal threshold and sampling) was repeated with MeanScore. As can be seen in Figure 6, both methods correctly identified the GAL4 and RAP1 sites as active, but only the activity score succeeded in identifying additional transcription factors which are known to be functionally associated with carbohydrate metabolism (MIG1, ADR1 and more).

5.2 Effect of motif localization on activity

We tested the utility of incorporating motif location distribution into the PWM by re-analyzing TRANSFAC's known PWMs on the same dataset using different positioning windows. We tested the activity of each PWM restricted to windows of 100bp starting at -500 with steps of 50bp and ending at 0. Some of the motifs (but not all) show higher activity in specific windows (Figure 6). Using LPWM can improve the specificity of TF-gene association and refine our understanding of the relation between DNA and TFs.

5.3 Discovery of active motifs in the galactose system

To infer a transcription program for the galactose system we first applied our active motif discovery algorithm to the dataset of [10]. 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 5.1. 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 with scores above the noise level. The active motifs are listed in Table 1. We accurately rediscovered de-novo the known motifs GAL4, GCN4 and ADR1/MIG1. We also discovered three additional motifs (CANCCCC, TT(9N)CCCC, CCG(5N)CCG) which achieve scores above or equal to ADR1/MIG1. Note that this test does not use any prior knowledge on transcription factor motifs and scans a space of about 240,000 possibilities, so the high specificity is encouraging.

Figure 6 shows the activity profiles of all inferred motifs. We observe four transcriptional sub-programs: a) GAL4 regulated genes are repressed in many of the tested conditions, most dominantly when galactose is absent from the media or when upstream regulators (GAL3) are manipulated. b) GCN4 repression is limited to GAL4 active conditions, but seems to be independent of the level of galactose (active in GAL3,4,7,10 knockouts, both when galactose is present or absent). This may be explained by the fact that GCN4 regulates amino acid biosynthesis pathways that in part depend on the yield of the TCA cycle, which in turn may be dependent on the galactose pathway. c) MIG1/ADR1,U1 and U3 are active together and may represent a more general cellular response to stress. U1 and U3 have a common sub-motif (CCCC) which also appears in MIG1 complement. When comparing the gene sets associated with each of the three motifs, only the sets of U1 and U3 have statistically significant intersection (30% of the genes). All three motifs may be related to the STRE site (CCCCT-AGGGG), but the STRE motif is not active itself. A particularly high activity of these motifs is observed in the gal1-gal10- double mutant, in which GAL4 response is weak. Interestingly, this activity seems to be a combination of the activity in the single mutants gal1- and gal10-. d) The U2 motif contains a fragment of the GAL4 sequence, but its gene set does not significantly intersect that of GAL4. Its activity does not manifest either galactose dependent or galactose independent pattern.

5.4 A dose-activity response model for the galactose system

We have used the six active motifs discovered above to build a model explaining the gene expression of all the variable genes in the 23 conditions in the data set of [10]. We selected 310 genes with significant expression change in the data set (at least 3 conditions with over 2-fold change in expression). We assigned affinity levels based on TF binding location data from [14] for GAL4 and GCN4. For the four other motifs, including MIG1, we used the LPWM matching likelihood. (The location data on MIG1 in [10] could not be used, since that experiment took place in normal growth, and in such condition MIG1 has low activity). We calculated initial TF doses using the activity scores of the motifs in each condition. We sorted affinities and doses of each TF and used the ranking as inputs for the inference algorithm. For model optimization, site activity and gene expression were assumed to have three possible values. We optimized DARs using consistency [22] for the topology score. Figure 6 shows the results of the optimization with consistency scores.

The optimized transcription model also constructs, of course, for each site, a concrete affinity level and concrete activity levels under each condition, as well as a specific regulation function for each gene. A fragment of the model is shown in Figure 6. We selected for easy visualization only those genes that are regulated by GAL4 according to the model. A gene is considered regulated by a TF if its affinity level is above the threshold selected to maximize the

TF's activity2. The figure shows TF-gene interactions, where genes with an identical set of regulators are grouped together. The graph represents the predicted combinatorial regulation schemes involving GAL4. For example, the important regulatory gene GAL3 is combinatorially regulated by GAL4, MIG1 and GCN4, PUT4, a proline permease, is controlled by GAL4, U1 and U2 . Interestingly, there are 19 genes regulated by GAL4 alone, 34 genes regulated by two TFs, 23 genes regulated by 3 TFs and only 3 regulated by four TFs. No gene is regulated by five or all six. This indicates both the power of combinatorial regulation to generate complex control using a few regulators, and its efficiency in terms of the number of regulators used.

