BIOINFORMATICS

Vol. 00 no. 00 2004
Pages 1–8

Inferring Quantitative Models of Regulatory Networks From Expression Data

I. Nachman [a], A. Regev,[b] N. Friedman [a]

[a] School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel, [b] Center for Genomics Research, Harvard University, Cambridge MA 02138, USA

ABSTRACT

Motivation: Genetic networks regulate key processes in living cells. Various methods have been suggested to reconstruct network architecture from gene expression data. However, most approaches are based on qualitative models that provide only rough approximations of the underlying events, and lack the quantitative aspects that are critical for understanding the proper function of biomolecular systems.

Results: We present fine-grained dynamical models of gene transcription and develop methods for reconstructing them from gene expression data within the framework of a generative probabilistic model. Unlike previous works, we employ *quantitative* transcription rates, and simultaneously estimate both the kinetic parameters that govern these rates, and the activity levels of unobserved regulators that control them. We apply our approach to expression data sets from yeast and show that we can learn the unknown regulator activity profiles, as well as the binding affinity parameters. We also introduce a novel structure learning algorithm, and demonstrate its power to accurately reconstruct the regulatory network from those data sets.

Keywords: transcription regulation, parameter learning, structure learning, regulatory networks

Contact: nir@cs.huji.ac.il

1 INTRODUCTION

Understanding the organization and function of gene regulatory networks is a key experimental and computational challenge in molecular biology. Recent studies (Guet *et al.*, 2002; Kitano, 2002) indicate that network function depends on both qualitative and quantitative aspects of network organization. For example, Guet *et al.*, 2002 show how differences in quantitative reaction rates have drastic effects on the function of circuits with identical qualitative properties such as connectivity and logic.

Various methods have been developed to reconstruct regulation networks from high-throughput data, including genomic sequences, expression profiles and transcription factor location assays (Ong *et al.*, 2002; Pe'er *et al.*, 2001; Segal *et al.*, 2002; Simon *et al.*, 2001; Spellman *et al.*, 1998; Tavazoie *et al.*, 1999). However, these methods are based on coarse grained qualitative models, and cannot provide a realistic and quantitative view of regulatory systems. Alternative methods were recently suggested (Ronen *et al.*, 2002) to estimate the quantitative parameters of more biologically realistic network models. However, these approaches are limited to networks of known, simple architecture and cannot be generalized to more complex architectures or unknown structures.

In this paper, we present a novel framework for the reconstruction of quantitative, realistic, fine-grained, dynamical models of gene regulatory networks. Given a dataset of gene transcription rates, our algorithm reconstructs the structure of a regulatory network, the quantitative kinetic parameters of transcription regulation, and the unobserved activity levels of regulator proteins. This focus on learning unobserved regulator activity levels is crucial, as activity levels are the result of a large variety of upstream biochemical events, such as RNA and protein expression, biochemical modifications, degradation rates, and changes in sub-cellular localization. However, neither activity levels nor most of the events that regulate them are measured today on a genomic scale. Thus, our models handle activity levels as unobserved variables, that indirectly encompass upstream regulatory events, without directly modeling these events. In particular, unlike previous work, we do not use the expression levels of regulators, and can thus identify the results of post-transcriptional events.

Our framework is based on a generative probabilistic model, dynamic Bayesian networks, that accounts for the processes that generated the data, handles noise in a principled way, and incorporates our prior biological knowledge into the solution. We model the underlying biochemical reaction equations and the sources of noise that can affect the dynamics of the system. The model is flexible, and can accommodate networks of realistic complexity, including activators, repressors, combinatorial regulation, and cooperative and competitive interactions between regulators.

Our approach can handle networks of either known, partially known, or unknown architecture. In particular, we introduce a novel learning algorithm for reconstructing the network structure. This algorithm reassigns regulators to

genes as well as detects when additional regulators should be added to the network, thus both improving existing structures and learning regulatory networks *ab initio*. We applied our methods to data from the yeast cell cycle regulation system, and were able to recover both regulator activity profiles and accurate parameters for networks of known architecture, as well as successfully learn a complex regulatory network *ab initio*. Overall, our approach combines for the first time the network reconstruction capabilities of qualitative approaches with the biochemical detail of quantitative ones, into a single framework for the reconstruction of realistic complex models of gene regulation from gene expression data.

2 TRANSCRIPTIONAL REGULATION MODEL

To develop a quantitative realistic probabilistic model of gene regulatory networks, we start by examining networks of known structure and unknown kinetic parameters (learning unknown network structure will be addressed in Section 3). First, we derive a kinematic model of how the transcription rate of a single gene depends on its regulators. We then consider how to model the behaviour of multiple genes over time, and how to learn the model parameters from actual measurements, including transcriptional rates and the unobserved activity levels of regulators.

