Herbach *et al. BMC Systems Biology* (2017) 11:105
DOI 10.1186/s12918-017-0487-0

BMC Systems Biology

RESEARCH ARTICLE Open Access

Inferring gene regulatory networks from single-cell data: a mechanistic approach

Ulysse Herbach[1,2,3], Arnaud Bonnaffoux[1,2,4], Thibault Espinasse[3] and Olivier Gandrillon[1,2*]

Abstract

Background: The recent development of single-cell transcriptomics has enabled gene expression to be measured in individual cells instead of being population-averaged. Despite this considerable precision improvement, inferring regulatory networks remains challenging because stochasticity now proves to play a fundamental role in gene expression. In particular, mRNA synthesis is now acknowledged to occur in a highly bursty manner.

Results: We propose to view the inference problem as a fitting procedure for a mechanistic gene network model that is inherently stochastic and takes not only protein, but also mRNA levels into account. We first explain how to build and simulate this network model based upon the coupling of genes that are described as piecewise-deterministic Markov processes. Our model is modular and can be used to implement various biochemical hypotheses including causal interactions between genes. However, a naive fitting procedure would be intractable. By performing a relevant approximation of the stationary distribution, we derive a tractable procedure that corresponds to a statistical hidden Markov model with interpretable parameters. This approximation turns out to be extremely close to the theoretical distribution in the case of a simple toggle-switch, and we show that it can indeed fit real single-cell data. As a first step toward inference, our approach was applied to a number of simple two-gene networks simulated in silico from the mechanistic model and satisfactorily recovered the original networks.

Conclusions: Our results demonstrate that functional interactions between genes can be inferred from the distribution of a mechanistic, dynamical stochastic model that is able to describe gene expression in individual cells. This approach seems promising in relation to the current explosion of single-cell expression data.

Keywords: Single-cell transcriptomics, Gene network inference, Multiscale modelling, Piecewise-deterministic Markov processes

Background

Inferring regulatory networks from gene expression data is a longstanding question in systems biology [1], with an active community developing many possible solutions. So far, almost all studies have been based on population-averaged data, which historically used to be the only possible way to observe gene expression. Technologies now allow us to measure mRNA levels in individual cells [2–4], a revolution in terms of precision. However, the network reconstruction task paradoxically remains more challenging than ever.

The main reason is that the variability in gene expression unexpectedly stands at a large distance from a trivial, limited perturbation around the population mean. It is now clear indeed that this variability can have functional significance [5–7] and should therefore not be ignored when dealing with gene network inference. In particular, as the mean is not sufficient to account for a population of cells, a deterministic model – e.g. ordinary differential equation (ODE) systems, often used in inference [8, 9] – is unlikely to faithfully inform about an underlying gene regulatory network. Whether such a deterministic approach could still be a valid approximation or not is a difficult question that may require some biological insight into the system under consideration [10]. Another key aspect when considering individual cells is that they generally have to be killed for measurements: from a statistical point of view, temporal single-cell data therefore should

*Correspondence: olivier.gandrillon@ens-lyon.fr
[1] Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie Site Jacques Monod, F-69007 Lyon, France
[2] Inria Team Dracula, Inria Center Grenoble Rhône-Alpes, Lyon, France
Full list of author information is available at the end of the article

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

not be seen as a set of time series, but rather *snapshots*, i.e. independent samples from a time series of distributions.

On the other hand, single-cell data give the opportunity of moving one step further toward a more accurate physical description of gene expression. Molecular processes of gene expression are overall now well understood, in particular transcription, but precisely how stochasticity emerges is still somewhat of a conundrum. Harnessing variability in single-cell data is expected to allow for the identification of critical parameters and also to provide hints about the basic molecular processes involved [11, 12]. Moreover, the variability arising from perturbations in cell populations is often crucial for network reconstruction to succeed [13, 14] as the deterministic inference problem suffers from intrinsic limitations [15]. From this point of view, the same information is expected to be contained in the variability between cells in single-cell data. Some of the few existing single-cell inference methods follow this path, for example using asynchronous Boolean network models [16] or generating pseudo time series [9, 17]. In this article, we use a mechanistic approach in the sense that every part of our model has an explicit physical interpretation. Importantly, mRNA observations are not used as a proxy for proteins since both are explicitly modeled.

Besides, mechanistic models that are accurate enough to describe gene expression at the single-cell level usually do not consider interactions between genes. For example, the so-called "two-state" (aka random telegraph) model has been successfully used with single-cell RNA-seq data [18], but the joint distribution of a set of genes contains much more information than the marginal kinetics of individual genes: our aim is to exploit this information while keeping the mechanistic point of view.

Namely, we propose to view the inference as a fitting procedure for a mechanistic gene network model. Whereas the goal here is not to achieve global predictability performances (e.g. as in [19]), our framework makes it possible to explicitly implement many biological hypotheses, and to test them by going back and forth between simulations and experiments. The main point of this article is to show that a tractable statistical model for network inference from single-cell data can be derived through successive relevant approximations. Finally, we demonstrate that our approach is capable of extracting enough information out of in silico-simulated noisy single-cell data to correctly infer the structures of various two-gene networks.

Methods

In this part, we aim at deriving a tractable statistical model from a mechanistic one. We will use the two-state model for gene expression to build a "network of two-state models" by making the promoter switching rates depend on protein levels. Then, successive relevant simplifications will lead to an explicit approximation of a statistical likelihood.

A simple mechanistic model for gene regulatory networks

Basic block: stochastic expression of a single gene

Our starting point is the well-known two-state model of gene expression [20–23], a refinement of the model introduced by [24] from pioneering single-cell experiments [25]. In this model, a gene is described by its promoter which can be either active (on) or inactive (off) – possibly representing a transcription complex being "bound" or "unbound" but it may be more complicated [26] – with mRNA being transcribed only during the active periods. Translation is added in a standard way, each mRNA molecule producing proteins at a constant rate. The resulting model (Fig. 1) can be entirely defined by the set of chemical reactions detailed in Table 1, where chemical species G, G^*, M and P respectively denote the inactive promoter, the active promoter, the amount of mRNA and proteins. The mathematical framework generally assumes stochastic mass-action kinetics [27] for all reactions, since they typically involve few molecules compared to Avogadro's number. In this fully discrete setting, one can use the master equation to compute stationary distributions: for mRNA the exact distribution is a Beta-Poisson mixture [28], and an approximation is available for proteins when they degrade much more slowly than mRNA [29]. In addition, the time-dependent generating function of mRNA is known in closed form [30] and can be inverted in some cases to obtain the transient distribution [28].

In practice, the formulas involve hypergeometric series that are not straightforward to use in a statistical inference framework. Besides, these series essentially arise from the fact that such a discrete model has to enumerate all potential collisions between molecules (the stochastic mass-action assumption in the master equation). It is therefore natural to consider keeping only the most important source of noise, that is, keeping a molecular representation for rare species but describing abundant

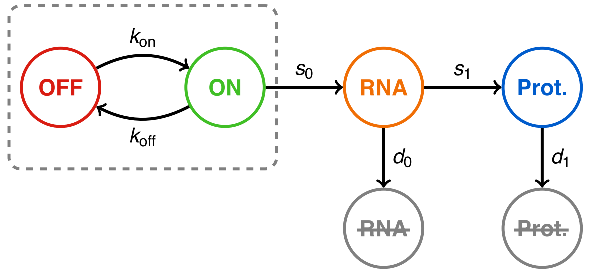


Fig. 1 Scheme of the two-state model of gene expression. We use it as the basic block of our network model

Table 1 Chemical reactions defining the two-state model. The rate constants are usually abbreviated to *rates* as they correspond to actual reactions rates when only one molecule of reactant is present. In the stochastic setting, these rates are in fact propensities, i.e. probabilities per unit of time

Reaction	Rate constant	Interpretation
$G \to G^*$	k_{on}	gene activation
$G^* \to G$	k_{off}	gene inactivation
$G^* \to G^* + M$	s_0	transcription
$M \to M + P$	s_1	translation
$M \to \varnothing$	d_0	mRNA degradation
$P \to \varnothing$	d_1	protein degradation

species at a higher level where molecular noise averages out to continuous quantities. A quick look at reactions in Table 1 indicates that the only rare species are G and G^*, with quantities $[G]$ and $[G^*]$ being equal to 0 or 1 molecule and satisfying the conservation relation $[G]+[G^*]= 1$. The other two, M and P, are not conserved quantities in the model and reach a much wider range in biological situations [31], meaning that saturation constants s_0/d_0 and s_1/d_1 are much larger than 1 molecule.

Hence, letting $E(t)$, $M(t)$ and $P(t)$ denote the respective quantities of G^*, M and P at time t, we consider a hybrid version of the previous model, where E has the same stochastic dynamics as before, but with M and P now following usual rate equations:

$$\begin{cases} E(t) : 0 \xrightarrow{k_{\text{on}}} 1, \;\; 1 \xrightarrow{k_{\text{off}}} 0 \\ M'(t) = s_0 E(t) - d_0 M(t) \\ P'(t) = s_1 M(t) - d_1 P(t) \end{cases} \tag{1}$$

This system simply switches between two ordinary differential equations, depending on the value of the two-state continuous-time Markov process $E(t)$, making it a Piecewise-Deterministic Markov Process (PDMP) [32]. From a mathematical perspective, model (1) rigorously approximates the original molecular model when s_0/d_0 and s_1/d_1 are large enough [33, 34] and interestingly, it has already been implicitly considered in the biological literature [22, 23]. Note also that the stationary distribution of mRNA is a scaled Beta distribution that is exactly the one of the Beta-Poisson mixture in the discrete model [28]. Similarly to a recent approach for a two-gene toggle switch [35], we will use (1) as a basic building block for gene networks.

