Steady-state distributions of nascent RNA for general initiation mechanisms

Steady-state distributions of nascent RNA for general initiation mechanisms

Juraj Szavits-Nossan* and Ramon Grima†
School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, United Kingdom

(Received 1 April 2022; accepted 15 December 2022; published 31 January 2023)

Fluctuations in the number of nascent RNA accurately reflect transcriptional activity. However, mathematical models predicting their distributions are difficult to solve analytically due to their non-Markovian nature stemming from transcriptional elongation. Here we circumvent this problem by deriving an exact relationship between the steady-state distribution of nascent RNA and the distribution of initiation times, which can be computed for any general initiation mechanism described by a set of first-order reactions. We test our theory using simulations and live cell imaging data.

DOI: 10.1103/PhysRevResearch.5.013064

I. INTRODUCTION

Transcription in single cells occurs in bursts whose size and timing is random [1]. Intuitively, a burst originates from rapid mRNA transcription when a promoter briefly switches on. The experimental distribution of mature mRNA numbers can often be fitted using the exact solution of a simple Markov model of gene expression, called the telegraph model [2–5]. This model describes the promoter switching between two states of activity and inactivity, the production of mature mRNA occurring one molecule at a time whilst in the active state, and the degradation of mature mRNA via a first-order reaction. However, in recent years, doubts have arisen about the validity of this kinetic description, principally because mature (cellular) mRNA does not provide a direct read-out of transcription [6,7]—the fluctuations of mRNA numbers in the cell are strongly influenced by various post-transcriptional events such as splicing, nuclear export, and DNA replication.

In order to circumvent such issues, it has been proposed that gene expression can be more accurately understood by studying nascent RNA fluctuations, i.e.. variability in the number of RNA polymerase (RNAP) molecules that are actively involved in the elongation phase of transcription. By its very definition, this is a direct read-out of transcription. The numbers of actively transcribing RNAPs can be estimated directly from nascent single-cell sequencing methods [8] or more commonly using single molecule fluorescence *in situ* hybridization (sm-FISH) where intronic probes specifically label (nonspliced) nascent RNA [7,9]. Fitting of these numbers using stochastic models that account for RNAP dynamics can help us more accurately understand which regulatory steps in transcription are tuned to achieve required mRNA expression levels [10,11].

Real-time observation of transcription in *vivo* has revealed that initiation is a stochastic process, whereas elongation and termination are fairly deterministic [12]. Transcription can be thus modelled by a stochastic, multistep process describing initiation, followed by a deterministic, single-step reaction describing elongation and termination. Despite this simplification, models of this type are particularly difficult to solve because they are inherently non-Markovian. So far, only two such models have been solved analytically: one in which the promoter is always active [13,14], and the other in which the promoter switches between two states of activity and inactivity [Fig. 1(b)] [6]. A few other, more realistic models have been studied, but only the first two moments of the nascent RNA distribution have been obtained analytically [15–17]. None of these models, however, account for the complex multistep process of initiation that has been elucidated by decades of biochemical research [18].

In this paper, we develop a general framework that allows us to find nascent RNA distribution for models with stochastic, multistep initiation, under the assumption of deterministic elongation and termination [Fig. 1(a)]. The significance of this result is that it applies to any initiation process that produces nascent RNA at time intervals that are independent and identically distributed (iid) random variables with an arbitrary probability density function $f(t)$. This is true for any stochastic process described by first-order reactions with time-independent kinetic rates, some of which are shown in Figs. 1(b)–1(d). We argue that while transcription initiation includes invariably many bimolecular steps, these are well approximated by pseudo first-order reactions because transcriptional machinery is generally abundant (the range of transcription factor copy numbers in bacteria is between 1–10^3 copy numbers per cell, whereas in mammalian cells it is between 10^3–10^6 [19]). Furthermore, we also ignore any variations of the rates over timescales that are much longer than the elongation time, e.g., initiation rates that change over time due to their dependence on the cell volume [20]. For such processes, we solve a first passage time problem to get an