Kinematic Model of Regulation. Our regulation model (Figure 1) is based on a *regulation function* that describes the *transcription rate* of a target gene (number of RNA molecules transcribed per unit of time per cell) as a function of the *concentration of active regulator(s)* (number of proteins in active form in nucleus per cell). Consistent with the input expression profiles that are typically measured on cell populations, we assume that we are examining a large population of cells, and that the derived transcription rates are actually average rates over this population. We assume that the change in concentration of the regulator H is much slower than the kinetics of reactions described, and that at each time point the system is nearly at an equilibrium. Thus, we model binding and disassociation reactions at steady-state. Finally, we assume that the number of active regulator molecules in each cell is much larger than the number of its target sites, thus neglecting any possible competition between different target genes on the same regulator.

We start with a single regulator, and then generalize to multiple regulators. In the simplest case (Figure 1a), of a single activator, the regulation function takes the familiar, non-linear Michaelis-Menten form:

$$g(H:\beta,\gamma) = \beta\frac{\gamma H}{1+\gamma H} \tag{1}$$

where H denotes the concentration of active regulator protein, β is the maximum transcription rate the gene can achieve, and γ is κ_b/κ_d the ratio of association and disassociation constants (Figure 1b-c).

To illustrate the general case of multiple regulators we consider a gene that has two regulators, with activity levels H_1 and H_2. Depending on whether no regulator, H_1, H_2, or both are bound to the promoter, we distinguish four possible binding site fractions, denoted $S^{-,-}$, $S^{H_1,-}$, S^{-,H_2}, and S^{H_1,H_2}. Solving the steady-state equations, we get the different binding state distribution:

$$\begin{array}{llll} S^{-,-} &= 1/Z & S^{-,H_2} &= \gamma_2 H_2/Z \\ S^{H_1,-} &= \gamma_1 H_1/Z & S^{H_1,H_2} &= \gamma_1 H_1 \gamma_2 H_2/Z \end{array}$$

where $Z = (1+\gamma_1 H_1)(1+\gamma_2 H_2)$ is a normalizing constant. We now can define a regulation function for two regulators as a generic weighted sum over all possible binding states:

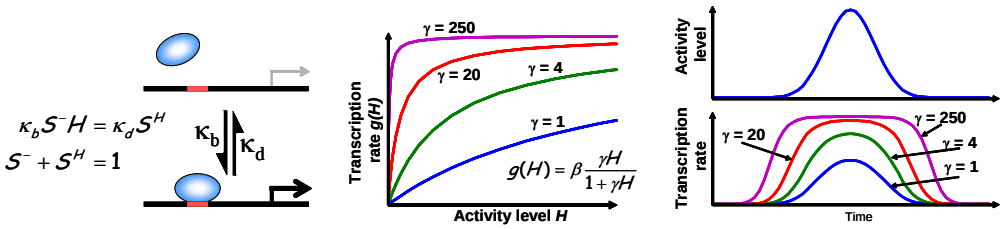
$$\begin{aligned} &g(H_1,H_2:\vec{\alpha},\beta,\gamma_1,\gamma_2) \quad = \\ &\beta(\alpha^{-,-}S^{-,-}+\alpha^{-,H_2}S^{-,H_2}+\alpha^{H_1,-}S^{H_1,-}+\alpha^{H_1,H_2}S^{H_1,H_2}) \end{aligned} \tag{2}$$

where $\vec{\alpha}$ is the vector of α parameters indicating the "productive" binding states that lead to transcription. Here, we focus on cases where α take binary values. For example, for two non-cooperative activators we set $\alpha^{H_1,-}, \alpha^{-,H_2}, \alpha^{H_1,H_2}$ to 1, and $\alpha^{-,-}$ to 0, reflecting that transcription occurs whenever at least one regulator is bound. To model general biological models, where different productive states may result in different rates, we can allow α to take real values.

Note, that this approach is general and is easily extended to more than two regulators by introducing additional binding states with different associated probabilities and transcription rates. It can also be extended to handle more complex scenarios, such as competitive or cooperative interactions between different regulators or a single regulator with two binding sites, each with a different effect on transcription. In this initial study we focus on the simple variants of the model.

Temporal Modeling of Regulons Using Dynamic Bayesian Networks. To model a regulatory network, we need to consider not only multiple regulators, but also multiple target genes and their *temporal* behaviour. Since regulators typically regulate multiple targets in the same regulon (Lee *et al.*, 2002; Shen-Orr *et al.*, 2002), the same activity levels of a regulator H can be used in the regulation functions of all of its targets. However, the functions themselves are gene specific. Consider a simple system of n genes that are regulated by the same regulator H where we measure transcription rates at T time points. Is it possible to reconstruct the values of H at different times, and the gene specific reaction constants? Since we have $n \times T$ observations, and we assume that these can be explained by T values of H and $2n$ parameters (different β and γ for each gene), we have an over-constrained problem when $n > 2$ and $T > 2$. Thus, such a reconstruction is feasible in principle.