To test the utility of our integrative approach for TF binding location and gene expression data we compared the results of the DAR optimization process when given the experimental TF location profiles for GAL4 and GCN4 and when using the LPWM matching likelihood instead. Overall we obtain improved score for the complete model (using all six TFs) when using the experimental location profiles (model score without location data: score=5657, with GAL4: score=5687, with GCN4: score=5687, with both: score=5717).

To further test the robustness of our framework, we repeatedly selected one location profile and one TF, and used that profile as the genes affinities for that TF. The rest of the model was initialized with LPWM matching likelihood and we recorded the score after optimizing the complete model. This was repeated for all 660 possible TF-profile combinations (for 6 TFs and 110 profiles). We found that the score reached for the only experimental profiles that were available (GAL4 and GCN4) was not matched by any of the other experiments (p=0.001). This shows the superiority of location profiles over computed ones, and also demonstrates the robustness of our methodology. Further experimentation based on our model may lead to improved understanding of the way in which the transcription program of galactose related genes respond to different stimulations.

6 Acknowledgments

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

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

[3] D. Y. Chiang, P. O. Brown, and M. B. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, 17(Suppl 1):S49–55, 2001.

[4] R. Durbin, S. Eddy, A. Krough, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[5] M.L. Fredman and R.E. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34:596–615, 1987.

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

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

[8] I. Gat-Viks and R. Shamir. Chain functions and scoring functions in genetic networks. to appear in Proc. ISMB, 2003.

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

[10] 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.

[11] T.E. Ideker, V. Thorsson, and R.M. Karp. Discovery of regulatory interaction through perturbation: inference and experimental design. In *Proceedings of the 2000 Pacific Symposioum in Biocomputing (PSB 00)*, pages 305–316, 2000.

[12] 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.

[13] 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.

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

[15] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Proc. Pacific Symposium on Biocomputing*, pages 18–29. World Scientific, 1998.

[16] 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.

[17] 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.

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

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

[20] 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.

[21] 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.

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

[23] A. Tanay and R. Shamir. Modeling transcription programs: Inferring binding site activity and dose-response model optimization. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 03)*, pages 267–279. ACM Press, 2003.

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

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

[26] M.P. Washburn, A. Koller, G. Oshiro, R.R. Ulaszek, D. Plouffe, C. Deciu, E. Winzeler, and J. R. 3rd. Yates. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *PNAS*, 100(6):3107–12, 2003.

[27] 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.

LPWM Consensus	AS Score	Remark
CGG(11N)CCG	30.8	GAL4 consensus
TGACTCAWT	22.57	GCN4 consensus
TGGGGTA	22.03	ADR1/MIG1 consensus
CANCCCC	26.92	Unknown, denoted U1
CCG(5N)CCG	26.6	Unknown, denoted U2
TT(9N)CCCC	22.03	Unknown, denoted U3

Table 1: De-novo identification of active motifs in the galactose system. Using our novel framework for active motif discovery, we identified three known and three putative binding site motifs with statistically significant activity. The results match the literature in predicting activity of GAL4, GCN4 and MIG1. Putative sites represent predictions that extend the model of galactose related transcription program.

Figure 1: The three components of a transcription program. Left: TFs are transcriptionally and post translationally regulated to attain specific doses. Center: target genes promoters contain signals that enhance chromatin modifications and bind specific TFs. Left: the relative positions of TFs in their binding sites and their conformational properties induce a combinatorial regulation scheme for each gene.

Figure 2: Transcription program models. Common gene network models (left) study a set of genes and the interactions among them. Genes and their proteins are indistinguishable and mRNA expression rates are identified with the active protein doses. We provide a more detailed model that uses a bipartite topology (right). The model represents transcription factors and their concentrations, promoter binding sites and their affinities (edges in the graph; affinity is schematically indicated by the edge thickness) and the regulated genes and their expression rate. The structure of our model allows for integration of diverse experimental data types (expression, TF binding location) and for clear, mechanistic interpretation of the results.

