BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 6 2006, pages 747–754
doi:10.1093/bioinformatics/btk034

Systems biology

Using a state-space model with hidden variables to infer transcription factor activities

Zheng Li, Stephen M. Shaw, Matthew J. Yedwabnick and Christina Chan*

Department of Chemical Engineering and Material Science, Michigan State University, East Lansing, MI 48824, USA

Received on November 30, 2005; revised on December 27, 2005; accepted on December 30, 2005
Advance Access publication January 10, 2006
Associate Editor: Steen Knudsen

ABSTRACT

Motivation: In a gene regulatory network, genes are typically regulated by transcription factors (TFs). Transcription factor activity (TFA) is more difficult to measure than gene expression levels are. Other models have extracted information about TFA from gene expression data, but without explicitly modeling feedback from the genes. We present a state-space model (SSM) with hidden variables. The hidden variables include regulatory motifs in the gene network, such as feedback loops and auto-regulation, making SSM a useful complement to existing models.

Results: A gene regulatory network incorporating, for example, feed-forward loops, auto-regulation and multiple-inputs was constructed with an SSM model. First, the gene expression data were simulated by SSM and used to infer the TFAs. The ability of SSM to infer TFAs was evaluated by comparing the profiles of the inferred and simulated TFAs. Second, SSM was applied to gene expression data obtained from *Escherichia coli K12* undergoing a carbon source transition and from the *Saccharomyces cerevisiae* cell cycle. The inferred activity profile for each TF was validated either by measurement or by activity information from the literature. The SSM model provides a probabilistic framework to simulate gene regulatory networks and to infer activity profiles of hidden variables.

Availability: Supplementary data and Matlab code will be made available at the URL below.

Contact: krischan@egr.msu.edu

Supplementary information: http://www.chems.msu.edu/groups/chan/ssm.zip

INTRODUCTION

Cells respond to environmental and physiological changes through an extensive transcriptional regulatory network, which is composed of transcription factors (TFs) and genes. These transcription factors bind to the promoter regions of specific genes to either positively or negatively regulate expression. High throughput technologies, such as cDNA microarray, allow the measurement of expression data of the whole genome; however, genome-wide measurement of the regulatory signals, i.e. transcription factor activities (TFAs), remains a challenge. Clustering has been applied to gene expression data to identify co-regulated genes (Bar-Joseph *et al*., 2002; Eisen *et al*., 1998; Ramoni *et al*., 2002) and Bayesian network analysis has been applied to infer regulatory networks (Friedman *et al*., 2000). The objective of this paper is to infer TFAs from gene expression data. The advent of the genome-wide binding assay to measure protein–DNA interactions has helped to uncover the network structure describing the connections between TFs and genes in *Escherichia coli K12* (Salgado *et al*., 2001) and *Saccharomyces cerevisiae* (yeast) (Lee *et al*., 2002). Given the regulatory network structure and gene expression profiles, the TFAs can be inferred with mathematical modeling.

Several methods have been developed to infer TFAs from gene expression data. A kinetic-based approach (Nachman *et al*., 2004), which modeled mRNA transcription and decay, did not include feedback from genes to TFs. Network component analysis (NCA) (Liao *et al*., 2003; Kao *et al*., 2004), which assumed a log-linear relationship between a gene's expression and its regulatory signals, i.e. TFA, modeled the gene regulatory network as multiple-input motifs. Feedback from genes to TFs within network structures, such as in auto-regulation, feed-forward loops, the regulator chain or the interaction between TFs, is modeled as a 'closed-loop' from the TF to the genes, without explicitly modeling the feedback (Tran *et al*., 2005).

To complement existing approaches, we have developed a state-space model (SSM) with hidden variables that explicitly models feedback in gene regulatory networks to infer the regulatory signals from the gene expression profile. SSM is a subclass of dynamic Bayesian network (DBN). DBN has been applied to infer the transcriptional regulatory network from gene expression profiles, e.g. T-cell activation (Rangel *et al*., 2004; Beal *et al*., 2005). Other models, such as Hidden Markov model (HMM), the Boolean network, and linear and non-linear auto-regression models, are also subclasses of DBN (Murphy and Mian, 1999). SSM assumes the existence of state variables that produce observations that are measurable, as well as hidden variables, which are state variables that do not produce an observation. This feature of SSM is attractive for modeling gene regulatory networks. As illustrated in Figure 1, a gene regulatory network consisting of TFs and genes can be represented by an SSM. The state of each gene produces observations, such as expression profiles, that can be measured with cDNA microarray. The state of each TF is hidden, and thus, does not produce measurable observations. The structure of the connections can be deduced from measurements of protein–DNA interactions.