When both $k_{\text{on}} \ll k_{\text{off}}$ and $d_0 \ll k_{\text{off}}$, mRNA is transcribed by *bursts*, i.e. during short periods which make the mRNA quantity stay far from saturation. Hence, the amount transcribed within each burst is approximately proportional to the burst duration, whose mean is $1/k_{\text{off}}$ by definition: this justifies the quantity s/k_{off} often being called “burst size” or “burst amplitude”. Furthermore, promoter active periods are much shorter than inactive ones so they can be seen as instantaneous, justifying the name “burst frequency” for the inverse of the mean inactive time k_{on}. We place ourselves in this situation as it often occurs in experiments [22, 23, 36–38]. Note however that these two notions are not clearly defined when relations $k_{\text{on}} \ll k_{\text{off}}$ and $d_0 \ll k_{\text{off}}$ do not hold.

Adding interactions between genes: the network model

Now considering a given set of n genes, a natural way of building a network is to assume that each gene i produces specific mRNA M_i and protein P_i, and to define a version of model (1) with its own parameters:

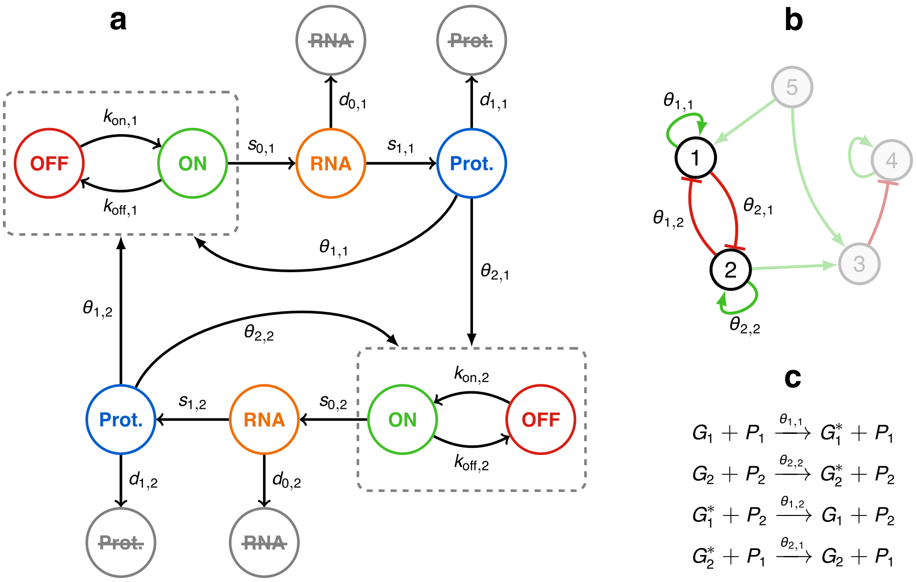
$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{\text{on},i}} 1, \;\; 1 \xrightarrow{k_{\text{off},i}} 0 \\ M_i{}'(t) = s_{0,i} E_i(t) - d_{0,i} M_i(t) \\ P_i{}'(t) = s_{1,i} M_i(t) - d_{1,i} P_i(t) \end{cases} \tag{2}$$

Still, genes have static parameters and do not interact with each other. To get an actual network, we need to go one step further: reactions $G_i \to G_i{}^*$ and $G_i{}^* \to G_i$ are not assumed to be elementary anymore, but rather represent complex reactions involving proteins so that promoter parameters $k_{\text{on},i}$ and $k_{\text{off},i}$ now depend on proteins (Fig. 2a), and a fortiori on time. Our network model will correspond to the explicit definition, for all gene i, of functions $k_{\text{on},i}(P_1, \ldots, P_n)$ and $k_{\text{off},i}(P_1, \ldots, P_n)$. These functions shall also depend on network-specific parameters quantifying the interactions, thus making the link between “fitting a chemical model” and “inferring a network”. As a toy example, consider replacing $G_i \to G_i{}^*$ with two parallel elementary reactions

$$G_i \xrightarrow{\theta_{i,0}} G_i{}^* \quad \text{and} \quad G_i + P_j \xrightarrow{\theta_{i,j}} G_i{}^* + P_j \tag{3}$$

for which applying the law of mass action directly gives $k_{\text{on},i}(P_1, \ldots, P_n) = \theta_{i,0} + \theta_{i,j} P_j$. In a regulatory network (Fig. 2b), it would correspond to adding a directed edge from gene j to gene i, with $\theta_{i,0}$ the basal parameter of gene i, and $\theta_{i,j}$ the strength of activation of gene i by protein P_j. We emphasize that the action of P_j on the promoter G_i is not necessarily direct. For example, P_j can instead indirectly modulate the amount/activity of a transcription factor: we suppose in this article that such hidden reactions are fast enough regarding gene expression dynamics so that protein P_j is a relevant proxy for the transcription factor. Moreover, although we assume here that interactions can only happen at the level of $k_{\text{on},i}$ and $k_{\text{off},i}$, mainly for identifiability purposes, it is also possible to make $d_{1,i}$ and $s_{1,i}$ depend on proteins without fundamentally changing the mathematical approach (e.g. see [39, 40]).

In order to simplify notations, we normalize model (2) into a dimensionless equivalent model: we rewrite it in



$$G_1 + P_1 \xrightarrow{\theta_{1,1}} G_1^* + P_1$$
$$G_2 + P_2 \xrightarrow{\theta_{2,2}} G_2^* + P_2$$
$$G_1^* + P_2 \xrightarrow{\theta_{1,2}} G_1 + P_2$$
$$G_2^* + P_1 \xrightarrow{\theta_{2,1}} G_2 + P_1$$

Fig. 2 Different views of the network model. **a** Two genes interacting with each other, forming a network. Interactions are assumed to arise from the dependence of promoter dynamics on protein quantities. **b** A higher level of abstraction leads to the traditional gene regulatory network representation. **c** A toy example of reactions defining the interactions between genes 1 and 2, making the link between representations (**a**) and (**b**)

terms of new variables $\overline{M}_i = \frac{d_{0,i}}{s_{0,i}} M_i$ and $\overline{P}_i = \frac{d_{0,i}d_{1,i}}{s_{0,i}s_{1,i}} P_i$, which have values between 0 and 1, and report this scale change in the definition of $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ (see section 1.1 of Additional file 1 for details). In the remainder of this article, the new variables will still be denoted by M_i and P_i as no confusion arises. The resulting normalized model is:

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{\mathrm{on},i}} 1, \quad 1 \xrightarrow{k_{\mathrm{off},i}} 0 \\ M_i{}'(t) = d_{0,i}\,(E_i(t) - M_i(t)) \\ P_i{}'(t) = d_{1,i}\,(M_i(t) - P_i(t)) \end{cases} \tag{4}$$

still omitting the dependence of $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ on $(P_1(t), \ldots, P_n(t))$ for clarity. This form enlightens the fact that $s_{0,i}$ and $s_{1,i}$ are just scaling constants: given a path $(E_i, M_i, P_i)_i$ of system (4), one can go back to the physical path by simply multiplying M_i by $(s_{0,i}/d_{0,i})$ and P_i by $(s_{0,i}/d_{0,i}) \times (s_{1,i}/d_{1,i})$.

Therefore, we get a general network model where each link between two genes is directed and has an explicit biochemical interpretation in terms of molecular interactions. The previous example is very simplistic but one can use virtually any model of chromatin dynamics to derive a form for $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$, involving hit-and-run reactions, sequential binding, etc. [41]. Such aspects are still far from being completely understood [42–45] and this simple network model can hopefully be used to assess biological hypotheses. In the next part, we will introduce a more sophisticated interaction form based on an underlying probabilistic model, which is both "statistics-friendly" and interpretable as a non-equilibrium steady state of chromatin environment [43].

Some known mathematical results

Thanks to some recent theoretical results [40, 46], simple sufficient conditions on $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ ensure that the PDMP network model (4) is actually well-defined and that the overall joint distribution of $(E_i, M_i, P_i)_i$ converges as $t \to +\infty$ to a unique stationary distribution, which will be the basis of our statistical approach. Namely, we assume in this article that $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ are continuous functions of $(P_1, \ldots, P_n)$ and that they are greater than some positive constants. These conditions are satisfied in most interesting cases, including the above toy example (3) when $\theta_{i,0} > 0$.

Contrary to creation rates $s_{0,i}$ and $s_{1,i}$, degradation rates $d_{0,i}$ and $d_{1,i}$ play a crucial role in the dynamics of the system. Intuitively, the ratios $(k_{\mathrm{on},i} + k_{\mathrm{off},i})/d_{0,i}$ and $d_{0,i}/d_{1,i}$ respectively control the buffering of promoter noise by mRNA and the buffering of mRNA noise by proteins. A common situation is when promoter and mRNA dynamics are fast compared to proteins, i.e. when $d_{0,i} \gg d_{1,i}$ with $(k_{\mathrm{on},i} + k_{\mathrm{off},i})/d_{0,i}$ fixed. At the limit, the promoter-mRNA noise is fully averaged by proteins and model (4) simplifies into a deterministic system [47]:

$$P_i{}'(t) = d_{1,i}\left(\frac{k_{\mathrm{on},i}(\mathbf{P}(t))}{k_{\mathrm{on},i}(\mathbf{P}(t)) + k_{\mathrm{off},i}(\mathbf{P}(t))} - P_i(t)\right) \tag{5}$$

where $\mathbf{P}(t) = (P_1(t), \ldots, P_n(t))$. The diffusion limit, which keeps a residual noise, can also be rigorously derived [48]. Unsurprisingly, one recovers the traditional way of modelling gene regulatory networks with Hill-type interaction functions. Equation 5 is useful to get an insight into the behaviour of the system (4) for given $k_{\text{on},i}$ and $k_{\text{off},i}$, yet it should be used with caution. Indeed, the $d_{0,i}/d_{1,i}$ ratio has been shown to span a high range, averaging out to the value $d_{0,i}/d_{1,i} \approx 5$ in mammalian cells [31], for which taking the limit $d_{0,i} \gg d_{1,i}$ is not obvious. This is consistent with recent single-cell experiments showing a high variability of both mRNA and protein levels between cells [37]. In that sense, the PDMP model is much more robust than its deterministic/diffusion counterpart while keeping a similar level of mathematical complexity, which motivates our approach.

