On observability and reconstruction of promoter activity statistics from reporter protein mean and variance profiles

Eugenio Cinquemani

On observability and reconstruction of promoter activity statistics from reporter protein mean and variance profiles

Eugenio Cinquemani

Inria Grenoble – Rhône-Alpes, Montbonnot, 38334 St. Ismier CEDEX, France,
Email: eugenio.cinquemani@inria.fr,
Homepage: https://team.inria.fr/ibis/eugenio-cinquemani/

Abstract. Reporter protein systems are widely used in biology for the indirect quantitative monitoring of gene expression activity over time. At the level of population averages, the relationship between the observed reporter concentration profile and gene promoter activity is established, and effective methods have been introduced to reconstruct this information from the data. At single-cell level, the relationship between population distribution time profiles and the statistics of promoter activation is still not fully investigated, and adequate reconstruction methods are lacking.
This paper develops new results for the reconstruction of promoter activity statistics from mean and variance profiles of a reporter protein. Based on stochastic modelling of gene expression dynamics, it discusses the observability of mean and autocovariance function of an arbitrary random binary promoter activity process. Mathematical relationships developed are explicit and nonparametric, i.e. free of a priori assumptions on the laws governing the promoter process, thus allowing for the decoupled analysis of the switching dynamics in a subsequent step. The results of this work constitute the essential tools for the development of promoter statistics and regulatory mechanism inference algorithms.

Keywords: Gene regulation, Doubly stochastic process, Spectral analysis

1 Introduction

A common experimental technique to monitor gene expression is the use of reporter proteins [9], i.e. fluorescent or luminescent proteins that are synthesized upon expression of the gene of interest. Light intensity measurements collected at different points in time are proportional to the amount of reporter molecules. This provides a quantitative, however indirect, readout of the activity of the gene, since reporter abundance depends on gene activation via its own transcription and translation dynamics.

When cellular populations are observed as a whole, such as in automated microplate readers, an average reporter profile is obtained. An estimate of the

average gene activation over the population of cells may thus be obtained by regularized inversion of the reporter synthesis dynamics [21]. Provided accurate knowledge of the latter, reconstruction of the promoter activity allows one to investigate gene expression regulatory mechanisms, a crucial step toward inference of gene regulatory networks [18].

When individual cells are observed, for instance via flow-cytometry or fluorescence videomicroscopy, a statistical distribution of gene expression levels over a sample of the population (often called a population snapshot [6]) is obtained at several points in time (reporter traces for individual cells can also be obtained by suitable experimental setups and image processing techniques [20], but we will not analyze this case here). In many cases of interest, this crucially reveals variability of gene expression levels across cells that can be explained in terms of the stochasticity of the gene regulation and expression process [16, 19, 13]. Reporter statistics thus contains information about the stochastic laws governing gene activation. However, recovering the relevant information from the data is less trivial than in the population average case, and no satisfactory methods exist to date.

With reference to population snapshot data, in [2, 3], we have started addressing the problem of estimating promoter activity statistics (the biological information of interest in gene expression reporting) from reporter mean and variance profiles. In [2], parametric models of stochastic gene activation have been considered, and the identifiability of promoter switching rates that are fixed over time and across cells has been analyzed. However, due to a priori unknown regulatory mechanisms, switching rates may fluctuate over time and/or across cells (extrinsic noise). To cope with this, in [3], a nonparametric method, i.e. avoiding assumptions on the regulatory mechanisms behind the expression of the gene of interest, has been proposed for the special case of irreversible activation. A rather extensive account of relevant research literature is also contained in these works.

Following up from the developments in [2, 3], for the general case of unmodelled stochastic (possibly time varying) gene expression regulation, we address here the problem of reconstructing second-order statistics of the promoter activity process from reporter mean and variance profiles. The importance of this problem lies in the fact that, in analogy with linear stochastic processes [12], cross-correlation of promoter activity at different points in time (i.e. the autocorrelation function) contains information about the time dynamics of activation and deactivation. Reconstruction of these statistics from data is thus the crucial step for the understanding of the gene regulatory mechanisms at the level of single cells, where stochastic variability offers more to discover than traditional population analysis [13, 14].

The contribution of this paper is the development of explicit relationships between the unknown (first- and) second-order promoter activity statistics and the experimentally measurable reporter mean and variance profiles. Crucially, these relationships rely on nonparametric models of gene activation, i.e. no a priori assumption is made except the absence of stochastic feedback from re-

porter abundance to the regulation of the gene itself, a hypothesis that agrees well with the biochemistry of reporter systems. Based on analytic investigation and examples, we show that these relationships are essentially linear, whence invertible in a tractable manner, and allow for the discrimination among different promoter activity regulatory laws. On these basis, the implementation of algorithms for the actual estimation of the statistics of interest is left for future work. For ease of reading, all mathematical proofs are deferred to Appendix A. Appendix B, instead, summarizes results from [2] that are used in this work.

2 Background material

Gene expression monitoring over time is commonly operated by the use of fluorescent or luminescent reporter proteins (see [9] and references therein). In essence, synthesis of a reporter protein is placed under the control of the promoter of the gene of interest by engineering its coding sequence onto the DNA at an appropriate place. When the gene is expressed, transcription and subsequent translation leads to the formation of new reporter protein molecules. Whether luminescent or fluorescent, reporter protein molecules can be quantified at any time by measuring light intensity at the relevant wavelength, thus providing a dynamical readout of the activity of the gene. To do so, time-lapse microscopy, flow cytometry, microplate reading, or other experimental techniques are used, depending whether single-cell measurements, population histograms, or population-average profiles are sought. Synthesis of reporter proteins is often completed by a maturation step, that takes immature proteins into their mature, visible form.

2.1 Stochastic gene expression modelling

Gene expression is commonly described in terms of the synthesis and degradation reactions for mRNA and protein molecules

$$\mathscr{R}_1 : \; F \xrightarrow{k_M} F + M \qquad\qquad \mathscr{R}_2 : \; M \xrightarrow{d_M} \emptyset \tag{1}$$

$$\mathscr{R}_3 : \; M \xrightarrow{k_P} M + P \qquad\qquad \mathscr{R}_4 : \; P \xrightarrow{d_P} \emptyset \tag{2}$$

[5, 10] where M and P denote mRNA and protein species, respectively, and F represents the active promoter species. In the context of this paper, P is the fluorescent or luminescent reporter protein. We will not distinguish between immature (invisible) and mature (visible) protein molecules. If necessary (e.g. for slow, stochastic maturation), an additional first-order reaction $P \to P_{mature}$ can be included in the model (along with $P_{mature} \to \emptyset$) to account for protein maturation (and mature protein degradation).

