BIOINFORMATICS Bioinformatics Advance Access published September 11, 2006

Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities

Guido Sanguinetti[a], Neil D. Lawrence[a] and Magnus Rattray[b]

[a] Department of Computer Science, Regent Court, 211 Portobello Road, Sheffield, S1 4DP, U.K.,
[b] School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, U.K.

Associate Editor: John Quackenbush

ABSTRACT

Motivation Quantitative estimation of the regulatory relationship between transcription factors and genes is a fundamental stepping stone when trying to develop models of cellular processes. Recent experimental high-throughput techniques such as Chromatine Immunoprecipitation provide important information about the architecture of the regulatory networks in the cell. However, it is very difficult to measure the concentration levels of transcription factor proteins and determine their regulatory effect on gene transcription. It is therefore an important computational challenge to infer these quantities using gene expression data and network architecture data.

Results We develop a probabilistic state space model that allows genome-wide inference of both transcription factor protein concentrations and their effect on the transcription rates of each target gene from microarray data. We use variational inference techniques to learn the model parameters and perform posterior inference of protein concentrations and regulatory strengths. The probabilistic nature of the model also means that we can associate credibility intervals to our estimates, as well as providing a tool to detect which binding events lead to significant regulation. We demonstrate our model on artificial data and on two yeast data sets in which the network structure has previously been obtained using Chromatine Immunoprecipitation data. Predictions from our model are consistent with the underlying biology and offer novel quantitative insights into the regulatory structure of the yeast cell.

Availability MATLAB code is available from
`http://umber.sbs.man.ac.uk/resources/puma`.

1 INTRODUCTION

Quantitative modelling of the regulatory network of the cell is one of the grand challenges of bioinformatics. Although recent techniques such as Chromatine Immunoprecipitation (ChIP) (Lee et al., 2002; Harbison et al., 2004) have uncovered much information about the architecture, or *connectivity*, of the networks, any quantitative model would require the knowledge of both the concentration of transcription factor proteins at a given time and the intensity with which they can promote or repress transcription of their target genes. Experimental estimation of these variables meets with insurmountable obstacles: measuring protein concentrations is a notoriously difficult task, and little help can be gleaned from knowledge of transcription factor gene expression levels. Transcription factors are often post-transcriptionally regulated and have low and noisy expression levels. Furthermore, the effect a transcription factor has on a target gene depends greatly on the experimental conditions (Harbison et al., 2004; Papp and Oliver, 2005), making experimental estimation of the strength of regulatory relationships a difficult task.

An idea that has gained a lot of interest in recent years has been to infer information about regulatory activity from the expression levels of target genes. Using information from ChIP experiments or from sequence analysis about the connectivity of the network and genome-wide microarray data for the expression levels of the targets, it should be possible to gain insights on the activity of the transcription factors.

This idea has been pursued by a number of research groups (Liao et al., 2003; Gao et al., 2004; Boulesteix and Strimmer, 2005). Most methods aim to infer a matrix of transcription factor activities (TFAs), which are supposed to sum up in a single number the concentration of the transcription factor at a certain experimental point and its binding affinity to its target genes. The techniques used are modified forms of linear regression, where the TFAs are obtained as regression coefficients. These models were able to obtain results in broad accordance with the existing biological knowledge, and have the advantage of being fast and practical for *genome-wide* analysis. However, a major limitation of these methods is that TFAs inferred are constant across genes, *i.e.* they can only infer the mean influence of a transcription factor on its target genes[1]. Also, none of these methods is probabilistic and therefore it is difficult to see how credibility intervals can be obtained, as well as how the models can be made robust against false positives (a notorious problem of ChIP data).

A more sophisticated approach was taken in Nachman et al. (2004). Here, dynamical Bayesian networks were used to model the concentrations of transcription factor proteins. The binding of a transcription factor protein to a target gene was then modelled using a binomial distribution, allowing for gene-specific effects to be considered. Although the model is elegant and more realistic than the regression based models, its computational costs ruled out *genome-wide* investigation.

In a recent study, we proposed a probabilistic extension of the linear regression model to describe gene and environment specific effects (Sanguinetti et al., 2006). We allowed a different regression matrix for each gene, and rendered the model identifiable by using a prior distribution on the gene-specific TFAs that was shared across all genes. The temporal structure of the data was captured by requiring the prior distribution to form a stationary Markov chain. Bayesian inference was exactly tractable with this model and the posterior estimates for the gene-specific TFAs proved to be consistent with known biological regulatory relationships. However,

[1] Gao et al. (2004) introduces gene-specific regulatory strengths by considering correlations across different conditions, but their method is incapable of obtaining gene-specific results from single condition data.

although useful, the model lacked in interpretability. While one could argue that the TFAs obtained by regression are monotonically linked to the actual protein concentrations, it was hard to separate the effect of high protein concentrations from the effect of high regulatory strength. This is a serious problem when trying to use the results of the model for further analysis, as it is impossible to distinguish between what is a time varying quantity (the protein concentration) from what should be considered a condition dependent parameter (the strength with which a transcription factors regulate one of its targets).

Here, we propose a modified model that separately models concentrations and regulatory strengths. While this makes the model no longer exact, it has the advantage of providing probabilistic estimates for the intensities of the regulatory relationships, hence allowing the genome-wide quantitative reconstruction of the dynamical process of transcriptional regulation. We model the temporal structure of the data by using a Markov chain, and we develop an efficient variational EM algorithm for the estimation of the model parameters and posterior statistics.