We demonstrate that SSM can be applied to represent gene regulatory networks of known structures and to infer the TFA from the gene expression profile. We first applied SSM to learn the TFA from data simulated for the gene regulatory network illustrated in Figure 1. Then we applied the model to experimental data from *E.coli* transitioning its carbon source from glucose to acetate

*To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

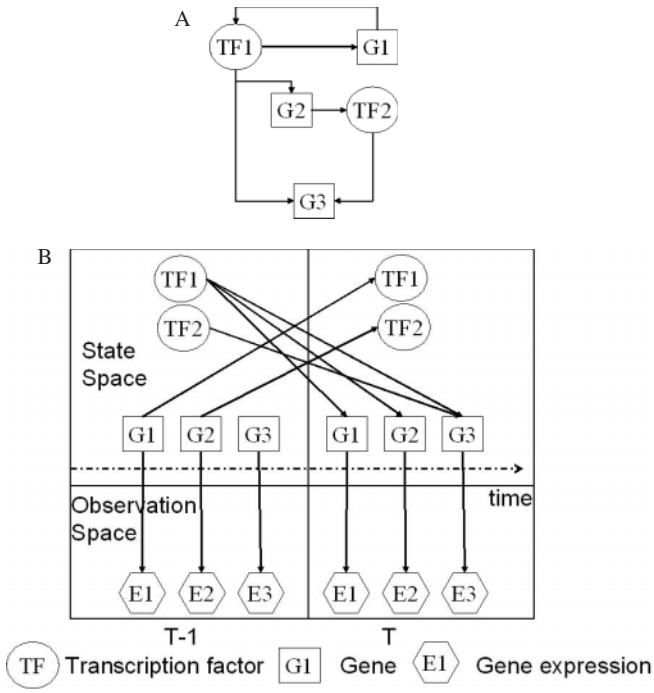

Fig. 1. (**A**) An example of a gene regulatory network with two transcription factors, TF1 and TF2, and three genes, G1, G2 and G3. TF1 regulates G1 and G1 encodes TF1, which formulates auto-regulation between TF1 and G1. TF1 regulates G2, G2 encodes TF2, both TF1 and TF2 regulate G3, which formulates a feed-forward loop between TF1, G2, TF2 and G3. (**B**) A SSM representation of the gene regulatory network illustrated in (A). The transcription factors and genes make up the state space, whereas the observation space is comprised of gene expression data.

(Kao *et al*., 2004) and to cell cycle data from *S.cerevisiae* (Spellman *et al*., 1998). The inferred activity profile for each TF was validated either by a physical measurement (if available) or activity information from the literature. Finally, further extensions to improve the current SSM model are discussed.

MATERIALS AND METHODS

State-space model

State and observation In SSM, a sequence of observations $(O_1, O_2, \ldots, O_T)$ is generated from a sequence of states $(S_1, S_2, \ldots, S_T)$ with the following model:

$$S_t = AS_{t-1} + W_t, \quad (1)$$

$$O_t = BS_t + V_t, \quad (2)$$

where A defines the state transition probability $P(S_t \mid S_{t-1})$, i.e. how the state at time point T can be determined from the state at time point $T-1$, and B defines the observation probability $P(O_t \mid S_t)$, i.e. how the observation at time point at T can be determined from the state at time point T. $W \sim N(0, Q)$ and $V \sim N(0, R)$ define the Gaussian noise of the state and observation, respectively. For convenience of notation, parameters A, B, W and V were combined into a single parameter vector $\theta = (A, B, W, V)$. In the SSM model, the structure is time invariant and the parameters are also time invariant, i.e. the parameters that determine the transition from $T-1$ to T are the same as the parameters that govern the transition from T to $T+1$. However, the time-scale for each loop is not assumed to be the same. For example, in the yeast dataset (Spellman *et al*., 1998), the time-scale for each loop is the same (~7 min). However, in the *E.coli* (Kao *et al*., 2004) example, the 10 time points were taken at 0, 5, 15, 30, 60, 120, 180, 240, 300 and 360 min. Therefore, the time-scale for each loop using the first five data points is not the same as for the last five data points.

Parameter learning $\theta = (A, B, W, V)$ defines an SSM and is learned from N samples of observation data $O = (O_{1:T}(1), \ldots, O_{1:T}(\mathrm{N}))$ by maximizing the likelihood of the observation. An Expectation-Maximization (EM) algorithm is used to learn the parameters. Starting with an initial guess of θ, we perform the E (expectation) step at iteration k to estimate the value of states S_hat with θ_κ and O using inference; then, we perform the M (maximization) step to maximize the likelihood of the conditional probability $P(O,\mathrm{S_hat} \mid \theta)$, such that $\theta_{k+1} = \mathrm{argmax}(P(O,\mathrm{S_hat} \mid \theta))$. For details of the parameter learning, see (Murphy and Mian, 1999). We used Bayes Net Toolbox (Murphy, 2001) for the model computation.

State inference After the parameters are learned with the EM algorithm from the observation data, the value of the state variables, including the hidden variables, can be recursively inferred with the Bayes rule:

$$P(S_t \mid O_{1:t}) = P(O_t \mid S_t) \sum_{O_{t-1}} P(S_t \mid S_{t-1}) P(S_{t-1} \mid O_{1:t-1}). \quad (3)$$

Sampling method to generate simulated data

If the structure and parameters of an SSM are defined, data can be generated with sampling methods such as Gibbs sampling. The function of sample_bnet in Bayes Net Toolbox (Murphy, 2001) was used to generate the simulated gene expression data and TFA profiles.