Specifically, we use the language of *dynamic Bayesian networks* (DBNs) (Friedman *et al.*, 1998) to model the evolution of a stationary Markovian stochastic system over discrete time points. Our model combines a *regulation diagram* (*e.g.*



(a) Two promoter states (b) Regulation as function of H (c) Temporal behaviour

Fig. 1. A kinematic model of transcription regulation by a single activator. (a) An active regulator protein, H, may bind to and disassociate from a target gene's promoter, with rate constants κ_b and κ_d, respectively. In a population of cells, fractions S^- and S^H of cells have free and bound promoters, respectively and satisfy the steady-state reaction equations. The bound gene is transcribed with rate $g(H) = \beta S^H$. (b) The regulation equation, describing the transcription rate as a function of the active regulator concentration H, is in the Michaelis-Menten form. The transcription rate is a non-linear function of the activity level of H that depends on $\gamma = \kappa_b/\kappa_d$; In particular, at high levels of H, the transcription rate saturates. (c) Temporal behaviour of a single activator and the transcription rates of genes it regulates with different kinematic parameters.

Figure 2(a)) that summarizes the regulation topology between two types of attributes: the activity of regulators $H_1, H_2, \ldots$ and the transcription rates of target genes $R_1, R_2, \ldots$. The state of the system at time point t is described by random variables $H_1^{(t)}, H_2^{(t)}, \ldots$ and $R_1^{(t)}, R_2^{(t)}, \ldots$ that denote the values of all the system's attributes at time t.

The model describes relations between variables at the same time point and at consecutive time points. First, to represent the behaviour of the regulator activity attribute, we assume that $H_i^{(t+1)}$ depends on $H_i^{(t)}$. We model this dependence with the *persistence equation*:

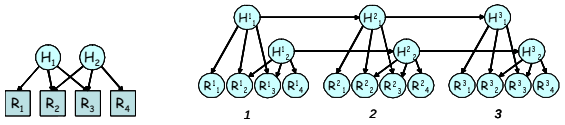
$$H_i^{(t+1)} = H_i^{(t)} + \epsilon_{h_i}^{(t+1)} \tag{3}$$

where $\epsilon_{h_i}^{(t+1)}$ is a normally distributed noise variable with zero mean and variance σ_i. By modeling the magnitude of change, our model prefers a smoother sequence of values H_i. Second, the transcription rate of each target gene depends on the instantaneous activity levels of the regulators that control it, as encoded by the regulation diagram. For example, if R_k depends on two regulators H_1 and H_2, then

$$R_k^{(t)} = g(H_1^{(t)}, H_2^{(t)} : \vec{\alpha}_k, \beta_k, \gamma_{k,1}, \gamma_{k,2}) \left(1 + \epsilon_{r_k}^{(t)}\right) \tag{4}$$

where $g(.)$ is the regulation function given by (2), and $\epsilon_{r_k}^{(t)}$ is a Gaussian noise variable with zero mean and variance σ_k. Note that the noise level for R_k depends on its expected value given the regulator activity levels. This stems from the fact that the transcription rate is a result of a sum of stochastic events, such as DNA binding, transcription initiation, and elongation (McAdams & Arkin, 1997). The higher the rate, the more events are involved, resulting in a higher variance.

Figure 2(b) illustrates the Bayesian network structure that corresponds to the regulation diagram of Figure 2(a) for three consecutive time points. The DBN model for time range



(a) Regulation Diagram (b) Unrolled network

Fig. 2. Schematic representation of a DBN model for temporal gene regulation. (a) A regulation diagram with 2 regulators and 4 targets. (b) An example of the Bayesian network induced by this diagram for three time points.

$1, \ldots, T$ defines a joint distribution over all the random variables in these T time points. The joint density of an assignment to all the variables is the product of the densities of the values of error variables $\epsilon_{h_i}^{(t)}$ and $\epsilon_{r_k}^{(t)}$ that achieve equality in Eq. 3 and 4.

Parameter Estimation. Once we define the DBN, we can learn the kinetic parameters and the hidden activity levels of regulators from observations. We consider an observed set $\boldsymbol{E}$ of transcription rates of n genes in T time points and try to optimize for the most likely assignment of parameters and levels. Thus, assuming a fixed regulation diagram G, we want to find parameters that maximize the likelihood

$$\ell(\boldsymbol{h}, \boldsymbol{\theta} : G, \boldsymbol{E}) = \log P(\boldsymbol{E}, \boldsymbol{h} \mid \boldsymbol{\theta}, G)$$

where $\boldsymbol{h}$ are the values of the unobserved regulator activity levels at different times, and $\boldsymbol{\theta}$ are the kinetic and variance parameters of the model. According to the DBN definition, the term $\log P(\boldsymbol{E}, \boldsymbol{h} \mid \boldsymbol{\theta}, G)$ is a log probability of error variables. To optimize the likelihood function, we use gradient ascent on the joint space of $\boldsymbol{h}$ and $\boldsymbol{\theta}$.