Figure 3: Transcription program models. A: The 3-layered mechanistic model of transcription regulation. The transcription factor proteins (top) attain their active states in particular doses d_i. Sites in the promoter region of the target gene g (s_i, center) allow binding of the TFs. Each site's activity is determined by the TF dose and the affinity α_i of the TF and the gene, via the dose-affinity-response function δ_j. Transcription rate e_g is then determined by these site activities via the regulation logic f_g. The site affinity is a constant value for each particular TF-gene pair. The doses - and hence the activities - vary depending on the experimental conditions. B: A typical DAR function. This function has three activity levels, denoted 0, 1, and 2. The activity level is determined as a function of the TF dose and the site's affinity. Note that the function is monotone increasing in each coordinate. DAR functions are specific to each TF, and are applied to all the TF-gene sites that include it.

Figure 4: Effects of affinity on activity. X axis - different experimental conditions from [10]. Y axis - mean expression log fold change. Left: gcn4 bounded genes show variable response as a function of binding strength. We sorted all yeast genes by gcn4 binding affinity from [14]. 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 [14] 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.

Figure 5: Single DAR optimization for binary activity levels. The grid graph represents the two dimensional dose-affinity plain of the DAR function. Arcs are directed such that paths may not go upward or to the left. A path in the graph is equivalent to a monotone DAR function, where nodes on or above the path have activity level 1 (black) and nodes below have activity 0 (white). Arc costs are set so that the longest path through the graph corresponds to an optimal DAR function.

Figure 6: LPWM optimization algorithm.

Figure 7: Different sources for TF affinities. We plot the mean expression of several groups of genes in galactose related conditions from [10]. X axis: different conditions. Y axis: Mean log expression ratio. The ChIP group consists of the 30 genes with strongest binding to gcn4, according to location analysis [14]. The Transfac PWM group consists of the 30 genes with highest matching likelihood of their promoter to Transfac PWM M00038. The consensus group consists of all yeast genes with the exact motif TGACTCA in their promoter. Since all groups manifest significant co-expression compared to the global mean, all these sources of information may be used as affinities to some extent. However the direct physical measurements outperform other methods.

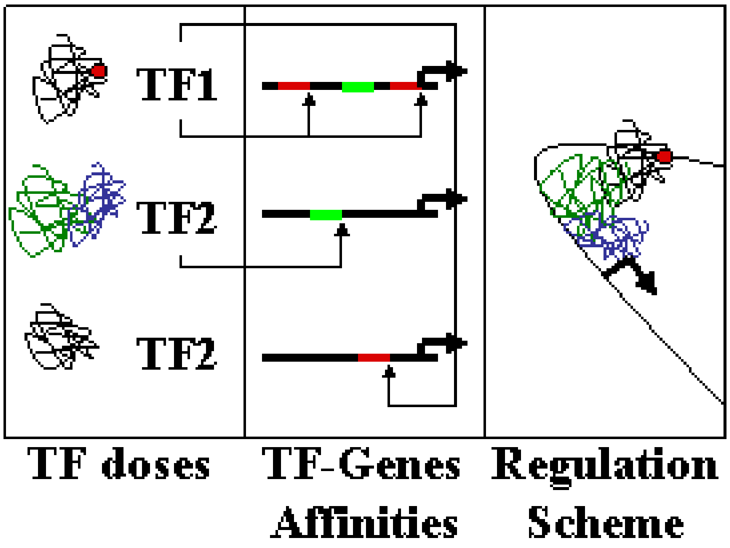
Figure 8: Comparing the activity score to mean expression. 32 yeast PWMs from Transfac were scored against a combined carbohydrate data set. Left: Mean scores. X axis: Mean score. Only RAP1, GAL4 and possibly GCN4 and CBF1 are detectable above the noise level (vertical bold line). Right: Activity scores. X axis: Activity score. Additional relevant active motifs (ADR1, MIG1, STRE, AP-1) are detected.

