BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 23 2008, pages 2748–2754
doi:10.1093/bioinformatics/btn527

Systems biology

Stochastic dynamics of genetic networks: modelling and parameter identification

Eugenio Cinquemani*, Andreas Milias-Argeitis, Sean Summers and John Lygeros

Automatic Control Laboratory, ETH, 8092 Zurich, Switzerland

Received on June 13, 2008; revised on September 17, 2008; accepted on October 7, 2008

Advance Access publication October 9, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Identification of regulatory networks is typically based on deterministic models of gene expression. Increasing experimental evidence suggests that the gene regulation process is intrinsically random. To ensure accurate and thorough processing of the experimental data, stochasticity must be explicitly accounted for both at the modelling stage and in the design of the identification algorithms.

Results: We propose a model of gene expression in prokaryotes where transcription is described as a probabilistic event, whereas protein synthesis and degradation are captured by first-order deterministic kinetics. Based on this model and assuming that the network of interactions is known, a method for estimating unknown parameters, such as synthesis and binding rates, from the outcomes of multiple time-course experiments is introduced. The method accounts naturally for sparse, irregularly sampled and noisy data and is applicable to gene networks of arbitrary size. The performance of the method is evaluated on a model of nutrient stress response in *Escherichia coli*.

Contact: cinquemani@control.ee.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Mounting experimental evidence suggests that gene expression, both in prokaryotes and eukaryotes, is an inherently stochastic process. Stochasticity can be attributed to the randomness of the transcription and translation processes (intrinsic noise), as well as to fluctuations in the amounts of molecular components that affect the expression of a certain gene (extrinsic noise) (Elowitz *et al.*, 2002; Longo and Hasty, 2006; McAdams and Arkin, 2002; Paulsson, 2005). The behaviour of gene regulatory networks also displays stochastic characteristics which, in several cases, can lead to significant phenotypic variation in isogenic cell populations (Ozbudak *et al.*, 2002). In Samad *et al.* (2005), stochastic modelling of genetic regulatory networks is reviewed along with numerical simulation methods and is compared to deterministic modelling. A related model of random dynamics of gene networks is discussed in Hespanha and Singh (2005). In practice, the stochastic dynamics of a regulatory network must be inferred from experiments. To this aim, deterministic models are not suitable, since they are unable to capture the randomness of the network. On the other hand, currently available stochastic models are typically too detailed and hardly tractable by analytic means.

In this work, we present a stochastic approach to modelling and parameter identification of gene regulatory network dynamics and test it on a model for a procaryotic cell. The aim of our modelling framework is to provide a convenient tradeoff between model accuracy and analytic tractability that is typically not offered by more complex models. In Cinquemani *et al.* (2008), we presented *ad hoc* solutions for the identification of a simple biological model composed of a chain of genes coupled to macroscopic (population-level) dynamics. In Koutroumpas *et al.* (2008) the same example was used to investigate identifiability of the gene network parameters from macroscopic data by way of randomized optimization methods. Both these contributions were Taylor-made for the specific form of the example. Here, we extend the concepts to establish a general genetic network modelling and identification methodology.

In Section 2.1, we introduce a model that accounts for the stochastic nature of gene regulation dynamics by describing the binding of transcription factors (TFs) as discrete random events. In contrast, the description of the transcription and translation processes is simplified by assuming that they can be approximated by deterministic first-order kinetics. A similar approach is taken in Zeiser *et al.* (2008) for the analysis and numerical simulation of basic transcriptional network modules. The co-existence of discrete and continuous-type events as well as of deterministic and stochastic dynamics makes the model a stochastic hybrid one (Kouretas *et al.*, 2006).

Based on this modelling formalism, genetic network identification is addressed in Section 2.2. The overall problem can be seen as a sequence of three tasks, each posing its own challenges: (1) identification of the network of interactions; (2) estimation of the unknown parameters; (3) validation of the model. Here, we concentrate on Task 2, assuming that Task 1 has been accomplished. Parameter estimation for regulatory networks has traditionally been studied in a deterministic setting. The literature on identification of stochastic regulatory network models is quite recent. In Reinker *et al.* (2006), an approximate maximum likelihood identification method is developed for a discrete Markov chain model of biochemical reaction networks. A similar approach is taken in Tian *et al.* (2007), where the likelihood function is evaluated by simulation. A Markov chain Monte Carlo method relying on an approximate diffusion model is considered in Golightly and Wilkinson (2005). Based on our modelling framework, we develop a method for the estimation

*To whom correspondence should be addressed.

 © The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

of the unknown parameters of the model from observations of the evolution of protein concentrations in a single cell. The method deals naturally with sparse and noisy observations. It reduces the identification problem to several subproblems. In each subproblem, the estimation of the dynamics of a single gene is performed based on the concentration profiles of the proteins that regulate its expression.

To demonstrate the effectiveness of the proposed identification algorithm, we apply it to the estimation of the parameters of a stochastic model of the *Escherichia coli* carbon starvation response network. The model is inferred from Ropers *et al.* (2006) and described in Section 3.1. Estimation is performed on simulated data from the same model with realistic parameter values (G.Ferrari-Trecate *et al.*, 2007, personal communication). Results are discussed in Section 3.2. In light of the current rapid progress of single-cell protein level measurement techniques (Cai *et al.*, 2006; Golding *et al.*, 2005), the next step will be applying the method to protein concentration profiles drawn from real-world experiments.