To avoid over fitting of the model to the data, we match the model complexity to the amount of available data. When data is scarce, we fix some of the parameters in advance, whereas

when the amount of data grows, we attempt to learn more parameters. In the current study, we preset the α parameters according to biological knowledge, keeping the number of free parameters low. For example, in the experiment described in section 4 we optimize between 3 to 4 parameters per each of the G target genes. This number is much lower than the number of observations (G times T, the size of the time series).

Transcription Rates While our regulation model is based on mRNA transcription rates, time series expression profiles typically provide us only with mRNA abundance levels. [1] To recover transcription rates, we consider how the mRNA expression level depends on both transcription and degradation, and use a simple gene-specific mRNA decay model:

$$\frac{d}{dt}e_k^{(t)} = r_k^{(t)} - \delta_k e_k^{(t)} \tag{5}$$

where $e_k^{(t)}$ is the expression level of gene k at time t and δ_k is the mRNA decay rate of gene k. We assume that mRNA decay rates may be gene-specific, but remain constant in time. Given the decay rate δ_k, and the expression measurements, we recover $r_k^{(t)}$ (up to a gene-specific multiplicative factor) by solving the differential equation (5). Such actual decay rates have been measured experimentally under specific conditions by several recent genome-wide studies (Holstege *et al.*, 1998; Wang *et al.*, 2002).

3 RECOVERING REGULATION DIAGRAMS

Structure Selection So far we have assumed that a defined regulation diagram is known. The key question of inferring this diagram *ab initio* from data is a *structure learning* (Friedman *et al.*, 1998) problem in the framework of DBNs: given a rate matrix $\boldsymbol{E}$, find the regulation diagram that is most likely to have generated $\boldsymbol{E}$. Note, that the likelihood of different models is not an appropriate score here, since richer models with more regulation relations will provably have better likelihood than simpler ones. A standard solution is to use the *BIC score* (Friedman *et al.*, 1998; Schwarz, 1978), which penalizes the likelihood term with a a structure complexity penalty term:

$$\text{score}(G : \boldsymbol{E}) = \max_{\boldsymbol{h},\boldsymbol{\theta}} \ell(\boldsymbol{h},\boldsymbol{\theta} : G, \boldsymbol{E}) - \frac{N_{\text{param}}}{2} log(T)$$

where N_{param} is the number of parameters in the model, and T is the number of time points. Once we define the score, structure selection is posed as an optimization problem over the discrete space of all possible regulation diagrams.

[1] Our methods are applicable to time series measurements of mRNA abundance from both oligonucleotides chips and from cDNA microarrays. For cDNA arrays, where values are relative to a common reference condition M_0, we can reconstruct the expression level up to a gene specific multiplicative constant that involves the level of the gene in the reference sample and probe-specific issues such as hybridization efficiency.

The typical approach to learn structure is by a heuristic search, such as greedy hill climbing, that explores local moves (*e.g.*, all legal edge additions and deletions) in the space of regulation diagrams. To evaluate a specific regulation diagram, we need to perform parameter learning to find a good reconstruction of the regulators for the proposed diagram. Since such evaluations are costly, this approach is infeasible in domains that involve many genes.

Instead, we propose an efficient algorithm to search for the regulation diagram, based on two main ideas. First, to allow efficient evaluation of proposed moves, we use a *structural EM* approach (Friedman *et al.*, 1998), and employ the regulator activity profiles from one diagram to approximate the score of modified diagrams. Second, rather than blindly evaluating all possible moves, we use the mathematical form of the regulation functions to focus on a small subset of promising moves. As we show below, this also allows us to detect when new regulators should be added to the model.

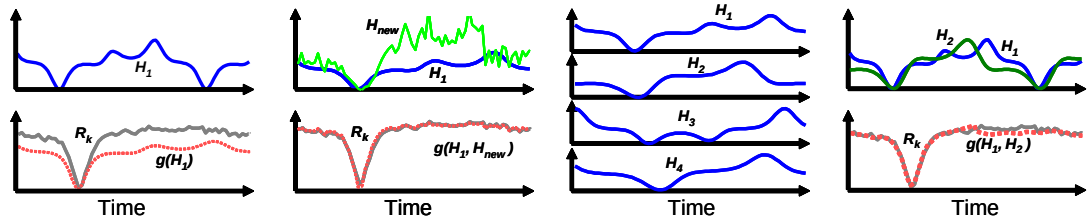
Ideal Regulators Profiles To illustrate the concept that will allow us to efficiently propose moves and detect new regulators, suppose we have a current network G, and we have maximized the parameters and regulator activity profiles with respect to this network. Now consider a gene R_k that is regulated by single regulator H_1 in the model (Figure 3(a)). Since H_1 predicts the transcription rates of multiple targets (and since in reality R_1 may be regulated by additional regulators), H_1 does not provide a perfect prediction of R_1's transcription rate. The best reduction of this error is by finding an "ideal" second regulator for R_1 that together with H_1 eliminates all predictions errors (Figure 3(b)). We now use this ideal profile to search against the current set of regulator activity profiles. If we find a regulator H_2 that is highly correlated with the ideal one, we evaluate it as a second regulator for R_k by searching for parameters that maximize the likelihood of R_k given H_1 and H_2 (Figure 3(d)).

