

A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription

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

^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.

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. This task, however, is difficult for a number of reasons: transcription factors' expression levels are often low and noisy, and many transcription factors are post-transcriptionally regulated. It is therefore useful to infer the activity of the transcription factors from the expression levels of their target genes.

Results We introduce a novel probabilistic model to infer transcription factor activities from microarray data when the structure of the regulatory network is known. The model is based on regression, retaining the computational efficiency to allow genome-wide investigation, but is rendered more flexible by sampling regression coefficients independently for each gene. This allows us to determine the strength with which a transcription factor regulates each of its target genes, therefore providing a quantitative description of the transcriptional regulatory network. The probabilistic nature of the model also means that we can associate credibility intervals to our estimates of the activities. We demonstrate our model on two yeast data sets. In both cases the network structure was obtained using Chromatine Immunoprecipitation data. We show how 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

One of the grand challenges of modern molecular biology is to understand quantitatively the mechanisms regulating mRNA transcription in cells. However, while it is relatively easy to measure the output of transcription using high throughput techniques, it is experimentally very difficult to measure the protein concentration levels of transcription factors and their environment specific chemical affinity to the promoter regions of genes. Moreover, transcription factors are often regulated at the post-transcriptional level, meaning that the mRNA expression levels of transcription factor genes is an unreliable proxy for their actual protein concentration levels and binding affinities.

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. New experimental techniques have allowed biologists to obtain information about the structure of the transcriptional regulatory network for yeast (Lee et al. [2002] using Chromatine Immunoprecipitation (ChIP)) and more recently for

higher organisms (Xie et al. [2005] using motif conservation information). These types of data, which we will collectively call *connectivity* data following Liao et al. [2003], give information about whether a certain transcription factor can bind the promoter region of a gene (in the case of motif data) or whether it binds it in a specific experimental condition (in the case of ChIP).

There has been a wealth of research on integrating connectivity and microarray data in recent years. 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 regression. For example, Liao et al. [2003] introduced “Network Component Analysis”, a dimension reduction technique which takes account of the connectivity information by imposing algebraic constraints on the factors; Alter and Golub [2004] used a different dimension reduction technique (SVD) to achieve the same aim; Gao et al. [2004] used multivariate regression plus backward variable selection to identify active transcription factors; Boulesteix and Strimmer [2005] estimate TFAs using partial least squares.

A major limitation of these methods is that TFAs do not contain any information about the strength (and the sign) of the regulatory interactions between the transcription factor and its different target genes. These can be expected to vary greatly from gene to gene depending on the experimental conditions and on the binding of different transcription factors to the same gene. Also, all these models are not fully 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 connectivity data).

Furthermore, it is known that the structure of the regulatory network of the cell can change dramatically in response to changes in environmental conditions or during different cellular processes (see Harbison et al. [2004], Luscombe et al. [2004]). It is hard to see how regression-based methods can track these changes.

A different approach is taken by Nachman et al. [2004]. They build a probabilistic model, using the framework of dynamic Bayesian networks, which separately models protein concentrations and binding affinities using discrete random variables. This leads to a more realistic model; however, the computational complexity introduced by the discrete variables means that genome-wide analysis can become prohibitively expensive.

We propose a probabilistic model that extends the linear regression model of Liao et al. [2003] to model the full probability distribution of each transcription factor acting on each gene. We

model the temporal changes in the gene-specific TFAs for time-series gene expression data using a Markov chain model. The covariance structure of the transcription factors, however, is shared among all genes, leading to a manageable parameter space and useful information about the correlation of TFAs.

We demonstrate our model on two data sets: the classical yeast cell cycle data set of Spellman et al. [1998], which was studied in all the models above and hence provides a source of useful comparisons, and the yeast metabolic cycle data set of Tu et al. [2005]. In both cases the connectivity data used was ChIP data, (Lee et al. [2002], Harbison et al. [2004]). The results obtained are consistent with known biology, but also allow us to infer previously unknown quantities such as the relative importance of the various transcription factors regulating a single gene. A posterior estimate also allows us to determine whether a target gene is significantly regulated by a certain transcription factor in a specific condition. As well as allowing us to determine whether a transcription factor is active under the given experimental conditions, this also provides some robustness against false positives in the connectivity data.

2 METHODS

2.1 Probabilistic model

The logged gene expression measurements are collected in a design matrix $\mathbf{Y} \in \mathbb{R}^{N \times d}$, where N is the number of genes and d the number of experiments. The connectivity measurements are collected in a binary matrix $\mathbf{X} \in \mathbb{R}^{N \times q}$, where q is the number of transcription factors; element (i, j) of \mathbf{X} is one if transcription factor j can bind gene i , zero otherwise.