Denote with $X_1 \in \mathbb{N}$ and $X_2 \in \mathbb{N}$ the number of copies of M and P, in the same order, and with $X_3 \in \{0, 1\}$ the state of the promoter, i.e. $X_3 = 0$ when the promoter is inactive (absence of F) and $X_3 = 1$ when it is active (presence

of F). Switching promoter dynamics (responsible of mRNA synthesis bursts in single cells) are formally captured by two additional reactions,

$$\mathscr{R}_5 : \ \emptyset \xrightarrow{\lambda_+ \cdot (1 - X_3)} F, \qquad\qquad \mathscr{R}_6 : \ F \xrightarrow{\lambda_-} \emptyset, \tag{3}$$

representing in the order activation with propensity $\lambda_+ \cdot (1 - X_3)$ (only enabled if $X_3 = 0$), and deactivation with propensity $\lambda_- \cdot X_3$ (only enabled if $X_3 = 1$). Overall, this is a system of $m = 6$ chemical reactions over $n = 3$ different species.

The kinetics of this biochemical reaction system can be expressed in terms of stoichiometry matrix S and reaction rate vector $a(x)$ given by

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad a(x) = \begin{bmatrix} k_M x_3 \\ d_M x_1 \\ k_P x_1 \\ d_P x_2 \\ \lambda_+(1 - x_3) \\ \lambda_- x_3 \end{bmatrix},$$

where, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, $S_{i,j}$ denotes the net change in molecule number of species i when reaction $\mathscr{R}_j$ occurs. At the level of a single cell, $X = [X_1 \ X_2 \ X_3]^T$ is a stochastic process and, for every j, $a_j\big(x\big)$ is interpreted as the infinitesimal probability that reaction $\mathscr{R}_j$ occurs in an infinitesimal time period when $X = x$ molecules of the different species are present in the reaction volume [16]. For constant rates λ_+ and λ_-, Eq.(1)–(3) together constitute the so-called random telegraph model [16]. In general, however, these rates might themselves depend upon the amount of transcription factors regulating the expression of the gene, which one may write as $\lambda_+(X_?)$ and $\lambda_-(X_?)$, with $X_?$ denoting the amount of some unspecified species.

The question we are going to investigate is what can be said about the statistics of F, given mean and variance profiles of the amounts of protein P across a population of cells. In practice, fluorescence or luminescence measurements proportional to the actual amount of protein are measured and are possibly affected by error. In this paper, however, we are not concerned with the details of the measurement model, and assume that mean and variance of X_2 are observed directly.

2.2 Propagation of moments

Consider an arbitrary biochemical reaction system with n reactants, m reactions, stoichiometry matrix S and reaction rates $a(x, u)$ possibly depending on a deterministic input u. Let $X(t)$ be the corresponding random state vector at time t, and define $\mu(t) = \mathbb{E}[X(t)]$ and $\Sigma(t) = \mathrm{Cov}\big(X(t)\big) = \mathbb{E}\big[\big(X(t) - \mu(t)\big)\big(X(t) - \mu(t)\big)^T\big]$. It can be shown (see e.g. [8]) that μ and Σ obey the so-called moment equations

$$\dot{\mu} = S\mathbb{E}[a(X, u)], \tag{4}$$

$$\dot{\Sigma} = S\mathbb{E}[a(X, u)(X - \mu)^T] + \mathbb{E}[(X - \mu)a^T(X, u)]S^T + S\mathrm{diag}(\mathbb{E}[a(X, u)])S^T. \tag{5}$$

Above and in the sequel, time t is omitted from notation where no confusion may arise. If rates are affine in the state, i.e. $a(x,u) = W(u)x + w_0(u)$ for some $W(u)$ and $w_0(u)$, these equations simplify to

$$\dot{\mu} = SW(u)\mu + Sw_0(u), \tag{6}$$

$$\dot{\Sigma} = SW(u)\Sigma + \Sigma W^T(u)S^T + S\mathrm{diag}\big(W(u)\mu + w_0(u)\big)S^T. \tag{7}$$

This system of differential equations is closed in the sense that it does not depend on unmodelled moments. If in addition W does not depend on u, then the system is linear in the input (and the initial conditions).

For the system (1)–(3), Eqs. (6)–(7) apply in the case of constant rates λ_+ and λ_-. In the general case of regulated switching rates $\lambda_+(X_?)$ and $\lambda_-(X_?)$, one may instead interpret (4)–(5) as the moment equations for the augmented state composed of X and $X_?$. Since the laws regulating $X_?$ are unspecified, the full system cannot be spelled out, but one may still work out the equations for the evolution of the moments of X_1, X_2 and X_3. Define

$$\left[z_{MP}^T \mid z_\times^T \mid z_F^T\right] = \left[\mu_M\ \mu_P\ \sigma_{MM}\ \sigma_{PP}\ \sigma_{MP} \mid \sigma_{MF}\ \sigma_{PF} \mid \mu_F\ \sigma_{FF}\right],$$

(vertical bars denoting vector blocks) where of course $\mu_\bullet$ and $\sigma_{\bullet\bullet}$ are the mean and covariance of the states corresponding to the species in subscript (identical subscripts denoting variance). From an engineering viewpoint, z_{MP} is the state of the dynamical sensor for the statistics of F, with sensor output given by the elements $[\mu_P\ \sigma_{PP}]^T$ of z_{MP}. Then one gets

$$\dot{z}_{MP} = A_{MP} \cdot z_{MP} + A_{MP,\times} \cdot z_\times + A_{MP,F} \cdot z_F, \tag{8}$$

$$\dot{z}_\times = A_\otimes \cdot z_\times + z_\otimes + A_{\times,F} \cdot z_F. \tag{9}$$

for matrices A_{MP}, $A_{MP,\times}$, $A_{MP,F}$, $A_\otimes$ and $A_{\times,F}$ depending solely on $\theta_{MP} = (k_M, d_M, k_P, d_P)$ (see Appendix A), i.e. the parameters of the sensing system. Note that, for every fixed t, $F(t)$ is a Bernoulli random variable. Then $\sigma_{FF}(t) = \mu_F(t)\big(1 - \mu_F(t)\big)$ for all t (as a consequence, (8)–(9) are somewhat redundant).