2 METHODS

2.1 Genetic network modelling

Gene expression in prokaryotes is mainly regulated at the stage of transcription (Wagner, 2000). Of the many steps that transcription comprises, gene activation and inactivation is one of the key events that contribute to random fluctuations of protein production. The main molecular causes for this 'switching' are dissociation of repressors and association of activators to the operator of the gene. Collectively called TFs, repressors and activators are composed of proteins that may be produced by other genes, or by the regulated gene itself, thus creating feedback loops among genes.

Consider a network with n genes, each encoding one protein. Let $x_i(t)$ denote the concentration of protein i at time t. For $i=1,\dots,n$, we describe the evolution of $x_i(t)$ by the following discrete-time model: for a given $T>0$ and all $t\in T\times\mathbb{Z}$ (integer multiples of T),

$$x_i(t+T)=\lambda_i x_i(t)+g_i(t), \tag{1}$$

where $\lambda_i\in[0,1]$ is a degradation rate and $g_i(t)\geq 0$ is a variable synthesis rate associated with the activation state of gene i at time t. More specifically, g_i is modelled as follows:

$$g_i(t)=\sum_j\left(b_i^j\prod_{k\in\ell(i,j)}u_{i,k}(t)\right), \tag{2}$$

where the b_i^j are fixed synthesis rates, the $u_{i,k}(t)$ are variables taking on values 0 or 1 and $\ell(i,j)$ is a subset of $\{1,\dots,n\}$. Each $u_{i,k}(t)$ indicates whether TF k is bound ($u_{i,k}(t)=1$) or not bound ($u_{i,k}(t)=0$) to the operator site of gene i at time t. Products of variables $u_{i,k}$ correspond to the requirement that all TFs listed in the index set $\ell(i,j)$ be simultaneously present for gene activation/inhibition. In other words, if a certain k belongs to $\ell(i,j)$, then TF k is involved in the regulation and hence $u_{i,k}(t)$ is included in the product; if $\ell(i,j)$ is empty, we assume by convention that the product is equal to 1. Finally, different summation terms (i.e. different values of j) may describe alternative regulation paths. This model may encode quite complicated activation rules and is illustrated in Figure 1.

More generally, one may think of (2) as abstract rules governing the expression of a gene. In this case, variable $u_{i,k}(t)$ may not express the binding of TF k to the operator of gene i, but a different discrete event whose outcome is the regulation of gene i by TF k, such as formation of complexes, translocation, etc.

Let $x_{\ell(i)}$ be the subvector of x collecting the concentrations of the proteins that act as TFs on gene i. We assume that the evolution in time of each $u_{i,k}$ is stochastic and is governed by the laws of a discrete-time Markov chain

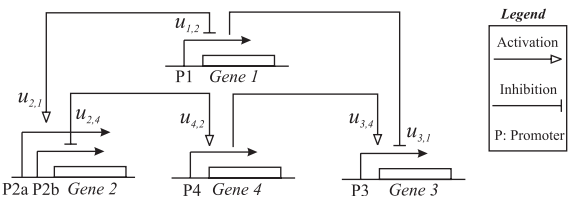

Fig. 1. A regulatory network with $n=4$ genes. Graphical conventions follow (Kohn, 2001). Expression of gene 4 is activated by TF2. Then, $g_4(t)=b_4^1u_{4,2}(t)$ (when TF2 is bound to the operator of gene 4, the gene is expressed and the encoded protein is synthesized at rate b_4^1). Conversely, expression of gene 1 is inhibited by TF2. Then, $g_1(t)=b_2^1\big(1-u_{1,2}(t)\big)$ (the protein encoded by gene 1 is synthesized at rate b_2^1 when TF2 is not bound to the operator). Expanding this product one gets $g_1(t)=b_2^1+b_2^2u_{1,2}(t)$, with $b_2^2=-b_2^1$, which is in the form of (2). Gene 3 has a single promoter but is controlled simultaneously by the activating TF4 and by the inhibiting TF1. Then, $g_3(t)=b_3^1u_{3,4}(t)\big(1-u_{3,1}(t)\big)$ (the gene is expressed whenever TF4 is bound and TF1 is not bound to the promoter site). Finally, gene 2 has two independent promoters, one for the activating TF1 and one for the inhibiting TF4. Then, $g_2(t)=b_2^1u_{2,1}(t)+b_2^2(1-u_{2,4}(t))$ (the synthesis of the protein encoded by gene 3 occurs at rate b_2^1 if TF1 is bound and at rate b_2^2 if TF4 is not bound; if both conditions hold simultaneously, the resulting synthesis rate is the sum of the two).

with transition probabilities

$$p_{i,k}(z)=\mathbb{P}[u_{i,k}(t)=1|u_{i,k}(t-T)=0,x_{\ell(i)}(t)=z],$$
$$q_{i,k}(z)=\mathbb{P}[u_{i,k}(t)=0|u_{i,k}(t-T)=1,x_{\ell(i)}(t)=z],$$

where z is arbitrary and the notation $\mathbb{P}[A|B]$ denotes the conditional probability of event A given event B. The probability that $u_{i,k}$ remains at 0 [respectively at 1] is fixed to $1-p_{i,k}(z)$ [respectively to $1-q_{i,k}(z)$]. The following is a basic underlying assumption of our model framework.

ASSUMPTION 1. *For different values of i and k, the random variables $u_{i,k}(t)$ are mutually conditionally independent given the $x_{\ell(i)}(t)$ and their previous values $u_{i,k}(t-T)$*

For ease of exposition, we shall also assume the following.

