Parameter inference with analytical propagators for stochastic models of autoregulated gene expression

Frits Veerman[1], Nikola Popović[2,*], and Carsten Marr[3]

[1]Universität Heidelberg, Institute of Applied Mathematics, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[2]University of Edinburgh, School of Mathematics, James Clerk Maxwell Building, King's Buildings, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom
[3]Helmholtz Zentrum München, Institute of Computational Biology, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
[*]Nikola.Popovic@ed.ac.uk

April 10, 2021

Abstract

Stochastic gene expression in regulatory networks is conventionally modelled via the Chemical Master Equation (CME). As explicit solutions to the CME, in the form of so-called propagators, are oftentimes not readily available, various approximations have been proposed. A recently developed analytical method is based on a separation of time scales that assumes significant differences in the lifetimes of mRNA and protein in the network, allowing for the efficient approximation of propagators from asymptotic expansions for the corresponding generating functions. Here, we showcase the applicability of that method to simulated data from a 'telegraph' model for gene expression that is extended with an autoregulatory mechanism. We demonstrate that the resulting approximate propagators can be applied successfully for parameter inference in the non-regulated model; moreover, we show that, in the extended autoregulated model, autoactivation or autorepression may be refuted under certain assumptions on the model parameters. These results indicate that our approach may allow for successful parameter inference and model identification from longitudinal single cell data.

Keywords: asymptotic analysis; parameter inference; propagator; stochastic gene expression.

1 Introduction and background

Gene expression in regulatory networks is an inherently stochastic process [8]. Mathematical models typically take the form of a Chemical Master Equation (CME), which describes the temporal evolution of the probabilities of observing specific states in the network [31]. Recent advances in single-cell fluorescence microscopy [5, 11, 16, 29, 34] have allowed for the generation of experimental longitudinal data, whereby the fluorescence intensity of mRNA or protein abundances in single

cells is measured. While most common models assume the availability of protein abundance data, abundances of mRNA may equally be of interest, depending on the model [19]. Here, we focus on abundances of protein, which we assume to be measured at regular sampling intervals Δt. A typical data set, denoted by Q, thus consists of protein abundances n_i at $N+1$ different points in time; see Figure 1A. We can group these abundances into transitions $n_i \to n_{i+1}$; cf. Figure 1B. A model-derived propagator $P_{n_{i+1}|\,n_i}(\Delta t, \Theta)$ allows for the calculation of the probabilities of such transitions for some set of model parameters Θ. Summing over all these probabilities for all N observed transitions, we can calculate the log-likelihood $L(\Theta)$ of that particular parameter set as

$$L(\Theta) = \sum_{i=0}^{N-1} \log\, P_{n_{i+1}|\,n_i}(\Delta t, \Theta), \tag{1.1}$$

which can be evaluated over a range of values for the model parameters to yield a 'log-likelihood landscape', the maximum of which corresponds to the most likely parameter set Θ subject to the measured data set Q. Assessing the log-likelihood of a model by evaluating the associated propagator for the observed transitions in a time-lapse experiment is a feasible, established approach that has been successfully applied previously [28, 29, 9]. Due to the complex nature of the underlying regulatory networks, explicit expressions for $P_{n_{i+1}|\,n_i}$ are difficult to obtain in general. Hence, a variety of approximations have been proposed, which can be either numerical [33, 9] or analytical [26, 23], to cite but a few examples. Here, we apply the analytical method recently developed by the current authors [32], which was based on ideas presented by Popović, Marr & Swain [23], to obtain fully time-dependent approximate propagators; an outline of the method is given in Section 2.

Our aim in the present article is to demonstrate the applicability of these propagators, as well as to evaluate their performance in the context of parameter inference for synthetic data. Specifically, we showcase the resulting inference procedure for a family of stochastic gene expression models. First, in Section 3, we consider a model that incorporates DNA 'on'/'off' states ('telegraph model'); see also the work of Raj *et al.* [24] and Shahrezaei & Swain [28]. Subsequently, in Section 4, that model is extended with an autoregulatory mechanism, whereby protein influences its own production through an autocatalytic reaction. In Section 5, we summarise our results and present an outlook to future research; finally, in Appendix A, we collate the analytical formulae that underly our inference procedure for the family of models showcased here.

2 Method

2.1 Calculation of propagators

Our method [32] is based on an analytical approximation of the probability generating function that is introduced for analysing the CME corresponding to the given gene expression model. Propagators can be calculated from the generating function via the Cauchy integral formula, which implies

$$P_{n_{i+1}|\,n_i}(\Delta t, \Theta) = \frac{1}{2\pi i} \oint_\gamma \frac{F(z; \Delta t, n_i, \Theta)}{z^{n_{i+1}+1}}\, \mathrm{d}z; \tag{2.1}$$

here, $F(z; \Delta t, n_i, \Theta)$ is the generating function of the (complex) variable z, which additionally depends on the sampling interval Δt, the protein abundance n_i, and the model parameter set Θ.

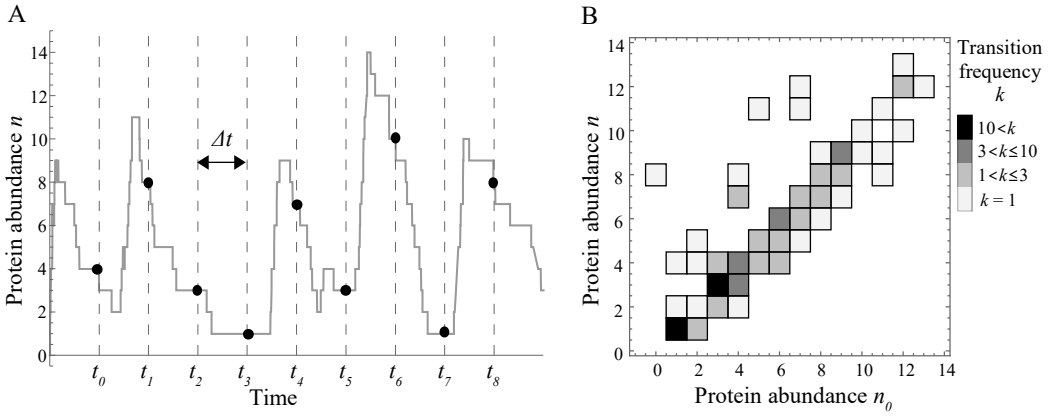

Figure 1: (A) Simulated time series of protein abundance n, with measurements at times t_i and sampling interval Δt. (B) Histogram of the frequency k_j of transitions $(n_0 \to n)_j$, inferred from a longer time series with 100 transitions.