Models with a Markov chain prior on continuous valued latent variables are special cases of dynamical Bayesian networks known as state space models, or Kalman filters (Kalman, 1960; Haykin, 2002), and these models have previously been used in microarray time-series analysis. For example, Beal et al. (2005) recently applied a state space model to learn about interactions between a subset of genes from highly replicated microarray experiments. However, they did not make use of prior knowledge about potential regulatory interactions to explicitly infer the activity of transcription factors. This knowledge greatly reduces the search space, allowing genome-wide applications to become feasible and reducing the need for substantial experimental replication.

Recently, Sabatti and James (2006) also presented an extension of the linear regression model that provides separate estimates of protein concentrations and regulatory intensities for regulatory networks of known connectivity. They use Markov chain Monte Carlo to perform approximate inference of the posterior distribution of the concentration and intensities, and provided an R package implementing it. While their model is in many ways similar to ours, in other ways the models differ substantially. A more detailed comparison of the two models is addressed in the Discussion section.

We validate the model both on synthetic data and real data from two yeast time series: the benchmark cell cycle data set of Spellman et al. (1998) and the recent metabolic cycle data set of Tu et al. (2005). The connectivity data we use in the cell cycle case is that obtained by Lee et al. (2002) using ChIP, while for the metabolic cycle data we combine the ChIP data of Lee et al. (2002) with the more recent data of Harbison et al. (2004).

Our results are largely confirmed in the biological literature, but, using the gene-specific nature of our model, we also manage to predict biologically plausible regulatory relationships which are not documented in the literature. The probabilistic nature of our model also means that we can identify false positives in the ChIP data as regulatory relationships below a certain significance threshold.

2 METHODS

The logged gene expression measurements are collected in a design matrix $\mathbf{Y} \in \Re^{N\times T}$, where N is the number of genes and T the number of time points in a time-series microarray experiment. The elements of $\mathbf{Y}$ are written $y_n(t)$ to denote the expression level of gene n at time t. The connectivity of the network is represented by a binary matrix $\mathbf{X} \in \Re^{N\times q}$, where q is the number of transcription factors; element (i,j) of $\mathbf{X}$ is one if transcription factor j binds gene i, zero otherwise.

2.1 Model

We propose a discrete time state space model (Kalman, 1960) which takes the form,

$$y_n(t) = \sum_{m=1}^{q} X_{nm} b_{nm} c_m(t) + \mu_n + \epsilon_{nt} , \quad (1)$$

$$c_m(t) = \gamma_m c_m(t-1) + \eta_{mt} . \quad (2)$$

Equation (1) describes a linear model of the effect that each transcription factor has on the expression level of each gene. Equation (2) describes the dynamics of the underlying transcription factor concentrations as a 1st order Markov chain. Elements of the *concentration matrix* $\mathbf{C} = [c_m(t)]$ represent the relative concentration of transcription factor protein m at time t. Elements of the *activity matrix* $\mathbf{B} = [b_{nm}]$ model the regulatory strength with which transcription factor m influences the target gene n. The mean vector $\boldsymbol{\mu} = [\mu_n]$ represents the baseline expression level for each gene, *i.e.* the expression level in the absence of any of the known transcription factors being bound. The variables ϵ_{nt} and η_{mt} each represent zero mean i.i.d. Gaussian noise on the measurements and underlying process respectively. The measurement noise has variance σ^2, $\epsilon_{nt} \sim \mathcal{N}(0,\sigma^2)$. The process noise $\eta_{mt} \sim \mathcal{N}(0, 1-\gamma_m^2)$ has a variance that ensures the Markov process governing the dynamics of the components of $c_m(t)$ is stationary with unit variance (we set $c_m(1) \sim \mathcal{N}(0,1)$). The parameter vector $\boldsymbol{\gamma} = [\gamma_m]$ has components $\gamma_m \in [0,1]$ that determine the temporal variability in the concentration of transcription factor m. Values of γ_m that are close to one indicate very little variability in time while lower values correspond to larger changes. Intermediate values of this parameter indicate a smoothly varying transcription factor concentration profile.

The assumption that the activities b_{nm} are independent of time is reasonable for time-series data when the conditions (*e.g.* pH, temperature, growth medium *etc.*) are kept relatively constant. Of course, large changes in the relative proportion of different transcription factors would eventually lead to the simple linear relationship in equation (1) breaking down, but by making this simplification the model remains sufficiently tractable for practical application to a genome-wide study.

An important feature of the model is that the connectivity matrix $\mathbf{X}$ is sparse, mirroring the biological fact that few transcription factors bind any single gene. This is crucial for the identifiability of our model: models of this high dimensionality (in both the measurement and latent space) with full matrices would require very large numbers of replicates to be identifiable, and even then would be unlikely to correctly identify the sparsity structure of the connectivity matrix. However, the presence of the matrix $\mathbf{X}$ ensures that only a few of the b_{nm} need to be estimated, making the task possible with the limited amount of data normally available in microarray time-series data. The degree of sparsity of the matrix $\mathbf{X}$ varies depending on the organism studied and the type of data used to build the network. Typically, regulatory networks for yeast constructed from ChIP data can be expected to lead to matrices with approximately only 1% nonzero entries, while networks for higher organisms constructed from motif data (see *e.g.* Xie et al., 2005) lead to matrices with approximately 20% nonzero entries, although these are expected to contain very many false positives.

The model is over-parameterised and we therefore use Bayesian methods for inferring the posterior probability of the activities in $\mathbf{B}$. This also provides us with a tool to determine which of these activities are significant. The b_{nm} are each given a zero mean spherical Gaussian prior distribution with variance α^2 which sets the typical scale of regulatory effects,

$$b_{nm} \sim \mathcal{N}\left(0, \alpha^2\right) . \quad (3)$$

The baseline expression level vector $\boldsymbol{\mu}$ is also given a spherical Gaussian prior with zero mean and unit variance.