Represent gene regulatory motifs within SSM framework

Different motifs, such as auto-regulation, feed-forward loops, multiple-inputs and single input, have been identified with protein–DNA interaction measurements (Lee *et al*., 2002). Protein–DNA interaction measurements identify the DNA sequences of a gene to which a TF will bind. Once a TF binds to a DNA sequence (i.e. binding site) of a gene, the TF regulates the transcription, and in turn the level of expression of that gene. Some TFs bind to similar DNA sequences that may exist on many genes, while other TFs bind to specific DNA sequences present in only one or a few genes. TFA is defined as the concentration of the active conformation of a TF that is capable of DNA binding. In the current SSM representation of a gene regulatory network, each TF has one TFA node and its TFA is the same for all binding sites. Therefore, in the model the level of activity of the transcription factor (i.e. the TFA) indicates only whether a transcription factor is activating (either positively or negatively) its genes or not. The likelihood that a gene is activated by a transcription factor is inferred from the data as conditional probabilities. For example, transcription factor TF1 binds to genes G1 and G2 at the same (or different) binding sites. The activity level of TF1 is assumed to be the same for G1 and G2, however, the probabilities that G1 or G2 is activated by TF1 are different as defined by $P(\mathrm{G1} \mid \mathrm{TF1})$ and $P(\mathrm{G2} \mid \mathrm{TF1})$. In addition, the SSM assumes that there is a time delay between binding and transcription.

Here we demonstrate that SSM is able to model these motifs, i.e. auto-regulation, feed-forward loops, multiple-inputs and single input. As shown in Figure 1, SSM is a dynamic model composed of two parts: states and observations. State variables generate observations that are measurable, whereas state variables that do not generate observations are called hidden variables. A static gene regulatory motif can be represented dynamically by connecting genes at time point $T-1$ to the TFs they encode at time point T, and connecting the TFs at time point $T-1$ to the genes that they regulate at time point T. Each TF has a TFA node. Each gene has a state node and an observation node. Duplicating the state and observable variables of a gene will increase the computational load. However, this limitation has an intrinsic advantage. Having the state variable (gene) and the observation variable (gene expression) as separate entities allows the SSM approach