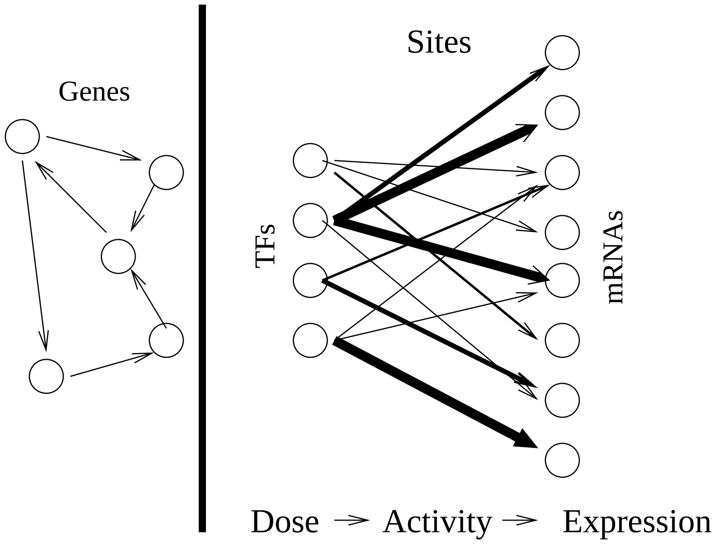
Figure 9: Effect of motif position on activity. Plots of activity as a function of the range window. X axis: distance of the left endpoint of the 100-bases window from promoter TSS. Y axis: AS score. The left figure shows the behavior of several known motifs: RAP1 is strongly biased to -400 bases from the TSS, GAL4 to -350, GCN4 to -200. In contrast, ADR1 is active in a broader range. The right figure shows the global and localized score for PHO4. The peak of the local score at -300 bases exceeds the detection level, while the global score is below it.

Figure 10: Activity profiles of galactose-related motifs. Rows represent the six binding site profiles that we detected. Columns represent different experimental conditions from [10] and are labeled using the naming convention of that study. Numbers in the cells are the activity scores. Darker cell shades indicate higher scores, and blank cells represent lack of significant activity. Note that high activity may be due to either induction or repression with respect to the wild type state. The direction of activity is not shown in the figure (but can be easily checked using the original expression values of the genes regulated by each motif.)

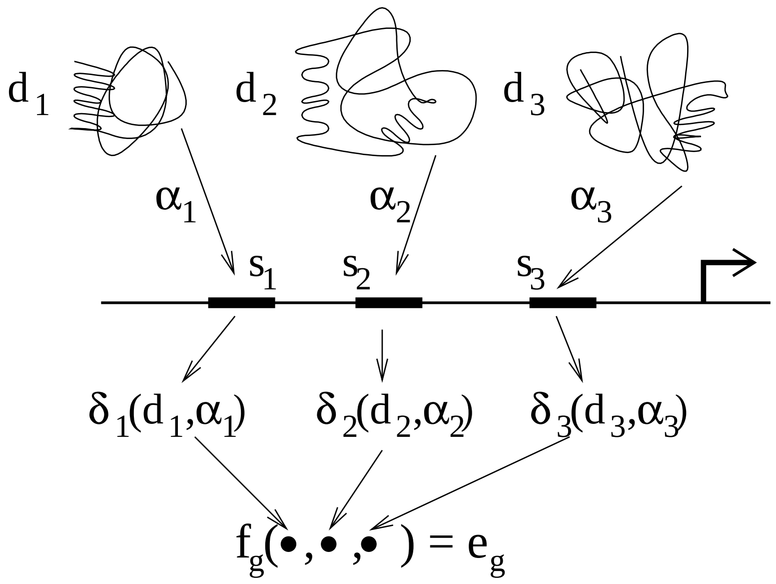
Figure 11: Transcription program model DAR's for the galactose TFs. X axis: ranked gene affinity. Y axis: ranked TF dose. The colors represent the response (or derived TF activity). Darker color indicates stronger response.