Simulation

We propose a simple algorithm to compute sample paths of our stochastic network model (4). It consists in a hybrid version of a basic ODE solver, making it efficient enough to perform massive simulations on large scale networks involving arbitrary numbers of molecules, which would be intractable with a classic molecule-based model (Fig. 3). The deterministic part of the algorithm is a standard explicit Euler scheme, while the stochastic part is based on the transient promoter distribution for single genes: this can be justified by the fact that during a small enough time interval, proteins remain almost constant so genes behave as if $k_{\text{on},i}$ and $k_{\text{off},i}$ were constant. We therefore use Bernoulli steps, in a similar way of a diffusion being simulated using gaussian steps.

After discretizing time with step δt, the numerical scheme is as follows. Starting from an initial state $\left({E_i}^0, {M_i}^0, {P_i}^0\right)_i$, the update of the system from t to $t+\delta t$ is given by:

$$\begin{cases} {E_i}^{t+\delta t} \sim \mathcal{B}\left(\pi_i^t\right) \\ {M_i}^{t+\delta t} = (1-d_{0,i}\delta t){M_i}^t + d_{0,i}\delta t {E_i}^t \\ {P_i}^{t+\delta t} = (1-d_{1,i}\delta t){P_i}^t + d_{1,i}\delta t {M_i}^t \end{cases} \tag{6}$$

where the Bernoulli distribution parameter π_i^t is derived by locally solving the master equation for the promoter [49], i.e.

$$\pi_i^t = \frac{a_i^t}{a_i^t + b_i^t} + \left({E_i}^t - \frac{a_i^t}{a_i^t + b_i^t}\right) e^{-(a_i^t + b_i^t)\delta t}$$

with the notation $a_i^t = k_{\text{on},i}({P_1}^t, \ldots, {P_n}^t)$ and $b_i^t = k_{\text{off},i}({P_1}^t, \ldots, {P_n}^t)$. Intuitively, the algorithm is valid when $\delta t \ll 1/\max_i \left\{K_{\text{on},i}, K_{\text{off},i}, d_{0,i}, d_{1,i}\right\}$ where $K_{\text{on},i}$ and $K_{\text{off},i}$ denote the maximum values of functions $k_{\text{on},i}$ and $k_{\text{off},i}$.

Deriving a tractable statistical model

We will now adopt a statistical perspective in order to deal with gene network inference, considering a set of observed cells. If they are evolving in the same environment for a long enough time, we can reasonably assume that their mRNA and protein levels follow the stationary distribution of an underlying gene network: this distribution can be used as a statistical likelihood for the cells. Furthermore assuming no cell-cell interactions (which may of course depend on the experimental context), we obtain a standard statistical problem with independent samples. Since the stationary distribution of the stochastic network model (4) is well-defined but a priori not analytically tractable, we will derive an explicit approximation and then reduce our inference problem to a traditional likelihood-based estimation. We will do so in two cases: when there is no self-interaction, and for a specific form of auto-activation.

Separating mRNA and protein timescales

It is for the moment very rare to experimentally obtain the amount of proteins for many genes at the single-cell level. We will therefore assume here that only mRNAs are observed. To deal with this problem, we take the protein timescale as our reference by fixing $d_{1,i}$ and assume that promoter dynamics are faster than proteins, i.e. $(k_{\text{on},i} +$

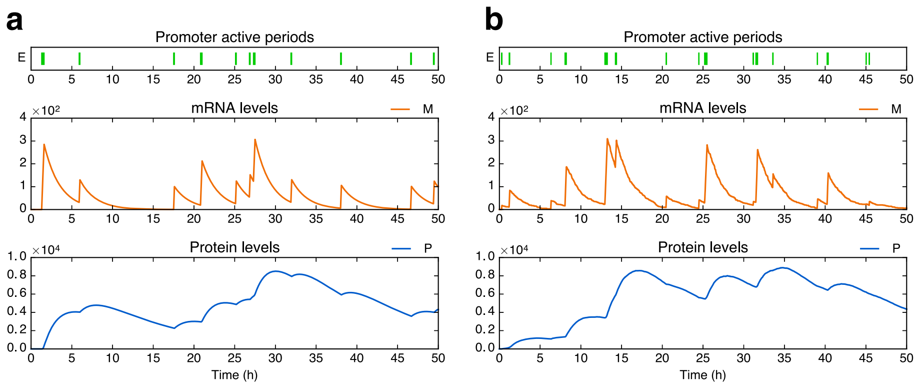


Fig. 3 Simulations of the two-state model for a single gene. **a** Sample path of the PDMP model using our hybrid numerical scheme (computation time ≈ 0.05 s). **b** Sample path of the classic model using exact stochastic simulation [27] (computation time ≈ 10 s). Parameters values are $k_{\text{on}} = 0.34$, $k_{\text{off}} = 10$, $s_0 = 10^3$, $s_1 = 10$, $d_0 = 0.5$ and $d_1 = 0.1$ (in h^{-1})

$k_{\mathrm{off},i}) \gg d_{1,i}$ in a biologically relevant way, say $(k_{\mathrm{on},i} + k_{\mathrm{off},i})/d_{1,i} > 10$ (thus the deterministic limit (5) does not necessarily hold). Furthermore, in line with several recent experiments [37, 50], we assume that $d_{0,i}$ is sufficiently larger than $d_{1,i}$ so that the correlation between mRNAs and proteins produced by the gene is very small: model (4) then can be reduced by removing mRNA and making proteins directly depend on the promoters (see section 1.2 of Additional file 1). The result is

$$\begin{cases} E_i(t) : 0 \xrightarrow{k_{\mathrm{on},i}} 1, \;\; 1 \xrightarrow{k_{\mathrm{off},i}} 0 \\ {P_i}'(t) = d_{1,i}\,(E_i(t) - P_i(t)) \end{cases} \tag{7}$$

which still admits the deterministic limit (5). Since mRNA dynamics are faster than proteins, one can also assume that, given protein levels $\mathbf{P} = (P_1, \ldots, P_n)$, each mRNA level M_i follows the quasi-steady state distribution

$$M_i \mid \mathbf{P} \sim \mathrm{Beta}\left(\frac{k_{\mathrm{on},i}(\mathbf{P})}{d_{0,i}}, \frac{k_{\mathrm{off},i}(\mathbf{P})}{d_{0,i}}\right) \tag{8}$$

corresponding to the single-gene model [28, 39] with constant parameters $k_{\mathrm{on},i}(\mathbf{P})$ and $k_{\mathrm{off},i}(\mathbf{P})$. Numerically, this approximation works well even for moderate values of $d_{0,i}$, such as $d_{0,i} = 5 \times d_{1,i}$ (see the "Results" section).

Biologically, Eqs. (7) and (8) suggest that correlations between mRNA levels may not directly arise from correlations between promoters *states* (which in fact are weak because of $(k_{\mathrm{on},i} + k_{\mathrm{off},i}) \gg d_{1,i}$), but rather originate from correlations between promoter *parameters* $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$, which themselves depend on the protein joint distribution.

Table 2 sums up the successive modelling steps introduced so far. From now on, we will always assume the form (8) for the mRNA distribution, and thus our model is reduced to Eq. (7) which only involves proteins.

Table 2 Successive dynamical models introduced in this article. We recall for each step the main feature and the form of the mRNA stationary distribution. The full network model (step 3) is used for simulations, while the simplified one (step 4) is used to derive the approximate statistical likelihood

1	*Single-gene, discrete* [29]	
	◇ All molecules are discrete	
	◇ mRNA distribution: Beta-Poisson	↓ Abundant species treated continuously
2	*Single-gene, PDMP* (1)	
	◇ Only the promoter is discrete	
	◇ mRNA distribution: Beta	↓ Introduction of interactions via k_{on}, k_{off}
3	*Network* (2), *normalized version* (4)	
	◇ Both accurate and fast to simulate	
	◇ mRNA distribution: unknown	↓ Timescale separation of Protein/mRNA ($d_0 \gg d_1$)
4	*Simplified network* (7)	
	◇ mRNA is removed from the network	
	◇ Conditional mRNA distribution: Beta (8)	

Hartree approximation

In this section, we present the Hartree approximation principle and provide an explicit formula in the particular case of no self-interaction. The simplified model (7) is still not analytically tractable, but it is now appropriate for employing the *self-consistent proteomic field* approximation introduced in [51, 52] and successfully applied in [53, 54]. More precisely, we will use its natural PDMP counterpart, which will be referred to as "Hartree approximation" since the main idea is similar to the Hartree approximation in physics [51]. It consists in assuming that genes behave as if they were independent from each other, but submitted to a common "proteomic field" created by all other genes. In other words, we transform the original problem of dimension 2^n into n independent problems of dimension 2 that are much easier to solve (see section 2 of Additional file 1 for details).