The integration contour γ is a closed contour in the complex plane around $z = 0$. The choice of contour is arbitrary; however, it can have a significant effect on computation times and on the accuracy of the resulting integrals [2]. Here, we choose γ to be a regular 150-sided polygon approximating a circle of radius 0.8 that is centred at the origin in the complex plane, which results in a 'hybrid analytical-numerical' procedure for the evaluation of $P_{n_{i+1}|\,n_i}$.

2.2 Parameter inference

The parameter inference procedure proposed here can be divided into the following steps:

1) Data binning. The simulated data Q is presented as a time series $\{n_i\}$, $0 \leq i \leq N$, which yields N transitions $n_i \to n_{i+1}$. Generically, some of these transitions occur more than once. For computational efficiency, we bin the data accordingly to create a binned data set $\widehat{Q} = \big\{\big(k_j, (n_0 \to n)_j\big)\big\}$, with $1 \leq j \leq \widehat{N}$ for $\widehat{N} \leq N$, where k_j is the frequency of the transition $(n_0 \to n)_j$; see also Figure 1. Here, $\widehat{N}$ denotes the number of different transitions observed in the data, that is, the size of the binned data set $\widehat{Q}$. We emphasise that binning can substantially accelerate parameter inference, in particular for near-stationary processes.

2) Marginalisation. Frequently, some of the involved species in a model are not observed, and hence have to be marginalised over. In the models discussed in Sections 3 and 4, we assume that protein is measured, while mRNA remains unobserved. Marginalisation over unobserved species is usually carried out on the transition probabilities in (2.1). However, since the marginalisation procedure is linear, it commutes with the Cauchy integral. Introducing the linear 'marginalisation operator' $\mathbb{M}$, we may write

$$\mathbb{M}\, P_{n_{i+1}|\,n_i}(\Delta t, \Theta) = \frac{1}{2\pi i} \oint_\gamma \frac{\mathbb{M}\, F(z; \Delta t, n_i, \Theta)}{z^{n_{i+1}+1}}\, \mathrm{d}z, \tag{2.2}$$

where $\mathbb{M}$ now acts on the generating function F. Therefore, given the analytical approximation for F resulting from our method [32], we define

$$\widehat{F}(z;\Delta t, n_i, \widehat{\Theta}) = \mathbb{M}\, F(z;\Delta t, n_i, \Theta), \tag{2.3}$$

where $\widehat{\Theta} \subset \Theta$ is the subset of parameters that remain after the marginalisation procedure has been applied. Note that $\widehat{F}$ is still a fully analytical, general expression which depends on the as yet unspecified values of its arguments.

3) Evaluation. We choose a set $\widehat{\Theta}_0$ of numerical values for the parameters in $\widehat{\Theta}$. Moreover, we specify the integration contour γ, which we discretise as described in 2) to approximate the Cauchy integral in (2.1) by a finite sum. Suppose that the contour γ is discretised as $\{\zeta(l)\}$, with $0 \leq l \leq L$ and $\zeta(0) = \zeta(L)$; then, the integral of a function G along γ is approximated as

$$\oint_\gamma G(z)\,\mathrm{d}z \approx \sum_{l=0}^{L-1} G(\zeta(l))\,\Delta\zeta(l), \quad \text{with } \Delta\zeta(l) = \zeta(l+1) - \zeta(l). \tag{2.4}$$

Now, for every transition $(n_0 \to n)_j$ in the binned data set $\widehat{Q}$, we evaluate $\widehat{F}$, as given in (2.3), for the chosen parameter values $\widehat{\Theta}_0$ along the discretised contour. We hence obtain the array

$$\left\{ \frac{1}{2\pi i} \frac{\widehat{F}(\zeta(l);\Delta t, (n_0)_j, \widehat{\Theta}_0)}{\zeta(l)^{(n)_j+1}} \Delta\zeta(l) \right\} \quad \text{for } 0 \leq l \leq L-1 \text{ and } 1 \leq j \leq \widehat{N}, \tag{2.5}$$

which we sum over l to find

$$p_j(\widehat{\Theta}_0, \Delta t) = \sum_{l=0}^{L-1} \frac{1}{2\pi i} \frac{\widehat{F}(\zeta(l);\Delta t, (n_0)_j, \widehat{\Theta}_0)}{\zeta(l)^{(n)_j+1}} \Delta\zeta(l) \tag{2.6}$$

as the approximate value of the propagator for the transition $(n_0 \to n)_j$.

4) Calculation of the log-likelihood. To calculate the log-likelihood of the parameter subset $\widehat{\Theta}_0$, we substitute the approximate propagators p_j, as defined in (2.6), into (1.1) to obtain

$$L(\widehat{\Theta}_0) = \sum_{j=1}^{\widehat{N}} k_j \log\, p_j(\widehat{\Theta}, \Delta t). \tag{2.7}$$

3 Showcase 1: The telegraph model

To demonstrate our parameter inference procedure, we consider a stochastic gene expression model that incorporates DNA 'on'/'off' states ('telegraph model') [24, 28]:

$$\begin{array}{rcll} D & \underset{k_1}{\overset{k_0}{\rightleftharpoons}} & D^* & \text{(DNA activation/deactivation)},\\ D^* & \overset{\nu_0}{\to} & D^* + M & \text{(transcription of DNA to mRNA)},\\ M & \overset{\nu_1}{\to} & M + P & \text{(translation of mRNA to protein)},\\ M & \overset{d_0}{\to} & \emptyset & \text{(decay of mRNA)},\\ P & \overset{d_1}{\to} & \emptyset & \text{(decay of protein)}. \end{array} \tag{3.1}$$

In recent work [32], we presented an analytical method for obtaining explicit, general, time-dependent expressions for the generating function associated to the CME that arises from the model in (3.1). A pivotal element of the application of that method to (3.1) is the assumption that the protein decay rate d_1 is notably smaller than the decay rate d_0 of mRNA, which implies that the parameter $\varepsilon := \frac{d_1}{d_0}$ is small; hence, the associated generating function is *approximated* to a certain order $O = k$, corresponding with a theoretical accuracy that is proportional to ε^k. For more details on the resulting approximation, we refer to Appendix A.

