Statistical reconstruction of transcription factor activity using Michaelis-Menten kinetics

R. Khanin*,[1] V. Vinciotti*,[2] V. Mersinias,[3] C.P. Smith[3] and E. Wit*[4]

[1]Department of Statistics, University of Glasgow,Glasgow, G12 8QW, UK

[2]School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, UB8 3PH, UK

[3]School of Biomedical and Molecular Sciences, University of Surrey, Guildford, GU2 7XH, UK

[4]Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4FY, UK

Summary.

The basic building block of a gene regulatory network consists of a gene encoding a transcription factor and the gene(s) it regulates. Considerable efforts have been directed recently at devising experiments and algorithms to determine transcription factors and their corresponding target genes using gene expression and other types of data. The underlying problem is that the expression of a gene coding for the transcription factor provides only limited information about the activity of the transcription factor, which can also be controlled post-transcriptionally. In the absence of a reliable technology to routinely measure the activity of regulators, it is of great importance to understand whether this activity can be inferred from gene expression data. We here develop a statistical framework to reconstruct the activity of a transcription factor from gene expression data of the target genes in its regulatory module. The novelty of our approach is that we embed the deterministic Michaelis-Menten model of gene regulation in this statistical framework. The kinetic parameters of the gene regulation model are inferred together with the profile of the transcription factor regulator. We also obtain a goodness-of-fit test to verify the fit of the model. The model is applied to a time series involving the *Streptomyces coelicolor* bacterium. We focus on the transcriptional activator *cdaR*, which is partly responsible for the production of a particular type of antibiotic. The aim is to reconstruct

*These authors equally contributed to the work

the activity profile of this regulator. Our approach can be extended to include more complex regulatory relationships, such as multiple regulatory factors, competition and cooperativity.

KEY WORDS: Gene regulation, Michaelis-Menten kinetics, maximum likelihood estimation, *Streptomyces coelicolor*

1. Introduction

Linking transcription factors (TFs) to their targets is a central problem in post-genomic biology. While genes regulated by the same TFs tend to be co-expressed, the relationship between the gene expression profiles of the TFs and their regulated genes can be quite complicated, often exhibiting time-shifted or inverted behaviour (Yu et al., 2003). This could be due to the fact that changes in the expression of a TF are subtle and its activity is often controlled at levels other than expression, e.g. via post-transcriptional modifications. Therefore, the expression of a gene coding for TF generally provides only limited information on the true transcription factor activity (TFA). The situation becomes even more complex in the presence of cooperativity or competition between two or more TFs that regulate a target gene.

New computational methods have been proposed to infer TFAs from the gene expression data under the assumption that the two are not necessarily the same. Zhou et al. (2005) propose to validate TFA through cross-platform integration of expression data. Kao et al. (2004) and Boulesteix and Strimmer (2005) estimate the TFAs by setting the problem in a (partial) least squares framework and by using algebraic matrix decomposition to deal with the high-dimensionality issue. Both assume a linear additive model of gene regulation. Gao, Foat and Bussemaker (2004) suggested a multivariate regression analysis, using the ChIP occupancy log-ratios for the TFs as a response and the genes as predictors. The coefficients of the regression express the changes in TFA. Regulated genes are those that are correlated with the TFA profile. In all of the above models, the data on the connectivity comes from outside sources, like ChIP-chip data or *a priori* knowledge.

In this paper we develop a statistical framework to model regulatory pairs of TFs and their target-genes using Michaelis-Menten kinetics for gene regulation. The Michaelis-Menten

(MM) model has been successfully used in various biological applications, including the regulation of a gene by a TF (Bolouri and Davidson, 2002; Mangan and Alon, 2003). Nachman et al. (2004) were the first to incorporate the quantitative MM regulation model into the generative Bayesian probabilistic model. These authors attempted to estimate simultaneously the structure of regulatory modules as well as the kinetic parameters of the MM regulation model and the levels of ideal regulators that control them. They considered multiple TFs and multiple targets in their model, as well as a dynamic temporal behaviour. They applied their Bayesian learning algorithms to yeast cell cycle expression datasets. In contrast, we develop a frequentist approach to find the parameters of the MM model of regulation by embedding this model in the statistical framework.

In Section 2 we introduce a model for observed gene expressions within a general network motif. Then in Section 3 we focus on the special case of a single input motif, for which we can obtain an explicit expression of the MM ordinary differential equation. In Section 4, we show how conjugate gradient methods can be used to estimate the kinetic MM parameters and the TFA of the regulator can be estimated via maximum likelihood. Finally, in Section 5 our statistical framework is applied to a 10-point time-course datasets for a wild type and mutant type *Streptomyces coelicolor*. We obtain some interesting biologically results and show that the model we propose has good fit to the data.

2. Model for TF-initiated gene transcription

In this section, we present a general gene expression model that takes into account (i) transcription rate, (ii) decay rate, (iii) network structure and (iv) stochastic effects.