When $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ do not depend on P_i (i.e. no self-interaction), this approach results in approximating the protein stationary distribution of model (7) by the function

$$u(y) = \prod_{i=1}^{n} \frac{{y_i}^{a_i(y)-1}(1-y_i)^{b_i(y)-1}}{\mathrm{B}(a_i(y), b_i(y))} \tag{9}$$

where $y = (y_1, \ldots, y_n) = (P_1, \ldots, P_n) = \mathbf{P}$, $a_i(y) = k_{\mathrm{on},i}(y)/d_{1,i}$, $b_i(y) = k_{\mathrm{off},i}(y)/d_{1,i}$ and B is the standard Beta function. Note that promoter states have been integrated out since they are not required by Eq. (8).

The function u is a heuristic approximation of a probability density function. It is only valid when interactions are not too strong, that is, when $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ are close enough to constants, and it becomes exact when they are true constants. Besides, it does not integrate to 1 in general. However, this approximation turns out to be very robust in practice and it has the great advantage to be fully explicit (and significantly simpler than in the non-PDMP case), thus providing a promising base for a statistical model.

When $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ depend on P_i, one can still explicitly compute the Hartree approximation in many cases: we will give an example in the next section. Alternatively, it is always possible to use formula (9) even with self-interactions, giving a correct approximation when the feedback is not too strong, as for other proteins.

An explicit form for interactions

We now propose an explicit definition of functions $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$. Recent work [36, 55, 56] showed that apparent increased transcription actually reflects an increase in burst frequency rather than amplitude. We therefore decided to model only $k_{\mathrm{on},i}$ as an actual function and to keep $k_{\text{off},i}$ constant. In this view, the activation frequency of a gene can be influenced by ambiant proteins, whereas the active periods have a random duration that is dictated only by an intrinsic stability constant of the transcription machinery.

Our approach uses a description of the molecular activity around the promoter in a very similar way as Coulon et al. [42]. Accordingly, we make a quasi-steady state assumption to obtain $k_{\text{on},i}$. This idea based on thermodynamics was also used in the DREAM3 in-Silico Challenge [57] to simulate gene networks. However, only mean transcription rate was described (instead of promoter activity in our work), which is inappropriate to model bursty mRNA dynamics at the single-cell level.

We herein derive $k_{\text{on},i}$ from an underlying stochastic model for chromatin dynamics. We first introduce a set of abstract chromatin states, each state being associated with one of two possible rates of promoter activation, either a low rate $k_{0,i}$ or a high rate $k_{1,i} \gg k_{0,i}$. More specifically, such chromatin states may be envisioned as a coarse-grained description of the chromatin-associated parameters that are critical for transcription of gene i. Second, we assume a separation of timescales between the abstract chromatin model and the promoter activity, so that the promoter activation reaction depends only on the quasi-steady state of chromatin. In other words, the effective $k_{\text{on},i}$ is a combination of $k_{0,i}$ and $k_{1,i}$ which integrates all the chromatin states: its value depends on the probability of each state and a fortiori on the transitions between them. We propose a transition scheme which leads to an explicit form for $k_{\text{on},i}$, based on the idea that proteins can alter chromatin by hit-and-run reactions and potentially introduce a memory component. Some proteins thereby tend to stabilize it either in a "permissive" configuration (with rate $k_{1,i}$) or in a "non-permissive" configuration (with rate $k_{0,i}$), providing notions of *activation* and *inhibition*. A more precise definition and details of the derivation are provided in section 3 of Additional file 1.

The final form is the following. First, we define a function of every protein but P_i,

$$\Phi_i(y) = \exp(\theta_{i,i}) \prod_{j \neq i} \frac{1 + \exp(\theta_{i,j})(y_j/s_{i,j})^{m_{i,j}}}{1 + (y_j/s_{i,j})^{m_{i,j}}}$$

which may represent the external input of gene i. Then, $k_{\text{on},i}$ is defined by

$$k_{\text{on},i}(y) = \frac{k_{0,i} + k_{1,i}\Phi_i(y)(y_i/s_{i,i})^{m_{i,i}}}{1 + \Phi_i(y)(y_i/s_{i,i})^{m_{i,i}}}. \tag{10}$$

Hence, when the input $\Phi_i(y)$ is fixed, $k_{\text{on},i}$ is a standard Hill function which describes how gene i is self-activating, depending on the Hill coefficient $m_{i,i}$ (Fig. 4). The neutral value is set to $\Phi_i(y) = 1$, so that for this particular value, $s_{i,i}$ is the usual dissociation constant. Moreover, if $\theta_{i,j} = 0$ for all $j \neq i$, then Φ_i becomes the constant function $\Phi_i(y) = \exp(\theta_{i,i})$, and thus $\theta_{i,i}$ may be seen as a "basal" parameter, summing up all potential hidden inputs. On the contrary, if some $\theta_{i,j} > 0$ (resp. $\theta_{i,j} < 0$), then Φ_i becomes itself an increasing (resp. decreasing) Hill-type function of protein P_j, where $m_{i,j}$ and $s_{i,j}$ again play their usual roles.

The $n \times n$ matrix $\theta = (\theta_{i,j})$ therefore plays the same role as the interaction matrix in traditional network inference frameworks [8]. For $i \neq j$, $\theta_{i,j}$ quantifies the regulation of gene i by gene j (activation if $\theta_{i,j} > 0$, inhibition if $\theta_{i,j} < 0$, no influence if $\theta_{i,j} = 0$), and the diagonal term $\theta_{i,i}$ aggregates the "basal input" and the "self-activation strength" of gene i. Note that self-inhibition could be considered instead, but the choice has to be made before the inference since the self-interaction form is notoriously difficult to identify, especially in the stationary regime. In the remainder of this article, we assume that parameters $k_{0,i}$, $k_{1,i}$, $m_{i,j}$ and $s_{i,j}$ are known and we focus on inferring the matrix θ.

A benefit of the interaction form (10) is to allow for a fully explicit Hartree approximation of the protein distribution (see section 3 of Additional file 1 for details). In particular, if $m_{i,i} > 0$ and $c_i = (k_{1,i} - k_{0,i})/(d_{1,i}m_{i,i})$ is a

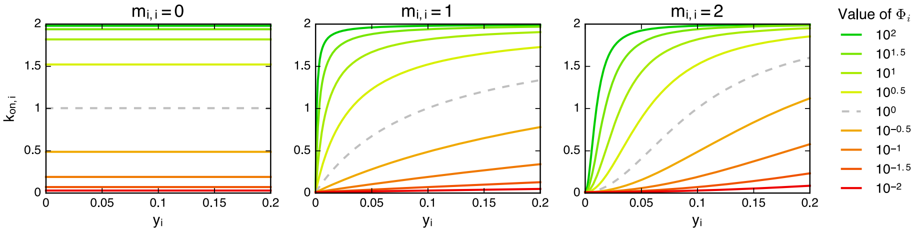


Fig. 4 Different auto-activation types in the network model. Each color corresponds to a fixed value of Φ_i in formula (10), and each curve represents $k_{\text{on},i}$ as a function of y_i for $m_{i,i} = 0$ (no feedback), $m_{i,i} = 1$ (monomer-type feedback) and $m_{i,i} = 2$ (dimer-type feedback). The neutral value $\Phi_i = 1$ is represented by a dashed gray line. Here $k_{0,i} = 0.01$, $k_{1,i} = 2$ and $s_{i,i} = 0.1$

positive integer for all i, the approximation is given by

$$u(y) = \prod_{i=1}^{n} \left(\sum_{r=0}^{c_i} p_{i,r}(y) \frac{{y_i}^{a_{i,r}-1}(1-y_i)^{b_i-1}}{\mathrm{B}(a_{i,r}, b_i)} \right) \tag{11}$$

with $a_{i,r} = ((c_i - r)k_{0,i} + rk_{1,i})/(d_{1,i}c_i)$, $b_i = k_{\mathrm{off},i}/d_{1,i}$ and

$$p_{i,r}(y) = \frac{\binom{c_i}{r}\, \mathrm{B}(a_{i,r}, b_i)(\Phi_i(y)/s_{i,i}^{m_{i,i}})^r}{\sum_{r'=0}^{c_i} \binom{c_i}{r'}\, \mathrm{B}(a_{i,r'}, b_i)(\Phi_i(y)/s_{i,i}^{m_{i,i}})^{r'}}.$$

In other words, the Hartree approximation (11) is a product of gene-specific distributions which are themselves mixtures of Beta distributions: for gene i, the $a_{i,r}$ correspond to "frequency modes" ranging from $k_{0,i}$ to $k_{1,i}$, weighted by the probabilities $p_{i,r}(y)$. It is straightforward to check that inhibitors tend to select the low burst frequencies of their target ($a_{i,r} \approx k_{0,i}$) while activators select the high frequencies ($a_{i,r} \approx k_{1,i}$). If $m_{i,i} = 0$ for some i, then $k_{\mathrm{on},i}$ does not depend on P_i so one just has to replace the i-th term in the product (11) with the single Beta form as in Eq. (9), which is equivalent to taking the limit $c_i \to +\infty$. Finally, when $m_{i,i} > 0$ but c_i is not an integer, using $\lceil c_i \rceil$ instead keeps a satisfying accuracy.