From (8)–(9) one observes that mean and variance of X_2, the observed elements of z_{MP}, are thus a dynamical transformation of those of F, i.e. z_F, plus a contribution from

$$z_\otimes = \mathrm{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} \lambda_+(X_?)(1 - X_3) \\ \lambda_-(X_?)X_3 \end{bmatrix}\right) \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

As it will become clear, $z_\otimes$ implicitly brings about a contribution from the correlation structure of F (see later Remark 1).

2.3 Marginalization of moments

From now on, abusing notation in favor of simplicity, we will refer to X_1, X_2 and X_3 by the symbols for the corresponding species, i.e. M, P and F, in the

same order. Let f be any possible outcome of F, and let

$$\mu_P(t) = \mathbb{E}[P(t)], \qquad \mu_P^f(t) = \mathbb{E}[P(t)|F = f],$$
$$\mathscr{M}_P(t) = \mathbb{E}[P(t)^2], \qquad \mathscr{M}_P^f(t) = \mathbb{E}[P(t)^2|F = f],$$

where, unlike the approach in [7], conditioning is intended over the whole history of F. By marginalization,

$$\mu_P = \mathbb{E}\big[\mathbb{E}[P|F]\big] = \int \mu_P^f d\mathscr{P}_F(f), \quad \mathscr{M}_P = \mathbb{E}\big[\mathbb{E}[P^2|F]\big] = \int \mathscr{M}_Y^f d\mathscr{P}_F(f), \tag{10}$$

with $\mathscr{P}_F$ the probability distribution of F over all possible binary switching sequences. Let us now state the following assuption.

Assumption 1 (Granger causality [12]) *There is no feedback from M and P to F, i.e., at any time t, the future of F is conditionally independent on the past of M and P given the past of F.*

This captures the idea that species M and P do not participate in the regulation of the promoter [3, 1], and corresponds well to all the reporter systems where reporter and regulatory proteins are physically different molecules. In the light of Assumption 1, the conditional moments μ_P^f and $\mathscr{M}_P^f$ are those of the reduced system (1)–(2) with f defining the state of species F at all times. Let

$$z_{MP}^f = \left[\mu_M^f \; \mu_P^f \; \sigma_{MM}^f \; \sigma_{PP}^f \; \sigma_{MP}^f\right]$$

be the vector of conditional moments of M and P. Working out the moment equations (6)–(7) for $X = \begin{bmatrix} M & P \end{bmatrix}^T$ and input $u = f$, one gets that

$$\dot{z}_{MP}^f = A_{MP} \cdot z_{MP}^f + (A_{MP,F})_1 \cdot f, \tag{11}$$

where $(A_{MP,F})_1$ denotes the first column of $A_{MP,F}$. Then μ_P^f and σ_{PP}^f follow from the solution of this system and $\mathscr{M}_P^f = \sigma_{PP}^f + (\mu_P^f)^2$, while marginalization (10) completes the computation of μ_P and $\mathscr{M}_P$. Note that, because of the relationship $\mathscr{M}_P = \sigma_{PP} + (\mu_P)^2$, we can equivalently consider (μ_P, σ_{PP}) or $(\mu_P, \mathscr{M}_P)$ to be the observed output quantities. We will often exploit this equivalence in the sequel without further notice.

Incidentally, notice that (11) represents a linear switching system with two alternating operational modes, $f = 0$ and $f = 1$.

3 The fixed rate promoter process

In order to investigate how statistics of F reflect into the observed profiles μ_P and $\mathscr{M}_P$, and how they may possibly be reconstructed from the output, we first focus on the fundamental case where switching rates λ_+ and λ_- are constant. Define $\alpha = \lambda_+ + \lambda_-$.

Proposition 1. *Mean* $\mu_F(t) = \mathbb{E}[F(t)]$ *and autocovariance function* $\rho_F(t,s) = cov\big(F(t), F(s)\big)$ *obey the equations*

$$\mu_F(t) = \mu_F(0)e^{-\alpha t} + \frac{\lambda_+}{\alpha}\left(1 - e^{-\alpha t}\right), \qquad t \geq 0, \quad (12)$$

$$\rho_F(t,\tau) = \left(\frac{\lambda_+}{\alpha} + \frac{(\alpha - \lambda_+)}{\alpha}e^{-\alpha(t-\tau)}\right) \cdot \mu_F(\tau) - \mu_F(t) \cdot \mu_F(\tau), \quad t \geq \tau. \quad (13)$$

In stationary conditions, with an abuse of notation for the arguments of ρ_F*,*

$$\mu_F = \frac{\lambda_+}{\alpha}, \qquad \rho_F(t-\tau) = \frac{\lambda_+(\alpha - \lambda_+)}{\alpha^2} \cdot e^{-\alpha(t-\tau)}. \quad (14)$$

Incidentally, the autocovariance in (14) is the same as that of an Ornstein-Uhlenbeck process [15].

It can be appreciated that, in transient conditions, the mean profile μ_F contains all the information about the statistics of F. Indeed, in this simple case, rates λ_+ and λ_- (or equivalently α), together with the initial condition $\mu_F(0)$, fully determine the laws of F. In turn, these three quantities have distinct effects on μ_F, i.e. they are distinguishable from a transient mean profile. In [2], it was shown that these and other model parameters, notably θ_{MP}, are also jointly distinguishable from the measured profiles μ_P and $\mathscr{M}_P$. The result is based on the specialization of (8)–(9) for the case of the fixed rate process, given by [2]

$$\begin{bmatrix} \dot{z}_{MP} \\ \dot{z}_\times \\ \dot{z}_F \end{bmatrix} = \begin{bmatrix} A_{MP} & A_{MP,\times} & A_{MP,F} \\ 0 & A_\times & A_{\times,F} \\ 0 & 0 & A_F \end{bmatrix} \cdot \begin{bmatrix} z_{MP} \\ z_\times \\ z_F \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ u_F \end{bmatrix}, \quad (15)$$

where $A_\times = -\alpha I + A_\otimes$, $u_F = \begin{bmatrix}\lambda_+ & \lambda_+\end{bmatrix}^T$ and A_F depends only on λ_+ and α as detailed in Appendix A. For known parameters θ_{MP}, we may easily show that λ_-, λ_+ and $\mu_F(0)$ are also distinguishable from the sole mean μ_P. For simplicity, we consider the case where M and P are identically 0 at time 0. By inspection of (15),

$$\begin{aligned} \dot{\mu}_F &= -\alpha\mu_F + \lambda_+, \\ \dot{\mu}_M &= -d_M\mu_M + k_M\mu_F, \\ \dot{\mu}_P &= -d_P\mu_P + k_P\mu_M \end{aligned} \quad (16)$$