We assume that the TFAs can be obtained by regressing the gene expressions using the connectivity information, giving the following linear model

$$\mathbf{y}_n = \mathbf{B}_n \mathbf{x}_n + \boldsymbol{\epsilon}_n. \quad (1)$$

Here $n = 1, \dots, N$ indexes the gene, $\mathbf{y}_n = \mathbf{Y}(n, :)^T$, $\mathbf{x}_n = \mathbf{X}(n, :)^T$ and $\boldsymbol{\epsilon}_n$ is an error term. The matrix \mathbf{B}_n has d rows and q columns, and models the gene-specific TFAs; each column contains the TFA of a certain transcription factor relative to gene n . The crucial difference between our model and other models previously proposed is that the regression coefficients (the TFAs) are allowed to be different from gene to gene. This is represented by the index n in equation (1).

Allowing different TFAs for each gene leads to an explosion in the number of model parameters. We can deal with this large parameter space through marginalization by placing a prior distribution on the rows of \mathbf{B}_n (the gene specific TFAs at a certain experimental point, denoted as \mathbf{b}_{nt}). We will make two physically plausible assumptions in selecting the prior distribution. Firstly, we assume that the gene specific TFA at time t depends solely on the gene specific TFA at time $t-1$ (mathematically, this means that the sequence \mathbf{b}_{nt} has the *Markov property*). This is a simplifying assumption but should be sufficient to capture the main correlations between time points. Secondly, we assume the prior distribution to be *stationary* in time. Mathematically, this amounts to requiring the distributions obtained by marginalising all but one of the time points to be the same.

There are two obvious limiting cases of prior distributions satisfying these conditions. The first is when all the \mathbf{b}_{nt} are assumed to be identical, so that

$$\begin{aligned} \mathbf{b}_{n1} &\sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \\ \mathbf{b}_{n(t+1)} &\sim \mathcal{N}(\mathbf{b}_{nt}, 0). \end{aligned} \quad (2)$$

This would be an appropriate model when the experimental data set consists of replicates of a condition. The second limiting case is when all the \mathbf{b}_{nt} are assumed to be independent and identically distributed,

$$\mathbf{b}_{nt} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma). \quad (3)$$

This is a static model which could be of use when the data set consists of independent samples drawn from conditions without a temporal order.

In general, we expect a realistic model of time-series data to be somewhere in between these two extremes. There are infinite possible choices for such a model; we will make the simplest possible choice of a linear combination of the two models, as this combines computational tractability with simplicity in interpreting the results. We therefore model the gene specific TFAs as

$$\mathbf{b}_{n(t+1)} \sim \mathcal{N}(\gamma \mathbf{b}_{nt} + (1 - \gamma) \boldsymbol{\mu}, (1 - \gamma^2) \Sigma) \quad (4)$$

for $t = 1, \dots, d-1$ and

$$\mathbf{b}_{n1} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

$\gamma \in [0, 1]$ is a parameter measuring the degree of temporal continuity of the TFAs: $\gamma = 1$ returns the replicates model (2), while $\gamma = 0$ returns the static model with all TFAs independent (3). For a given data set, the temporal continuity parameter, γ , is then learnt along with the other model parameters.

For the likelihood, we assume each gene to be independent of all others, given the TFAs, so that

$$p(\mathbf{Y}|\mathbf{B}, \mathbf{X}) = \prod_{n=1}^N p(\mathbf{y}_n|\mathbf{B}_n, \mathbf{x}_n),$$

where we take the noise to be Gaussian, $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, so that

$$p(\mathbf{y}_n|\mathbf{B}_n, \mathbf{x}_n) = \mathcal{N}(\mathbf{y}_n|\mathbf{B}_n \mathbf{x}_n, \sigma^2 \mathbf{I}).$$

The assumption that the noise is distributed according to a spherical Gaussian also allows us to factorize the likelihood along the experiments, giving

$$p(\mathbf{Y}|\mathbf{B}, \mathbf{X}) = \prod_{t=1}^d \prod_{n=1}^N p(y_{nt}|\mathbf{b}_{nt}, \mathbf{x}_n),$$

where

$$p(y_{nt}|\mathbf{b}_{nt}, \mathbf{x}_n) = \mathcal{N}(y_{nt}|\mathbf{b}_{nt}^T \mathbf{x}_n, \sigma^2). \quad (5)$$

We can now integrate out the TFAs to obtain a marginal likelihood for the observations

$$\begin{aligned} p(\mathbf{y}_n|\sigma, \Sigma, \boldsymbol{\mu}, \gamma, \mathbf{x}_n) &= \int \prod_{t=1}^d d\mathbf{b}_{nt} \mathcal{N}(y_{nt}|\mathbf{b}_{nt}^T \mathbf{x}_n, \sigma^2) \times \\ &\quad \left(\prod_{t=2}^d p(\mathbf{b}_{nt}|\mathbf{b}_{n(t-1)}) \right) \mathcal{N}(\mathbf{b}_{n1}|\boldsymbol{\mu}, \Sigma). \end{aligned} \quad (6)$$