2.1 *Kinetic model of gene transcription*

The gene expression of a regulated gene, $\mu(t)$, defined as the number of transcribed RNA molecules present at time t, changes due to gene transcription and the decay of RNA molecules. The average rate of change in expression of a target gene, $\dot{\mu}(t)$, is therefore described by the number of RNA molecules transcribed per unit of time and the number of decaying molecules

per unit time:

$$\dot{\mu}(t) = p(t) - \delta\mu(t), \tag{1}$$

where $p(t)$ is a production term, i.e. the rate of gene transcription, and δ is a linear degradation rate. Here $\mu(t)$ stands for the underlying expression of the regulated gene at time t. The general solution of the above linear differential equation is given by

$$\mu(t) = \mu_0 \mathrm{e}^{-\delta t} + \int_0^t \mathrm{e}^{-\delta(t-\tau)} p(\tau) d\tau. \tag{2}$$

Gene regulation is usually controlled by one or more TFs. The rate of gene transcription, $p(t)$, depends on the type of regulation, i.e. activation or repression, and on the type of regulation control, namely a single TF or multiple TFs. The transcription rate depends also on the so-called gate type in the case of multiple TF regulators. For example, the "AND" gate means that all TFs are required for regulation, while the "OR" gate implies that either of the TFs is sufficient to regulate the transcription of the target gene(s). In addition, the production term, $p(t)$, depends on gene-specific kinetics of regulation, θ. For example, the target genes can have different values for the maximal production rate. Also, the transcription of different targets could saturate at different levels of the TF regulator.

Gene regulation has commonly been described using a linear model: either the transcription rate of a target gene or its expression is assumed proportional to the level of the TFs that regulate this gene (Kao et al., 2004; Boulesteix and Strimmer, 2005). In this paper we model gene transcription with the so-called Michaelis-Menten (MM) kinetics. The MM kinetics have been used in modelling enzyme-mediated reactions and have also been applied to TF-initiated transcription (Bolouri and Davidson, 2002; Nachman et al., 2004). The MM kinetic model, unlike a linear model, is able to describe saturation effects, which are biologically plausible.

It is worth noting that the proposed statistical framework is by no means limited to a specific microarray platform. The model can equally be applied to both cDNA and oligonucleotide microarrays, as well as gene expression profiles obtained by other technologies, such as quantitative real-time PCR.

2.2 *Network motif*

The term network motif, coined by Milo et al. (2002), is defined as patterns of interconnections that recur in different parts of a network at frequencies much higher than those found in random networks. Several basic network motifs have been found in biological networks. Each network motif consists of several target genes regulated by one (single input motif) or several (multiple input motif, feed forward loop) TFs.

Within a network motif, the gene expression of a gene k at time t, $\mu^k(t)$, depends on the decay rate-constant δ^k and on the transcription rate, $p_k(t)$, as defined by equation (2). The transcription rate $p_k(t)$ depends on several gene-specific kinetic parameters, θ^k, as well as on the activity of its TF regulator(s), whose activity levels are denoted by $\eta_1(t), \ldots \eta_M(t)$ $(M \geq 1)$. The TFs are the common regulators to all the target genes, $\mu_k, k = 1 \ldots K$ in the network motif, while the kinetic parameters of gene regulation are likely to be target-dependent:

$$\mu^k(t) \equiv \mu^k(t; \theta^k, \eta_1, \ldots, \eta_M), \ k = 1, \ldots K.$$

It is biologically compelling to assume that the gene-specific parameters of the gene kinetic equation, θ^k, are the same between the different biological conditions, such as wild type and mutant. The only exception is the initial amount of gene expression, μ_0^k, which can be different due all sorts of external factors that affect gene transcription. In Section 3 we consider a Michaelis-Menten model implementation of the case of a *Single Input motif, (SIM)*, i.e. one regulator and many targets.

2.3 *Noise model*

As the MM kinetic model requires that we model the intensities on the original rather than log-transformed scale, it is important to find a suitable distribution for the noise process. In particular, it is unlikely if not impossible to have merely additive noise. As log-transformed intensity ratios have been found to be approximately normal (Lee et al., 2000), we use the log-normal distribution for the ratios of the intensities. Moreover, as every microarray measures the gene expression of a different biological sample due to destructive sampling, it is reasonable to assume that all observations are independent.