(the expression of $\dot{\mu}_F$ above coincides with the differential form of (12)). Thus, in terms of Laplace transform,

$$\mu_P(s) = \frac{\lambda_+ k_M k_P}{s(\alpha + s)(d_M + s)(d_P + s)} + \frac{\mu_F(0) k_M k_P}{(\alpha + s)(d_M + s)(d_P + s)},$$

and one may apply the method of [2] (also reported in Appendix B) to prove sensitivity of this solution (equivalently, the solution over time) to any change in the three unknown parameters, almost everywhere in the space of the remaining parameters. In practical terms, parameter values can be reconstructed either

from μ_F as obtained by deconvolution from μ_P, or by direct fit of (16) to an observed μ_P profile.

Now assume that F has reached stationarity. In this case, all relevant statistics of F are determined by λ_+ and λ_-. However, from Proposition 1, mean μ_F only conveys information about the ratio λ_+/α, and, because $\sigma_F^2 = \mu_F(1-\mu_F)$ at any point in time, no more information is contained in the variance. Specific contributions of the two parameters can instead be traced in the autocovariance function ρ_F. Indeed, multiplicative factor $\lambda_+(\alpha-\lambda_+)/\alpha^2$ and decay rate α have distinguishable effects on ρ_F (different choices of the two lead to different profiles $\rho_F(\cdot)$) and uniquely determine λ_+ and λ_-. The question arises whether ρ_F is observable from the measured profiles μ_P and $\mathscr{M}_P$ (i.e. whether λ_+ and λ_- are also distinguishable from the experimental output). In this section we provide a positive answer in terms of identifiability of λ_+ and λ_-, i.e. for processes F with fixed rates. A more general answer will be provided in the next section.

From Proposition 1, stationary conditions are achieved when μ_F is in steady state (i.e. when the factors of $\rho_F(t,\tau)$ involving μ_F no longer depend on τ). It then suffices to check identifiability of λ_+ and λ_- from the solution of (15) with stationary initial conditions $\mu_F(0) = \lambda_+/\alpha$ and $\sigma_F^2(0) = \lambda_+/\alpha(1-\lambda_+/\alpha)$. Using again the method of [2], one computes the Laplace response function of this system. The resulting equations are lengthy and not reported here. Then, it can be checked that the Laplace sensitivity condition also reported in Appendix B is verified, i.e. the time profiles of μ_P and $\mathscr{M}_P$ are sensitive to all possible changes of λ_+ and α, almost everywhere in the space of the parameters θ_{MP}.

Example 1. Refer to Figure 1. Statistics for two fixed-rate promoter activity

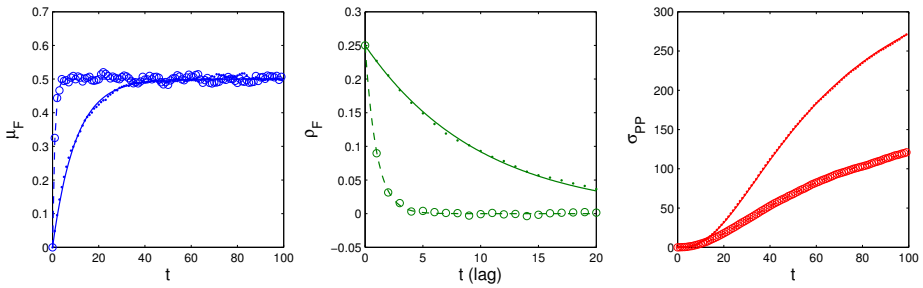

Fig. 1. Statistics for F (dashed lines and circles) and F' (solid lines and dots). Lines visualize analytic solutions, markers are for empirical statistics from Gillespie simulations. Gillespie simulations are performed using Stochkit [17] for the generation of 10^4 sample paths (i.e. simulated cells). Numerical calculations are performed in Matlab.

processes, F and F', are considered. F has $\lambda_+ = \lambda_- = 0.05$, while F' has the faster switching dynamics $\lambda_+ = \lambda_- = 0.5$. Starting from the non-stationary conditions $F = F' = 0$ at time 0, means μ_F and $\mu_{F'}$ converge both at 0.5 at different rates (Figure 1, left), thus resulting into different output profiles

μ_P (not shown). In other words, the two processes are distinguishable from the mean. In stationary conditions, instead, the means for F and F' are the same. Yet the stationary autocovariance functions ρ_F and $\rho_{F'}$ differ in the two cases (Figure 1, center). This results in different observed profiles of σ_{PP} (Figure 1, right). In other words, in stationary conditions, F and F' are distinguishable from the output variance.

Remark 1. Equations (15) are obtained from (8)–(9) by developing the expression of $z_{\otimes}$. This results in expressions depending on matrices $A_{\times}$ and A_F, which bring in the role of α, the decay rate of ρ_F, into the propagation of second-order moments from z_F to z_{MP}. This fact is indeed in agreement with the discussion of $z_{\otimes}$ at the end of Section 2.2.

To summarize, we have shown that constant switching rates, whence all statistics, of a promoter activity process F can be reconstructed from the output mean if F is not in stationary conditions. In stationary conditions, the promoter statistics cannot be determined from the output mean, but rather from the output variance since this reflects differences in the autocovariance function of F. Analytic expressions and a case study have been developed to support our arguments.

4 General promoter switching processes

We now wish to study how first- and second-order moments of switching process F reflect into outputs μ_P and $\mathscr{M}_P$, and how to possibly reconstruct the former from the latter, without a priori knowledge on F. In particular, we do not assume that switching rates λ_+ and λ_- are fixed. We only assume that F has continuous (mean and) autocovariance $\rho_F(t,s)$, and that Assumption 1 holds. For simplicity, we focus on the case where $z_{MP}(0)$ is null (M and P equal to zero at time zero).