More formally, suppose we are given a gene R_k that is regulated by H_1, with parameters β_k and $\gamma_{k,1}$. We want to find a regulator profile $\{h_{\text{new}}^{(t)} : t = 1, \ldots, T\}$ and binding affinity $\gamma_{k,\text{new}}$ such that $h_{\text{new}}^{(t)} = f^{-1}(r^{(t)})$ where

$$f(h) \equiv g(h_1^{(t)}, h : \vec{\alpha}_k, \beta_k, \gamma_{k,1}, \gamma_{k,\text{new}}) \tag{6}$$

Since the function f is generally invertible when $r_k^{(t)} > 0$, we can find this profile once we determine $\gamma_{k,\text{new}}$. However, examining the definition of g, it is easy to see that we can set $\gamma_{k,\text{new}}$ arbitrarily, as it only serves to scale the values of h_{new}. Thus, we get a regulator profile that is "ideal" for R_k and is unique up to rescaling.

Note, that the behaviour of the new ideal regulator can differ if we believe it is an activator or repressor and whether it works cooperatively with the current regulator(s), as indicated by the values (0 or 1) of $\vec{\alpha}_k$ in (6). Thus, rather than testing a single ideal regulator per gene, we construct a small



(a) Original model parameters (b) Reconstructed H_{new} (c) Comparison to regulators (d) New parameters

Fig. 3. Illustration of the ideal regulator approach. (a) A current regulator H_1 and its target's input transcription rate R_k (solid line) and predicted rates (dashed lines). (b) An "ideal regulator profile" H_{new} that together with H_1 predicts R_k without errors. (c) By computing the correlation between the profile of H_{new} and the activity profiles of each of the current regulators (based on the current network architecture), we can detect regulators that potentially regulate R_k, and evaluate how well these perform in predicting the transcription rates of R_k. (d) We reduce errors by introducing H_2, the best scoring candidate, as second regulator of R_k.

set of possible ones (with different $\vec{\alpha}_k$, representing different relevant logic), which we compare to the actual regulators. Once we find a good match, we may add a new edge from the existing regulator to R_k. In a similar way, we may also replace a current regulator by computing the set of possible ideal regulator profiles (with different $\vec{\alpha}_k$) in the absence of one of the current regulators of R_k, assuming that β and all other γ values are fixed.

Learning Algorithm The algorithm iteratively improves the network structure. It starts with some initial guess, which is either derived from prior biological knowledge (see below), or is simply the naïve network where all genes depend on a single regulator. Each iteration of the algorithm consists of two phases. In the first, we train parameters and regulator activity profiles to maximize the current model's likelihood function. We then use these to compute the ideal regulator profiles for each regulated gene, as described above. In the second phase, we propose possible modifications to the current structure (either addition or replacement of a regulatory connection), by computing correlations between the ideal profiles and the current regulator activity profiles. We explore each of these modifications suggested by correlations that exceed a fixed threshold, by training the local parameters β_k and $\gamma_{k,i}$ for the specific target k to maximize the score. This optimization is done without changing the regulator activity profiles, and hence changes the likelihood only locally in terms that involve R_k. Modifications that decrease the score are discarded, and the rest are applied at the end of the iteration. If there are several modifications that apply to the same R_k, we select only the one that leads to the biggest improvement in the score.

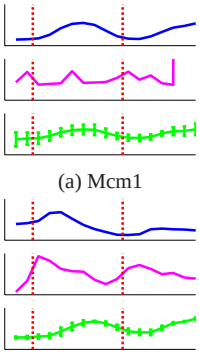
We may also remove an existing connection, thus correcting earlier mistakes We suggest an intuitive mechanism for selecting candidate connections for removal: if a certain regulatory connection has an associated affinity parameter which is much lower (*e.g.* by a factor of 10) from the other affinity parameters of that gene, we test this connection for removal. As with the addition steps, if the change yields a positive change in score - it is accepted.

Introducing New Regulators We can also introduce a new regulator into the network. This step is applied only if no other modification is accepted. To add a new regulator, we apply the CLUST algorithm (Ben-Dor *et al.*, 1999) to find clusters of ideal regulator profiles that are highly correlated (above 0.8), and may correspond to a new regulator of the genes for which these ideal profiles were generated. We evaluate each proposed new regulator by introducing it into the network, and then apply gradient ascent to find the best parameter values and regulators activity profiles for the modified network. We then choose the new regulator that leads to the biggest score improvement and add it and its target links to the current network. If no such regulator offers a positive improvement, no action is taken.