Let us denote by $g^k_{cr}(t)$ the observed gene expression of gene k at a time-point t for the replicate r under condition c. The condition c stands, for example, for wild type or mutant. We assume that the observed gene expressions of a target gene k are independent and log-normally distributed with location parameter $m^k_c(t)$ and scale parameter σ^2_k. This distribution takes into account the different variances associated with different amplitudes. The log-likelihood contribution of a single observation $g^k_{cr}(t)$ is given by

$$l(m^k_c(t), \sigma^2_k | g^k_{cr}(t)) = -\frac{1}{2}\left(\frac{\log[g^k_{cr}(t)] - m^k_c(t)}{\sigma_k}\right)^2 - \log(\sqrt{2\pi} g^k_{cr}(t)) - \log(\sigma_k). \tag{3}$$

Given the expectation of a log-normal distribution $E[g^k_{cr}(t)] = e^{m^k_c(t)+\frac{1}{2}\sigma^2_k}$, the relationship between the true gene expression under condition c, $\mu^k_c(t)$, and the location parameter of the log-normal distribution, $m^k_c(t)$, is given by

$$m^k_c(t) \equiv \log[\mu^k_c(t)] - \frac{1}{2}\sigma^2_k. \tag{4}$$

Therefore, the location parameter $m^k_c(t) \equiv m^k_c(t; \theta^k_c, \sigma^2_k, \eta_1 \dots \eta_M)$ implicitly depends on the kinetic parameters θ^k_c of the gene regulation model and on the TFs levels $\eta_1 \dots \eta_M$. The likelihood contribution in equation (3) can then be written as a function of the kinetic parameters of the gene regulation model as well as activities of TFs, namely $l(\theta^k_c, \sigma^2_k, \eta_1 \dots \eta_M | g^k_{cr}(t))$.

3. Michaelis-Menten model of a single input motif

3.1 *Single Input Motif*

We now apply our general methodology to a simple network architecture, called the Single Input Motif (SIM). It consists of a set of genes that are controlled by a single TF (Shen-Orr et al., 2002). All of the genes are under the same type of regulation (either all activated or all repressed), which presumably happen under a specific set of circumstances. None of these genes have additional transcriptional regulation. SIMs are potentially useful for coordinating a discrete unit of some biological function, such as a set of genes that code for the subunits of a biosynthesis apparatus or enzymes of a metabolic pathway (Lee et al., 2002). SIM is probably the simplest logical unit of a transcriptional regulatory network architecture that could serve as a starting point for the reconstruction of TFA.

There is compelling experimental evidence that SIMs frequently occur in biological systems (Lee et al., 2002; Shen-Orr et al., 2002). It is partly an open question as how to identify new SIMs, verify the targets and infer the activity of the regulators. The first source of information for SIMs is in databases such as RegulonDB, that were used by Shen-Orr et al. (2002) in their original study of network motifs. There is an increasing amount of ChIP-chip data, pioneered by Lee et al. (2002), which identify TF-and-target pairs. The use of such data together with statistical models such as (Bar-Joseph et al., 2003; Yu, 2004) helps to identify and verify SIMs.

Another rich source of data for identifying SIMs is contained in microarrays studies. For example, an experiment comparing a wild-type and a mutant, wherein the TF of interest is knocked out, yields a list of differentially expressed genes, which are potential targets of this TF. To identify whether these targets are primary or secondary, further experiments, such as data on binding sites, or *a priori* knowledge is required. In this paper, we identify a SIM for *Streptomyces coelicolor* by finding differentially expressed genes between a wild type and a mutant type (where the TF has been knocked-out) combined with biological knowledge on specific location of the TF and targets within the genome.

3.2 *Michaelis-Menten model*

When a gene is regulated by a single TF that binds to the promoter region of the regulated gene, the transcription rate $p(t)$ depends on the level of this TF, η and gene specific kinetic parameters. The Michaelis-Menten model of gene transcription activated by some TF states that production occurs in a saturating manner:

$$p(t) = \beta \frac{\eta(t)}{\gamma + \eta(t)} + \alpha. \tag{5}$$

Here β is the rate of production, γ is the half-saturation constant and α is the basal level of gene expression production. The general solution of the transcription equation (2) takes the form

$$\mu(t) = (\mu_0 - \frac{\alpha}{\delta})\mathrm{e}^{-\delta t} + \frac{\alpha}{\delta} + \beta \int_0^t \mathrm{e}^{-\delta(t-\tau)} \frac{\eta(\tau)}{\gamma + \eta(\tau)} \mathrm{d}\tau. \tag{6}$$

In a SIM, the same TF regulates more than one gene. The gene expression profile of gene k, $\mu^k(t)$, depends on several kinetic parameters that are gene specific, $\alpha^k, \beta^k, \gamma^k, \delta^k, \mu_0^k$, as well as the activity of the regulator, η, that is common for all targets in the SIM regulated by it. We use the following notation $\mu^k(t) \equiv \mu^k(t; \theta^k, \eta)$ and $\theta^k \equiv (\alpha^k, \beta^k, \gamma^k, \delta^k, \mu_0^k)$.