ASSUMPTION 2. $q_{i,k}=1-p_{i,k}$

The generalization of our methods to arbitrary choices of $q_{i,k}$ and $p_{i,k}$ is straightforward. Under Assumption 2, one finds that the $u_{i,k}(t)$ are independent of the $u_{i,k}(t-T)$ given the $x_{\ell(i)}(t)$ (Cinquemani *et al.*, 2008). We shall focus on the biologically relevant case where each transition probability $p_{i,k}$ is a sigmoidal function of x_k (*Hill function*). That is, dropping subscripts for simplicity, $p(x)=s^+(x;\eta,d)$ or $p(x)=s^-(x;\eta,d)$, where

$$s^+(x;\eta,d)=\frac{x^d}{x^d+\eta^d},\qquad s^-(x;\eta,d)=\frac{\eta^d}{x^d+\eta^d}$$

with $\eta>0$ and $d>0$. Function s^+ increases from 0 to 1 as concentration x increases from 0 to $+\infty$; the larger the concentration, the larger the probability that the TF will be bound to a target site. Parameter η is the value of x for which $s^+(x;\eta,d)=1/2$ and will be called the *threshold*. Parameter d defines how abruptly the transition from $s^+(x;\eta,d)<1/2$ to $s^+(x;\eta,d)>1/2$ occurs and will be called the *steepness*. Function $s^-=1-s^+$ is complementary, i.e. it decreases from 1 to 0, downcrossing $1/2$ at η with a steepness increasing with d. While this function cannot represent a binding probability, it is well suited to express the influence of a TF on the expression of a gene. For instance, if TF k inhibits the expression of gene i, function s^- says that larger concentrations of TF k imply smaller probabilities of the transcription of gene i. In more generality, the probability laws $s^\pm$ (meaning s^+ or s^-) quantify the probability of the discrete events

that the associated Markov chains describe. Examples of a decreasing and an increasing sigmoidal function can be found in Figure 4 (subfigures C2 and C4).

Let $\mathbb{E}[\cdot|x(t)]$ denote conditional expectation given state $x(t)$. For $i=1,\dots,n$, the expected evolution of (1) from $x(t)$ is

$$\mathbb{E}[x_i(t+T)|x(t)] = \lambda_i \mathbb{E}[x_i(t)|x(t)] + \sum_j b_i^j \prod_{k\in\ell(i,j)} \mathbb{E}[u_{i,k}(t)|x(t)]$$

$$= \lambda_i x_i(t) + \sum_j b_i^j \prod_{k\in\ell(i,j)} s^{\pm}(x_k(t);\eta_{i,k},d_{i,k}),$$

where the conditional expectation commutes with the products thanks to Assumption 1, and the expectation of each $u_{i,k}(t)$ follows from Assumption 2 and the definition of the $p_{i,k}$. This equation suggests a link with commonly used deterministic models of gene regulatory networks in the form of reaction-rate equations, such as those reviewed in de Jong (2002), where sigmoidal functions are used to model the binding of TFs on the operator DNA (Becskei *et al.*, 2001; Keller, 1995; Yang *et al.*, 2007). It shows that the expected evolution of the system from a given state is in agreement with standard ordinary differential equation models. However, due to randomness, the actual next state may differ from the expected one. This possibility is excluded by deterministic modelling.

The use of a Markov chain for modelling the changing states of the operator can be traced back to one of the first stochastic models for gene induction (Ko, 1991). Here, we further assume that the switching rate of a certain gene depends in a non-linear way on the concentration of the TFs that affect it. This assumption is based on experimental evidence suggesting that TFs directly affect the probability of formation of the transcription complex, while the rate of protein production once the gene is active remains independent of TF concentration (Walters *et al.*, 1995). Along the same lines of Ko (1991), we lump together all the individual steps between gene activation and mRNA production, by assuming that RNA polymerase produces a constant amount of RNA transcripts per unit time (once the gene is active). On average, the number of these transcripts will be proportional to the average number of protein molecules produced by translation [a commonly made assumption, which is also supported by recent experiments (Golding *et al.*, 2005)].

2.2 Parameter identification algorithm

Consider an experiment where the evolution of the protein concentration levels is observed at time instants $\tau_0<\tau_1<\tau_2<\cdots<\tau_L$. We model the observations as follows:

$$y(\tau_l) = x(\tau_l) + n(\tau_l), \tag{3}$$

where $x(\tau_l)$ is the vector of protein concentration levels at time τ_l and $n(\tau_l)$ is independent, identically distributed (i.i.d.) measurement noise with mean zero and covariance matrix $\Sigma=\mathrm{diag}(\sigma_1^2,\dots,\sigma_n^2)$. For simplicity, we assume that τ_l is an integer multiple of T for all l. We address the problem of estimating the parameters λ_i, b_i^j, $\eta_{i,k}$ and $d_{i,k}$ from the observations (3). Establishing the identifiability of these parameters analytically is an extremely difficult task, out of the scope of this article. A qualitative discussion of identifiability issues will be given in the case study of Section 3.1.

The identification method we propose is inspired by the Prediction Error Methods used for the parameter identification of linear stochastic systems (Ljung, 1999). For $i=1,\dots,n$, we perform simultaneous estimation of all parameters relevant to the i-th equation (1), namely $\theta_i=\{\lambda_i\}\cup\{b_i^j\colon \forall j\}\cup\{\eta_{i,k},d_{i,k}\colon \forall k\}$, from the observations of $x_{\ell(i)}$ and of x_i.