Figure 12: A part of the transcription program model reconstructed by our methodology for the galactose system. Having optimized the DAR model, only genes that have affinity above the activity threshold for a particular TF are considered to be regulated by it. Regulation is indicated by an edge. Genes with identical set of regulators are grouped together and shown within the same oval. Thicker edges underline larger groups of genes. The figure presents only those genes that are regulated by the TF GAL4. Filled ovals: TFs. White ovals: genes. The large set of genes in the oval at the right bottom contains some of the genes that are regulated only by GAL4, according to the model.


TF1
TF2
TF2
TF doses
TF-Genes
Affinities
Regulation
Scheme

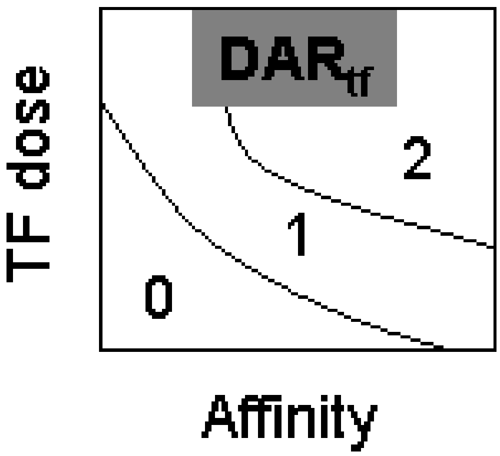


Genes
Sites
TFs
mRNAs
Dose → Activity → Expression



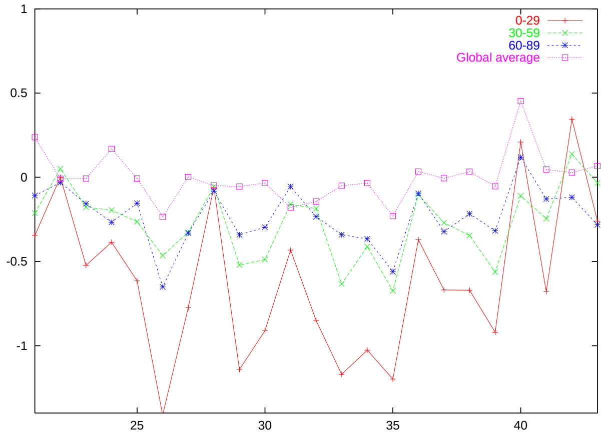
d 1
d 2
d 3
α1
α2
α3
s1
s2
s3
δ 1(d 1,α1)
δ 2(d 2,α2)
δ 3(d 3,α3)
fg(•,•,•) = eg

A

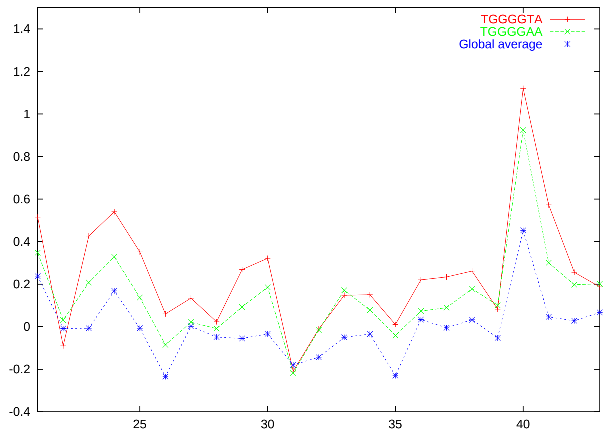


DARtf
TF dose
Affinity
0
1
2

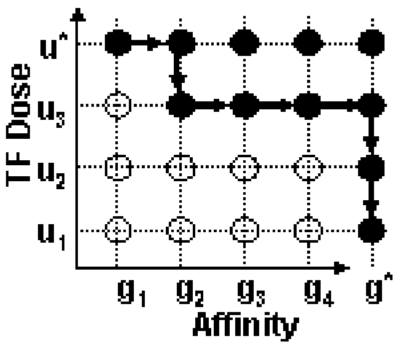
B



0-29
30-59
60-89
Global average



TGGGGTA
TGGGGAA
Global average



TF Dose
u*
u3
u2
u1
g1
g2
g3
g4
g*
Affinity

AS-EM(P_0, w^c_{ij}, s_j):

$k = 0$

while$(AS(P_k) > AS(P_{k-1}))$

  calculate $AS(P_k)$, extract optimal $J^{P_k}_{T_k}$

  For each $j \in J^{P_k}_{T_k}$ calculate $x_j = AS(J^{P_k}_{T_k}) - AS(J^{P_k}_{T_k} \setminus j)$.