4. Parameter estimation

4.1 *Likelihood*

The kinetic parameters of the MM model, θ^k, and the variance of the log-normal distribution, σ_k^2, for a single gene, k, can be estimated by an approximate maximum likelihood procedure. The likelihood for a gene k, regulated by one TF, η_c, given all observations, $g_{cr}^k(t)$, across all time-points, t, conditions, c, and replicates, r, is given by

$$l_k(g^k(t); \theta_k, \sigma_k^2, \eta) = \sum_{ctr} l(\theta_c^k, \sigma_k^2, \eta_c | g_{cr}^k(t)). \tag{7}$$

The likelihood of the whole SIM, wherein the TF with activity level $\eta(t)$, regulates several target-genes, can be written as

$$l_{\text{SIM}}(\Theta, \Sigma^2, \eta | G) = \sum_{k=1}^{K} l_k(\theta^k, \sigma_k^2, \eta | g^k(t)). \tag{8}$$

Here $G = \{g_1, \ldots, g_K\}$ is the set of K target genes; Θ represents all the kinetic parameters of the MM model, θ_k, for all genes in the SIM and Σ^2 stands for all the scale-parameters of the log-normal distribution, σ_k^2, that are also assumed to be gene-specific.

4.2 *Transcription Factor Activity*

A common approach (Bar-Joseph et al., 2003; Qian et al., 2003; Segal et al., 2003) assumes that the transcription of the gene coding for the TF represents its activity reasonably well. Therefore, the observed gene expression values for the TF (TFX) are used as a proxy for TFA. A biologically more plausible model suggests that the TFA is not equal or not necessarily even correlated with the TFX (Gao et al., 2004; Nachman et al., 2004) due to the processes of translation and post-translational modifications. In this case, the TFA, $\eta(t)$, can be thought of as an unknown parameter. The idea is that $\eta(t)$ can be reconstructed from the expression data of the genes that are known to be regulated by it. In a SIM, where a given TF regulates

several gene-targets, the TFA profile, $\eta_c(t)$, is the same for all target-genes in the regulatory module as all target genes become activated (or repressed) by the master TF regulator under a specific set of conditions, $c \in C$. The kinetic parameters of regulation are gene-specific for each of the K genes with profiles $\mu^1(t), \ldots, \mu^K(t)$. These kinetic parameters as well as the TFA profile can be found by maximizing the likelihood (8) for a given set of genes.

4.3 *MM model constraint*

The true expression of a target gene k at time t depends on a continuous integral of the TFA values (6). Without any further constraints, it is clear that the function $\eta(t)$ is unidentifiable. We therefore assume that the TFA can be approximated by a piecewise constant step-function $\bar{\eta}$ on the intervals (t_j, t_{j+1}), where t_j are the sampling points ($j = 0, \ldots, N-2$). Given this constraint, the integral in (6) can be approximated by a sum,

$$\int_0^t \mathrm{e}^{-\delta(t-\tau)} \frac{\eta(\tau)}{\gamma + \eta(\tau)} \approx e^{-\delta t} \frac{1}{\delta} \sum_{j=0}^{N-2} (\mathrm{e}^{\delta t_{j+1}} - \mathrm{e}^{\delta t_j}) \frac{\bar{\eta}_j}{\gamma + \bar{\eta}_j}.$$

yielding the full general solution of the gene transcription equation (6)

$$\mu(t) = (\mu_0 - \frac{\alpha}{\delta})\mathrm{e}^{-\delta t} + \frac{\alpha}{\delta} + \beta e^{-\delta t} \frac{1}{\delta} \sum_{j=0}^{N-2} (\mathrm{e}^{\delta t_{j+1}} - \mathrm{e}^{\delta t_j}) \frac{\bar{\eta}_j}{\gamma + \bar{\eta}_j}. \tag{9}$$

This approximation is used for each of the target $k = 1, \ldots, K$ in the SIM. The parameter $\bar{\eta} = (\bar{\eta}_0, \ldots, \bar{\eta}_{N-2})$ is $N-1$ dimensional, but due to its collinearity of β on the one hand and γ on the other in equation (9), it can only be identified up to a multiplicative constant. Therefore, without loss of generality we can fix $\bar{\eta}_0 = 1$. Computational details on maximizing likelihood by conjugate gradient are given in supplementary materials.

5. Application

The model described above has been applied to two 10-point time-series of two *Streptomyces coelicolor* strains grown on solid medium, one wild type and one mutant type for which a transcriptional regulator *cdaR* (SCO3217) has been knocked-out. Each time-point of the two time-courses is replicated twice using independent biological samples, as the sampling mechanism is destructive.

The importance of the genus *Streptomyces* results from the bacterium's production of over two-thirds of naturally derived antibiotics in current use, as well as many anti-tumour agents and immunosuppressants. *Streptomyces coelicolor* produces at least four chemically distinct antibiotics (Bibb, 1996). The genes responsible for the synthesis of each of the four antibiotics have been found to be clustered in distinct locations (Bentley et al., 2002). Here we study genes in the cluster responsible for the production of calcium-dependent antibiotics (CDA) (Hojati et al., 2002). This cluster of 40 genes (SCO3210-SCO3249) contains at least two genes encoding the transcriptional regulators, CdaR and AbsA2, whose specific roles in the regulation of antibiotic biosynthesis have not been characterized in detail. Only 34 genes from the 40-gene *cdaR* cluster are present on the arrays, so only these genes have been considered in the current study. The *cdaR* gene product is known to positively regulate genes for CDA biosynthesis (A.E., Hayes, P.P. Chong, Z. Hojati, V. Mersinias, F. Flett, C.P. Smith, unpublished results), while AbsA2 acts as an inhibitor, repressing CDA promoters, perhaps in competition with CdaR (Ryding et al., 2002; Sheeler et al., 2005). At the same time, the *cdaR* gene appears to be expressed independently of *absA* (Ryding et al., 2002). The current experimental and modelling study focusses on analyzing the role of the *cdaR* gene product in regulating the expression of the *cdaR* gene cluster. The details on data preprocessing can be found in supplementary materials.