Let $\mathscr{Y}_i(\tau_l)=\{y_{\ell(i)}(\tau_h),y_i(\tau_h)\colon h=1,\dots,l\}$ be the observations of $x_{\ell(i)}$ and of x_i up to time τ_l. We consider the optimal predictor $\hat{x}_i(\tau_{l+1},\tau_l;\theta_i)\triangleq \mathbb{E}[x_i(\tau_{l+1})|\mathscr{Y}_i(\tau_l),\theta_i]$ associated with model (1)–(3), and draw an estimate $\hat{\theta}_i$ of θ_i by solving the following optimization problem:

$$\hat{\theta}_i = \arg\min_{\theta_i} \sum_{l=0}^{L-1} \big(y_i(\tau_{l+1}) - \hat{x}_i(\tau_{l+1},\tau_l;\theta_i)\big)^2. \tag{4}$$

The idea is that the closer the estimated model is to the real model, the better it can predict the evolution of the state, i.e. the sum of the prediction errors should be as small as possible. For any value of l, assume that $\hat{x}_i(\tau_l,\tau_l;\theta_i)$ is known. Then, $\hat{x}_i(\tau_{l+1},\tau_l;\theta_i)$ may be computed by iterating

$$\hat{x}_i(t+T,\tau_l;\theta_i) = \lambda_i \hat{x}_i(t,\tau_l;\theta_i) + \mathbb{E}[\bar{g}\big(x_{\ell(i)}(t)\big)|\mathscr{Y}_i(\tau_l),\theta_i]$$

for $t=\tau_l,\tau_l+T,\dots,\tau_{l+1}-T$, where

$$\bar{g}\big(x_{\ell(i)}(t)\big) = \sum_j b_i^j \prod_{k\in\ell(i,j)} s^{\pm}(x_k(t);\eta_{i,k},d_{i,k}).$$

Although the rightmost expectation cannot be computed in closed form, in practice one can use the approximation

$$\mathbb{E}[\bar{g}\big(x_{\ell(i)}(t)\big)|\mathscr{Y}_i(\tau_l),\theta_i] \simeq \bar{g}\big(x^*_{\ell(i)}(t)\big), \tag{5}$$

where $x^*_{\ell(i)}(t)\simeq x_{\ell(i)}(t)$ is chosen based on the data. If $x^*_{\ell(i)}(t)$ was the optimal estimate of $x_{\ell(i)}(t)$ given the data, approximation (5) would just follow from the linearization of $\bar{g}(\cdot)$ about $x^*_{\ell(i)}(t)$ itself (Jazwinski, 1970). However, the computation of the optimal estimate of $x_{\ell(i)}(t)$ involves additional unknown parameters $\theta_{i'}$, with $i'\neq i$, and requires the simultaneous observation of several (possibly all) proteins of the network, which is impractical. Fortunately, for realistic values of the steepness parameters $d_{i,k}$, the choice of $x^*_{\ell(i)}(t)$ turns out not to be critical. Therefore, we still use approximation (5) but interpolate $x^*_{\ell(i)}(t)$ from its neighbouring measurements by simply setting $x^*_{\ell(i)}(t)\triangleq[y_{\ell(i)}(\tau_{l+1})-y_{\ell(i)}(\tau_l)]/2$ for all $t\in[\tau_l,\tau_{l+1})$. This choice will be shown to perform well in our numerical experiments.

It remains to discuss the computation of $\hat{x}_i(\tau_l,\tau_l;\theta_i)$. If $\sigma_i=0$ (noiseless observations), then $\hat{x}_i(\tau_l,\tau_l;\theta_i)=x_i(\tau_l)$ and hence

$$\hat{x}_i(\tau_l,\tau_l;\theta_i) = y_i(\tau_l). \tag{6}$$

In general, the optimal estimate of $x(\tau_l)$ given $\mathscr{Y}_i(\tau_l)$ may be computed as a balance between the predicted value $\hat{x}_i(\tau_l,\tau_{l-1};\theta_i)$ and the new observation $y_i(\tau_l)$; the larger σ_i, the smaller the weight attributed to y_i (Jazwinski, 1970). For the sake of simplicity, we shall use the estimate (6). Again, the validity of this choice will be apparent from the simulation results.

This method can be immediately generalized to exploit data from several independent experiments. It is sufficient to reformulate the optimization problem (4) as follows:

$$\hat{\theta}_i = \arg\min_{\theta_i} \sum_{m=1}^{M}\sum_{l=0}^{L-1} \big(y_i^{(m)}(\tau_{l+1}) - \hat{x}_i^{(m)}(\tau_{l+1},\tau_l;\theta_i)\big)^2, \tag{7}$$

where superscript '$^{(m)}$' denotes data and predictions from the m-th experiment.

The algorithm performs separate identification of the dynamics of every gene in the network from the proteins that act on it as TFs. This guarantees that the complexity of the method scales well with the size of the network (the dimension of the search space for the i-th optimization problem is equal to the number of unknown parameters that enter the laws of g_i). In addition, it allows one to identify portions of a larger network based on a convenient subset of all proteins in the network. In principle, there is no guarantee that the cost function in (7) is convex, nor that it has a unique local minimum. Therefore, numerical minimization may be challenging. In absence of prior information on the unknown parameters, global optimization methods such as those reviewed in Moles *et al.* (2003) may be advisable. In this article, we shall not investigate numerical optimization strategies in detail. A discussion of numerical optimization for our case study is given in Section 3.2.