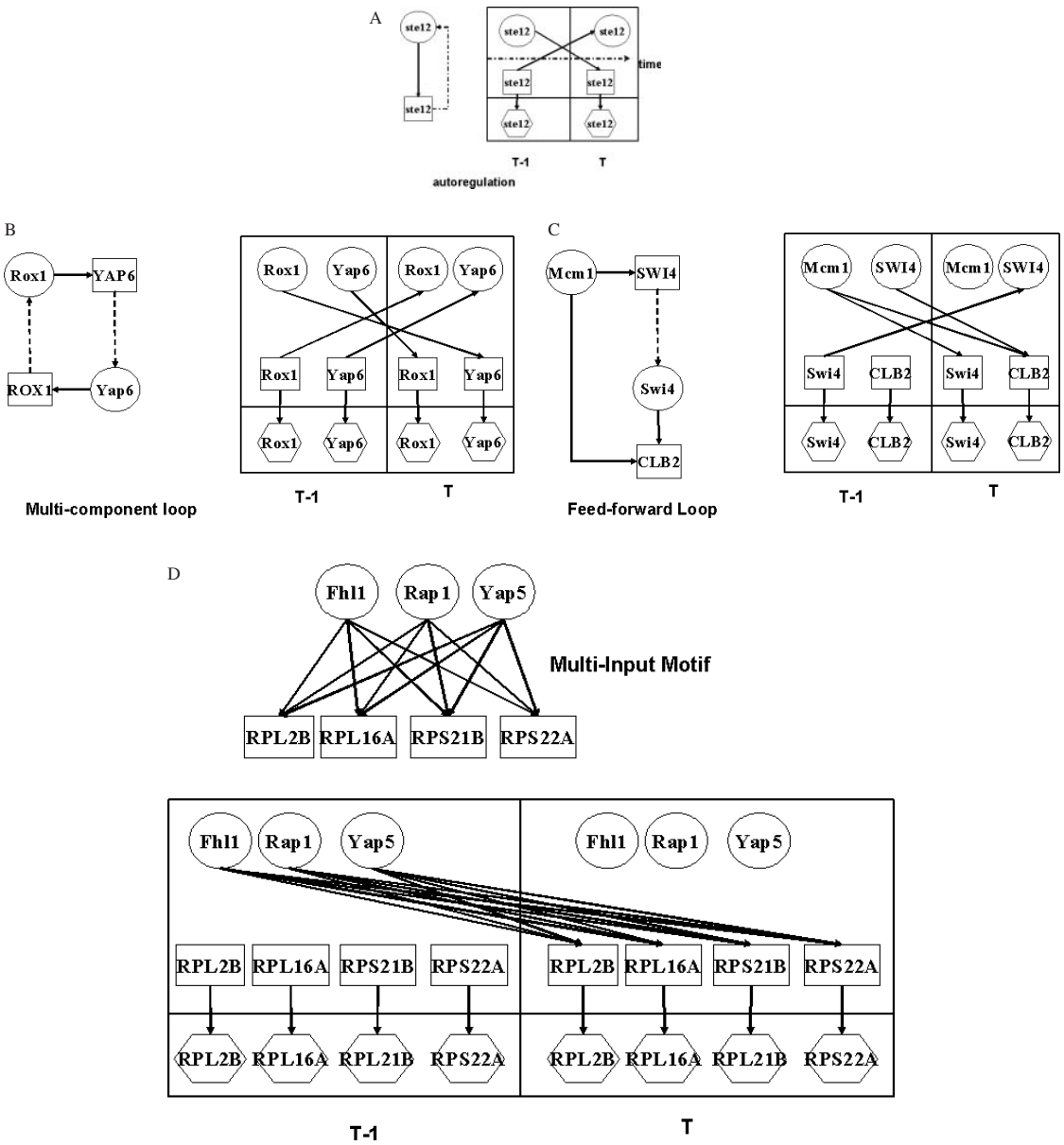

Fig. 2. An example and the SSM representation of the gene regulatory motifs of (**A**) auto-regulation – transcription factor Ste12 (protein) binds to the promoter sequence of gene STE12, which encodes for the transcription factor protein Ste12. (**B**) multi-component loop – transcription factor Rox1 (protein) binds to the promoter sequence of gene YAP6, which encodes for the transcription factor protein Yap6. The Yap6 protein binds to the promoter sequence of gene ROX1, which encodes for the transcription factor protein Rox1. (**C**) feed-forward loop – transcription factor Mcm1 (protein) binds to the promoter sequence of gene SWI4, which encodes transcription factor protein Swi4. Transcription factors Swi4 and Mcm1 bind to the promoter sequence of gene CLB2. (**D**) multiple-input motif – transcription factors Fhl1, Rap2 and Yap5 all bind to the promoter sequences of genes RPL2B, RPL16A, RPS21B and RPS22A.

to take into account potential effects, such as post-transcriptional effects (such as RNA decay) and measurement noise. If RNA decay and measurement noise did not occur then the state and observation variables could be combined into one variable. Under those circumstances a gene that is transcriptionally activated would be measured at the mRNA (gene expression) level to be activated, thus $\Pr(E = 1 \mid G = 1) = 1$. However, the gene expression level is determined by the net effect of mRNA transcription and degradation. Effects, such as RNA decay (post-transcriptional modification) and noise in the microarray measurement, could result in conditional probabilities $\Pr(E = 1 \mid G = 1) < 1$. In other words, a gene that is transcriptionally activated may be measured at the gene expression level to be inactivated if mRNA decay dominates over transcription (Wang *et al*., 2002). Figure 2 illustrates a graphical representation of the auto-regulation, feed-forward loop, multiple-input and multi-component loop motifs.

Threshold determination

We used a function with a definable threshold Th to discretize the gene expression data. Any gene that showed a change larger than Th, based upon $\log_2$ ratio, was assigned a discrete value of 1, or otherwise was assigned a value of 0. Thus, a threshold of 1 indicates that a 2-fold (21) change in the expression of a gene, relative to its initial state, is significant.

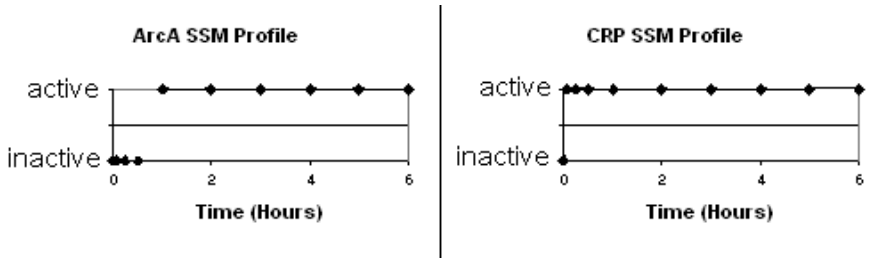

Fig. 3. The results of using an SSM to analyze an *E.coli* system. The SSM predicts that CRP is inactive initially and then active from the second time point onwards, as expected from the cAMP measurement (Kao, 2004). The SSM predicts that ArcA is inactive for the first four time points, but is activated for the last six time points. Catabolite repressor protein (CRP) is a well-known regulator involved in carbon source transition, its activation requires binding of cAMP (Kao, 2004). The CRP level responds to the transition in carbon source by regulating uptake and metabolic genes in preparation for a less favorable carbon source (Kao, 2005). The activation of ArcA has been proposed to reduce the production of electron donors and the level and activity of aerobic respiratory apparatus during the carbon source transition by downregulating the TCA cycle enzymes (Kao, 2004; Nystrom 1996).

SSM requires an optimal threshold in order to obtain reliable results. To find an optimal threshold, we evaluate the TFA data for each TF over various thresholds between -1 and 1. The optimal threshold, which gives the most appropriate profile for each TF, is determined by comparing the TFA results with known (e.g. measured or literature) values. For example, the gene expression data that we used for *S.cerevisiae* were taken over 18 time points that spanned two cell cycles (Spellman *et al*., 1998). The TFs we studied in *S.cerevisiae* are known to have phase-specific activity during the cell cycle (Aerne *et al*. 1998; Baetz *et al*., 1999; Kovacech *et al*., 1996; Lee *et al*., 2002; MacKay *et al*., 2001; McInerny *et al*., 1997; Oehlen *et al*., 1996; Spellman *et al*., 1998). Therefore, we assumed that the activity profile of each TF during the first cell cycle should be repeated in the second cell cycle. We identified the optimal threshold as the one that predicted this cyclic behavior for all the transcription factors.

RESULTS

Inferring TFA from simulated data

Before testing the SSM with experimental data where the transcription factor activities are unknown, we applied it to a simulated system. We created a simple regulatory network (Fig. 1A) with the feed-forward loop and auto-regulation motifs containing two TFs and three genes. From the known network structure, we constructed a SSM representation of the network (Fig. 1B) and predefined the θ parameter, which is shown in the Supplementary Table 1a. Using the network reconstructed by the SSM, we simulated the dynamic profiles of TFA and gene expression levels with the sampling method discussed in the Materials and methods section. Given the simulated gene expression profile (shown in Supplementary Table 1c) and the network structure, SSM learned the parameters θ and the TFA profile. The learned parameters are shown in Supplementary Table 1b. They match closely to the predefined parameters in Supplementary Table 1a. The learned or inferred TFA was compared with the simulated TFA, as shown in Supplementary Figure 1. The learned TFA matched 95% of the time points (38 out of 40) of the simulated TFA. The results of this simulation provided confidence that the SSM may be able to analyze a real system. The simulations were performed on Matlab R13, Windows XP, on a PC with Celeron CPU 2.4 GHZ and RAM 512 MB. The simulation time was 328.5 s.