Notice that the experimental points are no longer independent once the TFAs are marginalised; the dynamics are now important. However, we can still obtain a useful factorization of the marginal likelihood (6)

$$p(\mathbf{y}_n|\sigma, \Sigma, \boldsymbol{\mu}, \gamma, \mathbf{x}_n) = \mathcal{N}(k_1|\mathbf{x}_n^T \boldsymbol{\mu}, \phi_1) \prod_{t=2}^d \mathcal{N}(y_{t-1}|k_t, \phi_t) \quad (7)$$

where the quantities k_t and ϕ_t can be computed recursively from the data and model parameters. Define $\alpha_d = \sigma^{-2}$ and

$$k_d = \frac{y_d - (1 - \gamma) \boldsymbol{\mu}^T \mathbf{x}_n}{\gamma}.$$

Then we have

$$\begin{aligned} \alpha_{t-1}^2 &= \sigma^{-2} + \gamma^2 \left(\alpha_t^{-2} + (1 - \gamma^2) \mathbf{x}_n^T \Sigma \mathbf{x}_n \right)^{-1} \\ \frac{k_{t-1}}{\alpha_{t-1}^{-2}} &= \frac{y_{t-1}}{\sigma^2} + \gamma^2 \frac{k_t}{\alpha_t^{-2} + (1 - \gamma^2) \mathbf{x}_n^T \Sigma \mathbf{x}_n} \\ \phi_t &= \left(\sigma^2 + \frac{\alpha_t^{-2} + (1 - \gamma^2) \mathbf{x}_n^T \Sigma \mathbf{x}_n}{\gamma^2} \right) \\ \phi_1 &= \left(\alpha_1^{-2} + \mathbf{x}_n^T \Sigma \mathbf{x}_n \right) \end{aligned} \quad (8)$$

for $t = 2, \dots, d$. For details of the derivation we refer the reader to Sanguinetti et al. [2006].

2.2 Likelihood optimisation

The key observation when optimising the marginal likelihood (6) is that its gradient w.r.t. Σ is sparse. This can be seen by observing in equation (8) that Σ enters the likelihood only through the combination $\mathbf{x}_n^T \Sigma \mathbf{x}_n$, leading to the ij -th entry of the gradient to be proportional to the dot product of rows i and j in the connectivity matrix. This means that the elements of the gradient can be different from zero if and only if transcription factors i and j have some common target genes. The connectivity data is insufficient to determine the value of other correlations. We therefore set those elements of Σ to zero. This reflects the intuitively pleasing notion that two TFAs cannot be correlated if the corresponding transcription factors do not co-regulate any gene.

Parameter optimisation in probabilistic models such as (1) is often performed using an EM approach. However, it is hard to see how the inherent sparsity of the matrix Σ can be exploited in an EM algorithm; this will lead to having to explore a much larger (potentially q^2) dimensional parameter space, with consequent speed and convergence problems (overly redundant parameterisations can lead to very flat regions in the likelihood and spurious results). We therefore chose to optimise the likelihood (6) using a scaled conjugate gradient algorithm implemented in *Netlab* (Nabney [2002]).

The sparsity structure allows us to reduce the number of parameters in the prior covariance by representing Σ as $D + \hat{X} \hat{X}^T$, where D is a diagonal matrix and $\hat{X} \hat{X}^T$ has the same sparsity structure of $X X^T$.

Gradients w.r.t. the model parameters γ, σ, μ and Σ can then be obtained exploiting the factorisation (7) and the recursive relations (8). A full derivation is given in Sanguinetti et al. [2006].

2.3 Estimating the TFAs

Once the model parameters have been optimised, gene-specific TFAs can now be estimated from the posterior distribution over the \mathbf{b}_n s. This is obtained by applying Bayes' rule and has the form

$$p\left(\left[\mathbf{b}_{n1}^T, \dots, \mathbf{b}_{nd}^T\right]^T \mid \sigma, \gamma, \mu, \Sigma, \mathbf{X}, \mathbf{Y}\right) = \mathcal{N}\left(\bar{\mathbf{b}}_n, \Sigma_{\mathbf{b}_n}\right) \quad (9)$$

where the posterior covariance is given by

$$\Sigma_{\mathbf{b}_n} = \begin{pmatrix} A_1 & B & 0 & 0 \\ B & A & \dots & 0 \\ 0 & B & \dots & B \\ 0 & 0 & \dots & A_d \end{pmatrix}^{-1} \quad (10)$$

where

$$\begin{aligned} A_1 &= A_d = \sigma^{-2} \mathbf{x}_n \mathbf{x}_n^T + (1 - \gamma^2)^{-1} \Sigma^{-1} \\ A &= \sigma^{-2} \mathbf{x}_n \mathbf{x}_n^T + (1 + \gamma^2) (1 - \gamma^2)^{-1} \Sigma^{-1} \\ B &= -\gamma (1 - \gamma^2)^{-1} \Sigma^{-1}, \end{aligned}$$

and the posterior mean is given by

$$\bar{\mathbf{b}}_n = \Sigma_{\mathbf{b}_n} \begin{bmatrix} \sigma^{-2} y_1 \mathbf{x} + \frac{1}{1+\gamma} \Sigma^{-1} \mu \\ \sigma^{-2} y_2 \mathbf{x} + \frac{1-\gamma}{1+\gamma} \Sigma^{-1} \mu \\ \vdots \\ \sigma^{-2} y_d \mathbf{x} + \frac{1}{1+\gamma} \Sigma^{-1} \mu \end{bmatrix}.$$

Notice that the posterior mean is a dq dimensional vector and the posterior covariance a $dq \times dq$ matrix. These numbers for a genome-wide study are quite large (in the thousands) and inverting the matrix in equation (10) in a careless way can lead to severe computational costs. Fortunately, these can be greatly reduced by exploiting the block tri-diagonal structure of the matrix $\Sigma_{\mathbf{b}_n}^{-1}$ and the particular structure of the blocks (each diagonal block can be inverted efficiently using the Sherman-Morrison formula). Again the explicit calculations are performed in detail in Sanguinetti et al. [2006].