The statistical model in practice

Our statistical framework simply consists in combining the timescale separation (8) and the Hartree approximation (11) into a standard hidden Markov model. Indeed, conditionally to the proteins, mRNAs are independent and follow well-defined Beta distributions

$$v(x,y) = \prod_{i=1}^{n} \frac{{x_i}^{\widetilde{a}_i(y)-1}(1-x_i)^{\widetilde{b}_i(y)-1}}{\mathrm{B}(\widetilde{a}_i(y), \widetilde{b}_i(y))} \tag{12}$$

where $x = (x_1, \ldots, x_n) = (M_1, \ldots, M_n) = \mathbf{M}$, $\widetilde{a}_i(y) = k_{\mathrm{on},i}(y)/d_{0,i}$ and $\widetilde{b}_i(y) = k_{\mathrm{off},i}(y)/d_{0,i}$. Then one can use (11) to approximate the joint distribution of proteins. Hence, recalling the unknown interaction matrix θ, the inference problem for m cells with respective levels $(\mathbf{M}_k, \mathbf{P}_k)_{1 \leqslant k \leqslant m}$ is based on the (approximate) complete log-likelihood:

$$\begin{aligned} \ell &= \ell(\mathbf{M}_1, \ldots, \mathbf{M}_m, \mathbf{P}_1, \ldots, \mathbf{P}_m | \theta) \\ &= \sum_{k=1}^{m} \log(u(\mathbf{P}_k)) + \log(v(\mathbf{M}_k, \mathbf{P}_k)) \end{aligned} \tag{13}$$

where we used conditional factorization and independence of the cells.

The basic statistical inference problem would be to maximize the marginal likelihood of mRNA with respect to θ. Since this likelihood has no simple form, a typical way to perform inference is to use an Expectation-Maximization (EM) algorithm on the complete likelihood (13). However, the algorithm may be slow in practice because of the computation of expectations over proteins. A faster procedure consists in simplifying these expectations using the distribution modes: the resulting algorithm is often called "hard EM" or "classification EM" and is used in the "Results" section. Moreover, it is possible to encode some potential knowledge or constraints on the network by introducing a prior distribution $w(\theta)$. In this case, from Baye's rule, one can perform *maximum a posteriori* (MAP) estimation of θ by using the same EM algorithm but adding the penalization term $\log(w(\theta))$ to ℓ during the Maximization step (see section 4 of Additional file 1 and the "Results" section). Alternatively, a full bayesian approach, i.e. sampling from the posterior distribution of θ conditionally to $(\mathbf{M}_1, \ldots, \mathbf{M}_m)$, may also be considered using standard MCMC methods.

Taking advantage of the latent structure of proteins, we can also deal with missing data in a natural way: if the mRNA measurement of gene i is invalid in a cell k owing to technical problems, it is possible to ignore it by removing the i-th term in the conditional distribution of mRNAs (12). This only modifies the definition of v for cell k in Eq. (13), ensuring that all valid data is effectively used for each cell.

Results

In this part, we first compare the distribution of the mechanistic model (4) to the mRNA quasi-steady state combined with Hartree approximation for proteins, on a simple toggle-switch example. Then, we show that the single-gene model with auto-activation can fit marginal mRNA distributions from real data better than the constant-k_{on} model. Finally, we successfully apply the inference procedure to various two-gene networks simulated from the mechanistic model.

Relevance of the approximate likelihood

Starting from the normalized mechanistic model (4), two approximations were used to derive the final statistical likelihood (13): the quasi-steady state assumption for mRNAs given protein levels, and the Hartree approximation for the joint distribution of proteins. Crucially, this approximate likelihood has to be close enough to the exact one in order to preserve the equivalence between inferring a network and fitting the mechanistic model. To get an idea of the accuracy, we considered a basic two-gene toggle switch defined by $k_{\mathrm{on},i}$ following Eq. (10) with the interaction matrix given by $\theta_{1,1} = \theta_{2,2} = 4$ and $\theta_{1,2} = \theta_{2,1} = -8$ (full parameter list in section 6 of Additional file 1). By computing sample paths (Fig. 5), we estimated the stationary distribution and compared it with our approximation, which appeared to be very satisfying, both for proteins and mRNAs (Fig. 6).

Fitting marginal mRNA distributions from real data

A particularity of single-cell data is to often exhibit bursty regimes for mRNA (meaning $k_{\mathrm{on}} \ll k_{\mathrm{off}}$ and $d_0 \ll k_{\mathrm{off}}$) and potentially also for proteins (adding $d_1 \ll k_{\mathrm{off}}$), which

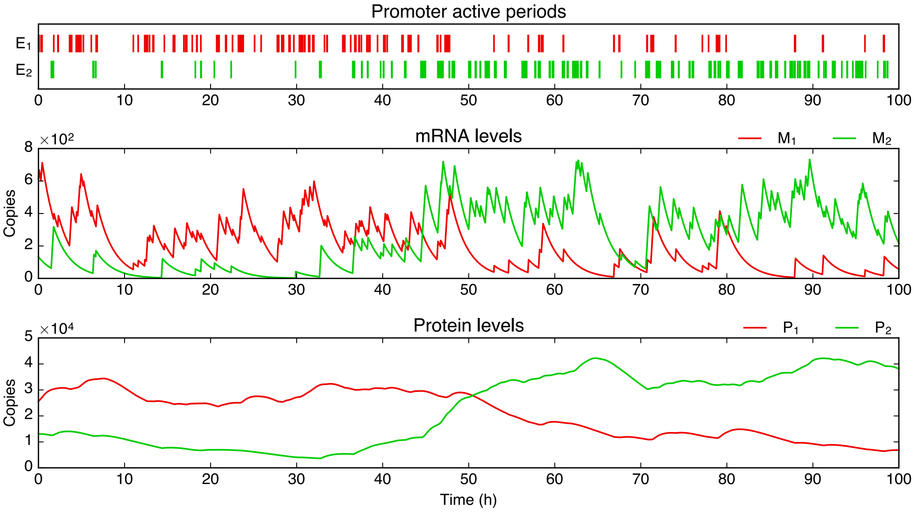


Fig. 5 Sample path of a two-gene toggle switch. The first gene is plotted in red and the second in green. While always staying in a bursty regime regarding mRNAs, genes can switch between high and low frequency modes (here at $t \approx 50$ h). From this example, it is clear that the overall joint distribution can contain correlations even if the bursts themselves are not coordinated

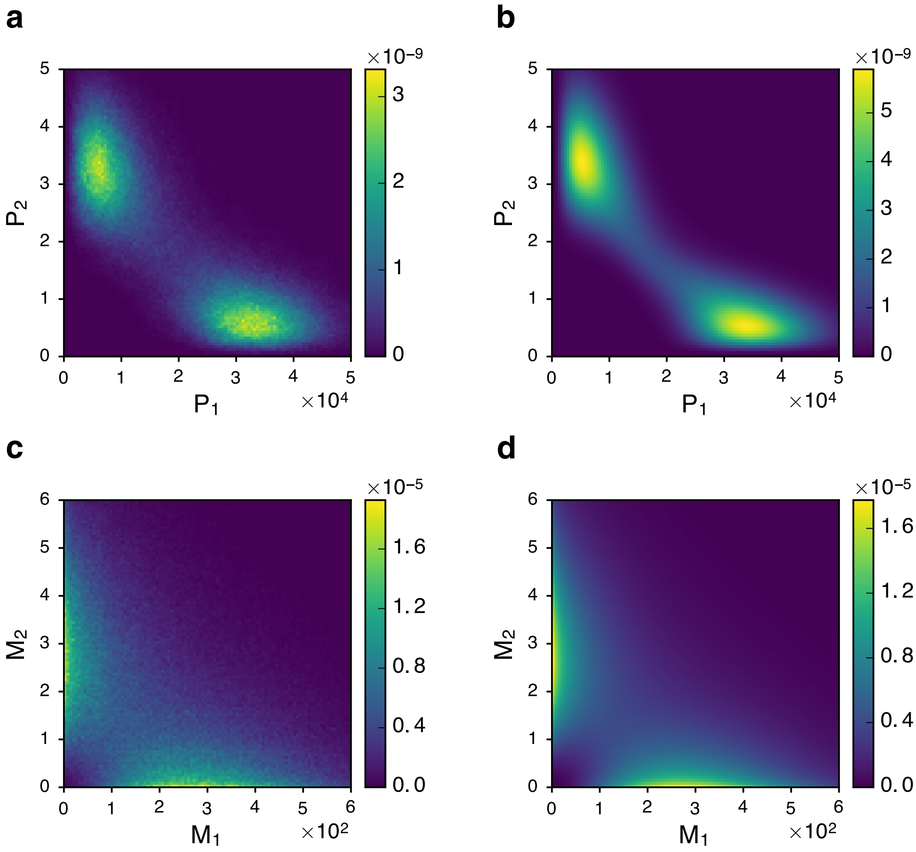


Fig. 6 Exact and approximate stationary distributions for the example of toggle switch. True distributions (left side) were estimated by sample path simulation, while approximations (right side) have explicit formulas. **a** True distribution of proteins. **b** Approximate distribution of proteins, from formula (11). **c** True distribution of mRNAs. **d** Approximate distribution of mRNAs, obtained by integrating the conditional distribution of mRNA (12) against (**b**)

are well fitted by Gamma distributions [37]. At this stage, it is worth mentioning that the Gamma distribution can be seen as a limit case of the Beta distribution. Intuitively, when $b \gg 1$ and $b \gg a$ (typically $a = k_{\mathrm{on}}/d_0$ and $b = k_{\mathrm{off}}/d_0$), most of the mass of the distribution $\mathrm{Beta}(a, b)$ is located at $x \ll 1$ so we have the first order approximation

$$\begin{aligned} x^{a-1}(1-x)^{b-1} &= x^{a-1}\exp((b-1)\log(1-x)) \\ &\approx x^{a-1}\exp(-bx) \end{aligned}$$

and thus $\mathrm{Beta}(a, b) \approx \gamma(a, b)$. This way, formulas (11) and (12) can be easily transformed into Gamma-based distributions. Parameters s_0 and k_{off} then aggregate in k_{off}/s_0 because of the scaling property of the Gamma distribution, so only this ratio has to be inferred: from an applied perspective, it simply represents a scale parameter for each gene. This remark leads to a possible preprocessing phase that can be used for estimating the crucial basal parameters of the network, without requiring the knowledge of such scale parameters (see section 5 of Additional file 1).