3 RESULTS AND DISCUSSION

One way to evaluate the accuracy of the estimates of the identification algorithm is to simulate identification *in silico*. The idea is to consider a model where the parameters are fixed to

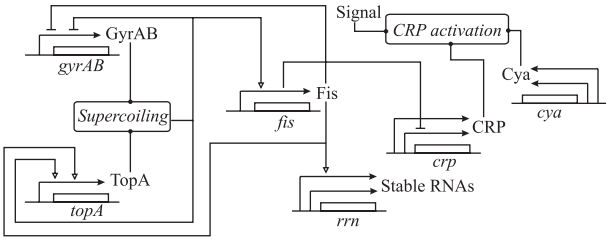

Fig. 2. Key global regulators and regulatory interactions taking place during the transition from stationary to exponential growth phase in *E.coli*. The model of the network used here is a reduced version of the original model by Ropers *et al.* (2006) for the case where the starvation signal is off. In the absence of starvation, the enzyme adenylate cyclase (*cya*) is not active. This leads to the deactivation of the global regulator CRP. The level of DNA supercoiling is controlled by the opposing effects of GyrAB and TopA and, in turn, regulates the expression of many genes in the cell, including *fis*. Note that the CRP activation and supercoiling boxes are used here to abstract sets of interactions. The genes encoding for stable RNAs (generally considered to represent cellular growth rate) are all regulated in the same way and are collectively represented in the model by *rrn*.

realistic values, and then estimate the same parameters from simulated data. This not only allows one to determine the reliability of the estimates under variable experimental conditions, but also provides hints on how to most profitably design the experiments. Our aim in this section is to establish a benchmark problem and to test the efficacy of our identification method on it.

3.1 Example: *E.coli* nutrients response

We consider the model of the *E.coli* carbon starvation response network discussed in Ropers *et al.* (2006), arguably one of the most sophisticated dynamical models of genetic regulatory networks found in the literature. Figure 2 shows all the various components of this network, but depicts only the regulatory interactions controlling the cell transition from the stationary phase (cells do not multiply) to the exponential growth phase (cells divide and the population grows exponentially). This happens only when the carbon starvation signal (acting as 'input' to the system) is off (lack of food is not detected). The model describes the concentration evolution of key global regulators in the network by means of ordinary differential equations, using sigmoidal activation functions. In Ropers *et al.* (2006) the reader can find the biochemical arguments leading from the graphical network representation to the derivation of the differential equations governing its behaviour.

In order to obtain a discrete-time stochastic hybrid model from Ropers *et al.* (2006), we replace the deterministic sigmoidal activation functions with random binary processes governed by sigmoidal probability laws. This leads to the following equations:

$$
\begin{aligned}
x_1^+ &= \lambda_1 x_1 + b_1^1 u_{1,3} + b_1^2 \\
x_2^+ &= \lambda_2 x_2 + b_2^1 \\
x_3^+ &= \lambda_3 x_3 + b_3^1 u_{3,3} + b_3^2 u_{3,3} u_{3,4} u_{3,5} \\
x_4^+ &= \lambda_4 x_4 + b_4^1 u_{4,3} + b_4^2 u_{4,3} u_{4,4} u_{4,5} \\
x_5^+ &= \lambda_5 x_5 + b_5^1 u_{5,3} u_{5,4} u_{5,5} \\
x_6^+ &= \lambda_6 x_6 + b_6^1 u_{6,3} + b_6^2
\end{aligned}
$$

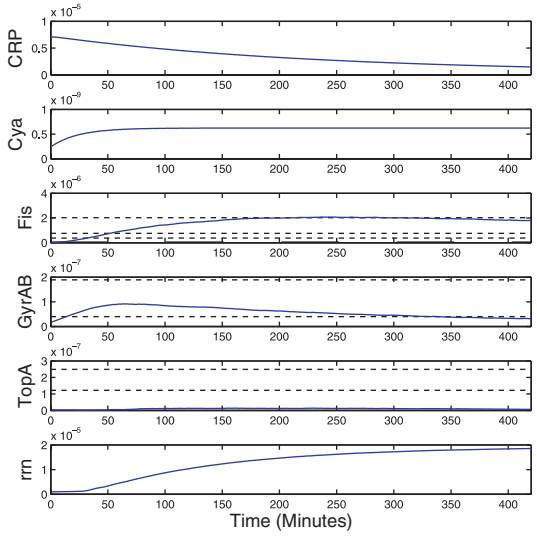

Fig. 3. Simulation of nutrient upshift response in *E.coli* cell (molar concentrations).

where x_i^+, x_i and $u_{i,k}$ are shorthand for $x_i(t+1)$, $x_i(t)$ and $u_{i,k}(t)$, respectively. The state variables x_1 and x_2 represent the concentrations of the global regulator cAMP (cyclic adenosine monophosphate) receptor protein (CRP) and the signalling enzyme adenylate cyclase (Cya), respectively. x_3 represents the concentration of the global regulator Fis, while x_4 and x_5 represent the concentrations of proteins GyrAB and TopA, controlling the level of DNA supercoiling. x_6 represents stable RNAs concentration. The probability laws of the binary variables $u_{i,k}$ are reported in the Supplementary Material. The resulting stochastic system has 16 rate parameters, 11 threshold coefficients and 11 steepness coefficients.