  Let $Pos = \{j \in J^{P_k}_{T_k} | x_j > 0\}$

  For each $j \in Pos$ calculate $o_j = argmax_{l<|s_j|} Pr(P_k, s_j, l)$.

  Let $X = \sum_{j \in Pos} x_j$

  Initialize P_{k+1} with uniform prior.

  for $l = 0$ to P_k's length and for each $j \in J^{P_k}_{T_k}$

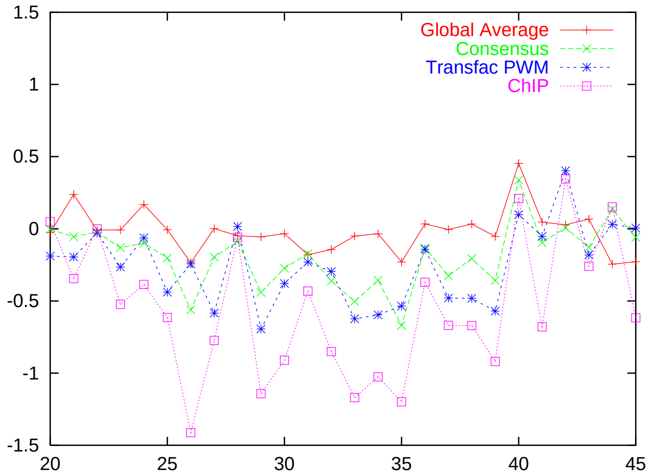
    $P_{k+1}(l, s_j(o_j + l)) = P_{k+1}(l, s_j(o_j + l)) + x_j/X$

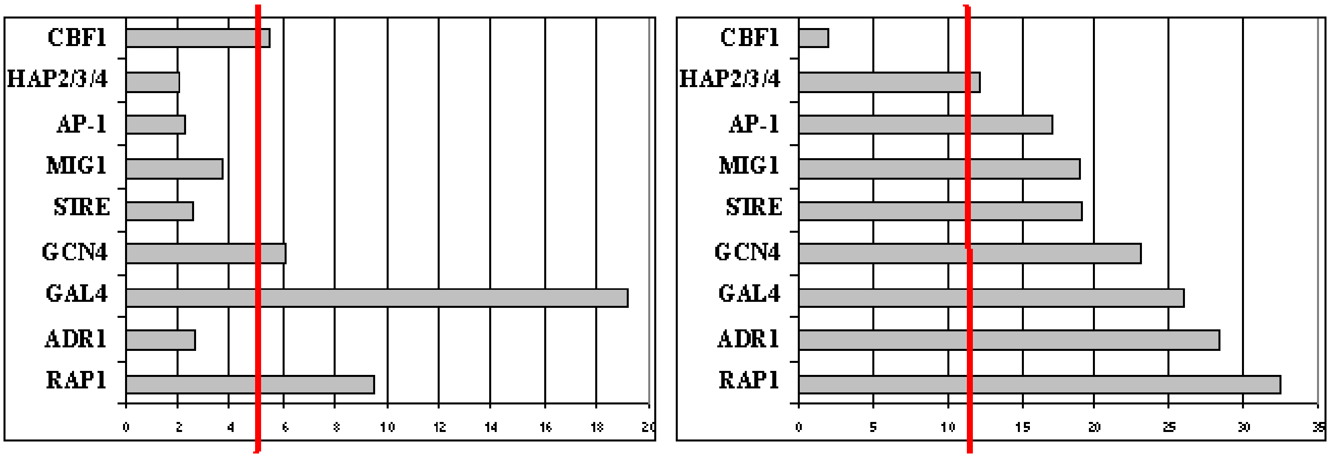
    $p_{k+1}(o_j) = p_{k+1}(o_j) + x_j/X$