4 RESULTS

We tested the power of our transcription regulation model and the effectiveness of the structure learning algorithm on a series of examples related to transcriptional regulation in the yeast cell cycle. First, we identify kinematic parameters and regulator activity profiles for a small transcriptional network operating at M phase, involving an activator and a repressor. Second, we study a curated model of the complex regulatory network of the entire cell cycle, and show that we can accurately identify activity levels of regulators based solely on our realistic modeling framework and the expression levels of their targets. Finally, we employ our structure learning algorithm to learn a regulatory network *ab initio*, based solely on expression data, and show the accuracy of both the resulting network topology and the reconstructed regulators and their activity profiles.

Two Regulator System Recent work shows that M phase-expressed genes in yeast can be distinguished into two subsets. A major set which is activated by Mcm1 and is expressed earlier in M phase, and a minor set which is activated by



(a) Mcm1

(b) Yox1

Fig. 4. Regulator reconstruction in an activator-repressor system. (a) The learned activity for Mcm1 (top), vs. its mRNA log expression levels (middle), and its target genes transcription rates (bottom). Vertical lines denote cell cycle start points (end of M/G1 transition). (b) Same for the repressor Yox1.

Mcm1 and repressed by Yox1, with delayed expression in late M phase. To evaluate the dynamics of this system, we built a model of this network, based on known Mcm1 and Yox1 targets (Simon *et al.*, 2001), and used cell-cycle mRNA expression data[2] (Spellman *et al.*, 1998) and experimentally derived decay rates (Wang *et al.*, 2002), to estimate transcription rates for these genes. We then applied our parameter learning methods on each of the time series and learned activity profiles for the two regulators. As seen in Figure 4, the reconstructed activity level of Yox1 peaks earlier than that of Mcm1, consistent with its documented repressive role, and explaining the subsequent shift in the peak transcription levels of its target genes compared to that of Mcm1-exclusive targets. Surprisingly, Yox1's reconstructed activity peak appears relatively early in the cell cycle, before M phase. This novel finding was also obtained on a separate time series (data not shown) and is corroborated by Yox1's expression profile (Figure 4b). Note, that the regulator expression profiles themselves are not used in the reconstruction. This allows us to recover the hidden activity levels of regulators that are themselves not transcriptionaly regulated. Thus, we accurately reconstruct the activity profile of Mcm1, which is not transcriptionaly regulated, with a clear peak at M phase.

Cell Cycle Regulation System We next turned to a large regulatory network of known topology, assembled from location data (Lee *et al.*, 2002) and biological databases (Costanzo *et al.*, 2001). In this network, seven different transcription factors control the expression of 141 genes throughout the cell cycle, alone or in pair-wise combinations. Using the alpha synchronization expression time series (Spellman *et al.*, 1998) shown in Figure 5(a), we learned activity profiles and kinematic parameters for this complex network. The predicted rates we learned for the 141 genes are shown in Figure 5(b).

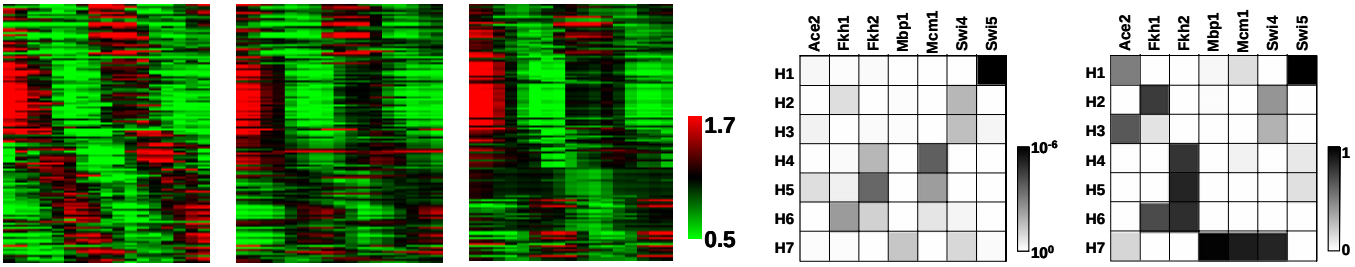
Figure 6 shows the learned activity profiles for the 7 modeled regulators, against their mRNA expression levels and their target genes behaviour. For all seven transcription factors, the model automatically reconstructs cyclic activity levels, that are consistent with their known activity based on molecular or genetic studies. For example, Swi5's activity peaks at late M/G1 and early G1, consistent with its previously reported activity (McBride *et al.*, 1999); Mbp1 and Swi4's activity levels peak at mid to late G1 consistent with their role in G1/S gene expression (Baetz & Andrews, 1999); and Fkh1 and Fkh2 peak at late S/G2 and G2/M respectively, consistent with their reported effects in genetic studies (Hollenhorst *et al.*, 2000). Thus, in many cases (*e.g.*, Swi5 and Swi4 or Fkh1 and Fkh2), the reconstructed activity levels distinguish between relatively subtle but important differences in true biological activities, the establishment of which has often required a large number of experiments. In some cases (*e.g.* Fkh2 or Swi5), our reconstructed activity profiles closely resemble the regulator's expression profile. More importantly, since our reconstruction does not use the expression levels of the regulators, we are able to accurately reconstruct their activity levels even if they are not regulated transcriptionaly (*e.g.*, Mcm1), or if their expression and activity profiles are shifted (*e.g.*, Ace2), highlighting the power of our approach.