The posterior estimates in (9) contain several interesting pieces of information. For example, by specialising on a particular gene, it is possible to

determine the relative weight of its regulators, or whether they are repressors or promoters. Furthermore, the posterior covariance provides an estimate of the uncertainty in the inferred gene-specific TFAs. This allows us to identify transcription factors which do not play a role in the specific conditions (at a certain level of significance), and provides potentially useful information about the correlations between the various TFAs.

2.4 Propagating uncertainty

If credibility intervals are provided with the microarray measurements, this information can be propagated through a probabilistic model as outlined in Sanguinetti et al. [2005]. We assume that the observations y_{ni} are obtained corrupting the underlying truth \bar{y}_{ni} with Gaussian noise of known variance η_{ni} . The unobserved true expression levels are integrated out when obtaining the marginal likelihood, leading to a modified version of equation (5)

$$p(y_{ni} \mid \Sigma, \mathbf{x}_n) = \mathcal{N}\left(\mathbf{b}_{ni}^T \mathbf{x}_n, \sigma^2 + \eta_{ni}\right).$$

This expression can be used to obtain modified versions of the posterior estimates for the gene specific TFAs, yielding

$$\bar{\mathbf{b}}_n = \Sigma_{\mathbf{b}_n} \begin{bmatrix} (\sigma^2 + \eta_{n1})^{-1} y_1 \mathbf{x}_n + \frac{1}{1+\gamma} \Sigma^{-1} \mu \\ (\sigma^2 + \eta_{n2})^{-1} y_2 \mathbf{x}_n + \frac{1-\gamma}{1+\gamma} \Sigma^{-1} \mu \\ \vdots \\ (\sigma^2 + \eta_{nd})^{-1} y_d \mathbf{x}_n + \frac{1}{1+\gamma} \Sigma^{-1} \mu \end{bmatrix},$$

where B is the same as in equation (10) and the A_i s are defined by adding η_{ni} to σ^2 in the definitions after equation (10).

3 RESULTS

3.1 Data sets

We tested our model on two yeast data sets, the cell cycle data set of Spellman et al. [1998] and the recent metabolic cycle data set of Tu et al. [2005]. Although the cell cycle data set is relatively old, it has been used in most studies on regulatory networks and is useful to compare our model with previous models. The connectivity data we used in both cases was obtained using ChIP: for the metabolic cycle data, we used the recent ChIP data of Harbison et al. [2004], while for the cell cycle data we chose to use the older ChIP data of Lee et al. [2002] for comparison purposes. The ChIP data is continuous, but, following the suggestion of Lee et al. [2002], we binarised it by giving a one value when the associated p -value was smaller than 10^{-3} . This seemed to provide an acceptable rate of false positives while retaining many regulatory interactions. For an in depth discussion of the robustness of our method to the choice of p -value, please see the Supplementary Material.

The ChIP data was obtained from cells grown in rich medium; these are fairly similar to both the growth conditions of Spellman et al. [1998] and those of the metabolic cycle data. We expect however that some regulatory relations predicted by the ChIP data may not be active in the specific data sets, and the model should identify these as false positives by assigning them a low signal to noise ratio.

Both data sets consist of time series; as the expression levels are mostly smoothly varying, the value of the temporal continuity parameter γ obtained by maximum likelihood is quite high in both cases, 0.96 for the cell cycle and 0.98 for the metabolic cycle. It is important to point out that the model can be easily applied to

data sets consisting of independent experiments, simply by setting $\gamma = 0$.

3.2 Cell cycle data

Spellman et al. [1998] used cDNA microarrays to monitor the gene expression levels of 6181 genes during the yeast cell cycle, discovering that over 800 genes are significantly cell cycle-regulated. Cells were synchronised using different experimental techniques. We selected the *cdc15* data set, consisting of 24 experimental points in a time sequence.

The connectivity data we used for this data set was that obtained by Lee et al. [2002]. In this study, ChIP was performed on 113 transcription factors, monitoring their binding to 6270 genes. We removed from the microarray data genes which were not bound by any of the transcription factors studied in Lee et al. [2002] and genes whose expression level was missing at more than five time points, leaving 1975 genes, and removed from the ChIP data the transcription factors which did not bind any gene studied by Spellman et al. [1998], leaving 104 transcription factors.