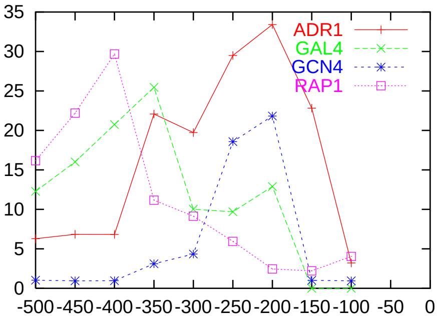
  smooth p_{k+1}

  next k

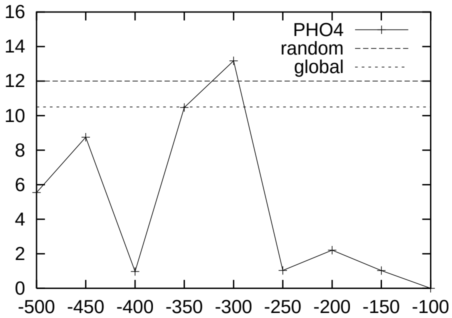
return P^k


Global Average
Consensus
Transfac PWM
ChIP
1.5
1
0.5
0
-0.5
-1
-1.5
20
25
30
35
40
45


CBF1
HAP2/3/4
AP-1
MIG1
STRE
GCN4
GAL4
ADR1
RAP1
0
2
4
6
8
10
12
14
16
18
20
CBF1
HAP2/3/4
AP-1
MIG1
STRE
GCN4
GAL4
ADR1
RAP1
0
5
10
15
20
25
30
35