From Equation (11), for some final time $T > 0$, the conditional moments $\mu_P^f(t)$ and $\sigma_{PP}^f(t)$ over $[0,T)$ are the output of a linear dynamical system with (zero initial conditions and) input f. We may then introduce linear operators, L_1 and L_2, and abstract the transformation from function f to μ_P^f and σ_{PP}^f as $\mu_P^f = L_1 f$ and $\sigma_{PP}^f = L_2 f$. When necessary, for any $t \in [0,T)$, we will write $\mu_P^f(t) = (L_1 f)(t)$ as $L_1^t f$ and $\sigma_{PP}^f(t) = (L_2 f)(t)$ as $L_2^t f$. Of course, for $k = 1$ and $k = 2$,

$$L_k^t f = \int_0^t d\tau\, \ell_k(t,\tau) f(\tau), \qquad \ell_k(t,\tau) = C_k e^{A_{MP}(t-\tau)} (A_{MP,F})_1,$$

with $C_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}$ (mean readout) and $C_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$ (variance readout).

4.1 Observability and reconstruction of the process mean

From the first equality in (10), one has that

$$\mu_P = \int (L_1 f) d\mathscr{P}_F(f) = L_1 \left(\int f d\mathscr{P}_F(f) \right) = L_1 \mu_F.$$

Not surprisingly at this point, μ_P thus follows from the linear dynamical transformation of μ_F already found in (8). Observability of μ_F from μ_P essentially depends on the spectrum of L_1. Since

$$\mu_P(s) = \frac{k_M k_P}{(d_M + s)(d_P + s)} \mu_F(s),$$

for strictly positive parameters θ_{MP}, the transformation is invertible over the whole spectrum, i.e μ_F can be perfectly reconstructed from μ_P. In practice, this amounts to a deconvolution problem of rather easy solution [2].

4.2 Observability and reconstruction of the process covariance

We begin with the following result.

Proposition 2. *For any time $t \in [0, T)$, it holds that*

$$\mathscr{M}_P(t) = L_2^t \mu_F + \mathbb{E}[(L_1^t F)^2]. \tag{17}$$

Clearly the autocovariance function of F plays a role in the term $\mathbb{E}[(L_1^t F)^2]$. To study this term further, consider the Karhunen-Loève decomposition [15] of process F, given by

$$F - \mu_F = \sum_{i=1}^{\infty} a_i \phi_i,$$

where the ϕ_i are the mutually orthogonal, unit norm eigenfunctions of the operator $K : \phi \mapsto \int d\tau \rho_F(\cdot, \tau)\phi(\tau)$, i.e. $K\phi_i = \sigma_i^2 \phi_i$, and the a_i are mutually uncorrelated, zero-mean random variables with variance equal to the eigenvalues σ_i^2 (function norm is in L^2 and the decomposition holds in the mean-square sense). Then

$$\rho_F(t, \tau) = \sum_{i=1}^{\infty} \sigma_i^2 \phi_i(t)\phi_i(\tau), \qquad \sigma_{FF}(t) = \sum_{i=1}^{\infty} \sigma_i^2 \phi_i^2(t).$$

Proposition 3. *It holds that $\mathbb{E}[(L_1^t F)^2] = (L_1^t \mu_F)^2 + \sum_{i=1}^{\infty} \sigma_i^2 (L_1^t \phi_i)^2$.*

In sums, from Propositions 2–3 and using the fact that $(\mu_P)^2 = (L_1 \mu_F)^2$,

$$\sigma_{PP}(t) = \mathscr{M}_P(t) - \mu_P^2(t) = L_2^t \mu_F + \sum_{i=1}^{\infty} \sigma_i^2 (L_1^t \phi_i)^2. \tag{18}$$

Comparing the expressions of σ_{PP} and σ_{FF} one notices that, besides term $L_2^t \mu_F$, the functions composing F and characterizing its autocovariance structure are transformed by L_1^t into contributions that make up the variance of P at time t. Were L_1^t an evaluation operator, i.e. $L_1^t \phi_i = \phi_i(t)$, then $\sigma_{PP}(t)$ would degenerate to $L_2^t \mu_F + \sigma_{FF}(t)$, i.e. information about the autocovariance structure of F would be lost. For every t, it is the integral nature of L_1^t that channels information about the whole $\rho_F(\cdot, \cdot)$ into $\sigma_{PP}(t)$. Another viewpoint on this is given in what follows.

Equation (18) explains the nature of the information transfer from ρ_F to σ_{PP}. For reconstruction purposes, however, we seek a more explicit relationship between σ_{PP} and ρ_F. The following result relies on the convolutional form of L_1.

Proposition 4. *It holds that*

$$\sigma_{PP}(t) = L_2^t \mu_F + \iint d\tau\, dv\, \ell_1(t,\tau)\ell_1(t,v)\rho_F(\tau,v). \tag{19}$$

Hence ρ_F undergoes itself a linear transformation H defined by

$$H^t\rho = (H\rho)(t) = \int_0^t d\tau \int_0^t dv\, \ell_1(t,\tau)\ell_1(t,v)\rho(\tau,v).$$

In particular, suppose that F is stationary. Then, by a change of variables,

$$H^t\rho_F = \int_0^t d\tau \int_0^t dv\, \ell_1(t,\tau)\ell_1(t,v)\rho_F(\tau - v) = \int_{-t}^t d\delta\, h(t,\delta)\rho_F(\delta)$$

with

$$h(t,\delta) = \int_{\max\{-\delta,0\}}^{\min\{t,t-\delta\}} dv\, \ell_1(t,v+\delta)\ell_1(t,v).$$

In the light of these results, the problem of the observability of ρ_F, or better the joint observability of ρ_F and μ_F from μ_P and σ_{PP}, is thus equivalent to that of the invertibility of the linear operator

$$(\mu_F, \rho_F) \mapsto (L_1\mu_F, L_2\mu_F + H\rho_F) \tag{20}$$

(with relevant simplifications if stationarity of F is hypothesized). We note that, besides term $L_2\mu_F$, the relationship between ρ_F and σ_{PP} is analogous to that pertaining linear transformations of second-order processes. In particular, using the fact that $\ell_1(t,\cdot)$ is the impulse response of a time-invariant dynamical system, the second term of (19) can be seen as the autocovariance of the output of a linear filter with response ℓ_1 fed with an input process with autocovariance ρ_F. It is then natural to frame observability analysis of ρ_F in the context of spectral analysis [11, 12]. This analysis is left for future work. Here we limit ourselves to the discussion of an illustrative example.