In addition, our network model is able to generate multiple modes while keeping such bursty regimes (Fig. 5), as noticeable in the stationary distribution (11). Interestingly, this feature has already been considered in the literature by empirically introducing mixture distributions [58, 59]. As a first step toward applications, we compared our model in the simplest case (independent genes with auto-activation) to marginal distributions of single-cell mRNA measurements from [38]. Our model was fitted and compared to the basic two-state model in the bursty regime, i.e. to a simple Gamma distribution: Fig. 7 shows the example of the LDHA gene. Although very close when viewed in raw molecule numbers, the distributions differ after applying the transformation $x \mapsto x^{\alpha}$ with $\alpha = 1/3$, which tends to compress great values while preserving small values. The data becomes bimodal, suggesting the presence of two bursting regimes, a "normal" one and a very small "inhibited" one: the auto-activation model then performs better than the simple Gamma, which necessarily stays unimodal for $0 < \alpha < 1$. Note that the RTqPCR protocol used in [38] was shown to be far more sensitive than single-cell RNA-seq in the detection of low abundance transcripts [60]. Since the data also contains small nonzero values, this tends to support a true biological origin for the peak in zero. Besides, the case of distributions that are not bimodal until transformed also arises for proteins [61].

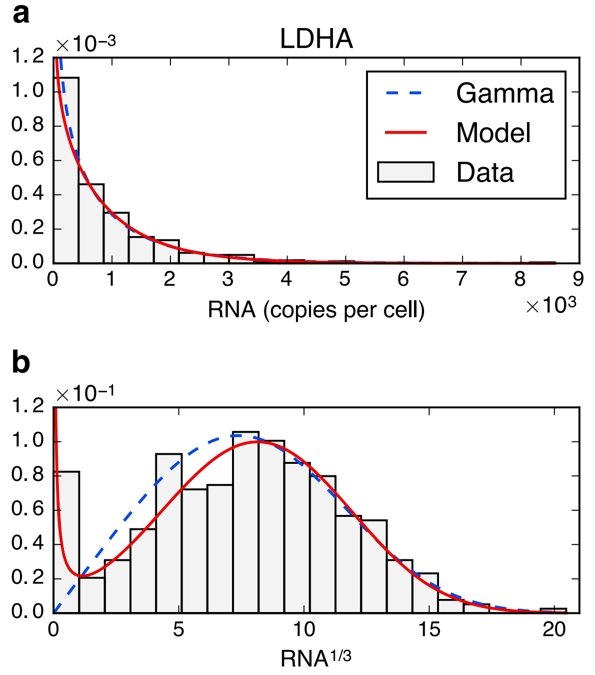


Fig. 7 Fitting marginal distributions from real single-cell data: example of the LDHA gene. The red curve is the stationary distribution associated with our interaction form (here a single gene with auto-activation), while the dashed blue curve corresponds to the basic two-state model in the bursty regime (Gamma distribution). **a** The raw data seems to be well fitted by the Gamma distribution, which in this view is close to our model. **b** Same fit viewed after applying the transformation $x \mapsto x^{1/3}$. The data becomes bimodal and the fit appears to be better with the auto-activation model

Application of the inference procedure

By construction of the mechanistic model, the interaction matrix θ can describe any oriented graph by explicitly defining causal quantitative links between genes, which is difficult to do within traditional statistical frameworks (e.g. bayesian networks or undirected Markov random fields). The logical downside is that identifiability issues seem inevitable. In a first attempt to assess this aspect, we implemented the inference method presented above and tested it on various two-gene networks, assuming auto-activation for each gene (i.e. $m_{i,i} > 0$) with Eq. (10) to maximize variability without considering perturbations of the system (parameter list in section 6 of Additional file 1).

We decided to investigate the worst case scenario in terms of cell numbers. We are fully aware of the existence of technologies allowing to interrogate thousands of cells simultaneously, but most of the recent studies still rely upon a much smaller number of cells. For each network, we therefore simulated mRNA snapshot data for 100 cells using the full PDMP model (4). We then inferred the matrix θ using a "hard EM" algorithm based on the likelihood (13), that is, alternatively maximizing the likelihood with respect to θ and with respect to the (unknown) protein levels of each cell. A lasso-like penalization term, corresponding to a prior distribution, was added to the $\theta_{i,j}$ for $i \neq j$ to obtain true zeros – so that the inferred network topology is clear – and to prevent keeping both $\theta_{i,j}$

and $\theta_{j,i}$ when one is significantly weaker (see section 4 of Additional file 1 for details of the penalization and the whole procedure).

We obtained highly encouraging results since every structure was inferred with a high probability of success (Fig. 8), meaning that the non-diagonal (i.e. interaction) terms of θ had the right sign and were nonzero at the right places. A list of the inferred values is provided in Additional file 1: Table S3. It is very important at that stage to emphasize that we are not trying to infer θ exactly: we only assess whether it has a zero or nonzero value and its sign. Although the results tend to support the identifiability of the full matrix θ in this simple two-gene case, one has to be aware that the quantity we maximize (an approximate likelihood) is a priori non convex and can have several local maxima (i.e. networks that are relevant

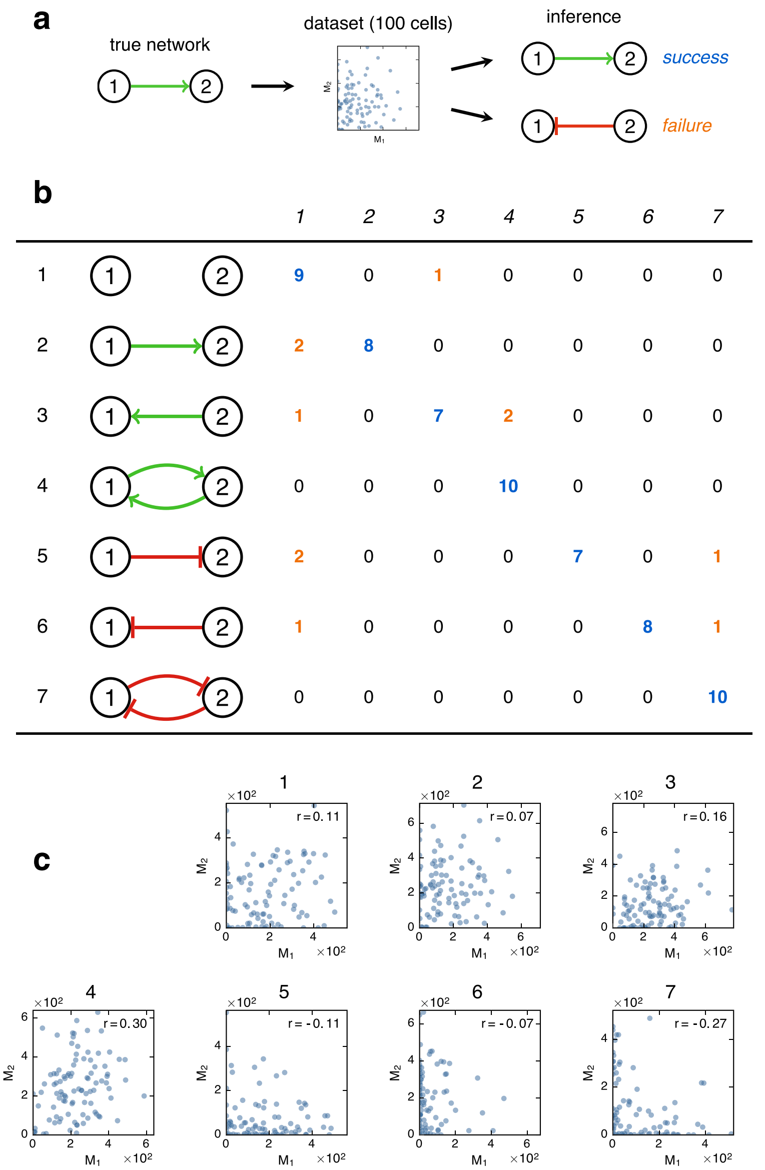


Fig. 8 Testing our inference method on simple networks. **a** For each network, numbered from 1 to 7, we simulated 100 cells using the full mechanistic model until the stationary regime was reached. Then we took a snapshot of their mRNA levels and inferred the parameters from this data. The result was called successful when the inferred structure (topology and nature of the links) was the same as the true network. **b** For each network (rows), 10 datasets were simulated and the results were reported by counting the number of inferred θ corresponding to each structure (columns), highlighting successes (blue) and failures (orange). The perfect inference would lead to 10 for all the diagonal terms and 0 everywhere else. **c** Examples of simulated mRNA datasets (one for each network). Although having coherent signs, Pearson's correlation coefficients (top right of each plot) would clearly be insufficient to distinguish between the different networks

candidates to explain the data). The result of the inference thus can depend on the starting point: in this first approach we chose the null matrix to be the starting point for θ, which corresponds to the – biologically relevant – expectation of "balanced" behaviors (e.g. we do not expect $\theta_{1,1} \ll \theta_{2,2}$). Alternatively, one can consider some probabilistic prior knowledge on θ to implement a (possibly rough) idea of parameter values from a Bayesian viewpoint: it is worth mentioning that any knockout information can be implemented this way in our model.