At a global level, one of the effects of our model is to provide a refinement of the connectivity data. It is well known that binding is only a necessary condition for regulation [see *e.g.* Martone et al., 2003], and that many true positives at the binding level can be expected to be false positives at the regulatory level. We can use our model to associate a confidence level to a regulatory relation by considering whether the changes in time of the gene-specific TFAs are significant compared with the associated posterior standard deviations (for example, a ratio greater than two between these quantities implies significance at 95% confidence level). Accordingly, out of 3656 potential regulatory relations (number of ones in the connectivity matrix), our model identified only 716 regulatory relations with an associated 95% confidence level. These involved 522 genes and 47 transcription factors, 40 of which were found to significantly regulate groups of two or more (up to 52) genes. 119 genes were found to be significantly regulated by two or three transcription factors. A more detailed discussion, including a list of the transcription factors involved and genes regulated, is available in the Supplementary Material.

3.2.1 Gene-specific TFAs As well as providing a tool for identifying false positives in the connectivity data, our model provides quantitative predictions on the strength of the regulatory activities through posterior estimation of the gene-specific TFAs. As an example, we considered the posterior gene-specific TFAs for the well studied transcription factor ACE2. More examples can be found in the Supplementary Material.

ACE2 is predicted from ChIP experiments to bind to 59 different genes. However, we found that *a posteriori* the regulation was significant only in 11 cases at a 95% confidence level and in five cases at 99% confidence level. Three of the four most regulated genes, *CTS1*, *YER124C* and *YHR143W*, belong to one of the clusters identified by Spellman et al. [1998], the *SIC1* cluster. Our results suggest that these genes are actually co-regulated and not only co-expressed. Their qualitative behaviour, peaking at time point 11 in the M/G1 phase of the cell cycle, is consistent with the biological role of ACE2 in the cell cycle (see Spellman et al. [1998]). TFA profiles with error bars¹ for two of these genes are shown in Figure 1.

¹ Error bars are quite similar for different experimental points, but not identical (even if the difference is hard to appreciate in these figures).

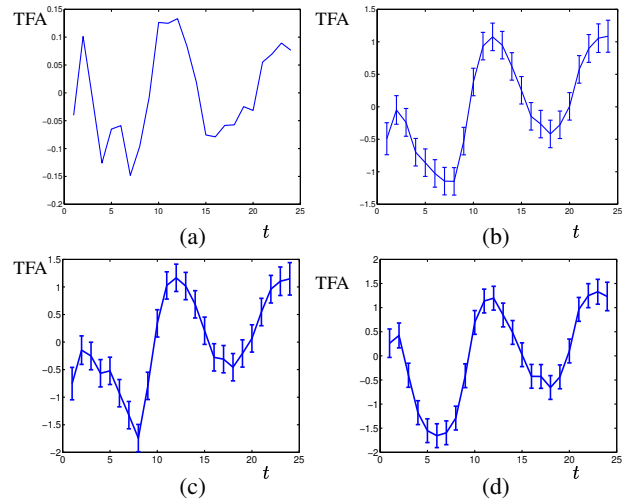


Figure 1. TFAs and gene-specific TFAs of ACE2: (a) TFAs obtained by standard multivariate regression (cf. Boulesteix and Strimmer [2005]); (b) precision weighted mean of gene-specific TFAs for the four most significantly regulated genes; (c) TFA for gene *YER124C* and (d) TFAs for *SCW11*, the two genes with highest signal to noise ratio in the gene-specific TFA.

For comparison, we also show a precision-weighted average of the gene-specific TFAs for the significantly regulated genes² and the TFA obtained by ordinary regression (see, for example, the similar figures in Boulesteix and Strimmer [2005]).

3.2.2 Genes with multiple regulators When a gene is regulated by more than one transcription factor, it is possible to use our model to determine *a posteriori* the relative weight the different transcription factors play in regulating the gene. This can be done, for example, by ranking the transcription factors according to the maximum gene-specific TFA, and the significance level can then be assessed using the posterior variance.

The results of applying this procedure to some example genes are shown in Table 1. The first two genes are two of the top four targets of ACE2 described in the previous section, *YER124C* and *YHR143W*. We can see in these cases that the activity of ACE2 explains the great majority of the expression of these two genes, and that the contribution of the other factors (*FKH1* and *FKH2*) cannot be considered statistically significant. The third row is an example of a gene which is significantly regulated by two transcription factors: while *NDD1* explains most of the expression of *PHO3*, a small but significant fraction is attributed by our model to the action of *FKH2*. The situation appears different in the case of *AGA1*, a member of the *MAT* cluster of genes involved in mating in the yeast cell [Spellman et al., 1998]. *AGA1*'s expression is chiefly explained by the activity of *MBP1*, but a significant role, both statistically and quantitatively, is also played by *SWI4*, while *MCM1*'s contribution appears to be insignificant.

At a more global level, one could consider patterns among the regulators of genes that are significantly regulated by more than one transcription factor. The most represented pair of transcription factors is *NDD1*/*FKH2* which significantly regulate six common target

² The qualitative trend of the precision-weighted average across all targets of ACE2 is very similar to the one obtained considering only the most significantly regulated genes, but the profile is more noisy.