Example 2. Refer to Figure 2. We consider a promoter activity process F with randomly regulated rates, and compare its statistics with those of relevant fixed-rate processes F' and F''. All processes are analyzed in stationary regime and have rate λ_- identically set to 0.5, i.e. their definition only differs in the activation rate. The activation rate of F is $\lambda_+(R) = 1 \cdot R$. Regulator R is another random binary process with switch-off rate equal to 0.1 and switch-on rate, equal to 0.2217, chosen so as to guarantee that the stationary mean of F is $\mu_F = 0.5$. Process F' is defined as in Example 1, i.e. it has $\lambda_+ = 0.5$, again resulting in $\mu_{F'} = 0.5$. Finally, process F'' has activation rate $\lambda_+ = \mathbb{E}[\lambda_+(R)] = 0.6892$, i.e.

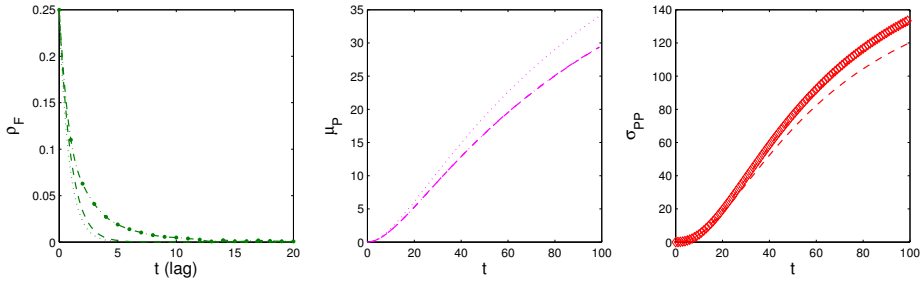

Fig. 2. Statistics for a random-rate promoter process F (dash-dotted lines) and relevant fixed-rate promoter processes F' (same as in Figure 1, dashed lines) and F'' (dotted lines). Left: autocovariance functions ρ_F (dots: estimates from Gillespie simulations; line: interpolation), $\rho_{F'}$ (from (14)) and $\rho_{F''}$ (from (14)); Center: Observed output mean μ_P for F (Gillespie simulation), F' (solution of (15)) and F'' (solution of (15)) – curves for F and F' are superimposed; Right: The observed output variance of P for F (diamonds: numerical computation of (19), based on the profile of ρ_F from Gillespie simulations; line: estimate from Gillespie simulation – diamonds and line are superimposed), F' (solution of (15)) and F'' (solution of (15)) – curves for F and F'' are superimposed. Gillespie simulations are performed using Stochkit [17] for the generation of 10^5 sample paths (i.e. simulated cells). Numerical calculations are performed in Matlab.

a switch-on rate equivalent on average to that of F. This results in a different mean, $\mu_{F''} = 0.5795$.

The autocovariance function of F (Figure 2, left) is markedly different from those of F' and F'', which are similar. Because $\mu_F = \mu_{F'} \neq \mu_{F''}$, F can be distinguished from F'', but not from F', from the output mean μ_P (Figure 2, center). However, because of the different autocovariance function, F can be distinguished from F' from the output variance σ_{PP}. Interestingly, the output variance profiles for F and F'' are quite similar, a sign that the differences between F and F'' in mean and autocovariance compensate each other in this case. This is possible since output variance depends not only on the autocovariance but also on the mean of the promoter activity process.

Finally, in the light of the linearity of (20), joint estimation of μ_F and ρ_F from possibly noisy and sampled measurements of μ_P and σ_{PP} can be seen as a linear inversion problem. Regularized solutions for both the stationary and the nonstationary case may be developed in accordance with the vast literature on the subject (see e.g. [4] and references therein). Note that, because μ_F can be reconstructed from the sole mean μ_P, the problem may also be reduced to that of estimating ρ_F from $\sigma_{PP} - L_2^t \mu_F$ via (regularized) inversion of H.

In summary, we have analyzed the relationship between second-order promoter activity statistics and mean and variance profiles of the reporter protein P in the case of promoter processes with randomly regulated rates. In particular, we have developed explicit relationships between the autocovariance function of

F and the readout variance profile of P, showing that this integral relationship is essentially linear. By this we provided the basis for a full spectral analysis of observability of ρ_F and its linear reconstruction from reporter protein mean and variance statistics. We also illustrated the relevance of our results by investigating the distinguishability of a random-rate and relevant fixed-rate processes on an example.

5 Conclusions

We have studied the relationships between second-order statistics of random promoter activity and the mean and variance profiles of gene expression reporter proteins typically observed in biological experiments. For both fixed and randomly regulated (thus also possibly time-varying) switching rates, we developed explicit mathematical formulas showing that these relationships are linear, and provided first results about the observability of the promoter process statistics from gene reporter data. Based on analytic considerations as well as on example case studies, we showed when and how analysis of second-order moments allows for discrimination of different promoter activation statistics.

This work provides the basis for an extensive observability analysis of promoter processes from gene reporter data at a single-cell level, and the development of promoter statistics reconstruction algorithms that are fully non-parametric, i.e. independent of a priori knowledge about the promoter activity laws. Our results show that both observability and estimation can be framed in the well-studied context of linear operators. Subsequent research work will henceforth focus on the application of the relevant spectral analysis and regularized linear inversion techniques. On these bases, we will then address a key challenge of this research effort, namely the identification and discrimination among alternative promoter activity regulatory mechanisms on the basis of the reconstructed promoter activation statistics and data from candidate regulators.