3.2 *In silico* simulation and identification

We consider a scenario where *E.coli* has undergone an initial growth phase followed by starvation. We simulate the re-entry into the exponential growth phase using realistic values for protein synthesis/degradation rates, sigmoid coefficients and initial concentrations after starvation. These parameter values were derived by manually fitting the model in Ropers *et al.* (2006) to experimental data and are reported in the Supplementary Material (G.Ferrari-Trecate *et al.*, 2007, personal communication). Figure 3 shows the evolution of the states during phase transition (blue solid line) as compared with the protein threshold values (dashed black lines). The narrow state evolution path limits the number of thresholds crossed, which may result in constant binary process values.

Parameter identification experiments were performed on the simulated data based on the same model but assuming that the model parameters are unknown. The assumption that the exact model structure is known clearly facilitates the identification task, as the presence of model mismatch would bias the parameter estimates. The stochastic nature of the system and the unobserved binding/unbinding activity of the regulation factors makes the identification problem challenging even if the dynamics that govern the network are perfectly known.

In each run of the identification algorithm, we considered data from 25 (and 100) *in silico* generated stochastic state trajectories.

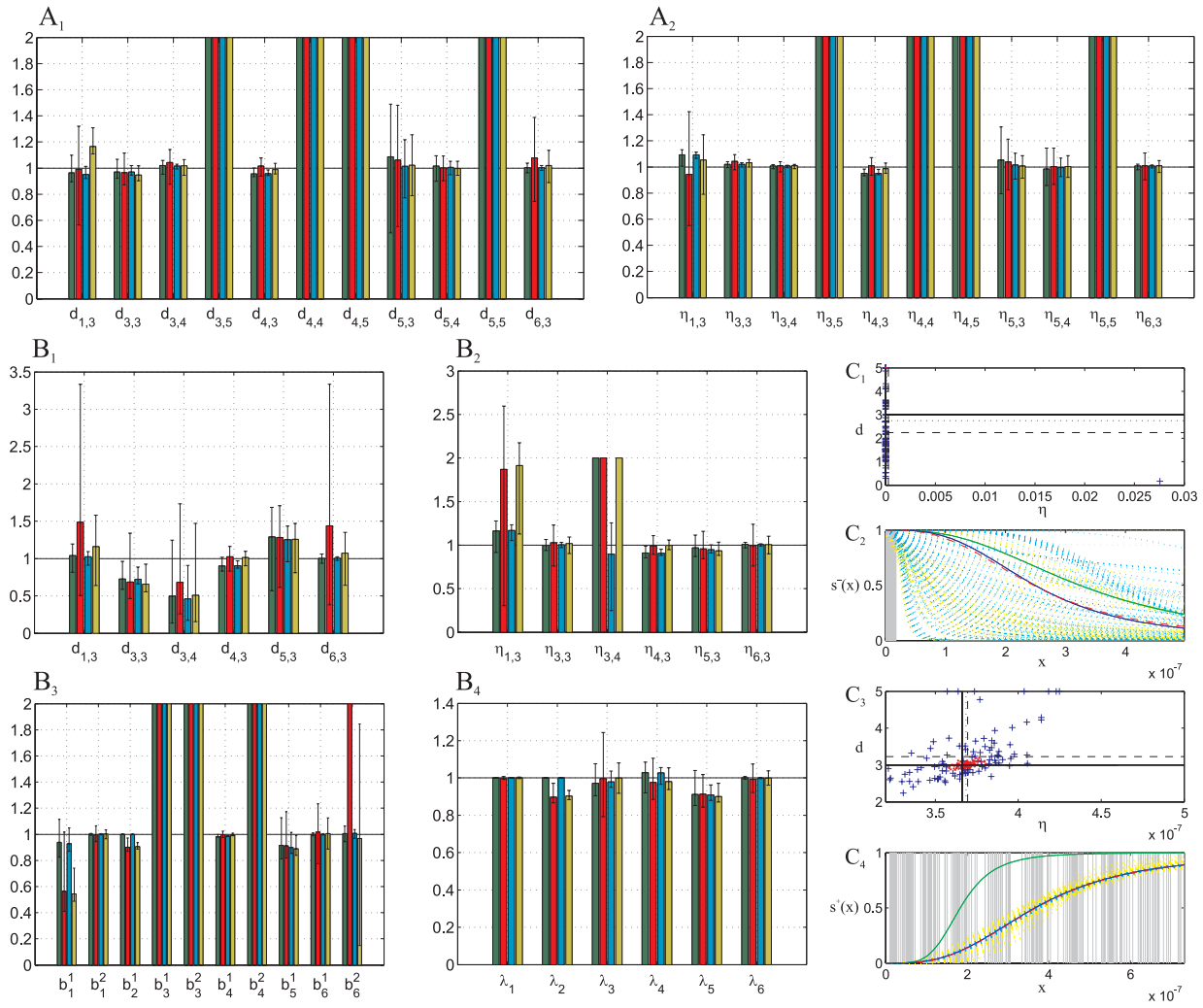

Fig. 4. The bar plots compare the estimation results from 100 Monte Carlo repetitions of the two independent identification tests under four different experimental conditions (25 trajectories with and without measurement noise and 100 trajectories with and without measurement noise). The normalized mean values (i.e. the mean estimate divided by the true parameter value) and a 90% confidence band (i.e. 90 out of 100 estimates fall within the interval) are visually conveyed by the bars and brackets, respectively. For each parameter estimate, the green bar represents results from 25 trajectories without measurement noise, red 25 trajectories with measurement noise, blue 100 trajectories without measurement noise and yellow 100 trajectories with measurement noise (from left to right). (**A**) Results from the first identification test. Steepness (**A1**) and threshold (**A2**) coefficients of all sigmoids are identified. Those coefficients deemed unidentifiable (see Supplementary Material) are saturated at 2. (**B**) Results from the second identification experiment. Steepness (**B1**) and threshold (**B2**) coefficients explored by the dataset are considered along with all synthesis (**B3**) and degradation (**B4**) rates. The parameters deemed unidentifiable (see Supplementary Material) are saturated at 2. (**C**) An example of scatter and sigmoidal estimation plots from one explored (**C3** and **C4**) and one unexplored (**C1** and **C2**) sigmoid in the first test case (only sigmoid coefficients identified) with a dataset of 25 trajectories. The scatter plots compare the sigmoid coefficient estimates with (blue crosses with mean represented by black dashed lines) and without (red dots with mean represented by black dotted lines) measurement noise to the true parameter values (black solid lines). The sigmoid curves visually convey the variance and mean of the estimates with (cyan dotted lines with mean represented by a red dashed line) and without (yellow dotted lines with mean represented by a red dotted line) measurement noise versus the true sigmoid curves (blue solid line) and the sigmoid curve representing the initial estimates (green solid line). Vertical gray lines indicate the location of the measurements.

The time scale for the model was set to $T = 12$ s and the initial state values for each run were randomly selected according to a Gaussian distribution such that 95% of the initial conditions fall within $\pm 10\%$ of the nominal initial values. Single cell protein concentration values were obtained at 5 min increments (significantly undersampled with respect to the dynamical sampling period T) over a 7 h transient period. Measurement noise was taken to be normally distributed and concentrated within 10% of the protein median value with 95% probability. A visual example of the intrinsic stochasticity of the process as well as the effect of noise and sparse sampling can be found in the Supplementary Material. The resulting measurement rate and the amount of data collected for subsequent processing appear to be compatible with the current experimental capabilities, as described e.g. in Golding *et al.* (2005)

and Cai *et al.* (2006). The simulation of the stochastic trajectories and the identification of the unknown parameters were performed in the MatLab environment. The optimization problem (7) was solved using the local (gradient-based) solver 'fmincon' under the assumption that the order of magnitude of the parameters is known. Several global methods were considered and a multi-start method was tested, but the numerical advantages were negligible.