SSM requires sufficient data to infer the parameters. We evaluated how many data points are needed with the simulation data by varying the number of time points from 6 to 40. Ten time points were needed to correctly infer 80% of the TFA profile and this percentage increased with the number of time points used. The sampling time needs to be small enough to capture the dynamic profile, which will vary with the biological system being modeled.

Escherichia coli We applied SSM to a model system: *E.coli* transitioning from glucose to acetate as a carbon source. The gene expression data were obtained from (Kao *et al*., 2004) and the regulatory information is available from the regulonDB database. The SSM model included two TFs, CRP and ArcA, and eight of the genes (shown in Supplementary Table 2) that are regulated by the two TFs. The dynamic profiles of these two TFs were learned from the gene expression levels of the eight genes. In the work by Kao *et al*. (2004), cAMP was measured to indirectly indicate the activity of CRP, since the activation of CRP requires the binding of cAMP. From the measurement of cAMP (Fig. 5 in Kao, 2004), we can see that the level decreased from its initial high but remained upregulated, around 10-fold above the basal level, from the second time point onward. Indeed, SSM inferred that CRP is active from the second point onwards (Fig. 3). For the first time point, however, CRP is predicted to be inactive. The expression level at the first time point is the reference; all subsequent expression levels are measured relative to the expression level at the first time point. SSM also identified that ArcA was inactive for the first four time points and active for the remaining 6 time points (Fig. 3). An ArcA measurement is not readily obtainable for comparison.

Saccharomyces cerevisiae Next, we applied SSM to model several of the common regulatory motifs in *S.cerevisiae*. According to Lee *et al*. (2002), 39 out of the 106 studied regulators were involved in feed-forward loops and 10 of the 106 were involved in auto-regulation. Lee *et al*. (2002) also found that in combinations of two or more of the 106 regulators, 295 were involved in multiple-input motifs. Therefore, we selected TFs with the aforementioned regulatory motifs, and whose activities could be verified by the literature. Namely, MCm1, Swi4, Swi5 and Swi6, which are well-studied and well-understood (Aerne *et al*., 1998; Baetz *et al*., 1999; Kovacech *et al*., 1996; Lee *et al*., 2002; MacKay *et al*., 2001; McInerny *et al*., 1997; Oehlen *et al*., 1996; Spellman *et al*., 1998) TFs. SSM was applied to analyze the system illustrated in Figure 4. The gene regulatory motifs, feed-forward (Mcm1 → SWI5 → Swi5 → YJL160C + PIR1 + PIR3), multiple-input

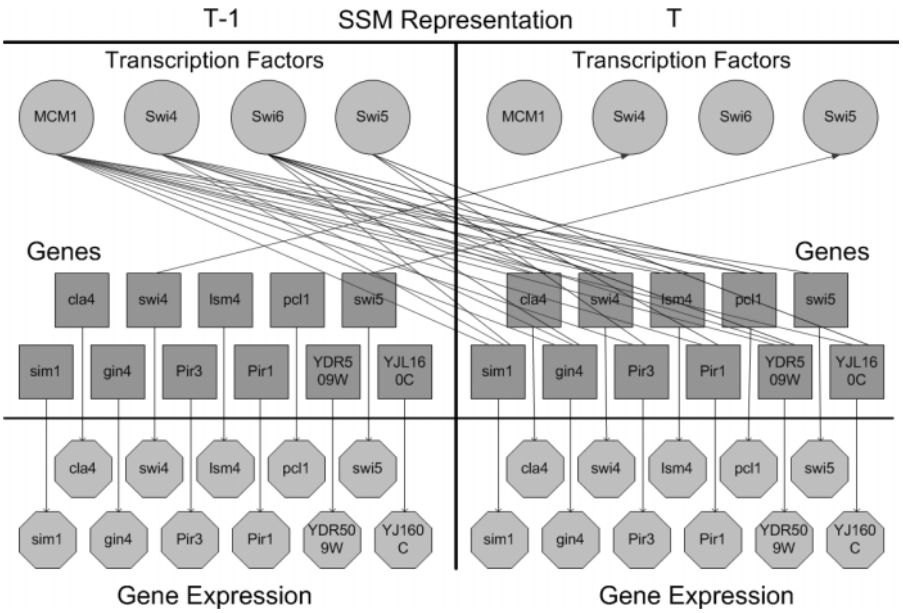

Fig. 4. An SSM representation of the *S.cerevisiae* (yeast) system studied. The gene regulatory motifs were taken from (Lee *et al*. 2002), including feed-forward (Mcm1 → SWI5 → Swi5 → YJL160C + PIR1 + PIR3), multiple-input (Mcm1 + Swi4 + Swi6 → CLA4 + SWI4 + LSM4 + PCL1 + SIM1 + GIN4 + YDR509W), auto-regulation (Swi5 → SWI5; Swi4 → SWI4).

(Mcm1 + Swi4 + Swi6 → CLA4 + SWI4 + LSM4 + PCL1 + SIM1 + GIN4 + YDR509W) and auto-regulation, (Swi5 → SWI5; Swi4 → SWI4) were obtained from (Lee *et al*. 2002).