A Definitions and proofs

Matrix definitions. A_{MP} $A_{MP,\times}$ $A_{MP,F}$ are given by

$$\begin{bmatrix} -d_M & 0 & 0 & 0 & 0 \\ k_P & -d_P & 0 & 0 & 0 \\ d_M & 0 & -2\,d_M & 0 & 0 \\ k_P & d_P & 0 & -2\,d_P & 2\,k_P \\ 0 & 0 & k_P & 0 & -d_M - d_P \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 2\,k_M & 0 \\ 0 & 0 \\ 0 & k_M \end{bmatrix}, \quad \begin{bmatrix} k_M & 0 \\ 0 & 0 \\ k_M & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

in the same order, while

$$A_{\otimes} = \begin{bmatrix} -d_M & 0 \\ k_P & -d_P \end{bmatrix}, \qquad A_{\times,F} = \begin{bmatrix} 0 & k_M \\ 0 & 0 \end{bmatrix}, \qquad A_F = \begin{bmatrix} -\alpha & 0 \\ \alpha - 2\lambda_+ & -2\alpha \end{bmatrix}.$$

Proof of Proposition 1. Process F is a homogeneous continuous-time binary Markov chain. Letting $p(t) = \big[\text{Prob}\{F(t)=0\}\ \text{Prob}\{F(t)=1\}\big]^T$, for any t and τ it holds that

$$p(t) = e^{Q(t-\tau)}p(\tau), \qquad\qquad Q = \begin{bmatrix} -\lambda_+ & \lambda_- \\ \lambda_+ & -\lambda_- \end{bmatrix}.$$

Mean $\mu_F = \text{Prob}\{F(t)=1\}$. Using the fact that $\dot{p} = Qp$, the differential equation for μ_F, the second element of p, is $\dot{\mu}_F = \lambda_+(1-\mu_F) - \lambda_-\mu_F = -\alpha\mu_F + \lambda_+$. The solution of this equation relative to $\mu_F(0)$ yields the expression in the statement. Covariance $\rho_F(t,\tau) = \text{Prob}\{F(t)=1, F(\tau)=1\} - \mu_F(t)\mu_F(\tau)$. By Bayes'law, $\text{Prob}\{F(t)=1, F(\tau)=1\} = \text{Prob}\{F(t)=1|F(\tau)=1\}\cdot\text{Prob}\{F(\tau)=1\}$. Second factor is equal to $\mu_F(\tau)$, while the first factor is given by the entry of row 2 and column 1 of $e^{Q(t-\tau)}$. Computing the matrix exponential thus yields the result. Stationary versions of μ_F and ρ_F are found simply by taking the limit of $\mu_F(t)$ as $t \to +\infty$ and replacing the result for $\mu_F(\tau)$ and $\mu_F(t)$ in the expression of $\rho_F(t,\tau)$.

Proof of Proposition 2. Starting from the second relation in (10),

$$\begin{aligned}
\mathscr{M}_P(t) = \mathbb{E}\big[\mathbb{E}[P^2|F]\big] &= \mathbb{E}\Big[\mathbb{E}\big[\big((P-\mathbb{E}[P|F]) + \mathbb{E}[P|F]\big)^2|F\big]\Big] = \\
&= \mathbb{E}\Big[\mathbb{E}\big[(P-\mathbb{E}[P|F])^2|F\big]\Big] + \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[P|F]^2|F\big]\Big] + \\
&\quad + 2\cdot\mathbb{E}\Big[\mathbb{E}\big[(P-\mathbb{E}[P|F])\cdot\mathbb{E}[P|F]|F\big]\Big] = \\
&= \mathbb{E}\Big[\mathbb{E}\big[(P-\mathbb{E}[P|F])^2|F\big]\Big] + \mathbb{E}\Big[\mathbb{E}[P|F]^2\Big] + \\
&\quad + 2\cdot\mathbb{E}\Big[\mathbb{E}\big[P-\mathbb{E}[P|F]|F\big]\cdot\mathbb{E}[P|F]\Big],
\end{aligned}$$

where the last row vanishes since $\mathbb{E}\big[P-\mathbb{E}[P|F]|F\big] = 0$. Then, using the definitions of μ_P^F and σ_{PP}^F, the chain of equalities continues with

$$= \mathbb{E}\big[\sigma_{PP}^F(t)\big] + \mathbb{E}\big[\big(\mu_P^F(t)\big)^2\big] = \mathbb{E}\big[L_2^t F\big] + \mathbb{E}\big[(L_1^t F)^2\big] = L_2^t\mu_F + \mathbb{E}[(L_1^t F)^2].$$

Proof of Proposition 3. The following chain of inequalities hold:

$$\begin{aligned}
\mathbb{E}[(L_1^t F)^2] = \mathbb{E}\bigg[\Big(L_1^t\Big(\sum_i a_i\phi_i\Big)\Big)^2\bigg] &= \mathbb{E}\bigg[\sum_{i,j} L_1^t(a_i\phi_i)L_1^t(a_j\phi_j)\bigg] = \\
&= \sum_{i,j}\mathbb{E}[a_ia_j](L_1^t\phi_i)(L_1^t\phi_j) = \sum_i \sigma_i^2(L_1^t\phi_i)^2, \quad (21)
\end{aligned}$$

where the latter equality follows from the mutual uncorrelation of the a_i.

Proof of Proposition 4. Expanding the last term of (17) one gets

$$\begin{aligned}\mathbb{E}[(L_1^t F)^2] &= \int d\mathscr{P}_F(f)(L_1^t f)^2 \\ &= \int d\mathscr{P}_F(f)\left(\int_0^t d\tau\, \ell_1(t,\tau) f(\tau)\right)\left(\int_0^t dv\, \ell_1(t,v) f(v)\right) \\ &= \int_0^t d\tau \int_0^t dv\, \ell_1(t,\tau)\ell_1(t,v)\left(\int d\mathscr{P}_F(f) f(\tau) f(v)\right) \\ &= \int_0^t d\tau \int_0^t dv\, \ell_1(t,\tau)\ell_1(t,v)\big(\rho_F(\tau,v) + \mu_F(\tau)\mu_F(v)\big)\end{aligned}$$

where the last integrand is of course the autocorrelation of F at τ and v. Therefore

$$\begin{aligned}\sigma_{PP}(t) &= \mathscr{M}_P(t) - \mu_P^2(t) \\ &= L_2^t \mu_F + \int_0^t d\tau \int_0^t dv\, \ell_1(t,\tau)\ell_1(t,v)\big(\rho_F(\tau,v) + \mu_F(\tau)\mu_F(v)\big) - \\ &\quad \left(\int_0^t d\tau\, \ell_1(t,\tau)\mu_F(\tau)\right)\left(\int_0^t dv\, \ell_1(t,v)\mu_F(v)\right),\end{aligned}$$

and the result follows by collecting integrals and simplifying.

B Laplace sensitivity method for the analysis of parameter identifiability

This section reports the identifiability analysis method of [2]. Let $\mathscr{Y}_\theta(t)$ be a vector function of $t \in \mathbb{R}$ depending on parameters θ. Typically $\mathscr{Y}_\theta(\cdot)$ is an observed response of a dynamical system defined in terms of θ.