To obtain synthetic data, we simulate the model in (3.1) using Gillespie's stochastic simulation algorithm (SSA) [13], for fixed values of the (rescaled) parameters

$$\kappa_0 := \frac{k_0}{d_1} = 1.3,\ \kappa_1 := \frac{k_1}{d_1} = 1.2,\ \lambda := \frac{\nu_0}{d_1} = 3.3,\ \mu := \frac{\nu_1}{d_0} = 2.85,\ \varepsilon := \frac{d_1}{d_0} = 0.1,\ \text{and } d_1 = 1 \quad (3.2)$$

on the time interval $0 \leq t \leq 10$, and we measure the protein abundance n with a fixed sampling interval Δt. As our method assumes that Δt is of order ε, cf. again Appendix A, we set $\Delta t = \varepsilon = 0.1$, which yields $N = 100$ transitions. Finally, we consider random initial states, with mRNA and protein numbers chosen uniformly between 0 and 10, and we assume DNA to be in the 'on' state with probability $\frac{\kappa_0}{\kappa_0+\kappa_1}$. Based on the simulated measurement data, we perform the parameter inference procedure described in Section 2. As the data consists of protein abundances only, and as propagators for the model in (3.1) depend on abundances of both mRNA and protein, we marginalise over mRNA, assuming a steady-state distribution reported by Raj *et al.* [24, Supporting Information, Protocol S1, Equation (1)]; that distribution coincides with the steady-state limit of the associated fully time-dependent distribution considered by Veerman, Marr & Popović [32]. We assume that the values of κ_0, κ_1, ε, and d_1 are known, and calculate the log-likelihood in (2.7) for varying λ and μ. We scan these two parameters in the range $\left\{10^{-3} \leq \lambda \leq 10^3,\ 10^{-3} \leq \mu \leq 10^2\right\}$, using a logarithmically spaced grid of 50×40 grid points. Figure 2 shows the resulting log-likelihood landscapes and, in particular, a comparison of the performance of the leading (zeroth) order approximation for the generating function, see Figure 2A, with that of the first order approximation in Figure 2B. We emphasise that our choice of λ and μ as the parameters to be inferred is not guided in any way by our analytical method – indeed, any other choice would have served our purpose, under the assumption that ε is sufficiently small. Rather, we chose λ and μ for illustrative purposes.

Moreover, it is important to note that, in the analytical derivation of the propagators used to produce Figure 2, all model parameters are assumed to be of order $\varepsilon^0 = 1$ [32, Assumption 3.2], which is reflected in the numerical values in (3.2). In particular, it follows that both the ratio of the transcription and translation rates $\frac{\nu_0}{\nu_1} = \varepsilon\frac{\lambda}{\mu}$ and the ratio of the transcription and mRNA decay rates $\frac{\nu_0}{d_0} = \varepsilon\lambda$ are assumed to be small. However, in Figure 2, the ranges over which λ and μ are scanned significantly exceed these estimates without causing apparent issues for our inference procedure. That could be taken as a sign that, while our analytical propagators were derived under certain assumptions on the order of model parameters, the resulting expressions may in fact be valid over a wider parameter range. On the other hand, it is worthwhile to note that noise-induced bistability is observed in a similar gene expression model if the above assumption on λ is abandoned, as considered with an alternative analytical approach in [7].

To quantify the performance of the method developed by Veerman, Marr & Popović [32] for parameter inference, we compare four different scenarios:

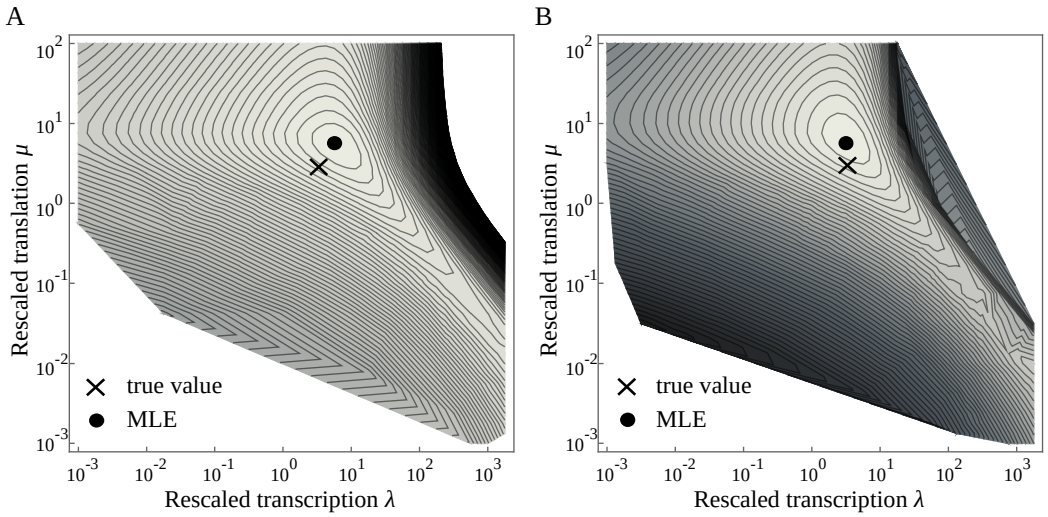

Figure 2: Log-likelihood landscapes inferred from a simulation of the telegraph model in (3.1) with $N = 100$ transitions and parameter values as in (3.2): true value (cross) versus maximum log-likelihood estimate (MLE; dot). (A) Leading (zeroth) order approximation. (B) First order approximation.

(a) Parameter values as in (3.2), with sampling interval $\Delta t = \varepsilon = 0.1$ on the time interval $0 \leq t \leq 10$, corresponding to $N = 100$ transitions, which is the original setup that yields the results shown in Figure 2.

(b) As in (a), with the time interval increased to $0 \leq t \leq 100$, which yields $N = 1000$ transitions.

(c) As in (a), with $\varepsilon = 0.01$: the sampling interval is decreased accordingly to $\Delta t = \varepsilon = 0.01$; measurements are taken on the time interval $0 \leq t \leq 1$, which yields $N = 100$ transitions.

(d) As in (a), with $\mu = 28.5$.