In the first test, we assume that all protein degradation and synthesis rates are known, and consider the identification of all sigmoid parameter values. This scenario is motivated by the fact that, in many cases of interest, protein synthesis and degradation rates do not depend on the structure of the network and can be estimated separately by means of dedicated experiments. The results from 100 Monte Carlo repetitions of the identification procedure are reported in Figure 4. They allow us to perform a statistical analysis of the performance of the identification method and provide information on the expected quality of the actual estimates. The mean and SD values for all parameter estimates are reported in the Supplementary Material.

The results reveal that the estimation accuracy depends crucially on the distribution of the data samples in space. Satisfactory parameter estimates from both 25 and 100 trajectories are obtained for the regulation functions that have been fully 'explored' by the system trajectories (i.e. the dataset includes sufficient information on both sides of the binary switching threshold value). In the presence of noise, the poor quality of the measurements contributes to estimation bias and increased variance. Additionally, as the number of trajectories is increased from 25 to 100 (with and without noise), the resulting mean values remain mostly unchanged while the SD reduces in magnitude.

On the contrary, estimation was inconsistent for those sigmoid probability laws that were only sampled near the values zero or one. In this case it is impossible to extrapolate the whole shape of the sigmoid (that is, the correct value of the parameters), as the randomness of the system and, when applicable, the measurement noise makes different parameter estimates agree with different datasets. This lack of robustness is not due to the identification method but to the data. This important observation suggests that, when it comes to parameter identification and experimental analysis, a good dataset can be just as important as a true model structure. When the true model structure is known, this difficulty can be ameliorated by designing experiments that guarantee that the activation functions are fully explored. In principle, this can be achieved by a convenient choice of the system's initial conditions. In practice, given that the initial protein concentration levels cannot be chosen freely, one could make the system follow different protein concentration profiles by controlled inhibition or enhancement of the expression of certain genes.

In a second test, we investigate the significantly more challenging problem of estimating protein synthesis and degradation rates simultaneously with the sigmoid coefficients. Since the synthesis rates b_i^j multiply the binary switching variables, their simultaneous estimation with the sigmoid parameters $\eta_{i,k}$ and $d_{i,k}$ may result in an ill-conditioned problem, especially if the value of the binary switching variable is nearly constant along the observed trajectories. In particular, it is clear that the identification of the sigmoid parameters that were not estimated correctly in the previous test is ruled out. Therefore, in addition to all degradation and synthesis rates, we only attempt to estimate the parameters of those regulation functions that are effectively explored by the data, while fixing the parameters of the unexplored sigmoids to their true values. The results are reported in Figure 4. They again demonstrate the strength of the identification method. The decay rates show excellent agreement for both the small (25 trajectories) and large (100 trajectories) datasets, with and without noise. In most instances the synthesis rate coefficients and the parameters of the sigmoids are estimated consistently, with slightly larger estimation variance in the presence of measurement noise. Yet, there are a few outstanding cases in which the estimation of certain parameters is weak or simply not possible, which would motivate additional experimentation. In the dynamical equation for x_1, the estimates of $\eta_{1,3}$, of b_1^1 and to some extent of $d_{1,3}$ are diluted in the presence of measurement noise. Contrary to the first test case, the combination of a large magnitude of noise and the additional requirement of identifying the synthesis rate b_1^1 led to a less than ideal statistical result. In the dynamical equation for x_3, the synthesis rates, b_3^1 and b_3^2, are lost under all conditions, while the estimates of $\eta_{3,4}$, $d_{3,3}$ and $d_{3,4}$ are considerably biased. Lastly, we consider the unidentified synthesis rate b_4^2 of the dynamical equation for x_4. Because the measurements are concentrated on one side of threshold $\eta_{4,5}$, the binary variable $u_{4,5}$ is essentially 0 during the entire trajectory. Since this variable multiplies the synthesis rate b_4^2, the effect of the latter is never felt. That is, the lack of identifiability of b_4^2 in this case is again due to the distribution of the data.