5.1 *Identification of cdaR regulatory module*

As there is not much a priori biological knowledge available, we use the data to inform us about which of the 34 available gene targets might be directly regulated by *cdaR*. We implement this by means of an ANOVA and checking the significance of the *knock-out effect* κ_c, $g_{ctr} = \mu + \kappa_c + \tau_t + \epsilon_{ctr}$ for each gene in the *cdaR* cluster separately (c = mutant, wild-type; $t = 1, \ldots 10$ time; $r = 1, 2$ replicates) accounting for a possible time effect. Apart from *cdaR* gene itself, another 17 genes within the *cdaR* cluster have been identified (with p-values < 0.01) as being differentially expressed between the two strains. Although performing 34 tests simultaneously, a p-value of 0.01 guarantees that it is unlikely that more than one of the

17 genes is falsely discovered. These 17 genes are therefore assumed to be activated directly by the transcriptional activator CdaR. Ten of these genes, SCO3235-39 (with SCO3238 absent from the array) and SCO3244-49, form two stretches of co-regulated genes probably belonging to the same operons, the latter extending from the fab operon (with known members SCO3245-49) that encodes the biosynthesis of the fatty acid moiety of CDA (Hojati et al., 2002). We further assume that this TF and its 17 target-genes constitute a SIM.

5.2 *Reconstruction of CdaR activity*

To reconstruct the activity profile of the CdaR regulator, the profiles of all 17 differentially expressed genes within this regulatory module are used. In other words, we consider a SIM with CdaR as its master regulator and the 17 genes as its targets. The maximum likelihood estimate of the activity profile $\bar{\eta}(t)$ for CdaR found by the conjugate gradient method using gene expression data for all 17 targets for wild type organism is shown in Figure 1.

[Figure 1 about here.]

The confidence bounds for $\bar{\eta}$-component were obtained via a classical Wilks procedure. Let L^* be the value of the maximum likelihood found with respect to all parameters, including $\bar{\eta}_j$. By perturbing each $\bar{\eta}_j + \triangle_j$, we obtain a value of likelihood $L_j^* = L(\bar{\eta}_0, \ldots, \bar{\eta}_j + \triangle_j, \ldots, \bar{\eta}_{N-2})$. The 95% confidence bound for $\bar{\eta}_j$ is found by finding $\triangle_j$ such that $(L^* - L_j^*)/2 = \chi^2_{1,0.95}$. Figure 1a shows the reconstructed CdaR activity profile as a piece-wise constant function (solid line). Dashed lines show upper and lower 95% Wilks confidence bounds for each $\bar{\eta}_j$. A CdaR profile smoothed over the reconstructed piece-wise profile is shown on Figure 1b (solid line). Smoothed profile was obtained by the cubic spline function (R-function `pspline`). Points (connected with dashed lines) represent the observed data for *cdaR* gene expression for the two independent biological replicates.

Because of the arbitrary scale of the expression data, the shapes of reconstructed $\bar{\eta}$ and the expression data for cdaR are of interest to us, rather than their absolute values. The Pearson correlation between inferred activity profile and the average expression profile is 0.45, suggesting that the regulator CdaR is modified post-translationally. Indeed, it is highly likely

that the activity of CdaR protein is influenced by its phosphorylation state. The deduced CdaR protein sequence has a putative ATP-binding site and it is known that the activity of related streptomycete antibiotic regulatory proteins, such as AfsR, is governed by protein phosphorylation. The data presented here would be consistent with such post-transcriptional modification. This indicates that it is not safe to substitute the activity of the regulator by the measured gene expression in the models of gene regulation.

Expression profiles of all 17 differentially expressed genes between wild-type and mutant has been used in the reconstruction of the transcription activity profile of their common regulator. To evaluate how sensitive the result is to the false positives among the targets, we performed the same analysis by iteratively leaving one of the putative targets out. The TFA profiles found for each of the 17 SIMs with 16 targets were compared to the TFA profile found for the original SIM with 17 targets. The results are shown in Figure 2.

[Figure 2 about here.]

The mean correlation between the original TFA and the ones found for SIMS with 16 targets is 0.872. It is clear from Figure 2 that some difference is noticeable on the first and last time intervals. However, in each case the reconstructed profile of the gene target that has been left out shows excellent fit with the expression data for this gene (not shown). This is not surprising, as each of the inferred TFA profiles has a high correlation with the original inferred profile.