For each scenario, we infer the most likely values of the parameters λ and μ, for increasing approximation order O. The inferred values of λ and μ are compared to the 'true' values λ_{true} and μ_{true}, where we consider relative errors to quantify the performance of our inference procedure. The results of that comparison are shown in Figure 3. The accuracy of inference for λ clearly increases when the approximation order O is increased from 0 to 1; the increase in accuracy from $O = 1$ to $O = 2$ is obfuscated by grid size effects. A ten-fold increase in the number of transitions (b) increases the accuracy of the leading order approximation, while a ten-fold increase in the value of μ_{true} (d) decreases the accuracy of the leading order approximation. For μ, there is no noticeable increase in accuracy with the approximation order O within the parameter grid used, which could indicate that the number of transitions remains the dominant limiting factor. However, the accuracy of inference for μ increases overall when N is increased (b), the small parameter ε is decreased (c), or the value of μ_{true} is increased (d). An alternative explanation for the apparent insensitivity of the error in μ to the approximation order can be found in the underlying mRNA marginalisation

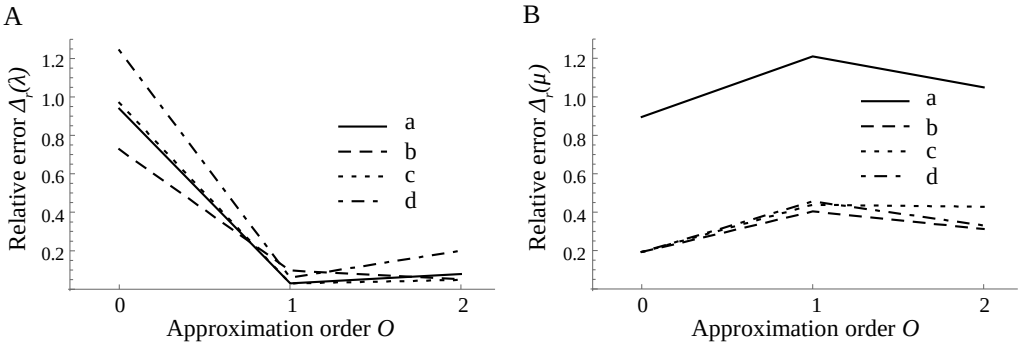

Figure 3: Relative error $\Delta_r(x) := \frac{x - x_{\text{true}}}{x_{\text{true}}}$ of the inferred parameters λ (A) and μ (B), for increasing approximation order O; for $O = k$, the propagators $P_{n_{i+1}|\,n_i}$ are approximated up to and including terms of order ε^k. (a) $N = 100$ transitions, $\varepsilon = 0.1$, $\lambda_{\text{true}} = 3.3$, and $\mu_{\text{true}} = 2.85$. (b) $N = 1000$ transitions, other parameters as in (a). (c) $\varepsilon = 0.01$, number of transitions N and other parameters as in (a). (d) $\mu_{\text{true}} = 28.5$, number of transitions N and other parameters as in (a).

procedure; however, one would expect that marginalisation to negatively influence the sensitivity of transcription (λ), rather than that of translation (μ).

4 Showcase 2: An autoregulated telegraph model

We extend the telegraph model in (3.1) with an autoregulatory mechanism, where the DNA activation rates are influenced by the presence of protein. Autoregulation is modelled in a catalytic manner, via one of the two following reactions:

$$D + P \xrightarrow{a_P} D^* + P \qquad \text{(autoactivation through protein)}, \tag{4.1a}$$

$$D^* + P \xrightarrow{r_P} D + P \qquad \text{(autorepression through protein)}. \tag{4.1b}$$

The above pair of autoregulation mechanisms was introduced by Hornos *et al.* [17], and implemented, e.g. by Iyer-Biswas & Jayaprakash [18]; see Section 5 for a discussion of the physical validity of these mechanisms. Propagators for the telegraph model incorporating both autoregulation through protein, as in (4.1), and autoregulation via mRNA were determined in [32]. The protein-autoregulated telegraph model is also discussed in [15], where it is called the 'full model'; cf. [15, Figure 1] for a comparison and discussion of several reductions of that model.

To assess the performance of our parameter inference procedure, we fix the parameter values as in (3.2). Again, we marginalise over mRNA, assuming a steady-state distribution; see Section 3. Note that the degree to which that steady-state assumption is valid directly depends on the magnitude of the autoregulation rates, as protein levels influence levels of mRNA through DNA activation rates. We generate six data sets, as follows:

(A) Simulate the model in (3.1) without autoregulation ('null model'; $a_P = 0 = r_P$) on the time interval $0 \leq t \leq 10$, which yields $N = 100$ transitions.

(B) As in (A), with the time interval increased to $0 \leq t \leq 100$, which yields $N = 1000$ transitions.

(C) Simulate the extended model {(3.1),(4.1a)} with autoactivation for $a_P\delta = 0.3$ on the time interval $0 \leq t \leq 10$, which yields $N = 100$ transitions.

(D) As in (C), with the time interval increased to $0 \leq t \leq 100$, which yields $N = 1000$ transitions.

(E) Simulate the extended model {(3.1),(4.1b)} with autorepression for $r_P\delta = 0.3$ on the time interval $0 \leq t \leq 10$, which yields $N = 100$ transitions.

(F) As in (E), with the time interval increased to $0 \leq t \leq 100$, which yields $N = 1000$ transitions.

Every data set consists of 10 runs of equal length.

Generating functions for the autoregulated extension, by (4.1), of the telegraph model in (3.1) have been derived in the theoretical companion article [32] to the current work, under the assumption that the autoregulation rate a_P or r_P is small compared to the protein decay rate d_1. That assumption implies that the ratios $\frac{a_P}{d_1} := \alpha_P\delta$ and $\frac{r_P}{d_1} := \rho_P\delta$ are small.

Parameter inference now proceeds as follows. We fix a data set, and take a single run from that set. For that run, we determine the log-likelihood of the autoactivated model in {(3.1),(4.1a)}, varying $0 \leq \alpha_P\delta \leq 1$; likewise, we determine the log-likelihood of the autorepressed model in {(3.1),(4.1b)}, varying $0 \leq \rho_P\delta \leq 1$. The log-likelihood L of the autoregulated extension is then compared with the log-likelihood L_0 of the non-regulated model in (3.1); as before, N denotes the number of data points, where L is defined as in (2.7), that is, we simply evaluate the probability of the observed transitions given the model under consideration, after marginalisation over mRNA. The log-likelihood difference $L - L_0$, which is equal to the logarithm of the likelihood ratio, quantifies the evidence for that model. We repeat the above procedure for all 10 runs in the data set, and we determine the mean and standard deviation; the outcome is illustrated in Figure 4. We observe that 1000 transitions suffice to correctly refute autorepression in (B,D), and to correctly refute autoactivation in (F). In the case of 100 transitions, no conclusion can be drawn from (A) and (C), while (E) correctly refutes autoactivation; however, the small difference between L and L_0, with $|L - L_0| < 1$, indicates low significance.

It is important to reiterate that the validity of our assumption that mRNA obeys a steady-state distribution, which was used in the marginalisation step, is coupled to the magnitude of the auroregulation rates. While a steady-state assumption can hence be argued to be approximately valid for small δ, an increase in autoregulation strength increases the protein level feedback on DNA activation rates, thereby indirectly, but dynamically, influencing levels of mRNA. We propose that the ensuing breakdown of the steady-state assumption is reflected in the behaviour of the log-likelihood difference shown in Figure 4. The assumption can be argued to induce a bias as $\alpha_P\delta, \rho_P\delta \to 1$, accordingly skewing the log-likelihood difference, which could also explain the absence of a clear maximum in Figures 4D and 4F.