The Markov process in equation (2) can be formulated as a Gaussian distribution on the vector $\boldsymbol{\kappa} = (\mathbf{c}(1), \ldots, \mathbf{c}(T))^{\mathrm{T}}$ where $\mathbf{c}(t) = [c_m(t)]$ is the row vector of concentrations at time t,

$$\boldsymbol{\kappa} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

with

$$\mathbf{K} = \begin{pmatrix} A_1 & B & 0 & 0 \\ B & A & \cdots & 0 \\ 0 & \cdots & A & B \\ 0 & 0 & B & A_d \end{pmatrix}^{-1} . \tag{4}$$

Here we have defined

$$A_1 = A_d = \left(\mathbf{I} - \Gamma^2\right)^{-1} , \quad A = \left(\mathbf{I} + \Gamma^2\right)\left(\mathbf{I} - \Gamma^2\right)^{-1} ,$$
$$B = -\Gamma\left(\mathbf{I} - \Gamma^2\right)^{-1} \quad \text{with } \Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_q) .$$

Having defined our model, we can now write a joint distribution for the observed and latent variables

$$\begin{aligned} p(\mathbf{Y}, \mathbf{B}, \mathbf{C}, \boldsymbol{\mu}) &= p\left(\mathbf{Y}|\mathbf{B}, \mathbf{C}, \boldsymbol{\mu}, \sigma^2\right) p\left(\mathbf{B}|\alpha^2\right) p(\mathbf{C}|\boldsymbol{\gamma})\, p(\boldsymbol{\mu}) \\ &= \left(\Pi_{n=1}^{N}\Pi_{t=1}^{T}\mathcal{N}\left(y_n(t)|\sum_{m=1}^{q} X_{nm} b_{nm} c_m(t), \sigma^2\right)\right) \times \\ &\left(\Pi_{n=1}^{N}\Pi_{m=1}^{q}\mathcal{N}\left(b_{nm}|0, \alpha^2\right)\right)\mathcal{N}(\boldsymbol{\kappa}|\mathbf{0}, \mathbf{K})\,\mathcal{N}(\boldsymbol{\mu}|\mathbf{0}, \mathbf{I}) . \end{aligned} \tag{5}$$

The standard approach would now be to marginalise the latent variables and apply type II maximum likelihood to obtain the values of the hyperparameters $\alpha, \sigma, \boldsymbol{\gamma}$. Estimates of the activities and concentrations can then be obtained by posterior estimation using Bayes' theorem. However, exact marginalisation of (5) is intractable and we have to resort to approximate techniques. We use a variational EM algorithm that exploits a factorisation assumption to achieve an efficient approximation of the log likelihood.

2.2 Variational Inference

The most common method for carrying out approximate Bayesian inference for models with many parameters is Markov chain Monte Carlo (MCMC). However, we prefer to use a variational Expectation Maximisation (EM) algorithm here. The advantage of using this approach over MCMC is that we can deal more efficiently with the huge number of parameters in the matrices $\mathbf{B}$ and $\mathbf{C}$. The deterministic nature of the resulting parameter optimisation is also attractive, as it is easier to assess convergence compared to MCMC. Variational EM algorithms have previously been applied to similar models with impressive results (Beal et al., 2005) although it is always important to validate approximate methods in a new application using simulated data.

Variational inference (see *e.g.* Jordan et al., 1999) approximates the intractable posterior probability distribution for the model parameters by using a simplified form, usually involving a factorised approximation. An EM algorithm can then be used to minimise the KL-divergence between this approximation and the true posterior disitrbution. The EM algorithm exploits Jensen's inequality to obtain the following bound on the log likelihood

$$\log\left(p(\mathbf{Y}|\boldsymbol{\theta})\right) \geq \left\langle \log p(\mathbf{Y}, \mathbf{B}, \mathbf{C}, \boldsymbol{\mu}|\boldsymbol{\theta})\right\rangle_{q(\mathbf{B},\mathbf{C},\boldsymbol{\mu})} + H(q) \tag{6}$$

where $\langle\rangle_q$ denotes expectation under the probability distribution q, $q(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu})$ is any probability distribution on the variables $\mathbf{B}$, $\mathbf{C}$ and $\boldsymbol{\mu}$ and H denotes the entropy of the distribution. For brevity, we use $\boldsymbol{\theta}$ to denote collectively the hyperparameters α, σ and $\boldsymbol{\gamma}$. It can be shown that the bound is saturated if and only if $q(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu})$ is the posterior distribution $p(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu}|Y, \boldsymbol{\theta})$.

Computing the posterior distribution is as intractable as computing the marginal likelihood. We will, however, construct a tractable distribution that approximates the posterior distribution. We take the approximating distribution to factorise over the hidden variables

$$q(\mathbf{B}, \mathbf{C}, \boldsymbol{\mu}) = q_1(\mathbf{B})\, q_2(\mathbf{C})\, q_3(\boldsymbol{\mu}) .$$

We can then construct the approximating distribution iteratively as follows: start with any distributions $q_2(\mathbf{C})$ and $q_3(\boldsymbol{\mu})$ (*e.g.* the prior distributions) and average the joint likelihood under them. Then the choice for $q_1(\mathbf{B})$ that maximises the bound in equation (6) can be computed exactly since we averaged over $\mathbf{C}$ and $\boldsymbol{\mu}$. However, the sufficient statistics of $q_1(\mathbf{B})$ will depend on the sufficient statistics of $q_2(\mathbf{C})$ and $q_3(\boldsymbol{\mu})$. We can now iterate computing similarly updated distributions for $q_2(\mathbf{C})$ and $q_3(\boldsymbol{\mu})$ until the algorithm is deemed to have converged. The approximation becomes exact if the random variables $\mathbf{C}$, $\mathbf{B}$ and $\boldsymbol{\mu}$ are independent *a posteriori*.