5.3 *Kinetic profiles of target genes*

For each of the 17 target DE genes, the mean gene expression profiles $\mu^k(t)$ and kinetic parameters $\theta^k,\ k = 1 \ldots 17$ of the MM model (6) were estimated given the reconstructed profile of the TF, $\bar{\eta}(t)$. Two representative gene profiles within the regulatory module for wild type are shown in Figure 3.

[Figure 3 about here.]

It is difficult to evaluate the estimates of the kinetic parameters, θ^k, as quantitative biological knowledge on gene transcription in general, and for *Streptomyces coelicolor* in particular, is very limited. It is nevertheless worthwhile mentioning some details about the kinetic parameters (see Web Appendix D).

5.4 *Goodness-of-fit*

As the observed gene expression data, $g^k_{cr}(t)$, are assumed to be log-normally distributed (**??**), $\log[g^k_{cr}(t)]$ is normally distributed. Therefore,

$$\frac{\log[g^k_{cr}(t)] - m^k_c(t)}{\sigma_k} \sim N(0,1),$$

where the location parameter m^k_c is given by formula (4). Whether the inferred data truly comes from a normal distribution can be tested by a Kolmogorov-Smirnov test by using, for example, the R-function `ks.test`.

Figure 4 shows a QQ-plot of the p-values from the Kolmogorov-Smirnov test for all 17 differentially expressed genes between wild type and *cdaR* mutant. This figure shows that the MM model combined with log-normal deviations displays a very good fit to the observed time-course gene expression data. The dashed line stands for an ideal fit of the data to the model. If the p-values fall below this line, the fit is poor. P-values above the line indicate some overfit of the model. However, the 95% confidence bounds of the uniform distribution (dotted line) show how most of the p-values might be higher than the line simply by chance, as they fall within the upper confidence bound.

[Figure 4 about here.]

To address concerns of overfitting, we compare the current model with gene-specific variances σ^2_k, with a model, wherein a common variance σ^2 is used for all genes. The maximum likelihood estimate for common σ^2 has been found by a grid-search between the smallest and largest values of σ^2_k. The likelihoods of the two models are compared using a χ^2-test with 16 degrees of freedom, i.e. the difference in the number of parameters. This yields a statistic of 153.17, which far exceeds the 95% cut-off of $\chi^2_{16,0.95} = 26.3$. This suggests that the model with

a gene-specific variance gives the best trade-off between the goodness-of-fit and the number of parameters in the model.

6. Discussion

In this paper we developed a statistical framework that embeds the deterministic Michaelis-Menten kinetics of gene regulation within a stochastic model of microarray measurement noise. As an alternative for direct experimental measurement of the activity profile of the TF, the model reconstructs this profile using the gene expression profiles of its targets within a Single Input Motif regulatory module. In addition, estimates of gene specific kinetic parameters of the gene regulation are found. We have shown that in the case of post-transcriptional modifications, such as is the case in the cdaR gene in *Streptomyces coelicolor*, the amount of mRNA of a regulator is not a good approximation for its protein activity levels and one cannot be substituted for the other in quantitative models of gene regulation.

Our statistical framework requires some knowledge of the structure of the regulatory module, which can be determined by experimental methods (ChIP technology), analytical (e.g. by finding differentially expressed genes) and available biological knowledge. Currently, in the absence of a reliable technology to routinely measure the TFA of regulators, it is of great importance to understand whether TFA can be inferred from the expression of its targets. A straightforward experimental verification of the results is to measure the phosphorylation profile of CdaR and compare it with the TFA, inferred by our model.

The statistical framework developed in this paper can be extended to include cooperativity and competitive regulation by two or more TFs with both *AND* and *OR* gate-types. It can be used to reconstruct the activity of TFs in known regulatory modules and to discriminate between the types of regulation (activation/inhibition; gate types) by using likelihood ratio and goodness-of-fit tests. The model can also be extended to search for the TFA and gene-specific kinetic parameters of regulation by combining different microarray datasets. Other types of data as well as available knowledge can be incorporated in the model.

7. Supplementary Materials

R-code and other supplementary materials are available under the Paper information link at the Biometrics website: `http://www.tibs.org/biometrics`.

Acknowledgements

The work on this project was funded by a BBSRC Exploiting Genomics initiative consortium grant "DNA microarray data analysis and modelling: an integrated approach".

References

Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N. and Yoo, J. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnoly* **21**, 1337–1342.

Bentley, S., Chater, K., Cerdeno-Tarraga, A.-M., Challis, G. and Thomson, N. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147.

Bibb, M. (1996). The regulation of antibiotic production in *Streptomyces coelicolor* A3(2). *Microbiology* **142**, 1335–1344.

Bolouri, H. and Davidson, E. (2002). Modelling transcriptional regulatory networks. *BioEssays* **24**, 1118–1129.

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**.

Gao, F., Foat, B. and Bussemaker, H. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 131–141.

Hojati, Z., Milne, C., Harvey, B., Gordon, L. and Borg, M. (2002). Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*. *Chemistry and Biology* **9**, 1175–1187.