The full power of our framework lies in its ability to learn not only accurate activity profiles and kinematic parameters, but also the full network architecture *ab initio*. We therefore ran our structure learning algorithm on a naïive network with the same 141 target genes all wired to a single activator. We allowed the algorithm to add more regulators and change regulatory connections until convergence, surprisingly resulting in a network with seven regulators. See Figure 5(c) for predicted rates with this model.

To evaluate the quality of our *ab initio* reconstructed network and identify the reconstructed regulators, we compared the topology of the learned network to that of the curated one (Figure 5(d)), and the learned activity profiles to those learned on the curated network (Figure 5(e)). In some cases, such as inferred regulator 1 and the known regulator Swi5 (Figure 5(d-e), top row), the correspondence in both targets and activity levels is striking. In others, a single inferred regulator corresponds to two separate factors with similar activity patterns (*e.g.*, regulator 7 and the G1/S factors Mbp1 and Swi4). Overall, since in the known network some targets are regulated by more than one factor and some factors have similar profiles, by combining both tests we can roughly identify most of our inferred profiles (regulators 1, 2/3, 4/5, 6 and 7) with known regulatory activities (Swi5, A G1 regulatory activity, the Fkh2/Mcm1 complex, Fkh1, and MBF/SBF). Thus, these tests indicate that the inferred regulators have both targets and activity levels strikingly similar to those in the known curated network, and highlight the success of our approach in learning both correct structure and parameters in the most stringent challenge.

Finally, despite their impressive correspondence, both the *ab initio* learned network and the curated model are likely only approximations of the true biological systems. Thus, we combined our curated network and our structure learning

[2] The data set consists of three time series that contain 17 to 23 time points, with time intervals of 7 to 10 minutes.



(a) Input rates (b) Curated model (c) *ab initio* model (d) Target intersection (e) Profile correlation

Fig. 5. Comparison of *ab initio* structure learning vs. parameter learning for the curated cell cycle regulatory diagram. (a) Measured transcription rates for 141 genes. (b) Predicted rates in the curated model after learning parameters. (c) Predicted rates after *ab initio* structure learning. (d) log p-value of target intersection groups between known and *ab initio* regulators. (e) Positive correlations between learned activity profiles of known and *ab initio* regulators.

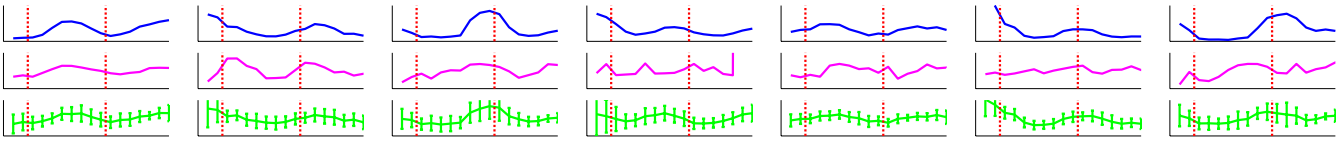


Fig. 6. Regulator activity profiles learned from the curated network diagram. Each profile (top) is plotted against the regulator's mRNA log expression levels (middle) and the average transcription rates of all its target genes (bottom).

approach, and used the curated network as a starting point for the structure learning algorithm, trying to improve the known structure. Indeed, this yielded a dramatic improvement in score (610 bits), by introducing changes in the connections for about 35 genes (despite not adding new regulators), primarily changing genes from SBF to MBF regulation and from Fkh1 to Fkh2. These modifications suggest novel hypotheses, potentially extending our partial biological knowledge.

5 DISCUSSION

In this paper, we examined the question of learning the dynamics of transcription networks, in terms of the temporal behaviour of regulators, as well as the kinetic parameters governing their effect on their targets. Our method provides a principled approach to handle a wide range of transcriptional network architectures and regulation functions. Unlike previous methods based on probabilistic models (Friedman *et al.*, 2000; Kim *et al.*, 2003; Ong *et al.*, 2002; Pe'er *et al.*, 2001), we addressed the fact that the relevant sizes - transcription rates and regulator activity levels - are usually not measured. This is done by preprocessing steps to extract transcription rates, and by the use of hidden variables to account for unobserved regulator activity levels. Several recent works (Battogtokh *et al.*, 2002; Liao *et al.*, 2003; Perrin *et al.*, 2003) use a fixed regulation diagram to reconstruct unobserved regulator activity profiles and parameters. This work is the first to introduce a network structure learning algorithm in this context. Our algorithm is based on the notion of "ideal" regulators, and we demonstrated its power on the cell cycle regulatory network.