*Juraj.Szavits.Nossan@ed.ac.uk
†Ramon.Grima@ed.ac.uk

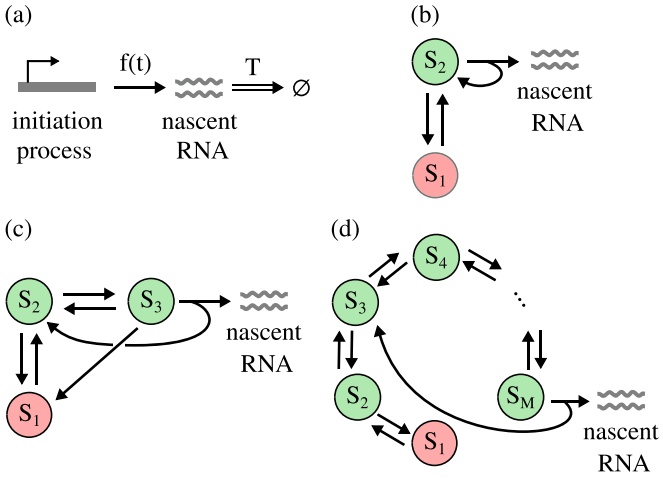

FIG. 1. (a) Nascent RNA is produced by an initiation process at time intervals that are independent and identically distributed (iid) random variables with an arbitrary probability density function $f(t)$. Elongation and termination are modelled deterministically, i.e., they take a fixed amount of time T to finish. [(b)–(d)] Examples of the initiation processes that can be studied by this framework, in the order of increasing complexity. S_1, S_2, etc. are gene states. Red circles denote the off states. (b) The telegraph process [2]. (c) A three-state process that accounts for the binding of RNAP [21]. (d) A stepwise process of eukaryotic transcription that accounts for the binding of general transcription factors and RNAP [18,22], and the promoter proximal pausing of RNAP in metazoans [23,24].

analytic expression for $f(t)$. The steady-state nascent RNA distribution is then computed using renewal theory, which generalizes the Poisson process by allowing for a nonexponential $f(t)$. We verify our theory by stochastic simulations and using experimental results for the transcription kinetics in *Escherichia coli* [25].

The paper is organized as follows. The waiting time distribution between two successive nascent RNA production events is computed in Sec. II for general initiation processes described by first-order reactions with time-independent rates. An exact relationship between the waiting time distribution and the nascent RNA distribution is derived and applied to examples of initiation processes in Sec. III. The theory is compared to experimental data from live cell imaging in *E. coli* in Sec. IV. Conclusions are presented in Sec. V.

II. WAITING TIME DISTRIBUTION BETWEEN SUCCESSIVE NASCENT RNA PRODUCTION EVENTS

Transcription is divided into initiation, elongation, and termination. Initiation is a complex multistep process, which starts by the binding of an RNAP at the promoter, together with several transcription factors, and ends when the RNAP escapes the promoter and starts productive elongation. During elongation, RNAP traverses the gene and copies its sequence into a nascent RNA. The termination occurs at the gene end, leading to a mature RNA transcript.

We consider a stochastic initiation process consisting of M gene states labeled by $S_1, \ldots, S_M$. The gene switches between the states according to first-order reactions

$$S_i \underset{k_{ji}}{\overset{k_{ij}}{\rightleftharpoons}} S_j,\ i, j = 1, \ldots, M,\ i \neq j,\ S_M \xrightarrow{k_M} S_K + N. \quad (1)$$

The last reaction describes production of nascent RNA (N), after which the initiation process starts again from the state S_K (K takes a fixed value between 1 and M). Equation (1) is a general description of the initiation process using first-order reactions. Particular cases are obtained by removing some reactions, which is equivalent to setting the rates of those reactions to zero.

We are interested in computing the probability density function (pdf) of the waiting time between successive events of nascent RNA production. This is a first passage time problem that can be solved by replacing the last reaction in Eq. (1) with $S_M \to A$, where A is an absorbing state, i.e., once the process reaches A, it stops. The pdf $f(t)$ is then equal to $k_M P_M(t)$, where $P_M(t)$ is the probability that the gene is in state M at time t, given that it was in state K at time 0, and $k_M dt$ is the probability that the transition $S_M \to A$ occurs in the time interval $[t, t + dt)$. The probability $P_M(t)$ can be found by solving the following master equation:

$$\frac{dP_i}{dt} = \sum_{j=1}^{M} k_{ji} P_j - \left(\sum_{j=1, j\neq i}^{M} k_{ij} + \delta_{i,M} k_M \right) P_i, \quad (2)$$

with the initial condition $P_i(0) = \delta_{i,K}$, where $\delta_{i,j}$ is the Kronecker delta. The pdf $f(t)$ is given by

$$f(t) = k_M (\exp(\mathbf{L}t))_{MK}, \quad (3)$$

where the matrix elements of $\mathbf{L}$ read

$$L_{ij} = k_{ji},\ i \neq j,\ L_{ii} = -\sum_{j=1, j\neq i}^{M} k_{ij} - \delta_{i,M} k_M. \quad (4)$$