We evaluated the ability of the SSM to infer TFA profiles during the cell cycle of *S.cerevisiae*. Lee *et al*. (2002) performed a genome-wide binding analysis to obtain the connectivity information between the TFs and genes in yeast. We coupled the connectivity information (Lee *et al*., 2002) with gene expression data taken from yeast cultures synchronized by α-factor arrest (Spellman *et al*., 1998). Of the four synchronization methods used by Spellman *et al*., we chose the data obtained with the α-factor method because it presented the least amount of missing data for the genes studied. With the α-factor arrest method, the data were sampled every 7 min, which captured approximately two cell cycles with the 18 time points (Spellman *et al*., 1998). This provided 9 time points (~63 min) for each cell cycle. Each phase (i.e. M, G1, S and G2) within a cell cycle takes ~15 min (Liao *et al*., 2003). In other words, one cell cycle is ~60 min. Therefore, a 7 min sampling time is small enough to capture the phase change profile within a cell cycle. The binding motifs and gene expression data were used by SSM to infer the TFAs.

The SSM predictions are consistent with the literature results. SSM inferred that MCm1 is active during the G2/M/G1 phases, Swi4 and Swi6 are active during the G1 and S phases, and Swi5 is active during the M and G1 phases. We confirmed the predictions (Fig. 5) made by SSM with the literature. Past studies found that MCm1 induced the expression of many genes during the G2/M/G1 phases. High transcription of both FAR1 and STE2 in the G2/M phases requires MCm1 (Oehlen *et al*., 1996). MCm1 is also known to induce the transcription of CLN3, SWI4 and CDC6 at the M/G1 boundary (Spellman *et al*., 1998). In contrast, Swi4 induced genes in the G1 and S phases. Swi4 is the DNA binding component of SBF (Baetz *et al*., 1999). Baetz *et al*. (1999) indicated that SBF promotes the induction of gene expression at the G1/S-phase transition of the mitotic cell cycle. MacKay *et al*. (2001) also showed that a complex containing Swi4 induces CLN1 and CLN2 transcription in the late G1 and drives the transition to S. Similarly, Swi6 induced genes during the G1 and S phases. The activity of Swi6 is very similar to that of Swi4, and these two factors are known to be connected (Baetz *et al*., 1999; MacKay *et al*., 2001). MacKay *et al*. (2001) showed that a Swi4-Swi6 complex induces CLN1 and CLN2 transcription in late G1 until S. Baetz *et al*. (1999) also suggested that the DNA binding domain of Swi4 is inaccessible in the full-length protein when not complexed with Swi6. In contrast, Swi5 was activated during the M phase and the M/G1 boundary. Kovacech *et al*. (1996) found that Swi5 is partially responsible for the peak in EGT2 expression during late M and early G1 phases. Aerne *et al*. (1998) found that Swi5 regulates the expression of PCL2, PCL9, and the SIC1 Cdk inhibitor in the late M phase.

In summary, SSM was applied to three systems, a simulated gene regulatory network and two experimental systems, *E.coli* and *S.cerevisiae*. A comparison of the three examples is presented in Table 1. The simulation study confirmed the ability of SSM to infer network parameters and state values from observational data. Application of SSM to the experimental systems illustrates that TFAs can be inferred from the gene expression data given the regulatory network structure.

DISCUSSION

SSM is a subclass of DBN. DBN has been applied to infer the structure of regulatory networks from temporal gene expression data (Rangel *et al*., 2004; Beal *et al*., 2005; Ong *et al*., 2002; Perrin *et al*., 2003; Nachman *et al*., 2004). Ong *et al*. (2002) explicitly included operons as hidden variables in the model to facilitate

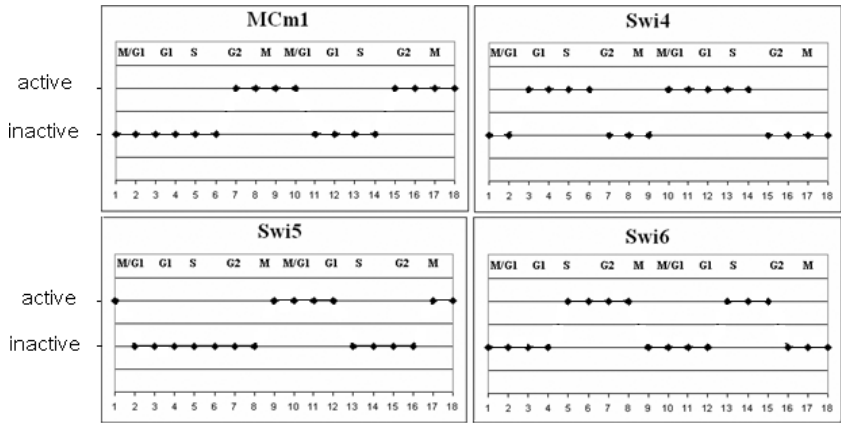

Fig. 5. The results of using an SSM to analyze a yeast system. The SSM predictions closely followed the experimental trends and phases in which each transcription factor is known to be active or inactive (Aerne *et al*., 1998; Baetz *et al*., 1999; Kovacech *et al*., 1996; Lee *et al*., 2002; MacKay *et al*., 2001; McInerny *et al*., 1997; Oehlen *et al*., 1996; Spellman *et al*., 1998). The SSM inferred that MCm1 is active during the G2/M/G1 phases, that Swi4 and Swi6 are active during the G1 and S phases, and that Swi5 is active during the M and G1 phases.