Kao, K., Yang, Y., Boscolo, R., Sabatti, C., Roychowdhury, V. and Liao, J. (2004). Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proceedings of the National Academy of Science* **101**, 641–646.

Lee, M., Kuo, F., Whitmore, G. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cdna hybridizations. *Proceedings of the National Academy of Sciences of the United St ates of America* **97**, 9834–9839.

Lee, T., Rinaldi, N., Robert, F., Odom, D. and Bar-Joseph, Z. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799–804.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Science* **100**, 11980–11985.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827.

Nachman, I., Regev, A. and Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* **20**, I248–I256.

Qian, J., Lin, J., Luscombe, N., Yu, H. and Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* **19**, 1917–1926.

Ryding, N., Anderson, T. and Champness, W. (2002). Regulation of the *Streptomyces coelicolor* calcium-dependent antibiotic by *absA*, encoding a cluster-linked two-component system. *Journal of Bacteriology* **184**, 794–805.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifing regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**, 166–176.

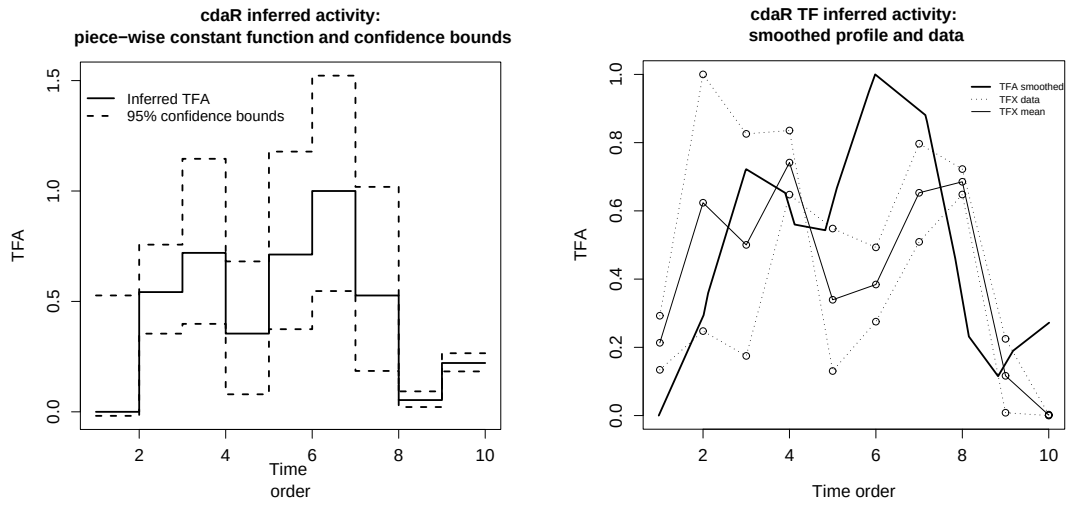
Sheeler, N., MacMillan, S. and Nodwell, J. (2005). Biochemical activities of the *absA* two-component system of *Streptomyces coelicolor*. *Journal of Bacteriology* **187**, 687–696.

Shen-Orr, S., Milo, R., Mangan, S. and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* **31**, 64–68.

Yu (2004). A mixed model approach to identify yeast transcriptional regulatory motifs via microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**.

Yu, H., Luscombe, N., Qian, J. and Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in Genetics* **19**, 422–427.

Zhou, X., Kao, M., Huang, H., Wong, A. and Nunez-Iglesias, J. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature Biotechnology* **23**, 238–243.



(a) (b)

Figure 1. Transcription Factor Activity of CdaR inferred from gene expression data. (a) the TFA is the piece-wise constant step function (solid line) together with its 95% confidence bounds (dashed lines). The inferred profile is re-scaled between zero and one. Corresponding confidence bounds are rescaled accordingly. (b) TFA vs TFX of cdaR for wild-type time-course. The TFA profile is smoothed using R `spline` function (solid line) from inferred piece-wise constant function. Points represent the observed data for two biological replicates for TFX (wild type) (dashed lines). Smoothed profile has been re-scaled between zero and one; data points have also been re-scaled independently to be between zero and one.