A central function in renewal theory is the Laplace transform of the pdf $f(t)$, denoted by $f^*(s)$. Formally,

$$f^*(s) = k_M P_M^*(s) = k_M (s\mathbf{I} - \mathbf{L})^{-1}_{MK}, \quad (5)$$

where $P_M^*(s)$ is the Laplace transform of $P_M(t)$ and $\mathbf{I}$ is the $M \times M$ identity matrix. Particular examples of $f(t)$ and $f^*(s)$ are discussed later in the text.

III. DISTRIBUTION OF NASCENT RNA IN THE STEADY STATE

A. Results for general $f(t)$

Our approach for solving the general problem in Fig. 1(a) is based on recognizing that the number of nascent RNA production events that occurred up to time t, denoted by $N_I(t)$, constitutes a renewal process [26]. A renewal process generalizes the Poisson process by allowing for nonexponentially distributed waiting times between successive events. As in the Poisson process, the waiting times in the renewal process are mutually independent. In transcription, this assumption is supported by experimental evidence [12].

We denote by T_i the time of the ith nascent RNA production event, measured from some reference time point $t = 0$. For simplicity, we assume that a nascent RNA production event

occurred immediately before $t = 0$. If t_i denotes the waiting time between ith and $(i-1)$th nascent RNA production events, $t_i = T_i - T_{i-1}$, where $T_0 = 0$, then

$$T_i = \sum_{j=1}^{i} t_j, \ i = 1, 2, \ldots. \tag{6}$$

Since t_j are iid random variables, the pdf of T_i is the i-fold convolution of f,

$$k_i(t) = f^{*i}(t). \tag{7}$$

The Laplace transform of $k_i(t)$ is given by

$$k_i^*(s) = \int_0^\infty dt\ \mathrm{e}^{-st} k_i(t) = [f^*(s)]^i. \tag{8}$$

The probability of $N_I(t) = n$ is equivalent to the probability of $T_n \leqslant t < T_{n+1}$, which is given by

$$\begin{aligned} P(N_I(t) = n) &= P(T_n \leqslant t < T_{n+1}), \\ &= K_n(t) - K_{n+1}(t), \end{aligned} \tag{9}$$

where $K_n(t) = \int_0^t dt'\ k_n(t')$ for $n \geqslant 1$ is the cumulative distribution function of T_n, and $K_0(t) = 1$. The mean number of initiations up to time t, which we denote by $H(t)$, is given by

$$\begin{aligned} H(t) = \langle N_I(t) \rangle &= \sum_{n=0}^{\infty} n P(N_I(t) = n) \\ &= \sum_{n=0}^{\infty} n(K_n(t) - K_{n+1}(t)) = \sum_{n=1}^{\infty} K_n(t). \end{aligned} \tag{10}$$

In renewal theory, $H(t)$ is known as the renewal function. Related to the renewal function is the renewal density $h(t)$, which is defined as the time derivative of $H(t)$,

$$h(t) = \frac{d}{dt} H(t) = \sum_{n=1}^{\infty} k_n(t). \tag{11}$$

The interpretation of $h(t)$ is that $h(t)dt$ is equal to the probability that a new renewal event occurs between t and $t + dt$. The Laplace transform of $h(t)$ is given by

$$h^*(s) = \frac{f^*(s)}{1 - f^*(s)}. \tag{12}$$

Using the small-s expansion $f^*(s) = 1 - \mu s + O(s^2)$, we get that the long-time limit of $h(t)$ is given by

$$\lim_{t\to\infty} h(t) = \lim_{s\to 0} h^*(s) = \frac{1}{\mu}. \tag{13}$$

In other words, the density of renewals in the steady state is uniform.