E-step*: posterior updates* In our model, the log of the joint probability has the following form

$$\begin{aligned} &\log\left[p(\mathbf{Y}, \mathbf{B}, \boldsymbol{\kappa}, \boldsymbol{\mu}|\boldsymbol{\theta})\right] = \\ &-\frac{1}{2}\sum_{n=1}^{N}\left\{T\log\left(\sigma^2\right) + \sum_{t=1}^{\mathrm{T}}\frac{1}{\sigma^2}\left[\left(y_n(t) - \mu_n\right)^2 - \right.\right. \\ &2\left(y_n(t) - \mu_n\right)\mathbf{b}_n^{\mathrm{T}}\chi_n\mathbf{c}_t + \mathrm{trace}\left(\mathbf{b}_n\mathbf{b}_n^{\mathrm{T}}\chi_n\mathbf{c}_t\mathbf{c}_t^{\mathrm{T}}\chi_n\right)\Big] + \\ &\left.+Tq\log\left(\alpha^2\right) + \alpha^{-2}\mathbf{b}_n^{\mathrm{T}}\mathbf{b}_n + \log|\mathbf{K}| + \boldsymbol{\kappa}^{\mathrm{T}}\mathbf{K}^{-1}\boldsymbol{\kappa} + \mu_n^2\right\} \end{aligned} \tag{7}$$

where χ_n is the diagonal matrix having the n-th row of $\mathbf{X}$ on the diagonal, $\mathbf{b}_n^{\mathrm{T}}$ is the nth row of $\mathbf{B}$, $\mathbf{c}(t) = [c_n(t)]$ and $\mathbf{K}$ and $\boldsymbol{\kappa}$ are defined in (4).

By inspection, one obtains that

$$q_1(\mathbf{B}) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{b}_n|\mathbf{m}_n, \Sigma_n)$$

with

$$\begin{aligned} \Sigma_n &= \left(\alpha^2\mathbf{I} + \frac{1}{\sigma^2}\sum_{t=1}^{T}\chi_n\left\langle\mathbf{c}_t\mathbf{c}_t^{\mathrm{T}}\right\rangle_{q_2}\chi_n\right)^{-1} , \\ \mathbf{m}_n &= \Sigma_n\left(\sum_{t=1}^{T}\frac{\left(y_n(t) - \langle\mu_n\rangle_{q_3}\right)}{\sigma^2}\chi_n\langle\mathbf{c}_t\rangle_{q_2}\right) , \end{aligned} \tag{8}$$

where $\langle\rangle_{q_2}$ and $\langle\rangle_{q_3}$ represent averaging under $q_2(\boldsymbol{\kappa})$ and $q_3(\boldsymbol{\mu})$ respectively. Notice that both the posterior inverse covariance and the posterior mean are sparse, mirroring the sparsity of the connectivity data.

Similarly we can compute the approximating distribution for $\mathbf{C}$

$$q_2(\boldsymbol{\kappa}) = \mathcal{N}\left(\boldsymbol{\kappa}|\boldsymbol{\nu}, \mathbf{K}'\right)$$

with

$$\begin{aligned} \mathbf{K}' &= \left(\mathbf{K}^{-1} + \mathbf{I}_T \otimes \frac{1}{\sigma^2}\sum_{n=1}^{N}\chi_n\left\langle\mathbf{b}_n\mathbf{b}_n^{\mathrm{T}}\right\rangle_{q_1}\chi_n\right)^{-1} , \\ \boldsymbol{\nu} &= \mathbf{K}'\left(\frac{\mathbf{y}_n - \langle\mu_n\rangle_{q_3}\mathbf{e}}{\sigma^2}\chi_n\langle\mathbf{b}_n\rangle_{q_1}\right) , \end{aligned} \tag{9}$$

where $\otimes$ denotes the matrix Kronecker product[2], $\mathbf{y}_n = [y_n(t)]$ and $\mathbf{e}$ is a T dimensional vector whose entries are all one.

Notice that the matrix $\mathbf{K}'$ is of size $Tq \times Tq$. For most genome-wide networks Tq is in the thousands, which makes a naïve inversion of the covariance matrix computationally prohibitive (the complexity and memory requirements for inverting a matrix of size Tq are $O\left((Tq)^3\right)$). Also, fast recursive methods to compute posterior expectations such as forward-backward algorithms are prone to numerical instability when the dimension

[2] The Kronecker product (also known as tensor product) of a $q \times q$ matrix A with a $T \times T$ matrix B is the $Tq \times Tq$ block matrix C whose ij-th block is given by $a_{ij}B$.

of the latent space becomes high (for a detailed review of recursive methods from the point of view of control theory, as well as for a discussion of their numerical problems, see Haykin, 2002). To avoid these problems, we used the recent inversion algorithm for banded block matrices proposed by Asif and Moura (2005). This computes the whole covariance matrix while reducing the computational complexity and memory requirements by a factor of order T (for most typical microarray applications T is between 10 and 40) and maintaining numerical stability.