Table 1. Comparison of three examples of SSM application

SSM application	Network motifs	Data input	Validation
Simulation	Auto-regulation	Simulated data by sampling a predefined network	Compared with sampled data
	Feed-forward		Compared with predefined parameters
E.coli	Multiple-input	Discretized experimental data	Compared with measurements or NCA results
S.cerevisiae	Multiple-input Feed-forward Auto-regulation	Discretized experimental data	Compared with literature

the incorporation of a priori biological knowledge of the co-expressed genes to improve the quality of the analysis. Here, we explicitly included TFs in the model and included the connections between TFs and genes obtained through binding analysis (Lee *et al*., 2002). SSM has the potential to infer unmeasurable, as well as unmeasured, connections and events.

The benefit of the SSM over existing models [e.g. NCA (Liao *et al*., 2003), kinetic modeling (Nachman *et al*., 2004) is that all gene regulatory motifs, including feedback from gene to TFs, are explicitly modeled. NCA implicitly models the feedback from gene to TFs (Tran *et al*., 2005). In kinetic modeling (Nachman *et al*., 2004), feedback from gene to TFs is not incorporated. In contrast, SSM can explicitly model the feedback from gene to TFs, such as the auto-regulatory motif. This facilitates the incorporation of domain or experimental knowledge. For example, if a TF is experimentally knocked–out or silenced, the SSM approach could easily incorporate this information for the auto-regulatory motif.

In SSM, the nonlinear relationship between the TFs and genes are quantified with conditional probabilities, i.e. $P(O_t \mid S_{t-1})$. By using conditional probabilities, SSM does not presuppose a relationship between the TFs and genes in the model, whereas NCA assumes a log-linear relationship and kinetic modeling assumes a form for the rate law, such as Michaelis–Menten kinetics. Conversely, NCA (Liao *et al*., 2003) and kinetic modeling (Nachman *et al*., 2004) can infer continuous profiles of TFA. Another advantage of kinetic modeling is the ability to explicitly model both mRNA transcription and decay. In the current application of SSM the gene expression data are discretized, thus allowing us to infer when the TFAs are active. This was sufficient to allow verification of the model predictions with the literature. For example, the model predicted the phase in the cell cycle in which the genes that are regulated by a TF are activated, which can be compared with the phase of the cell cycle in which the genes are known to be activated in the literature.

The SSM assumes there is a time delay between binding and transcription. As illustrated in Figure 1 in the Supplementary data, by considering the state values at time T as a function of the state values at time $T-1$, the SSM implicitly assumes a time delay for all TF effects on gene expression. In other words, although the binding of a TF to the DNA sequence of a gene may occur quickly (McAdams and Arkin, 1997), there is a time offset between the binding of the TF to the DNA sequence of a gene and the onset of transcription. This time offset ranges between minutes to hours (Kerszberg, 2004). It has been shown that incorporating a time delay in modeling gene regulatory networks is critical to inferring the oscillatory behavior of NF-κB (Monk, 2003). We further evaluated this assumption by allowing the genes to be regulated by the current TFA in the yeast dataset. Without the time delay, the cyclic TFA could not be inferred as illustrated in Figure 2 in the Supplementary data. This, in addition to the previous study (Monk, 2003), suggests that this biologically relevant time delay (Kerszberg, 2004) must be incorporated in the model to accurately infer the TFA profiles. In some cases, if the actual

time delay is on the order of minutes and the measurements are taken on the order of hours, then considerable error would be introduced. In those cases, it would be more appropriate to incorporate the connection between the TFs and genes in the same time slice.

In the simulation study, the TFA inferred by the SSM matched the simulated TFA well but not exactly. The mismatches may be due in part to the EM algorithm being a local optimization method (Ong *et al*., 2002), in other words, the algorithm cannot guarantee a TFA of (global) maximal likelihood. The optimization could be improved by either running EM multiple times from different starting points or using a global search algorithm, such as Markov Chain Monte Carlo (MCMC) (Murphy, 2001).

The SSM model determines an optimal threshold value for discretizing the gene expression data based upon a priori knowledge of the TFA. If no a priori knowledge is available for the TFA dynamics, this can be addressed one of two ways. In one approach, the TFAs could be estimated from other approaches, e.g. NCA, and the estimated TFA could be used to determine the threshold value. In the other approach, the optimal threshold could be determined in the SSM by including the threshold value (Th) as a part of the parameter learning process, i.e. in the parameters $\theta = (A, B, W, V, \mathrm{Th})$. How the observation data $O(\mathrm{Th}) = (O_{1:T}(1), \ldots, O_{1:T}(N))$ is discretized depends on the threshold value, e.g. a very high threshold value will set all the genes in the inactive state while a very low threshold value will set all the genes in the active state. The Expectation-Maximization (EM) algorithm could be used to learn the parameters. Starting with an initial guess of θ, we can perform the E (expectation) step at iteration k to estimate the value of states S_hat with θ_κ and $O(\mathrm{Th})$ using inference; then, we can perform the M (maximization) step to maximize the likelihood of the conditional probability $P(O(\mathrm{Th}),\mathrm{S_hat} \mid \theta)$, such that $\theta_{k+1} = \mathrm{argmax}(P(O(\mathrm{Th}),\mathrm{S_hat} \mid \theta))$. Therefore, the parameters, including the threshold value, can be inferred by maximizing the probability of the observation and state values given the inferred parameters (e.g. conditional probabilities).