Gene name	Regulators' activity
YER124C	ACE2=1.1±0.2, FKH2=0.03±0.04
YHR143W	ACE2=1.4±0.2, FKH2=0.03±0.04, FKH1=0.011±0.009
PHO3	NDD1=1.6±0.2, FKH2=0.06±0.02
AGA1	MBP1=1.5±0.4, SWI4=1.0±0.4, MCM1=0±0.003

Table 1. Four examples of genes regulated by multiple transcription factors. The first two genes are effectively regulated only by ACE2, while the other two genes genuinely have more than one active regulator

genes. This relationship is confirmed in the literature [Lee et al., 2002]. Other relationships suggested by our model and documented in the literature are MBP1/FKH2, MBP1/SWI4 and MBP1/SKN7. A pair of transcription factors sharing three common targets and not previously documented in the literature is DAL82/MTH1. Their shared targets are mostly low expressed genes, but the model still assigns fairly high confidence to the prediction.

3.2.3 Correlations among transcription factors ACE2 was shown in Lee et al. [2002] to be part of a group of eleven transcription factors which form an almost independent subnetwork in the cell cycle regulatory network. These are, besides ACE2: SWI4, SWI5, SWI6, STB1, MBP1, SKN7, FKH1, FKH2, NDD1 and MCM1. We therefore expect correlations between these transcription factors to be particularly significant.

Correlations are captured by our model in several ways. The matrix Σ models the *a priori* covariance between transcription factors across genes, while posterior estimation can give insight into how correlated two transcription factors regulating the same gene are and how correlated TFAs are at different time points.

To validate our model, we considered the normalised matrix of correlations $\hat{\Sigma}$ (the matrix Σ with each entry divided by the standard deviation of the corresponding transcription factors). We considered the row of $\hat{\Sigma}$ corresponding to ACE2. This has *a priori* 39 elements different from zero, as there are 38 transcription factors that share at least one target gene with ACE2; however, most of these are very close to zero as the relevant transcription factors do not significantly regulate any genes. We sorted them according to the absolute value of their correlation to ACE2, finding that 7 out of the 14 transcription factors most correlated with ACE2 are among the ten identified in Lee et al. [2002]. Of these seven transcription factors, two have large overlaps in their period of activity with ACE2 (SWI4 and MBP1), while two others (FKH2 and SWI5) are known to have important functional relationships with ACE2. The remaining three (SKN7, FKH1 and NDD1) are known to coregulate genes with FKH2 and probably their high correlation with ACE2 is obtained indirectly via the interaction with FKH2, rather than mirroring an actual interaction with ACE2.

Of the remaining three transcription factors identified by Lee et al. [2002], but not identified by our model, MCM1 is active only during the inactive phase of ACE2, SWI6's period of activity is approximately a quarter of a period out of phase with ACE2's, and STB1 is active only for a short interval approximately in the middle of the period of activity of ACE2, justifying a small correlation.

Some transcription factors were identified by our model as highly correlated with ACE2 even if they are not involved in the cell cycle. In some cases, there is an obvious biological link between these transcription factors and ACE2: YAP1, which is very positively correlated with ACE2, is known to be bound by ACE2 (Lee et al.

[2002]), while MSN4 is of the same functional type as ACE2 (zinc finger proteins). Also, it is possible that some transcription factors were active in other, collateral cellular processes which were taking place simultaneously with the cell cycle: for example, YAP1 is known (Tu et al. [2005]) to be active at the peak of the oxidative phase in the metabolic cycle of the yeast cell, shortly before the cell cycle can take place. Correlation between YAP1 and ACE2 could be supporting the hypothesis (Tu et al. [2005]) that the metabolic cycle and the cell cycle can be coupled.

It is interesting to notice that, of the five transcription factors with highest correlation to ACE2, two have high positive correlations, FKH1 and FKH2 (correlation 0.70 and 0.48 respectively), while three have large negative correlations, MBP1, SKN7 and SWI5 (correlation -0.59, -0.40 and -0.38 respectively). This suggests that these transcription factors play complementary or opposite roles in the regulation of their target genes, being active at mutually excluding times or being promoter-repressor pairs. It would be interesting to validate experimentally this prediction.

3.3 Metabolic cycle data

Tu et al. [2005] investigated the molecular origin of the glycolytic 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.

Raw data was processed with the multi-mgMOS algorithm of Liu et al. [2005], which provides uncertainties as well as expression levels for each gene and each experiment. We then used the modified model described in section 2.4 to propagate these uncertainties through the estimation of gene-specific TFAs.

The connectivity data used was recently obtained by Harbison et al. [2004], where the binding of 204 transcription factors to 6229 genes during growth in rich medium was monitored. By removing genes not bound by any transcription factor and transcription factors not binding any gene, we reduced the size of the data set to 2732 genes and 169 transcription factors.