Finally, we can compute the approximating distribution for $\boldsymbol{\mu}$, which is easily derived as

$$q_3(\boldsymbol{\mu}) = \prod_{n=1}^{N} \mathcal{N}\left(\mu_n | \zeta_n, \beta_n^2\right)$$

where

$$\begin{aligned} \zeta_n &= \frac{\sigma^{-2}}{1 + T\sigma^{-2}} \sum_{t=1}^{T} \left(y_n(t) - \mathbf{b}_n^{\mathrm{T}} \chi_n \mathbf{c}_t\right) , \\ \beta_n^2 &= \left(1 + \sigma^{-2}\right)^{-1} . \end{aligned} \tag{10}$$

Notice that, in the limit in which the observation noise σ goes to zero, equation (10) returns the temporal mean of the difference between the observed values and the predicted values, as expected.

The update equations (8), (9) and (10) can be iterated until convergence.

M-step*: hyperparameter updates* Having computed an approximation to the logarithm of the marginal likelihood, we can perform an M-step to optimise the hyperparameters. Fixed point update equations for α^2 and σ^2 are readily found as

$$\begin{aligned} \alpha^2 &= \sum_{n=1}^{N} \mathrm{trace} \left\langle \mathbf{b}_n \mathbf{b}_n^{\mathrm{T}} \right\rangle_{q_1} \\ \sigma^2 &= \sum_{n=1}^{N} \sum_{t=1}^{T} \Big[y_n(t)^2 - 2 y_n(t) \left\langle \mu_n \right\rangle_{q_3} + \left\langle \mu_n^2 \right\rangle_{q_3} - \\ & 2 \left(y_n(t) - \left\langle \mu_n \right\rangle_{q_3}\right) \left\langle \mathbf{b}_n^{\mathrm{T}} \right\rangle_{q_1} \chi_n \left\langle \mathbf{c}_t \right\rangle_{q_2} + \\ & \mathrm{trace} \left(\left\langle \mathbf{b}_n \mathbf{b}_n^{\mathrm{T}} \right\rangle_{q_1} \chi_n \left\langle \mathbf{c}_t \mathbf{c}_t^{\mathrm{T}} \right\rangle_{q_2} \chi_n \right) \Big] . \end{aligned} \tag{11}$$

Unfortunately, it is not possible to obtain fixed point equations for the hyperparameters $\boldsymbol{\gamma}$, since they appear both in the prior mean and in the prior covariance for the concentrations $\mathbf{C}$. We optimised them using a scaled conjugate gradient algorithm in the NETLAB implementation of Nabney (2002).

3 RESULTS

3.1 Artificial data

To check the consistency of our model, we first tested it on artificial data. We randomly generated data from the model with known parameters, and ran the EM algorithm from a random initialisation. To simulate more faithfully a real situation, we used a connectivity matrix obtained from the transcriptional regulatory network of the yeast cell.

We generated eight samples from a system simulating a cellular network with 649 genes and 19 transcription factors. The parameters obtained at convergence were generally in good agreement with the true ones. The inferred posterior expectations, apart from the obvious sign ambiguity (the sign of c_m and b_m can both be changed without altering the model)[3] are in accordance with the true ones. Figure 1 shows a plot of the original activities versus the inferred ones and a reconstructed concentration profile versus the true profile.

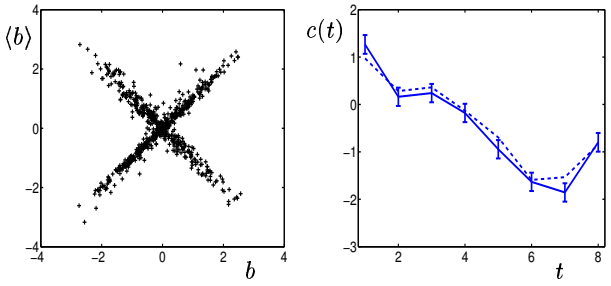

Figure 1. Experimental results on synthetic data: *left* shows the posterior estimates for the gene-specific activities versus the true ones; *right* shows a true concentration profile (dashed line) and the reconstructed posterior concentration profile. The true value of the temporal continuity γ was 0.7255 and the reconstructed value of γ was 0.7240

3.2 Cell cycle data

We then turned to the benchmark yeast cell cycle data set of Spellman et al. (1998). This consists of the expression profiles of 6181 genes measured at 24 equally spaced time points covering the yeast cell cycle. We integrated the microarray data with the connectivity described by Lee et al. (2002), who performed ChIP on 113 transcription factors, measuring their binding to 6270 genes. Although both these data sets are relatively old, they have been extensively studied in the literature on regulatory networks, thus providing an excellent benchmark for model validation and comparison.

The ChIP data is continuous, but, following the suggestion of Lee et al. (2002), we binarised it by considering only regulatory relationships which gave p-values smaller than 10^{-3}. This threshold was confirmed as providing a good compromise between retaining enough regulations without introducing to many false positive in our previous study (see supplementary material to Sanguinetti et al., 2006).

We removed from the data set genes which were not bound by any transcription factor and transcription factors not binding any gene. We also removed the expression data of genes with five or more missing values in the microarray data, leaving a network of 1975 genes and 104 transcription factors.

Figure 2 shows posterior estimates of the protein concentrations and gene specific activities, as well as expression profiles, of two of the most important regulators of the yeast cell cycle, ACE2 and SWI5.

The protein concentration profiles obtained by posterior estimation are similar to the ones obtained using regression methods (*cf.* Boulesteix and Strimmer (2005, Figure 4) and Figure 2 in the supplementary material). This is in accordance with the idea that TFAs obtained by regression are monotonically linked to protein concentrations and provide an estimate of the average effect of a transcription factor over its target genes.