5 Discussion

In the present article, we showcase a parameter inference procedure that is based on a recently developed analytical method [32] which allows for the efficient numerical approximation of propagators via the Cauchy integral formula on the basis of asymptotic series for the underlying generating

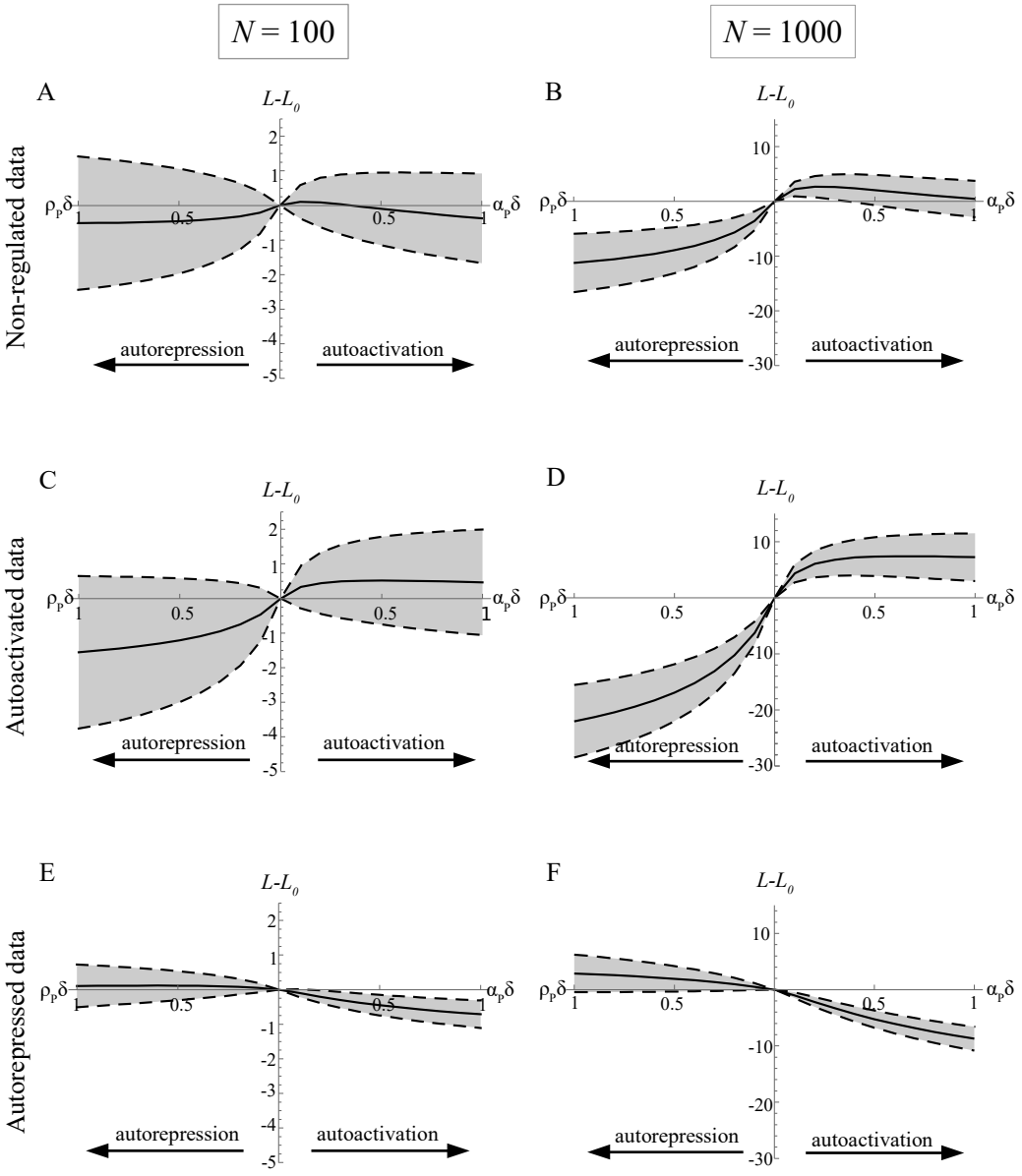

Figure 4: Parameter inference for the extended autoregulated model in {(3.1),(4.1)} on the basis of various types of synthetic data, where the performance of the model is quantified via the log-likelihood difference $L - L_0$. On the vertical axis, $L - L_0$ indicates the difference between the log-likelihood of the autoactivated or autorepressed model and that of the non-regulated model in (3.1); the higher the value of $L - L_0$, the more likely the associated model is. In each panel, the solid curve indicates the mean values based on 10 model runs; dashed curves indicate the uncertainty (one standard deviation). On the horizontal axis, the strength of autoregulation is measured by $\alpha_P\delta$ (increasing to the right) or $\rho_P\delta$ (increasing to the left). (A) Data generated from the null (non-regulated) model in (3.1), with $N = 100$ transitions. (C) Data generated from the model in (3.1), with autoactivation as in (4.1a), for $\alpha_P\delta = 0.3$ and $N = 100$ transitions. (E) Data generated from the model in (3.1), with autorepression as in (4.1b), for $\rho_P\delta = 0.3$ and $N = 100$ transitions. (B,D,F) as (A,C,E), but with $N = 1000$ transitions; note that the vertical axis has a different scaling. All other model parameters were assumed to be known.

functions. The resulting hybrid analytical-numerical approach reduces the need for computationally expensive simulations; moreover, due to its perturbative nature, it is highly applicable over relatively short time scales, such as occur naturally in the calculation of the log-likelihood in (1.1).

We simulate protein expression with a simple stochastic simulation algorithm (SSA) [13]; the resulting protein abundances are in the lower range of experimentally measured values [27, 22], which does not, however, limit our proof of principle. In addition, in the presented application of our approach, we assume that some of the model parameters are known, that protein abundances can be measured at regular sampling intervals, and that there is no noise. We then present results for synthetic data in a family of models for stochastic gene expression from the literature under the central assumption that lifetimes of protein are significantly longer than those of mRNA, which introduces a small parameter ε and, hence, a separation of scales. An extensive discussion of the validity of our assumption that ε is small can be found elsewhere [28, 32]; see also Section 3. The assumed smallness of ε is crucial to the underlying analytical method, as introduced by Veerman, Marr & Popović [32]. In addition, our approach is specifically attuned to the time variability of the expression process, in the sense that we assume the sampling interval Δt to be small, as well; cf. again Section 3. It is thus ideally suited to describing transient dynamics far from steady state, as is evident in the bursting behaviour seen in Figure 1A.