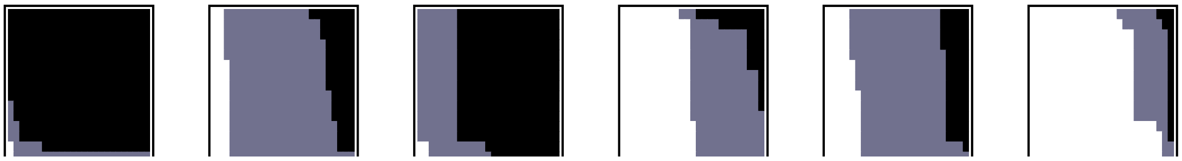


ADR1
GAL4
GCN4
RAP1
0
5
10
15
20
25
30
35
-500
-450
-400
-350
-300
-250
-200
-150
-100
-50
0

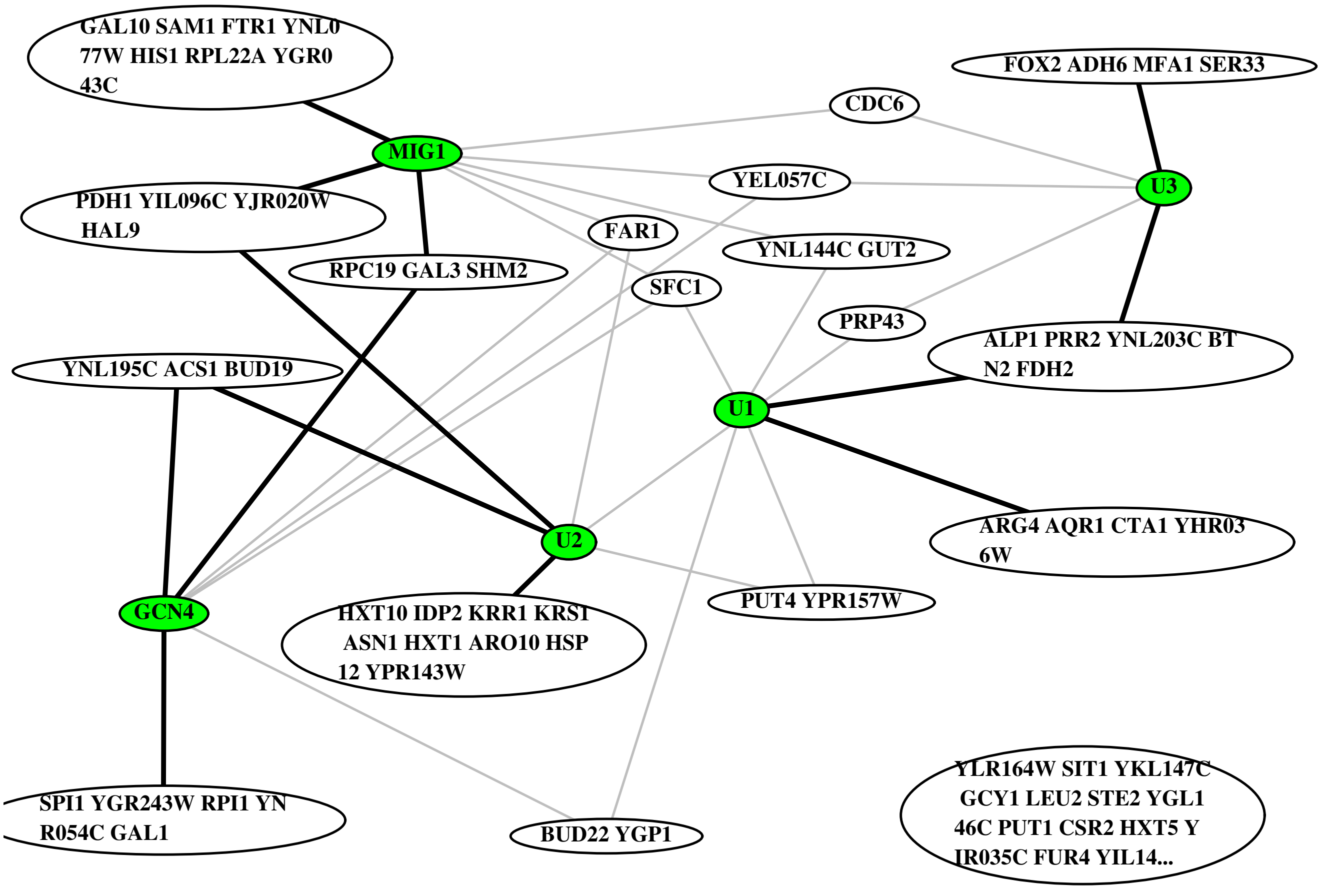


PHO4
random
global
0
2
4
6
8
10
12
14
16
-500
-450
-400
-350
-300
-250
-200
-150
-100

	wt−gal	wt+gal	gal1+gal	gal2+gal	gal3+gal	gal4+gal	gal5+gal	gal6+gal	gal7+gal	gal10+gal	gal80+gal	gal1−gal	gal2−gal	gal3−gal	gal4−gal	gal5−gal	gal6−gal	gal7−gal	gal10−gal	gal80−gal	gal1gal10+gal	gal2gal80−gal	gal4gal80−gal
GAL4	1.2				2.78	1.8			1.2	1.67		1.67	1.37	2.44	1.67	2.28	2.92	3.01	3.94		0.85		2.07
GCN4	1.02				1.71	1.5			2.12	1.56			2.17	1.26	1.89			0.7	2.04		1.53		
MIG1/ADR1	1.17		2.89	0.9	3.28				1.8	2.07				1.13		1.07	1.71			2.25	3.04		
U1	1.19		1.63	1.7	2.62					1.82				2.31		1.08	1.65	1.04	1.21	1.52	2.94		
U3	1.29		0.95	1.52	1.93					1.04				1.21		0.73	1.37			1.61	2.03		
U2	1.52		1.92	1.21	2.54				1.19	2.08			1.14	1.41				0.93	2.31	2.73	3.85		



GAL4 MIG1 GCN4 U1 U3 U2


GAL10 SAM1 FTR1 YNL077W HIS1 RPL22A YGR043C
MIG1
FOX2 ADH6 MFA1 SER33
CDC6
U3
YEL057C
PDH1 YIL096C YJR020W HAL9
FAR1
YNL144C GUT2
RPC19 GAL3 SHM2
SFC1
PRP43
ALP1 PRR2 YNL203C BTN2 FDH2
YNL195C ACS1 BUD19
U1
U2
ARG4 AQR1 CTA1 YHR036W
GCN4
HXT10 IDP2 KRR1 KRS1 ASN1 HXT1 ARO10 HSP12 YPR143W
PUT4 YPR157W
YLR164W SIT1 YKL147C GCY1 LEU2 STE2 YGL146C PUT1 CSR2 HXT5 YIR035C FUR4 YIL14...
SPI1 YGR243W RPI1 YNR054C GAL1
BUD22 YGP1