The main novelty of our model is represented by the third column of Figure 2. This plots as a histogram the ratios between the posterior gene specific activities of the two transcription factors and the associated posterior standard deviations for all of their target

[3] The sign ambiguity can easily be resolved for real data by comparing protein concentration profiles with expression data or using known regulatory relationships.

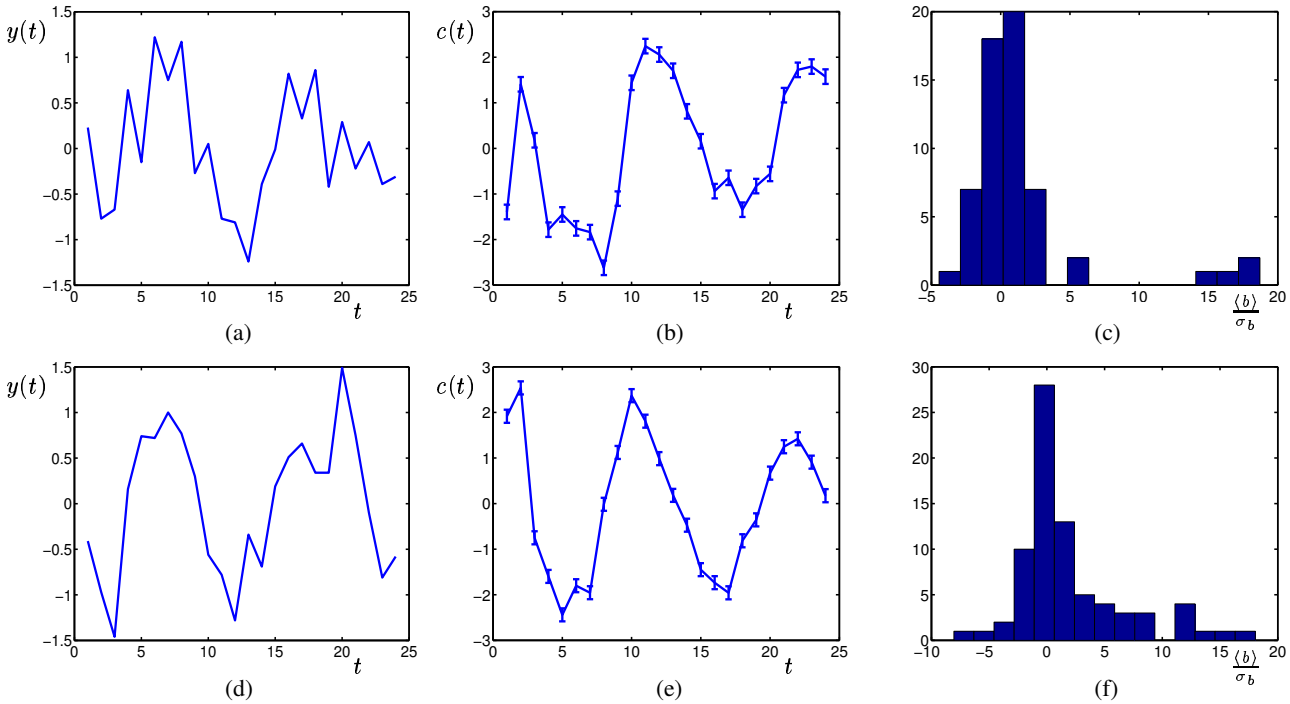

Figure 2. Expression profiles, inferred protein concentrations and gene-specific activities for ACE2 (*top*), SWI5 (*bottom*). (a) and (d) show the transcription profiles of ACE2 and SWI5, (b) and (e) show the posterior temporal profile for protein concentration levels with confidence intervals at each point and (c) and (f) show histograms of the significance levels of the regulatory interactions for each transcription factor acting on all of its target genes (ratios between the inferred mean regulatory strength $\langle b \rangle$ and its associated standard deviation σ_b).

genes. This can be viewed as a generalisation of the standard significance level where we also retain the sign of the interaction. In both cases, there is a large number of genes whose level of regulation is not statistically significant, which could then be considered as false positives in the ChIP data.

A detailed analysis of which genes are significantly regulated can be used for further validation of our model. For example, ACE2 exhibits a group of four genes very significantly promoted (signal to noise ratios between 14 and 19). These are SCW11, CTS1, DSE1 and DSE2. They were clustered with ACE2 in Spellman et al. (1998) and were also identified as the four main targets of ACE2 in our previous study (Sanguinetti et al., 2006). A search of the GO database [4] reveals that the functional annotation of these genes are very coherent: CTS1 is well known to be mediated by ACE2 and is required for cell separation, while both DSE1 and DSE2 are involved in degrading the cell wall causing the daughter cell to separate from the parent. The functional annotation for SCW11 is less clear but it is known that its protein is again localised at the cell wall.

At the other end of the spectrum, we find that our model predicts with reasonable confidence (signal to noise ratio 4.4) that ACE2 represses NCE4. Although this interaction is not documented in the literature as far as we know, NCE4 is known to be important in ensuring DNA stability during replication (Mullen et al., 2005). It is therefore not unreasonable that ACE2, whose main function is to terminate mitosis, should inhibit production of NCE4.

[4] `http://db.yeastgenome.org/`

Among the most significant targets of SWI5 we find SIC1 and PCL2 (signal to noise ratio 12.1 and 4.3 respectively) which were identified as main targets for SWI5 by Aerne et al. (1998), and EGT2 (signal to noise ratio 11.1). EGT2 was shown in Kovaceh et al. (1996) to be primarily regulated by SWI5, but also regulated, to a lesser extent, by ACE2. This is confirmed by our model which assigns a signal to noise ratio of 1.5 to the activity of ACE2 on EGT2.