In Section 3, we discuss a simple ('telegraph') model for gene expression without autoregulation, showing that our approach can successfully infer relevant model parameters. Unlike in previous work [10], the underlying implementation avoids potential bias due to zero propagator values and large initial protein numbers through the use of 'implicit' series expansions in ε; see Appendix A for an in-depth argument. Note that, in the simple telegraph model, transcription factors are assumed to act as single molecules. Our analytical approach would have to be extended to incorporate transcription factor dimer regulation, e.g. via the recently developed linear mapping approximation [3].

In Section 4, we perform a model comparison in an autoregulated extension of the telegraph model. We consider three types of gene regulation: autoactivation, autorepression, and no regulation of DNA activity (null model). For each type, we simulate protein abundance data with 100 and 1000 transitions, respectively. Throughout, we find that 100 data points are insufficient to reject model hypotheses with our approach. With 1000 data points, however, we can successfully reject the non-regulated and the autorepressed model for simulated data from an autoactivated model, in which case we can even infer the correct order of the autoactivation parameter. For simulated autorepression, we can reject the model with autoactivation, but not the non-regulated model. Our approach fails to identify the correct model for data from a non-regulated model for 1000 transitions, where the autoactivated model is clearly, but wrongly, favoured. We believe that more research is needed into the sources of these discrepancies in dependence on both model parameters and the order of our approximation.

In both showcases, we observe a trade-off between the accuracy of inference versus the required computation time. Computation times seem to increase exponentially with the approximation order, at least for the setup realised in this article. For practical purposes, we hence propose an algorithm whereby the fastest, leading order approximation is used to obtain an initial estimate for the underlying model parameters; that estimate can then be improved by including higher order corrections, resulting in a much more computationally efficient procedure.

It is insightful to compare our results with other recent work on parameter inference in regulated gene expression models. In work by Feigelman *et al.* [9], three models for regulated gene expression with a slightly different structure compared to the models studied in the present article were simulated and inferred via a particle filtering-based inference procedure that employs genealogical information of dividing cells. Interestingly, positive and negative autoregulation could be successfully rejected there for data that was simulated from a no-feedback model. However, the no-feedback model could not be rejected for data originating from the corresponding models with positive or negative feedback. From that comparison with Feigelman *et al.* [9], we conclude that the structure of the data, the intensity of regulatory feedback, and the chosen inference procedure together will influence the extent of insight which can be obtained from an approach that is based on stochastic models of gene expression.

We emphasise that our analytical method is not restricted to the specific models showcased here. Our aim in the present article is to demonstrate the applicability of the method, and to investigate its performance, rather than to assess the biological validity of a given model. Minimally, our approach can be extended to recent, physiologically more relevant modifications of the telegraph model with autoregulation [17] by Grima, Schmidt & Newman [14] and Congxin *et al.* [4]; another feasible alternative model can be obtained by introduction of a refractory state [35]. Ideally, we would like to test a variety of stochastic gene expression models against a given set of measurement data. Current computational approaches struggle to provide propagators for models with more than one regulated species, which can often only be approximated even in that simple scenario. The principal advantage of our hybrid approach is that propagators can be evaluated in a computationally efficient manner, via a combination of mathematical analysis and numerical integration [32]; other approaches rely either on the calculation of propagators based on direct numerical simulation of the underlying model – which is computationally demanding – or on the assumption that symbolic derivatives of the generating function are explicitly known, which only holds for specific models of relatively low complexity [23].

The input for our propagator-based approach is the abundance of the involved species, viz. of protein. Thus, we assume that absolute protein numbers are measured, which is in practice hampered by an unknown scaling factor between the observed fluorescence and the corresponding abundances, and by noise. While various suggestions for inferring that factor have been made [16, 25, 1], and while a linear scaling is customarily assumed [29, 33], an accurate experimental determination remains extremely challenging.

It is instructive to compare our propagator-based approach to alternative approaches to parameter inference in the context of stochastic gene expression, such as the linear noise approximation [30], a system size expansion with moment closure [12], or tensor-based methods for the corresponding Fokker-Planck equation [21]. Both the linear noise approximation and a system size expansion assume large system sizes and are only valid for large copy numbers, see also [30]; moreover, the impact of moment closure on the relevance of nonlinearities is potentially non-negligible [20]. The relatively recently developed tensor-based methods focus on steady state distributions for large system sizes; their validity away from the thermodynamic limit is as yet unclear, nor is it clear how results are influenced by the particular tensor decomposition method that is chosen. In contrast to these approaches, the propagator-based approach we have employed in this article not only performs well with low copy numbers, i.e. for small system sizes; our analytical propagators are also

explicitly *time-dependent*, in contrast to the usual steady state assumption. Details can be found in our theoretical companion article [32].

Finally, the showcases presented in this article are based on synthetic data that was generated *in silico*; in the future, we plan to consider experimental data, such as can be found in work by Suter *et al.* [29].

Acknowledgements

The authors thank Ramon Grima and Peter Swain (both University of Edinburgh), as well as two anonymous reviewers, for valuable comments and suggestions. This work has been supported by the Leverhulme Trust, through Research Project Grant RPG-2015-017 ('The nature of gene expression: model selection and parameter inference').

Author Contributions Statement

All authors conceived the experiments; F.V. conducted the experiments; F.V. and C.M. analysed the results; F.V. and N.P. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Competing interests

The authors declare no competing interests.

A Analytical details

The hybrid analytical-numerical approach developed by Veerman, Marr & Popović [32] introduces a probability generating function which transforms the CME corresponding to a given stochastic gene expression model into a system of partial differential equations. Explicit expressions for the solutions to these equations are obtained using dynamical systems techniques, in combination with perturbation theory. The values of the associated propagators are recovered through a numerically efficient implementation of the Cauchy integral in (2.1); see also Section 2.