Definition 1. *The parametric family (of functions)* $\{\mathscr{Y}_\theta : \theta \in \Theta\}$*, with* $\Theta \subseteq \mathbb{R}^N$*,* $N \in \mathbb{N}$*, is*

(a) locally identifiable at θ^* *if a neighborhood* $B_{\theta^*} \subseteq \Theta$ *of* θ^* *exists such that the implication holds* $\forall \theta \in B_{\theta^*}$*;*
(b) locally identifiable *if (a) holds for almost every (a.e.)* $\theta^* \in \Theta$*.*

For any given θ let $Y(s,\theta)$ be the Laplace transform of $\mathscr{Y}_\theta(\cdot)$. Let $\nabla Y(s,\theta) = \frac{\partial Y}{\partial \theta}(s,\theta) = \left[\frac{\partial Y}{\partial \theta_1} \;\cdots\; \frac{\partial Y}{\partial \theta_N}\right](s,\theta)$.

Proposition 5. *If, for some* $L \in \mathbb{N}$*, a set of points* $\mathscr{S}_L = \{s_1, \ldots, s_L\} \subseteq \mathbb{R}$ *(or* $\mathbb{C}$*) exists such that the matrix*

$$\Delta(\mathscr{S}_L, \theta^*) = \left[\nabla Y(s_1,\theta^*)^T \cdots \nabla Y(s_L,\theta^*)^T\right]^T$$

has full column rank, then $\{\mathscr{Y}_\theta : \theta \in \Theta\}$ *is locally identifiable at* θ^* *(in the sense of Definition 1(a)).*

Now assume that the elements of $Y(s,\theta)$ are ratios of polynomials in the entries of θ.

Corollary 1. *If, for a given set of points $\mathscr{S}_L$ and a given θ^*, matrix $\Delta(\mathscr{S}_L,\theta^*)$ is full column rank, then $\{\mathscr{Y}_\theta : \theta \in \Theta\}$ is locally identifiable (a.e. in the sense of Definition 1(b)).*

In the present paper, the Laplace transforms that are used to discuss identifiability belong to this last class (see [2]), whence Corollary 1 applies. In practice, these conditions can be easily checked by the use of the Matlab Symbolic Math Toolbox and evaluation of the rank conditions based on a finite set of heuristically chosen points $\mathscr{S}_L$ (see again [2]).

References

1. Bowsher, C.G., Voliotis, M., Swain, P.S.: The fidelity of dynamic signaling by noisy biomolecular networks. PLoS Comput Biol 9(3), e1002965 (2013)
2. Cinquemani, E.: Reconstruction of promoter activity statistics from reporter protein population snapshot data. In: 2015 54th IEEE Conference on Decision and Control (CDC). pp. 1471–1476 (Dec 2015)
3. Cinquemani, E.: Hybrid Systems Biology: Fourth International Workshop, HSB 2015, Madrid, Spain, September 4-5, 2015. Revised Selected Papers, chap. Reconstructing Statistics of Promoter Switching from Reporter Protein Population Snapshot Data, pp. 3–19. Springer International Publishing, Cham (2015)
4. De Nicolao, G., Sparacino, G., Cobelli, C.: Nonparametric input estimation in physiological systems: Problems, methods, and case studies. Automatica 33(5), 851 – 870 (1997)
5. Friedman, N., Cai, L., Xie, X.S.: Linking stochastic dynamics to population distribution: An analytical framework of gene expression. Phys. Rev. Lett. 97, 168302 (Oct 2006)
6. Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., Allgower, F.: Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics 12(1), 125 (2011)
7. Hasenauer, J., Wolf, V., Kazeroonian, A., Theis, F.J.: Method of conditional moments (MCM) for the Chemical Master Equation. Journal of Mathematical Biology 69(3), 687–735 (2014)
8. Hespanha, J.: Modelling and analysis of stochastic hybrid systems. Control Theory and Applications, IEE Proceedings 153(5), 520–535 (Sept 2006)
9. de Jong, H., Ranquet, C., Ropers, D., Pinel, C., Geiselmann, J.: Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. BMC Syst. Biol. 4(1), 55 (2010)
10. Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J.: Stochasticity in gene expression: From theories to phenotypes. Nat. Rev. Gen. 6, 451–464 (2005)
11. Koopmans, L.H.: The Spectral Analysis of Time Series. Probability and Mathematical Statistics, Academic Press, San Diego (1995)
12. Lindquist, A., Picci, G.: Linear stochastic systems – A geometric approach to modeling, estimation and identification. Springer Verlag, Berlin Heidelberg (2015)
13. Munsky, B., Trinh, B., Khammash, M.: Listening to the noise: Random fluctuations reveal gene network parameters. Mol. Syst. Biol. 5(318) (2009)

14. Neuert, G., Munsky, B., Tan, R., Teytelman, L., Khammash, M., van Oudenaarden, A.: Systematic identification of signal-activated stochastic gene regulation. Science 339(6119), 584–587 (2013)
15. Papoulis, A.: Probability, random variables, and stochastic processes. McGraw-Hill series in electrical engineering, McGraw-Hill, New York (1991)
16. Paulsson, J.: Models of stochastic gene expression. Phys. Life Rev. 2(2), 157 – 175 (2005)
17. Sanft, K.R., Wu, S., Roh, M., Fu, J., Lim, R.K., Petzold, L.R.: Stochkit2: Software for discrete stochastic simulation of biochemical systems with events. Bioinformatics 27(17), 2457–2458 (2011)
18. Stefan, D., Pinel, C., Pinhal, S., Cinquemani, E., Geiselmann, J., de Jong, H.: Inference of quantitative models of bacterial promoters from time-series reporter gene data. PLoS Comput. Biol. 11(1), e1004028 (01 2015)
19. Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., Koeppl, H.: Moment-based inference predicts bimodality in transient gene expression. PNAS 21(109), 8340–8345 (2012)
20. Zechner, C., Unger, M., Pelet, S., Peter, M., Koeppl, H.: Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nature Methods 11, 197–202 (2014)
21. Zulkower, V., Page, M., Ropers, D., Geiselmann, J., de Jong, H.: Robust reconstruction of gene expression profiles from reporter gene data using linear inversion. Bioinformatics 31(12), i71–i79 (2015)