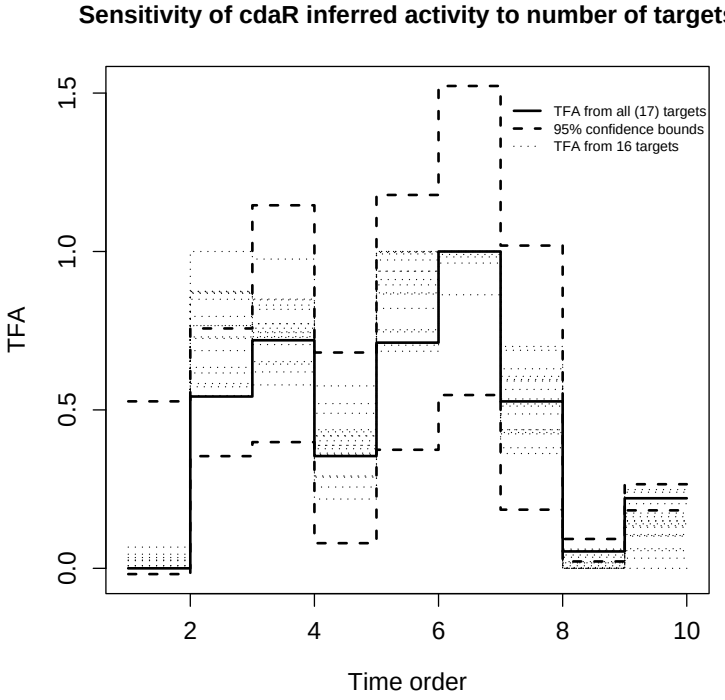
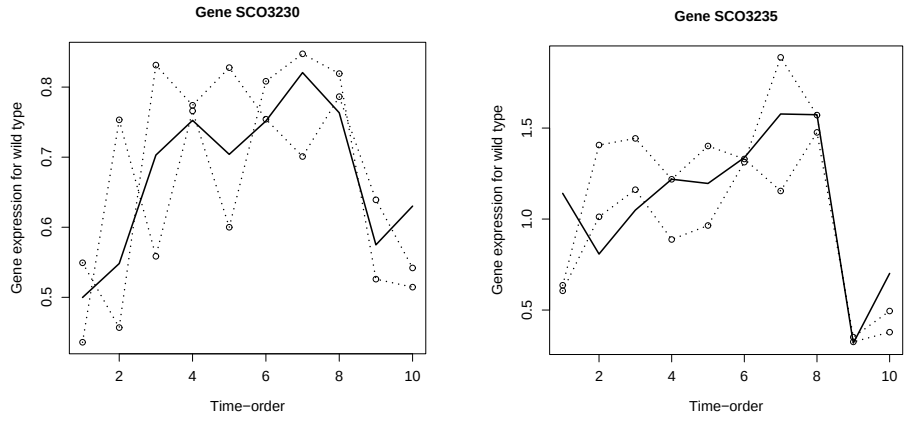


Figure 2. Sensitivity of the TFA to possible false positives among the targets. The TFA reconstructed from an original SIM with 17 targets (solid line); TFAs reconstructed for SIMs with 16 targets (leave-one-out) (dotted lines); 95% confidence bounds (dashed lines).



(a) (b)

Figure 3. Two representative profiles of target genes within the SIM regulatory module. The points connected by dotted lines stand for the observed data for the wild type (2 replicates). The solid line is for a gene profile fitted with the inferred TFA of CdaR regulator $\bar{\eta}$. (a) Gene SCO3230. ML estimates of kinetic parameters are $\beta = 168$, $\gamma = 569$, $\delta = 48$, $\alpha = 0.55$, $\sigma = 0.14$. (b) Gene SCO3235. ML estimates of kinetic parameters are $\beta = 265$, $\gamma = 516$, $\delta = 9.4$, $\alpha = 0.000001$, $\sigma = 0.31$.

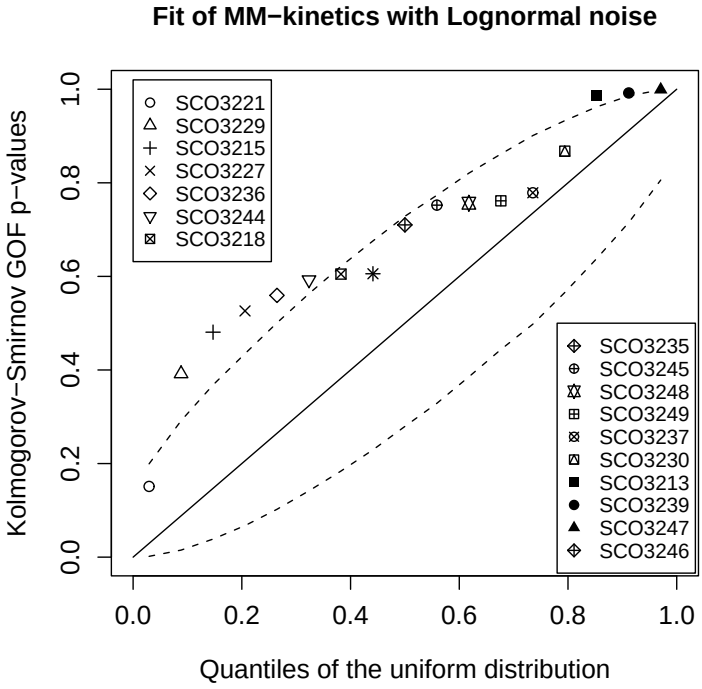


Figure 4. Fit of MM-kinetics with lognormal noise and ML estimate of η^A for 17 genes identified as differentially expressed between the wild type and the *cdaR*-mutant. The p-values from Kolmogorov-Smirnov test are shown versus the quantiles of the uniform distribution. Dashed line stands for an ideal fit of the data to the model.