A.1 Approximate generating functions

The generating functions used to approximate propagators in the present article, cf. Section 2, are derived via the analytical method presented by Veerman, Marr & Popović [32]. For the telegraph model in Section 3, the leading order generating function $\widehat{F}_0$ is given by

$$\widehat{F}_0(z;\Delta t,n_0,\varepsilon,\lambda,\mu)=\left[1-(1-z)\mathrm{e}^{-\Delta t}\right]^{n_0}{}_1\mathrm{F}_1\left[\kappa_0,\kappa_0+\kappa_1,\varepsilon\lambda\frac{1-f(z)}{f(z)}\left(1-\mathrm{e}^{-f(z)\frac{\Delta t}{\varepsilon}}\right)\right], \quad \text{(A.1)}$$

where ${}_1\mathrm{F}_1$ denotes the confluent hypergeometric function [6, Chapter 13] and $f(z)=1+\mu(1-z)\mathrm{e}^{-\Delta t}$. All parameters have been rescaled according to (3.2). The generating function has been marginalised

over mRNA, assuming a steady-state distribution reported by Raj et al. [24]. Analogously, the first order approximation $\widehat{F}_1$ of the generating function is given by

$$\begin{aligned}\widehat{F}_1(z;\Delta t,n_0,\varepsilon,\lambda,\mu,\chi) &= \left[1-(1-z)\mathrm{e}^{-\Delta t}\right]^{n_0}\\ &\times\left[1+\varepsilon(1-\chi)\lambda\frac{1-f(z)}{f(z)}\mathrm{e}^{-f(z)\frac{\Delta t}{\varepsilon}}\left(\frac{1-\mathrm{e}^{f(z)\frac{\Delta t}{\varepsilon}}}{f(z)}+\frac{\Delta t}{\varepsilon}\mathrm{e}^{f(z)\frac{\Delta t}{\varepsilon}}\right)\right]\\ &\times {}_1\mathrm{F}_1\left[\kappa_0,\kappa_0+\kappa_1,\varepsilon\lambda\frac{1-f(z)}{f(z)}\left(1-\mathrm{e}^{-f(z)\frac{\Delta t}{\varepsilon}}\right)\left\{1+\varepsilon\left[\frac{1}{f(z)^2}+\frac{1}{f(z)}\frac{\Delta t}{\varepsilon}+(1-f(z))\frac{\Delta t^2}{2\varepsilon^2}\right]\right\}\right];\end{aligned} \tag{A.2}$$

here, $\chi=\frac{\kappa_1}{\kappa_0+\kappa_1}$. For the autoregulated model discussed in Section 4, the same expressions for the generating functions are used; however, χ now depends on the autoregulatory mechanism according to

$$\chi=\begin{cases}\dfrac{\kappa_1}{\kappa_0+\kappa_1} & \text{(no autoregulation)},\\[2ex] \dfrac{\kappa_1}{\kappa_0+\kappa_1+\alpha_P\delta\,n_0} & \text{(autoactivation)},\\[2ex] \dfrac{\kappa_1+\rho_P\delta\,n_0}{\kappa_0+\kappa_1+\rho_P\delta\,n_0} & \text{(autorepression)}.\end{cases} \tag{A.3}$$

A.2 'Implicit' expansions

It is important to note that neither the leading order generating function in (A.1) nor the first order approximation given by (A.2) are expressed as asymptotic series in powers of ε, as would be expected on the basis of the perturbative approach taken by Veerman, Marr & Popović [32]. The underlying reasoning can be summarised as follows.

First, in the derivation of these generating functions, it was assumed that the sampling interval Δt was small, i.e. of order ε; note that this assumption is satisfied in all numerical simulations shown in the current article, where $\Delta t=\varepsilon$ throughout. Thus, we can write

$$\Delta t=\varepsilon\Delta s. \tag{A.4}$$

The accuracy of our series approximation depends on the asymptotic scaling of Δt; see [32, Remark 3.6]. With the above scaling, the approximation order is equal to the order to which the resulting propagators are accurate, in powers of ε.

With the scaling for Δt given in (A.4), an expansion of $\widehat{F}_0$ and $\widehat{F}_1$, as defined in (A.1) and (A.2), respectively, into asymptotic series in ε to the appropriate order yields

$$\widehat{F}_0=z^{n_0}, \tag{A.5}$$

$$\widehat{F}_1=z^{n_0}\left\{1+\varepsilon(1-z)\left[\frac{n_0\Delta s}{z}-\frac{(1-\chi)\lambda\mu}{1+\mu(1-z)}\left(\frac{\mu(1-z)\left\{1-\mathrm{e}^{[-(1+\mu(1-z)]\Delta s}\right\}}{1+\mu(1-z)}+\Delta s\right)\right]\right\}. \tag{A.6}$$

From (A.5), one can immediately conclude that

$$\oint_\gamma \widehat{F}_0=\delta_{n,n_0}, \tag{A.7}$$

where $\delta_{n,n_0} = 1$ if $n = n_0$ and $\delta_{n,n_0} = 0$ otherwise. From the series for $\widehat{F}_1$ in (A.6), we see that we can write

$$\widehat{F}_1(z) = z^{n_0}\Big\{1 + \varepsilon \sum_{k=-1}^{\infty} f_{1,k} z^k\Big\}, \tag{A.8}$$

with appropriately chosen coefficients $f_{1,k}$; hence, it follows that

$$\oint_\gamma \widehat{F}_1 = 0 \quad \text{if } n_0 > n + 1. \tag{A.9}$$

More generally, an expansion of the generating function to order k in ε will yield

$$\oint_\gamma \widehat{F}_k = 0 \quad \text{if } n_0 > n + k. \tag{A.10}$$

From these observations, we conclude that decreasing transitions ($n_i \to n_{i+1}$), where $n_i > n_{i+1} + k$, will be assigned a probability that is identically zero. Hence, if such transitions *are* present in the data, the model is ruled out immediately, as our perturbative approach excludes the possibility that decreasing transitions can occur. One can understand this phenomenon by considering the definition of the small parameter ε as the ratio of the protein decay rate d_1 over the mRNA decay rate d_0. A leading order approximation of $\varepsilon = 0$ is thus equivalent to taking $d_1 \to 0$ which, in turn, implies that protein does not decay at all, since (natural) protein decay is the only reaction in (3.1) that can decrease protein abundance. By the same reasoning, a straightforward expansion of the generating function to order ε^k will restrict the model to transitions ($n_i \to n_{i+1}$), where $n_{i+1} - n_i \geq -k$. It would follow that either the order O of the method would be limited from below by the data, leading to high-order expansions in ε and, hence, to increased computation times, or that the method could only be applied to a subset of the given data, which would introduce a bias.

Lastly, an asymptotic expansion such as (A.6) implicitly assumes that all parameters and variables in the model are of order 1 in ε. For the series expansion of $\widehat{F}_1$ in (A.6), that assumption would significantly restrict the range of λ; in comparison, in Figure 2, likelihood values for λ up to order ε^{-3} are calculated. More importantly, the above assumption would restrict the range of n_0, implying that only a subset of the data – with sufficiently low protein numbers – could be used as input for parameter inference.