4 CONCLUDING REMARKS

We have considered the problem of stochastic modelling and parameter identification of genetic regulatory networks in prokaryotes. We introduced a model where the discrete nature of the interactions between TFs and gene operators obey stochastic laws that depend on the protein concentrations in the network, whereas the evolution of protein concentration levels is modelled by simple first-order reaction dynamics. In our opinion, this model provides a convenient tradeoff between accuracy and tractability. Based on this model, we proposed an algorithm that performs estimation of the network parameters from the observation of protein concentration time profiles, under the assumption that the topology of the network and the nature of the interactions (activation/repression) is known. The identification algorithm was applied on a benchmark model of carbon stress response in *E.coli*. This allowed us to assess the efficacy of the method and to gain insight into experiment design issues. We believe that, in addition to parameter estimation, the stochastic modelling framework we presented is well suited to the development of tools for model validation and for the identification of the network of interactions. These aspects of genetic network identification are part of our current research activity.

ACKNOWLEDGEMENTS

The authors would like to thank G. Ferrari-Trecate, R. Porreca, H. de Jong and D. Ropers for providing realistic parameter values for the *E.coli* carbon starvation response model.

Funding: European Commission under project HYGEIA (NEST-004995).

Conflict of Interest: none declared.

REFERENCES

Becskei,A. *et al*. (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.*, **20**, 2528–2535.

Cai,L. *et al*. (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature*, **440**, 358–362.

Cinquemani,E. *et al*. (2008) Subtilin production by *Bacillus subtilis*: stochastic hybrid models and parameter identification. *IEEE Trans. Automat. Contr., Special Issue on Systems Biology*, **53**, 38–50.

de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 69–105.

Elowitz,M.B. *et al*. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Golding,I. *et al*. (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Golightly,A. and Wilkinson,D. (2005) Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, **61**, 781–788.

Hespanha,J. and Singh,A. (2005) Stochastic models for chemically reacting systems using polynomial stochastic hybrid systems. *Int. J. Robust Nonlinear Contr.*, **15**, 669–689.

Jazwinski,A. (1970) *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Keller,A.D. (1995) Model genetic circuits encoding autoregulatory transcription factors. *J. Theor. Biol.*, **172**, 169–185.

Ko,M.S.H. (1991) A stochastic model for gene induction. *J. Theor. Biol.*, **153**, 181–194.

Kohn,K. (2001) Molecular interaction maps as information organizers and simulation guides. *Chaos*, **11**, 84–97.

Kouretas,P. *et al*. (2006) Stochastic hybrid modelling of biochemical processes. Ch.g. In Cassandras,C. and Lygeros,J. (eds). *Stochastic Hybrid Systems*, vol. 24 of *Automation and Control Engineering Series*. CRC Press, Boca Raton, FL, USA.

Koutroumpas,K. *et al*. (2008) Parameter identification for stochastic hybrid systems using randomized optimization: a case study on subtilin production by *Bacillus subtilis*. *Nonlinear Anal. Hybrid Syst.*, **2**, 786–802.

Ljung,L. (1999) *System Identification – Theory For the User*. Prentice Hall, Upper Saddle River, N.J.

Longo,D. and Hasty,J. (2006) Dynamics of single-cell gene expression. *Mol. Syst. Biol.*, **2**, Article No. 64.

McAdams,H.H. and Arkin,A. (2002) It's a noisy business! genetic regulation at the nanomolar scale. *Trends Genet.*, **15**, 65–69.

Moles,C. *et al*. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.*, **13**, 2467–2474.

Ozbudak,E.M. *et al*. (2002) Regulation of noise in the expression of a single gene. *Nat. Genet.*, **31**, 69–73.

Paulsson,J. (2005) Models of stochastic gene expression. *Phys. Life Rev.*, **2**, 157–175.

Reinker,S. *et al*. (2006) Parameter estimation in stochastic biochemical reactions. *IET Syst. Biol.*, **153**, 168–178.

Ropers,D. *et al*. (2006) Qualitative simulation of the carbon starvation response in *Escherichia coli*. *Biosystems*, **84**, 124–152.

Samad,H.E. *et al*. (2005) Stochastic modeling of gene regulatory networks. *Int. J. Robust Nonlinear Contr.*, **15**, 691–711.

Tian,T. *et al*. (2007) Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, **23**, 84–91.

Wagner,R. (2000) *Transcription Regulation in Prokaryotes*. Oxford University Press, Oxford, UK.

Walters,M.C. *et al*. (1995) Enhancers increase the probability but not the level of gene expression. *Proc. Natl Acad. Sci. USA.*, **92**, 7125–7129.

Yang,H. *et al*. (2007) An analytical rate expression for the kinetics of gene transcription mediated by dimeric transcription factors. *J. Biochem.*, **142**, 135–144.

Zeiser,S. *et al*. (2008) Simulation of genetic networks modelled by piecewise deterministic Markov processes. *IET Syst. Biol.*, **2**, 113–135.