Finally, we assessed the inference behavior in the presence of dropouts, i.e. genes expressed at a low level in a cell that give rise to zeros after measurement [4]. Our first tests tend to indicate that our approach is robust regarding dropouts, in the sense that up to 30% of simulated dropouts does not drastically affect the estimation of θ once the other parameters have been estimated correctly (see Additional file 1: Table S4 for an example).

Discussion

In this paper, we introduce a general stochastic model for gene regulatory networks, which can describe bursty gene expression as observed in individual cells. Instead of using ordinary differential equations, for which cells would structurally all behave the same way, we adopt a more detailed point of view including stochasticity as a fundamental component through the two-state promoter model. This model is but a simplification of the complexity of the real molecular processes [42]. Modifications have been proposed, from the existence of a refractory period [23] to its attenuation by nuclear buffering [62]. In bacteria, the two states originate from the accumulation of positive supercoiling on DNA which stops transcription [63]. In eukaryotes, although its molecular basis is not quite understood, the two-state model is a remarkable compromise between simplicity and the ability to capture real-life data [18, 22, 36–38]. Our PDMP framework appears to be conceptually very similar to the *random dynamical system* proposed in [64] but it has two major advantages: time does not have to be discretized, and the mathematical analysis is significantly easier. We also note that a similar framework appears in [65, 66] and that a closely related PDMP – which can be seen as the limit of our model for infinitely short bursts – has recently been described in [67].

We then derive an explicit approximation of the stationary distribution and propose to use it as a statistical likelihood to infer networks from single-cell data. The main ingredient is the separation of three physical timescales – chromatin, promoter/RNA, and proteins – and the core idea is to use the self consistent proteomic field approximation from [51, 52] in a slightly simpler mathematical framework, providing fully explicit formulas that make possible the massive computations usually needed for parameter inference. From this viewpoint, it is a rather simple approach and we hope it can be adapted or improved in more specific contexts, for example in the study of lineage commitment [68]. Besides, the main framework does not necessarily has to include an underlying chromatin model and thus it can in principle also be used to describe gene networks in procaryotes.

Mechanistic modelling and statistical inference

An important quality of the PDMP network model is that the simulation algorithm is comparable in speed with classic ODE and diffusion systems, while providing an effective approximation of the "perfect", fully discrete, molecular counterpart [33, 35]. It is worth noticing that the PDMP – at least the promoter-mRNA system – naturally appears as an example of Poisson representation [28, 69], that is, not a simple approximation but rather the core component of the *exact* distribution of the discrete molecular model. Furthermore, such a simulation speed allowed us to compare our approximate likelihood with the true likelihood for a simple two-gene toggle switch, giving excellent results (Fig. 6). This obviously does not constitute a proof of robustness for every network: a proper quantitative (theoretical or numeric) comparison is beyond the scope of this article but would be extremely valuable. Intuitively, it should work for any number of genes, provided that interactions are not too strong.

Besides, some widely used ODE frameworks [8, 17, 57] can be seen as the fast-promoter limit of the PDMP model: this limit may not always hold in practice, especially in the bursty regime. In particular, Fig. 5 highlights the risk of using mRNA levels as a proxy for protein levels. It also explains why ordering single-cell mRNA measurements by pseudo-time may not always be relevant, as found in [38]. In [70], the authors use a hybrid model of gene expression to infer regulatory networks: it is very close to the diffusion limit of our reduced model (7) with the difference that the discrete component, called "promoter" by the authors, would correspond to the "frequency mode" in the present article, as visible for proteins in Fig. 5. From such a perspective, our approach adds a description of bursty mRNA dynamics that allows for fitting single-cell data such as in Fig. 7.

Finally, our method performed well for simple two-gene networks (Fig. 8), showing that part of the causal information remains present in the stationary distribution: this suggests that it is indeed possible to retrieve network structures with a mechanistic interpretation, even from bursty mRNA data.

Perspectives

We focused here on presenting the key ideas behind the general network model and the inference method: the logical next step is to apply it to real data and with a larger number of genes, which is the subject of work in progress in our group. In particular, we propose a functional preprocessing phase, detailed in section 5 of Additional file 1, that only requires the knowledge of the ratio $d_{0,i}/d_{1,i}$ to estimate all the relevant parameters before inferring θ. The ratio between protein and mRNA degradation rates (or half-lives) hence appears to be the minimum required for such a mechanistic approach to be relevant. Depending upon the species, mRNA and protein half-lives values can be found in the literature (see e.g. [31] for human proteins half-lives), or should be estimated from ad hoc experiments.

From a computational point of view, the main challenge is the algorithmic complexity induced by the fact that proteins are not observed and have to be treated as latent variables. There is a priori no possibility of reducing this without loosing too much accuracy, and therefore some finely optimized algorithms may be required to make the method scalable. Furthermore, the identifiability properties of the interaction matrix θ seem difficult to derive theoretically. In this paper we focused on the stationary distribution for simplicity: importantly, several aspects such as time dependence (computing the Hartree approximation in transitory regime) or perturbations (changing the cell's medium or performing knockouts [71], which can be naturally embedded in our framework) could greatly improve the practical identifiability.

From a biological point of view, our model does not really describe individual cells but rather a concatenation of trajectories obtained by following cells throughout divisions. Experiments suggest that it should be a relevant approximation, providing one considers mRNA and proteins levels in terms of concentrations instead of molecule numbers [72], which is made possible by the PDMP framework. In this view, the cell cycle results in increasing the apparent degradation rates – because of the increase in cell volume followed by division – and thus plays a crucial role for very stable proteins. However, at such a description level, many aspects of possible compensation mechanisms [73] and chromatin dynamics [74] remain to be elucidated. Regarding the latter aspect, our abstract chromatin states were not modeled from real-life data – chromatin composition for instance – but our approach is relevant in that partitioning into dual-type chromatin states as we did is now known as a pervasive feature of all eukaryotic genomes [75–78].

Conclusions

Protein and mRNA measurements in individual cells have revealed the importance of stochasticity in gene expression, which may potentially affect many aspects of gene regulation within cells. The traditional paradigm of gene network dynamics consisting in a deterministic structure plus an external noise – historically based on population-averaged data – should therefore be questioned, as such a noise appears to be itself part of the network structure and far from a small perturbation.

By modelling gene networks using piecewise-deterministic Markov processes, which are a simple way to introduce the minimum amount of mechanistic, non-diffusive stochasticity (corresponding to low molecule numbers), we derived a likelihood-based statistical model with interpretable parameters that successfully describes single-cell expression data. Our first results show that oriented interactions can indeed be inferred using such a method. Hence, this type of approach may take gene network inference to the next level by optimally exploiting single-cell data and improving the physical interpretability of inferred networks.

Additional file

Additional file 1: Additional file 1. Supplementary information. This document contains details of the theoretical derivations and all the parameter values used in the examples. (PDF 362 kb)

Acknowledgements
We thank Geneviève Fourel (LBMC/ENSL) for critical reading of the manuscript and Anne-Laure Fougères (ICJ) for her constant support. We also thank all members of the SBDM and Dracula teams for enlightening discussions, and BioSyL Federation and Ecofect Labex for inspiring scientific events.

Funding
This work was supported by funding from the Institut Rhônalpin des Systèmes Complexes (IXXI) and from the French agency ANR (ICEBERG; ANR-IABI-3096).

Availability of data and material
The data used to obtain Fig. 7 is available from [38]. The inference method was implemented in Scilab and the code is available upon request.

Authors' contributions
UH, AB, TE and OG designed the study. UH performed the theoretical derivations, implemented the algorithms and conceived/analyzed the examples. UH, AB, TE and OG interpreted the results and wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1] Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 allée d'Italie Site Jacques Monod, F-69007 Lyon, France. [2] Inria Team Dracula, Inria Center Grenoble Rhône-Alpes, Lyon, France. [3] Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-6962 Villeurbanne Cedex, France. [4] The CoSMo company, 5 passage du Vercors, 69007 Lyon, France.