We emphasise that none of these difficulties occur with the expressions in (A.1) and (A.2), where the expansion order in ε is expressed 'implicitly' in the respective functional forms of $\widehat{F}_0$ and $\widehat{F}_1$.

References

[1] E. Bakker and P.S. Swain. Estimating numbers of intracellular molecules through analysing fluctuations in photobleaching. *Nature Scientific Reports*, 9:15238, 2019.

[2] F. Bornemann. Accuracy and stability of computing high-order derivatives of analytic functions by Cauchy integrals. *Foundations of Computational Mathematics*, 11(1):1–63, 2011.

[3] Z. Cao and R. Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nature Communications*, 9:3305, 2018.

[4] L. Congxin, F. Cesbron, M. Oehler, M. Brunner, and T. Höfer. Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell Systems*, 6(4):409–423, 2018.

[5] M.M. Crane, I.B. Clark, E. Bakker, S. Smith, and P.S. Swain. A microfluidic system for studying ageing and dynamic single-cell responses in budding yeast. *PLoS One*, 9(6):e100042, 2014.

[6] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.1 of 2021-03-15. F.W.J. Olver, A.B. Olde Daalhuis, D.W. Lozier, B.I. Schneider, R.F. Boisvert, C.W. Clark, B.R. Miller, B.V. Saunders, H.S. Cohl, and M.A. McClain, eds.

[7] A. Duncan, S. Liao, Vejchodský, R. Erban, and R. Grima. Noice-induced multistability in chemical systems: Discrete versus continuum modeling. *Physical Review E*, 91(4):042111, 2015.

[8] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[9] J. Feigelman, S. Ganscha, S. Hastreiter, M. Schwarzfischer, A. Filipczyk, T. Schroeder, F.J. Theis, C. Marr, and M. Claassen. Analysis of cell lineage trees by exact bayesian inference identifies negative autoregulation of nanog in mouse embryonic stem cells. *Cell Systems*, 3(5):480–490, 2016.

[10] J. Feigelman, N. Popović, and C. Marr. A case study on the use of scale separation-based analytical propagators for parameter inference in models of stochastic gene regulation. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):164–173, 2015.

[11] A. Filipczyk, C. Marr, S. Hastreiter, J. Feigelman, M. Schwarzfischer, P.S. Hoppe, D. Loeffler, K.D. Kokkaliaris, M. Endele, B. Schauberger, O. Hilsenbeck, S. Skylaki, J. Hasenauer, K. Anastassiadis, F.J. Theis, and T. Schroeder. Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nature Cell Biology*, 17:1235–1246, 2015.

[12] F. Fröhlich, P. Thomas, A. Kazeroonian, F.J. Theis, R. Grima, and J. Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Computational Biology*, 12(7):e1005030, 2016.

[13] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[14] R. Grima, D.R. Schmidt, and T.J. Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *Journal of Chemical Physics*, 190:035104, 2012.

[15] J. Holehouse, Z. Cao, and R. Grima. Stochastic modeling of autoregulatory genetic feedback loops: a review and comparative study. *Biophysical Journal*, 118(7):1517–1525, 2020.

[16] P.S. Hoppe, M. Schwarzfischer, D. Loeffler, K.D. Kokkaliaris, O. Hilsenbeck, N. Moritz, M. Endele, A. Filipczyk, A. Gambardella, N. Ahmed, M. Etzrodt, D.L. Coutu, M.A. Rieger, C. Marr, M.K. Strasser, B. Schauberger, I. Burtscher, O. Ermakova, A. Bürger, H. Lickert, C. Nerlov,

F.J. Theis, and T. Schroeder. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*, 535:299–302, 2016.

[17] J.E.M. Hornos, D Schultz, G.C.P. Innocentini, J. Wang, A.M. Walczak, J.N. Onuchic, and P.G. Wolynes. Self-regulating gene: An exact solution. *Physical Review E*, 72:051907, 2005.

[18] S. Iyer-Biswas and C. Jayaprakash. Mixed Poisson distributions in exact solutions of stochastic autoregulation models. *Physical Review E*, 90:052712, 2014.

[19] S.M. Janicki, T. Tsukamoto, S.E. Salghetti, W.P. Tansey, R. Sachidanandam, K.V. Prasanth, T. Ried, Y. Shav-Tal, E. Bertrand, R.H. Singer, and D.L. Spector. From silencing to gene expression: real-time analysis in single cells. *Cell*, 116(5):683–698, 2004.

[20] C. Kuehn. Moment closure—a brief review. In E. Schöll, S.H.L. Klapp, and P. Hövel, editors, *Control of Self-Organising Nonlinear Systems*, pages 253–271. Springer International Publishing Switzerland, 2016.

[21] S. Liao, T. Vejchodský, and R. Erban. Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *Journal of the Royal Society Interface*, 12(108):20150233, 2015.

[22] R. Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, 35(12):1050–1055, 2013.

[23] N. Popović, C. Marr, and P.S. Swain. A geometric analysis of fast-slow models for stochastic gene expression. *Journal of Mathematical Biology*, 72(1):87–122, 2016.

[24] A. Raj, C.S. Peskin, D. Tranchina, D.Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):1707–1719, 2006.

[25] N. Rosenfeld, T.J. Perkins, U. Alon, M.B. Elowitz, and P.S. Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophysical Journal*, 91(2):759–766, 2006.

[26] D. Schnoerr, G. Sanguinetti, and R. Grima. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A*, 50(9):093001, 60, 2017.

[27] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[28] V. Shahrezaei and P.S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[29] D.M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[30] P. Thomas, H. Matuschek, and R. Grima. How reliable is the linear noise approximation of gene regulatory networks? *BMC Genomics*, 14(Suppl 4):S5, 2013.

[31] N. G. van Kampen. *Stochastic processes in physics and chemistry*, volume 888 of *Lecture Notes in Mathematics*. North-Holland Publishing Co., Amsterdam-New York, 1981.

[32] F. Veerman, C. Marr, and N. Popović. Time-dependent propagators for stochastic models of gene expression: an analytical method. *Journal of Mathematical Biology*, 77:261–312, 2018.

[33] C. Zechner, M. Unger, S. Pelete, M. Peter, and H. Koeppl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature Methods*, 11:197–202, 2013.

[34] D. Zenklusen, D.R. Larson, and R.H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15:1263–1271, 2008.

[35] B. Zoller, D. Nicolas, N. Molina, and Naef. F. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology*, 11(7):823, 2015.