Our DBN-based model to transcription rates and regulator activity levels allows us to handle these biologically relevant quantities despite the indirect measurement of the former and the lack of measurements of the latter. It also allows us to handle the inherently noisy measurement in a principled way, and provides a framework both for learning parameters and for structure learning. However, our model still abstracts away some of the explicit processes that generate the actual observed expression data. A more explicit modeling of these will provide a more principled treatment of different sources of noise in the data. Furthermore, our model does not handle directly any of the upstream events that affect regulator activity. In fact, the current model is an open loop system, such that the regulation of regulator activity is itself viewed as exogenous to the system. By developing a richer modeling language we may capture more complex reaction models, model the upstream regulation of activity levels, and learn systems that involve feedback mechanisms and signalling networks. Finally, such extensions open the possibility of incorporating additional types of data, such as binding sites models, transcription factor binding data or protein-protein interaction data. These could serve not only as additional sources for initialization or validation of models, but also as a primary source of observations for model learning, thus widening the molecular scope covered by our framework.

6 ACKNOWLEDGEMENTS

We thank Uri Alon for many stimulating discussions, and Rani Nelken for comments on the manuscript. A.R. and N.F. were supported by an NIGMS Center of Excellence grant. I.N. is supported by a Horowitz fellowship.

REFERENCES

Baetz, K. & Andrews, B. (1999) Regulation of cell cycle transcription factor swi4 through auto-inhibition of dna binding. *Mol Cell Biol*, **19** (10), 6729–41.

Battogtokh, D., Asch, D. K., Case, M. E. & Arnold, J. (2002) An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of *neurospora crassa*. *Proc Natl Acad Sci U S A*, **99** (26), 16904–9.

Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999) Clustering gene expression patterns. *J. Comp. Bio.*, **6** (3-4), 281–97.

Costanzo, M., Crawford, M., Hirschman, J., Kranz, J., Olsen, P., Robertson, L., Skrzypek, M., Braun, B., Hopkins, K., Kondu, P., Lengieza, C., Lew-Smith, J., Tillberg, M. & Garrels, J. (2001) Ypd, pombepd, and wormpd: model organism volumes of the bioknowledge™ library, an integrated resource for protein information. *Nuc. Acids Res.*, **29**, 75–9.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *J. Comp. Bio.*, **7**, 601–620.

Friedman, N., Murphy, K. & Russell, S. (1998) Learning the structure of dynamic probabilistic networks. In *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*. pp. 129–138.

Guet, C. C., Elowitz, M. B., Hsing, W. & Leibler, S. (2002) Combinatorial synthesis of genetic networks. *Science*, **296** (5572), 1466–70.

Hollenhorst, P. C., Bose, M. E., Mielke, M. R. & Fox, C. A. (2000) Forkhead genes in transcriptional silencing, cell morphology and the cell cycle. overlapping and distinct functions for fkh1 and fkh2 in saccharomyces cerevisiae. *Genetics*, **154** (4), 1533–48.

Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95** (5), 717–28.

Kim, S. Y., Imoto, S. & Miyano, S. (2003) Inferring gene networks from time series microarray data using dynamic bayesian networks. *Brief Bioinform*, **4** (3), 228–35.

Kitano, H. (2002) Computational systems biology. *Nature*, **420** (6912), 206–10.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D. & Young, R. (2002) Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Liao, J. C., Boscolo, R., Tran, L. M., Sabatti, C. & Roychowdhury, V. P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, **100** (26), 15522–7.

McAdams, H. H. & Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A*, **94** (3), 814–9.

McBride, H. J., Yu, Y. & Stillman, D. J. (1999) Distinct regions of the swi5 and ace2 transcription factors are required for specific gene activation. *J Biol Chem*, **274** (30), 21029–36.

Ong, I., Glasner, J. & Page, D. (2002) Modelling regulatory pathways in *e. coli* from time series expression profiles. *Bioinformatics*, **18** (Suppl 1), S241–248.

Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17** (Suppl 1), S215–24.

Perrin, B., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J. & D'Alche-Buc, F. (2003) Gene networks inference using dynamic bayesian networks. *Bioinformatics*, **19 Suppl 2**, II138–II148.

Ronen, M., Rosenberg, R., Shraiman, B. I. & Alon, U. (2002) Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, **99** (16), 10555–10560.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Segal, E., Barash, Y., Simon, I., Friedman, N. & Koller, D. (2002) From promoter sequence to expression: a probabilistic framekwork. In *RECOMB'02*.

Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002) Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, **31**, 64–8.

Simon, I., Barnett, J., Hannett, N., Harbison, C., Rinaldi, N., Volkert, T., Wyrick, J., Zeitlinger, J., Gifford, D., Jaakkola, T. & Young, R. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9** (12), 3273–97.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) Systematic determination of genetic network architecture. *Nat Genet*, **22** (3), 281–5.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. & Brown, P. O. (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A*, **99** (9), 5860–5.