Received: 12 May 2017 Accepted: 9 November 2017
Published online: 21 November 2017

References

1. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. BioSystems. 2009;96(1):86–103.
2. Kanter I, Kalisky T. Single cell transcriptomics: methods and applications. Front Oncol. 2015;5:53.
3. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82.
4. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34(11):1145–1160.
5. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. Development. 2009;136(23):3853–3862.
6. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010;467(7312):167–173.
7. Dueck H, Eberwine J, Kim J. Variation is function: Are single cell differences functionally important? Bioessays. 2015;38:172–180.
8. Mizeranschi A, Zheng H, Thompson P, Dubitzky W. Evaluating a common semi-mechanistic mathematical model of gene-regulatory networks. BMC Syst Biol. 2015;9(5):1–12.
9. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I. Scode: An efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. Bioinformatics. 2017;33(15):2314–2321.
10. Symmons O, Raj A. What's luck got to do with it: Single cells, multiple fates, and biological nondeterminism. Mol Cell. 2016;62(5):788–802.
11. Munsky B, Trinh B, Khammash M. Listening to the noise: random fluctuations reveal gene network parameters. Mol Syst Biol. 2009;5(1):1–7.
12. Zimmer C, Sahle S, Pahle J. Exploiting intrinsic fluctuations to identify model parameters. IET Syst Biol. 2015;9(2):64–73.
13. Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol. 2013;9(5):1–13.
14. Djordjevic D, Yang A, Zadoorian A, Rungrugeecharoen K, Ho JWK. How difficult is inference of mammalian causal gene regulatory networks? PLoS One. 2014;9(11):1–10.
15. Angulo MT, Moreno JA, Lippner G, Barabási A-L, Liu Y-Y. Fundamental limitations of network reconstruction from temporal data. J R Soc Interface. 2017;14(127):1–6.
16. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, Nishikawa S-I, Piterman N, Kouskoff V, Theis FJ, Fisher J, Göttgens B. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015;33(3):1–8.
17. Ocone A, Haghverdi L, Mueller NS, Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. Bioinformatics. 2015;31(12):89–86.
18. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 2013;14:7.
19. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9(8):796–804.
20. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science. 2004;304(5678):1811–1814.
21. Becskei A, Kaufmann BB, van Oudenaarden A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet. 2005;37(9):937–944.
22. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biology. 2006;4(10):1707–1719.
23. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332(6028):472–474.
24. Ko MSH. A stochastic model for gene induction. J Theor Biol. 1991;153: 181–194.
25. Ko MSH, Nakauchi H, Takahashi N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. EMBO J. 1990;9(9): 2835–2842.
26. Larson DR. What do expression dynamics tell us about the mechanism of transcription? Curr Opin Genet Dev. 2011;21(5):591–599.
27. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–2361.
28. Dattani J, Barahona M. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. J R Soc Interface. 2017;14(126):1–20.
29. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. PNAS. 2008;105(45):17256–17261.
30. Iyer-Biswas S, Hayot F, Jayaprakash C. Stochasticity of gene products from transcriptional pulsing. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;79:1–9.
31. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;495:337–342.
32. Davis MHA. Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. J R Stat Soc. 1984;46(3):353–388.
33. Crudu A, Debussche A, Radulescu O. Hybrid stochastic simplifications for multiscale gene networks. BMC Syst Biol. 2009;3(1):89.
34. Crudu A, Debussche A, Muller A, Radulescu O. Convergence of stochastic gene networks to hybrid piecewise deterministic processes. Ann Appl Probab. 2012;22(5):1822–1859.
35. Lin YT, Galla T. Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models. J R Soc Interface. 2016;13:1–11.
36. Viñuelas J, Kaneko G, Coulon A, Vallin E, Morin V, Mejia-Pous C, Kupiec J-J, Beslon G, Gandrillon O. Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. BMC Biol. 2013;11(1):15.
37. Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M, Tay S. Digital quantification of proteins and mrna in single mammalian cells. Mol Cell. 2016;61:914–924.
38. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, Guillemin A, Papili Gao N, Gunawan R, Cosette J, Arnaud O, Kupiec J-J, Espinasse T, Gonin-Giraud S, Gandrillon O. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. PLoS Biol. 2016;14(12):1–35.
39. Boxma O, Kaspi H, Kella O, Perry D. On/Off Storage Systems with State-Dependent Input, Output, and Switching Rates. Probab Eng Inf Sci. 2005;19:1–14.
40. Benaïm M, Le Borgne S, Malrieu F, Zitt P-A. Quantitative ergodicity for some switched dynamical systems. Electron Commun Probab. 2012;17(56):1–14.
41. Ong KM, Blackford, JA Jr, Kagan BL, Simons, SS Jr, Chow CC. A theoretical framework for gene induction and experimental comparisons. PNAS. 2010;107(15):7107–7112.
42. Coulon A, Gandrillon O, Beslon G. On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter. BMC Syst Biol. 2010;4:2.
43. Coulon A, Chow CC, Singer RH, Larson DR. Eukaryotic transcriptional dynamics: from single molecules to cell populations. Nat Rev Genet. 2013;14(8):1–13.
44. Friedman N, Rando OJ. Epigenomics and the structure of the living genome. Genome Res. 2015;25(10):1482–1490.
45. Bintu L, Yong J, Antebi YE, McCue K, Kazuki Y, Uno N, Oshimura M, Elowitz MB. Dynamics of epigenetic regulation at the single-cell level. Science. 2016;351(6274):720–724.
46. Benaïm M, Le Borgne S, Malrieu F, Zitt P-A. Qualitative properties of certain piecewise deterministic Markov processes. Annales de l'Institut Henri Poincaré - Probabilités et Statistiques. 2015;51(3):1040–1075.
47. Faggionato A, Gabrielli D, Crivellari MR. Non-equilibrium thermodynamics of piecewise deterministic markov processes. J Stat Phys. 2009;137:259–304.
48. Pakdaman K, Thieullen M, Wainrib G. Asymptotic expansion and central limit theorem for multiscale piecewise-deterministic Markov processes. Stoch Process Appl. 2012;122:2292–2318.
49. Peccoud J, Ycart B. Markovian Modelling of Gene Product Synthesis. Theor Popul Biol. 1995;48:222–234.
50. Li G-W, Xie XS. Central dogma at the single-molecule level in living cells. Nature. 2011;475(7356):308–315.
51. Sasai M, Wolynes PG. Stochastic gene expression as a many-body problem. PNAS. 2003;100(5):2374–2379.

52. Walczak AM, Sasai M, Wolynes PG. Self-consistent proteomic field theory of stochastic gene switches. Biophys J. 2005;88:828–850.
53. Kim K-Y, Wang J. Potential energy landscape and robustness of a gene regulatory network: toggle switch. PLoS Comput Biol. 2007;3(3):565–577.
54. Zhang B, Wolynes PG. Stem cell differentiation as a many-body problem. PNAS. 2014;111(28):10185–10190.
55. Senecal A, Munsky B, Proux F, Ly N, Braye FE, Zimmer C, Mueller F, Darzacq X. Transcription Factors Modulate c-Fos Transcriptional Bursts. Cell Rep. 2014;8:75–83.
56. Fukaya T, Lim B, Levine M. Enhancer control of transcriptional bursting. Cell. 2016;166(2):358–368.
57. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. PNAS. 2010;107(14):6286–6291.
58. Gu J, Gu Q, Wang X, Yu P, Lin W. Sphinx: modeling transcriptional heterogeneity in single-cell RNA-Seq. bioRxiv preprint. 2015.
59. Ghazanfar S, Bisogni AJ, Ormerod JT, Lin DM, Yang JYH. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. BMC Syst Biol. 2016;10:127.
60. Mojtahedi M, Skupin A, Zhou J, Castano IG, Leong-Quong RYY, Chang HH, Giuliani A, Huang S. Cell fate decision as high-dimensional critical state transition. PLOS Biol. 2016;14(12):1–28.
61. Sokolik C, Liu Y, Bauer D, McPherson J, Broeker M, Heimberg G, Qi LS, Sivak DA, Thomson M. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. Cell Syst. 2015;1:117–129.
62. Battich N, Stoeger T, Pelkmans L. Control of transcript variability in single mammalian cells. Cell. 2015;163(7):1596–1610.
63. Chong S, Chen C, Ge H, Xie XS. Mechanism of transcriptional bursting in bacteria. Cell. 2014;158(2):314–326.
64. Antoneli F, Ferreira RC, Briones MRS. A model of gene expression based on random dynamical systems reveals modularity properties of gene regulatory networks. Math Biosci. 2016;276:82–100.
65. Potoyan DA, Wolynes PG. Dichotomous noise models of gene switches. J Chem Phys. 2015;143(19):195101.
66. Hufton PG, Lin YT, Galla T, McKane AJ. Intrinsic noise in systems with switching environments. Phys Rev E. 2016;93(5):052119.
67. Pájaro M, Alonso AA, Otero-Muras I, Vázquez C. Stochastic modeling and numerical simulation of gene regulatory networks with protein bursting. J Theor Biol. 2017;421:51–70.
68. Teles J, Pina C, Edén P, Ohlsson M, Enver T, Peterson C. Transcriptional Regulation of Lineage Commitment - A Stochastic Model of Cell Fate Decisions. PLoS Comput Biol. 2013;9(8):1–13.
69. Schnoerr D, Grima R, Sanguinetti G. Cox process representation and inference for stochastic reaction-diffusion processes. Nat Commun. 2016;7:1–11.
70. Ocone A, Millar AJ, Sanguinetti G. Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. Bioinformatics. 2013;29(7):910–916.
71. Pinna A, Soranzo N, de la Fuente A. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. PLoS One. 2010;5(10):1–8.
72. Corre G, Stockholm D, Arnaud O, Kaneko G, Viñuelas J, Yamagata Y, Neildez-Nguyen TMA, Kupiec J-J, Beslon G, Gandrillon O, Paldi A. Stochastic Fluctuations and Distributed Control of Gene Expression Impact Cellular Memory. PLoS ONE. 2014;9(12):115574.
73. Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, Wu AR, Churchman LS, Singh A, Raj A. Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. Mol Cell. 2015;58(2):339–352.
74. Hathaway NA, Bell O, Hodges C, Miller EL, Neel DS, Crabtree GR. Dynamics and memory of heterochromatin in living cells. Cell. 2012;149(7):1447–1460.
75. Fourel G, Magdinier F, Gilson E. Insulator dynamics and the setting of chromatin domains. BioEssays. 2004;26(5):523–532.
76. Kueng S, Oppikofer M, Gasser SM. Sir proteins and the assembly of silent chromatin in budding yeast. Annu Rev Genet. 2013;47:275–306.
77. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–1680.
78. Obersriebnig MJ, Pallesen EMH, Sneppen K, Trusina A, Thon G. Nucleation and spreading of a heterochromatic domain in fission yeast. Nat Commun. 2016;7:1–11.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central