At a more global level, we can use the posterior distribution over the regulatory intensities to assess which binding events (non-zero entries in the connectivity matrix) result in significant regulation. For example, only 1238 out of 3656 bindings give regulations with a signal to noise ratio greater than 2 (95% significance level). Similarly, we get that only 86 transcription factors regulate at least one target at 95% significance, and that only 155 genes are regulated by more than one transcription factor at 95% significance level (while 792 are bound by more than one transcription factor according to the connectivity data). For a more detailed discussion of these global issues, as well as for more examples of inferred concentrations and regulatory intensities, we refer to the Supplementary Material.

3.3 Metabolic cycle data

Tu et al. (2005) investigated the molecular origin of the glycolitic and respiratory oscillations that constitute the yeast metabolic cycle. mRNA was prepared at regular intervals of approximately 25 minutes over three consecutive cycles. The study identified that, at 95%

significance, over half of the yeast genes (approximately 3500) display periodic behaviour, and can hence be assumed to be metabolic cycle-regulated.

In order to capture as many regulatory relationships as possible, we built our network by merging the results of two ChIP experiments (Lee et al., 2002; Harbison et al., 2004). After eliminating genes that were not bound by any transcription factors and transcription factors not binding any genes, this resulted in a very large network of 177 transcription factors and 3195 target genes. Although merging two ChIP experiments can be expected to result in an increased number of false positives, the probabilistic nature of our model means it can deal efficiently with false positives by assigning a low signal to noise ratio to regulatory relationships which are not consistent with the expression data. For example, while the ChIP data describes 7082 binding events, only 3150 of these where associated by our model with a regulatory intensity which was significant at 95% confidence level. These significant regulations involved 163 of the 177 transcription factors and 2264 of the 3178 genes present in the network.

A transcription factor that is of particular interest in the metabolic cycle of yeast is LEU3, which is involved in the metabolism of branched chain amino-acids. This has been the subject of a recent experimental study (Boer et al., 2005) in which LEU3's regulon was investigated through comparison of *in vitro* binding affinities, ChIP data and data from mutant strains. They identified nine target genes for LEU3 confirmed by all experimental techniques, plus several more putative targets that were confirmed by two experiments.

Figure 3 shows the expression profile, concentration profile and activities for LEU3. Figure 3 (c) shows the gene-specific activities of LEU3 on its target genes. Again, as we already noticed for the cell cycle data, most target genes are not significantly regulated, indicating false positives in the ChIP data. However, the model identifies a number of very highly significantly regulated genes. The three most significantly regulated genes (signal to noise ratio greater than 10) are OAC1, BAT1 and LEU1, which were all confirmed as targets of LEU3 by the *in vivo* experiments of Boer et al. (2005). At the other end of the spectrum, our model predicts significant downregulation for two genes, YLR356W and YHR209W. The functional annotation of these genes is very poor (for YHR209W the protein is unknown); a significant link with a well characterised transcription factor as LEU3 could be the start to a better understanding of these genes.

For more examples of transcription factors involved in the metabolic cycle see the Supplementary Material.

4 DISCUSSION

In this paper we introduced a novel probabilistic model to infer transcription factor protein concentration and regulatory strengths from microarray data when the structure of the regulatory network is known. The expression levels of target genes are modelled as sparse linear combinations of the transcription factor protein concentrations, where the coefficients represent the intensity of the regulatory relationship between a transcription factor and its targets. The regulatory intensities are given a spherical Gaussian prior distribution, while the protein concentrations are modelled as a discrete time state space model. Approximate posterior inference allows estimation of both the intensities and the protein concentration profiles with associated credibility intervals.

A key feature of our model is the way it exploits the natural sparsity of the regulatory network. State space models had previously been used to analyse microarray data (Beal et al., 2005), but the absence of a sparsity constraint meant that they could only be applied to small networks and highly replicated data. The sparse nature of the inference in our model means that we can successfully apply it in genome-wide studies of time courses.

The contribution that is closest to ours is the recent paper of Sabatti and James (2006). While we share many of their aims, there are some important differences between the two approaches: firstly, their approach is *static* and cannot account for the temporal structure of the data. While it is in principle possible to modify their algorithm to include dynamics, the authors themselves acknowledge that this may make the computational cost prohibitive (see Supplementary Material to Sabatti and James (2006)). Secondly, one of the aims of their approach is to be able to identify false positives in the network structure. To do this, they need a prior distribution over the binary connectivity matrix, which they obtain from sequence information using their Vocabulon method (Sabatti et al., 2005). It is not clear, however, how to obtain such a prior distribution for ChIP data. It would seem therefore that Sabatti and James's approach is perhaps most suitable for network structures derived from motif analysis, rather than ChIP data. Lastly, there are important differences at the algorithmic level in the choice of approximate inference techniques (MCMC *versus* variational EM) and in the optimisation of the hyperparameters.

We demonstrated our model both on artificial data and on two yeast data sets, the benchmark cell cycle data set of Spellman et al. (1998) and the more recent metabolic cycle data set of Tu et al. (2005), using network structure obtained by ChIP (Lee et al., 2002; Harbison et al., 2004).

While results on artificial data confirmed the identifiability of our model, results on biological data provided wide ranging predictions on the regulatory network of the yeast cell. Most of these predictions were confirmed by the existing biological literature. However, as in the case of ACE2 repressing NCE4, our model predicted regulatory relationships which are not documented in the biological literature but are consistent with the known function of both the transcription factor and the target gene.