The current SSM infers the most probable model, given the observed data, by approximating the underlying structure of the noise to be Gaussian (Murphy and Mian, 1999; Perrin *et al*., 2003). Gene expression is an inherently stochastic phenomenon (McAdams and Arkin, 1997). SSM modeled the regulatory networks stochastically using conditional probabilities. This probabilistic SSM may capture some of the stochastic nature of the gene regulatory network, but an accurate representation of the stochasticity requires further understanding of the underlying structure of the noise. Without knowing the structure of the noise, studies have assumed it to be Gaussian (Perrin *et al*., 2003).

The current SSM model could be extended to incorporate a step to learn the structure before inferring the TFAs by searching for a network that gives the maximal likelihood against the observation. The structural information obtained from the binding analysis could be used to construct the initial network as a starting point for the search (Nachman *et al*. 2004). Alternatively, the connections indicated by the interaction data could be used to define a priori probabilities of the connections in the network (Hartemink *et al*., 2002). Thus, the connections that are supported by the interaction measurements would have a higher likelihood of being valid connections in the network than the unsupported connections. By including a step to learn the structure, it could help refine the network by inferring the interactions that are unmeasurable or missing due to error (noise) in the measurement. A more accurate network provides more confidence to the inferred TFA profile. In this study, the model size was limited by the computational tool that was used, namely Bayes Net Toolbox (Murphy, 2001), which is on a Matlab platform. Future work will involve developing an executable SSM in a C^{++} version of Bayes Net Toolbox, using Probabilistic Networks Library (PNL) http://www.intel.com/technology/computing/pnl/ to handle larger model sizes.

ACKNOWLEDGEMENTS

This work is supported in part by the National Science Foundation (BES 0222747, BES 0331297 and 0425821), the Environmental Protection Agency and the Whitaker Foundation. Z.L. is supported in part by a deVlieg fellowship for computational engineering at Michigan State University.

Conflict of Interest: none declared.

REFERENCES

Aerne,B.L. *et al*. (1998) Swi5 controls a novel wave of cyclin synthesis in late mitosis. *Mol. Biol. Cell*, **9**, 945–956.

Baetz,K. and Andrews,B. (1999) Regulation of cell cycle transcription factor Swi4 through auto-inhibition of dna binding. *Mol. Cell Biol*., **19**, 6729–6741.

Bar-Joseph,Z., Gerber,G., Gifford,D., Jaakkola,T. and Simon,I. (2002) A new approach to analyzing gene expression time series data. In *Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology*, pp. 39–48.

Beal,M.J. *et al*. (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **21**, 349–356.

di Bernardo,D. *et al*. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol*., **23**, 377–383.

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

Friedman,N. *et al*. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol*., **7**, 601–620.

Hartemink,A.J. *et al*. (2002) Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput*., 437–449.

Kao,K.C. *et al*. (2004) Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc. Natl Acad. Sci. USA*, **101**, 641–646.

Kao,K.C. *et al*. (2005) A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. *J. Biol. Chem*., **280**, 36079–36087.

Kerszberg,M. (2004) Noise, delays, robustness, canalization and all that. *Curr. Opin. Genet. Dev*., **14**, 440–445.

Kovacech,B. *et al*. (1996) EGT2 gene transcription is induced predominantly by Swi5 in early G1. *Mol. Cell. Biol*., **16**, 3264–3274.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Liao,J.C. *et al*. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

MacKay,V.L. *et al*. (2001) Early cell cycle box-mediated transcription of CLN3 and SWI4 contributes to the proper timing of the G1-to-S transition in budding yeast. *Mol. Cell Biol*., **21**, 4140–4148.

McAdams,H. and Arkin,A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA*, **94**, 814–819.

McInerny,C.J. *et al*. (1997) A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. *Genes Dev*., **11**, 1277–1288.

Monk,N.A.M. (2003) Oscillatory expression of Hes1, p53 and NF-kB driven by transcriptional time delays. *Curr. Biol*., **13**, 1409–1413.

Murphy,K. and Mian,S. (1999) Modelling gene expression data using dynamic Bayesian networks. *Technical Report*, University of California, Berkeley.

Murphy,K. (2001) The Bayes net toolbox for Matlab. *Proceedings of Interface: Computing Science and Statistics*, p. 33.

Nachman,I. *et al*. (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, **20** (Suppl. 1), i248–i256.

Nystrom,T. *et al*. (1996) Bacterial defense against aging: role of the *Escherichia coli* ArcA regulator in gene expression readjusted energy flux and survival during stasis. *EMBO J*., **15**, 3219–3228.

Oehlen,L.J. *et al*. (1996) Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. *Mol. Cell Biol*., **16**, 2830–2837.

Ong,I. *et al*. (2002) Modelling regulatory pathways in *E.coli* from time series expression profiles. *Bioinformatics*, **18**, S241–S248.

Perrin,B.-E. *et al*. (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**, S138–S148.

Ramoni,M.F. *et al*. (2002) Cluster analysis of gene expression dynamics. *Proc. Natl Acad. Sci. USA*, **99**, 9121–9126.

Rangel,C. *et al*. (2004) Modelling T cell activation using gene expression profiling and state space models. *Bioinformatics*, **20**, 1361–1372.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Salgado,H. *et al*. (2001) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acid Res*., **29**, 72–74.

Tran,L.M. *et al*. (2005) gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab. Eng*., **7**, 128–141.

Wang,Y. *et al*. (2002) Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA*, **99**, 5860–5865.