3.3.1 Active transcription factors Tu et al. [2005] provide a list of 133 transcription factors which display periodic behaviour in their expression levels at 95% significance level. However, for all the reasons mentioned in the introductory section, it is doubtful whether periodicity in expression level can be considered strong evidence as to whether a transcription factor is taking part in the regulation of the metabolic cycle.

We therefore examined the gene-specific TFAs of all the 169 transcription factors actively binding targets, determining whether they significantly regulated any genes by examining signal to noise ratios. We obtained a list of 2410 regulatory relations involving 2167 genes and 151 transcription factors. We considered transcription factors regulating five or more genes and obtained 107 transcription factors which significantly regulated five or more genes. Sixty-six of these also belong to the list obtained in Tu et al. [2005]. However, 41 transcription factors which, according to our method, do significantly regulate periodic genes do not have significantly periodic expression levels (see Supplementary Material).

For example, the zinc-finger transcription factor LEU3 is not included in the list of Tu et al. [2005]; its expression profile, although

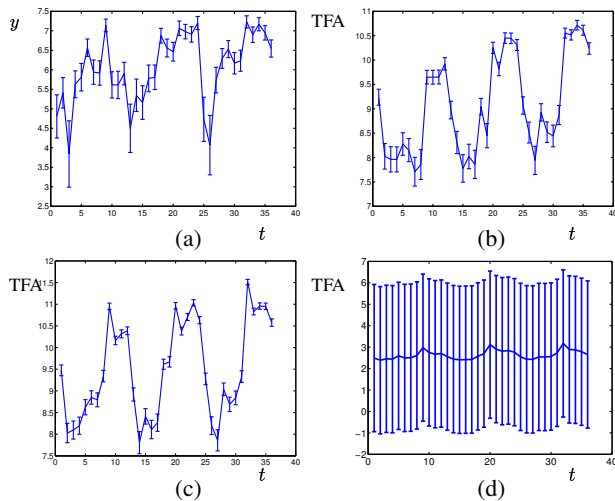


Figure 2. (a) Expression profile for LEU3. Gene-specific TFAs of LEU3 acting on (b) LEU1 and (c) BAT1. These genes were among a list of nine confirmed target genes identified by Boer et al. [2005] using a comparison of various experimental techniques. (d) Gene-specific TFA of LEU3 acting on YGR190C, an example of a binding that doesn't result in significant regulation.

approximately periodic, is rather noisy (see Figure 2 (a)). However, according to our method, it significantly regulates eight genes in the data set. Five of these, OAC1, LEU1, BAT1, RPS11B and POL1, are highly periodic according to Tu et al. [2005]. Three of these have recently been confirmed experimentally as targets of LEU3, and another significantly regulated (but non-periodic) gene, YOR271C, has been shown to be bound by LEU3 *in vitro* [Boer et al., 2005].

These results indicate that LEU3 should be considered as a significant player in regulating the yeast metabolic cycle. Notice that inference techniques which do not discriminate TFAs between genes would miss this fact: LEU3's average activity is indeed not periodic, as many of its targets are not periodic; only its gene-specific activity for some genes is periodic. Gene-specific TFAs of LEU3 in two periodic cases are shown in Figure 2 (b)-(c). Figure 2 (d) shows the gene-specific TFA of LEU3 acting on YGR190C. Notice the large error bars associated with the gene-specific TFA. YGR190C is, according to our model, a significantly regulated target of SW14.

A complete list of the active transcription factors and a comparison with the results of Tu et al. [2005] is given in the Supplementary Material.

3.3.2 Comparison between metabolic and cell cycle data set

One of the main features of our model is its environment specificity. Through its probabilistic nature, it can infer a different network structure in different experimental conditions. As an example, we again considered the transcription factor ACE2 and how it correlates to other transcription factors.

ACE2 is an active transcription factor in the metabolic cycle, both according to our model and according to Tu et al. [2005]. It significantly regulates 14 genes, five of which are highly periodic according to Tu et al. [2005]. However, its list of significantly regulated targets is very different from the list drawn from the cell cycle data: its four main targets are KRE32, YJR149W, YHL013C and LCP5. None of these was significantly regulated in the cell cycle

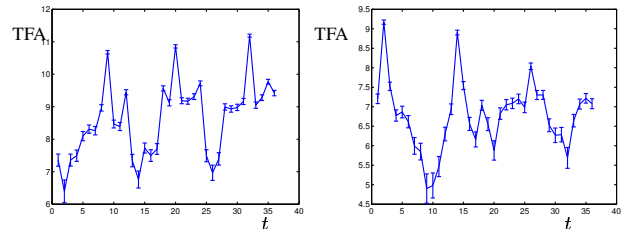


Figure 3. Gene-specific TFAs for ACE2 acting on KRE32 (left) and YJR149W (right).

data set³. Gene-specific TFAs for ACE2 on its most significantly regulated targets are shown in Figure 3. Notice that the two gene-specific TFAs are out of phase (correlation -0.5). It is hard to see how a regression method that assigns the same TFA to all target genes could have described this situation.