In this paper we made a number of simplifying assumptions. We considered all noise on microarray measurement to be explained by a spherical Gaussian term. This is not always a realistic assumption, and we are aware that the model's results, particularly on low expressed genes, could be negatively affected by large levels of noise. While in principle it is straightforward to propagate noise through a probabilistic model along the lines outlined in Sanguinetti et al. (2005), the computational costs of considering heteroscadistic models can be significant. Also, we assumed the regulatory intensity with which a transcription factor affects a target gene to be constant across time. This is not always the case in reality; however, making the regulatory intensities time-dependent would make the model less identifiable. In these cases it would perhaps be more appropriate to use a model that combines concentrations and intensities such as the one presented in Sanguinetti et al. (2006).

Perhaps the most glaring assumption we make is that an additive linear model is appropriate to describe a complex biological process such as transcription. While this is clearly not the case, a linear model should still capture the most prominent features of the system. Although nonlinear models do obtain better results (Nachman

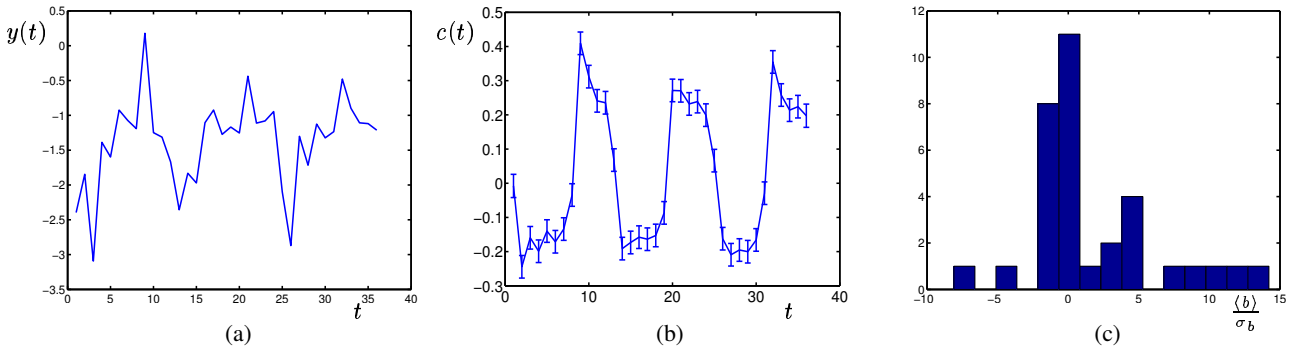

Figure 3. (a) expression profile for LEU3 during the metabolic cycle, (b) posterior protein concentration profile inferred by our model for LEU3 and (c) posterior mean to standard deviation ratios for the gene-specific activities of LEU3 on its target genes.

et al., 2004; Beer and Tavazoie, 2004), their computational complexity rules out inference on a genome-wide scale, thus providing a serious limit to their usefulness in exploratory studies.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from a BBSRC award "Improved processing of microarray data with probabilistic models".

REFERENCES

Aerne, B. L., Johnson, A. L., Toyn, J. H., and Johnston, L. H. (1998). Swi5 controls a novel wave of cyclin synthesis in late mitosis. *Molecular Biology of the Cell*, 9(4):945–956.

Asif, A. and Moura, J. M. F. (2005). Block matrices with L-banded inverse: Inversion algorithms. *IEEE Transactions on Signal Processing*, 53(2):630–643.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2005). A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356.

Beer, M. and Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, 117(2):185–198.

Boer, V. M., Daran, J.-M., Almering, M. J., de Winde, J. H., and Pronk, J. T. (2005). Contribution of the *Saccharomyces Cerevisiae* transcriptional regulator leu3p to physiology and gene expression in nitrogen- and carbon-limited chemostat cultures. *FEMS Yeast Research*, 5:885–897.

Boulesteix, A.-L. and Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med.Model.*, 2(23):1471–16582.

Gao, F., Foat, B. C., and Bussemaker, H. J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5(31):1471–16582.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannet, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104.

Haykin, S. (2002). *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, NJ.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, pages 35–45.

Kovaceh, B., Nasmyth, K., and Schuster, T. (1996). EGT2 gene transcription is induced predominantly by Swi5 in early G1. *Molecular and Cellular Biology*, 16(7):3264–3274.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Liao, J. C., Boscolo, R., Yang, Y.-L., Tran, L. M., Sabatti, C., and Roychowdhury, V. P. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences USA*, 100(26):15522–15527.

Mullen, J. R., Nallaseth, F. S., Lan, Y. Q., Slagle, C. E., and Brill, S. J. (2005). Yeast Rmi1/Nce4 controls genome stability as a subunit of the Sgs1-Top3 complex. *Molecular and Cellular Biology*, 25(11):4476–4487.

Nabney, I. T. (2002). *Netlab: Algorithms for Pattern Recognition*. Springer, London.

Nachman, I., Regev, A., and Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20:i248–i256.

Papp, B. and Oliver, S. (2005). Genome-wide analysis of the context dependence of regulatory networks. *Genome Biology*, 6(2).

Sabatti, C. and James, G. M. (2006). Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746.

Sabatti, C., Rohlin, L., Lange, K., and Liao, J. C. (2005). Vocabulon: a dictionary model approach for reconstruction and localizarion of transcription factor binding sites. *Bioinformatics*, 21(7):932–937.

Sanguinetti, G., Milo, M., Rattray, M., and Lawrence, N. D. (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21(19):3748–3754.

Sanguinetti, G., Rattray, M., and Lawrence, N. D. (2006). A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. To appear in *Bioinformatics*.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297.

Tu, B. P., Kudlicki, A., Rowicka, M., and L.McKnight, S. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5715):1152–1158.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434:338–345.