We also examined the correlations between transcription factors predicted by our model in the metabolic cycle conditions. Again, the picture looks quite different from the cell cycle case: some correlations are similar, for example FKH2 is still positively correlated (correlation 0.62), while MBP1 is negatively correlated (correlation -0.53). However, some correlations which were negligible in the cell cycle data set appear to be more important in the metabolic cycle. For example RME1, which was almost completely uncorrelated with ACE2 in the cell cycle data set, is now strongly positively correlated (correlation 0.50), while PHO4 and IME4 are now strongly negatively correlated (-0.46 and -0.51 respectively).

It would therefore seem that not only the structure of the regulatory network can change between different conditions, with some transcription factors changing their targets, but the way transcription factors work together can change dramatically in different conditions. This is consistent with biological knowledge, but, to our knowledge, our model is the first computational tool to provide a genome-wide quantitative estimate of this change.

4 DISCUSSION

In this paper we introduced a novel probabilistic model for integrating connectivity and microarray data. While most existing models infer transcription factor activities that are shared across genes, we are able to infer gene specific activities. The probabilistic nature of the model means that we can associate credibility intervals with our results, and hence determine which regulations are significant in a given experimental condition.

We validated our model on two yeast data sets, finding that the results of our model are in broad accordance with known biological facts about the yeast transcriptional regulatory network. To demonstrate the flexibility of our model, we have shown how it can be used to answer a number of important biological questions, such as: whether a transcription factor significantly regulates its target genes, what is the respective strength of transcription factors co-regulating a gene? And what are the correlations among transcription factors? The ability to provide quantitative answers to these questions, while retaining a computational efficiency allowing genome-wide investigations, is a unique feature of our model. It is a feature that

³ It must be borne in mind, though, that the ChIP data used was different. However, the same analysis conducted using the connectivity data of Lee et al. [2002] highlights similar differences between the cell cycle and the metabolic cycle.

we expect to be of great assistance to biologists interested in the structure of regulatory networks.

The model also provides new predictions that could shed light on some aspects of the regulatory mechanism of the cell: for example, anticorrelation between transcription factors suggests mutually exclusive protein interactions, and negative gene-specific TFAs might be taken as a sign that the transcription factor is repressing rather than promoting its target. These predictions could hopefully be validated by new biological data in the future.

A common problem when using connectivity data is its notorious unreliability. The probabilistic nature of our model means that false positives have a reduced influence on the results, as the model will associate them with high variances. However, false negatives or incomplete data may still be problematic. For example, it is possible that new transcription factors for yeast may be identified in the future⁴. Also, regulatory relationships are strongly environment dependent and the use of ChIP data obtained in conditions even slightly different from the ones of the microarray data can result in many false negatives in the connectivity data. This may lead to spurious results, as the model is forced to explain the effects of the unknown regulatory relations with the limited data available. There is no quick fix to this problem: while we are investigating treating the connectivity data as random variables, thus dealing with the associated uncertainty, it is probable that the computational complexity will prove prohibitive in a genome-wide application. For the time being, the best solution is to handle the results with some care, for example requiring that a sufficient number of genes are significantly regulated by a transcription factor before concluding that that transcription factor is active in the given condition. Ultimately, the final arbiter should always be experimental biological validation.

ACKNOWLEDGEMENTS

We thank Balasz Papp, Stephen Oliver, Alastair Goldman and Marta Milo for useful discussions. We thank Andrzej Kudlicki for kindly making available the CEL files for the yeast metabolic cycle data set, and Xuejun Liu for help using the mmgMOS package. The authors gratefully acknowledge support from a BBSRC award “Improved processing of microarray data with probabilistic models”.

REFERENCES

- O. Alter and G. H. Golub. Integrative analysis of genome-scale data using pseudoinverse projection predicts novel correlation between dna replication and rna transcription. *Proceedings of the National Academy of Sciences USA*, 101(47):16577–16582, 2004.
- V. M. Boer, J.-M. Daran, M. J. Almering, J. H. de Winde, and J. T. Pronk. 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, 2005.
- A.-L. Boulesteix and K. Strimmer. 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, 2005.
- F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5(31): 1471–16582, 2004.
- C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.
- T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.
- J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences USA*, 100(26):15522–15527, 2003.
- X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.
- N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.
- R. Martone, G. Euskirchen, P. Bertone, S. Hartman, T. E. Royce, N. M. Luscombe, J. L. Rinn, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. Distribution of nf- κ b-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences USA*, 100(21):12247–12252, 2003.
- I. T. Nabney. *Netlab: Algorithms for Pattern Recognition*. Springer, London, 2002.
- I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20:i248–i256, 2004.
- G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence. Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21(19):3748–3754, 2005.
- G. Sanguinetti, M. Rattray, and N. D. Lawrence. A probabilistic model to integrate chip and microarray data. Technical Report CS-06-02, University of Sheffield, 2006.
- P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.
- B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight. Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5715):1152–1158, 2005.
- X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. *Nature*, 434:338–345, 2005.

⁴ In the two years between Lee et al. [2002] and Harbison et al. [2004] 91 new transcription factors have been identified.