We denote by T the total time of elongation and termination. Since T is fixed in our model, the number of nascent RNA actively engaged in transcription at time t is equal to

$$N(t) = \begin{cases} N_I(t), & t \leqslant T, \\ N_I(t) - N_I(t-T), & t > T. \end{cases} \tag{14}$$

The probability $P(N(t) = n)$ for $t \leqslant T$ is given by Eq. (9). In order to find the probability distribution of $N(t)$ for $t > T$, we first need to find the probability density function $f_{t_0}(\tau)$ of the forward recurrence time τ, where τ is defined as the time until the next initiation event measured from a fixed time point t_0 (we will later set $t_0 = t - T$). Using the renewal density $h(t)$, the expression for $f_{t_0}(\tau)$ reads [26]

$$f_{t_0}(\tau) = f(t_0 + \tau) + \int_0^{t_0} dt' h(t_0 - t') f(t' + \tau). \tag{15}$$

The first term corresponds to having no renewals before t_0, whereas the second term corresponds to having the last renewal at some earlier time $t_0 - t'$, followed by a renewal at a time $t_0 - t' + t' + \tau$. In the steady state, $t_0 \to \infty$ and the expression for $f_{t_0}(\tau)$ simplifies to

$$f_\infty(\tau) = \lim_{t_0\to\infty} f_{t_0}(\tau) = \frac{1 - F(\tau)}{\mu}, \tag{16}$$

where $F(t) = \int_0^t dt'\ f(t')$ is the cumulative distribution function of the waiting times between successive nascent RNA production events, μ is the mean waiting time, and we assumed that $\lim_{t_0\to\infty} f_{t_0}(t_0 + \tau) = 0$. The Laplace transform of $f_\infty(\tau)$, which we will need in a moment, is given by

$$f_\infty^*(s) = \frac{1 - f^*(s)}{\mu s}. \tag{17}$$

We now have all the preliminaries to compute the probability $P(N = n)$ to find n nascent RNA molecules on the gene in the steady state. For $n = 0$,

$$P(N = 0) = \int_T^\infty d\tau\ f_\infty(\tau) = 1 - \int_0^T d\tau\ f_\infty(\tau). \tag{18}$$

The Laplace transform of $P(N = 0)$ with respect to T is therefore given by

$$\begin{aligned} P^*(0, s) &= \int_0^\infty dT \mathrm{e}^{-sT} P(N = 0) \\ &= \frac{\mu s - 1 + f^*(s)}{\mu s^2}. \end{aligned} \tag{19}$$

Because nascent RNA production events are mutually independent, the Laplace transform $P^*(n, s)$ of $P(N = n)$ for $n \geqslant 1$ is given by a product of $f_\infty^*(s)$ and the Laplace transform of $P(N_I(t) = n - 1)$,

$$\begin{aligned} P^*(n, s) &= \int_0^\infty dT\ \mathrm{e}^{-Ts} P(N = n) \\ &= \frac{[1 - f^*(s)]^2 [f^*(s)]^{n-1}}{\mu s^2}, \ n \geqslant 1. \end{aligned} \tag{20}$$

Equations (19) and (20) are our main result. They connect the Laplace transform of $P(N = n)$ to the Laplace transform of $f(t)$ for *all* initiation processes described by Eq. (2). For these processes, $f^*(s)$ is a rational function of s, meaning that $P(N = n)$ can be computed from Eqs. (19) and (20) by partial fraction decomposition, for which many methods are available [27].

The moments of $P(N = n)$ can be computed from the probability generating function,

$$G(z) = \sum_{n=0}^{\infty} z^n P(N = n), \tag{21}$$

whose Laplace transform $G^*(z,s)$ with respect to T is given by

$$G^*(z,s) = \frac{1}{s} + \frac{(z-1)[1-f^*(s)]}{\mu s^2[1-zf^*(s)]}. \tag{22}$$

For example, the mean number of nascent RNA $\langle N\rangle$ and the variance Var$[N]$ are given by

$$\langle N\rangle = \frac{T}{\mu}, \tag{23a}$$

$$\text{Var}[N] = \mathcal{L}^{-1}\left\{\frac{1+f^*(s)}{\mu s^2[1-f^*(s)]}\right\}(T) - \left(\frac{T}{\mu}\right)^2, \tag{23b}$$

where $\mathcal{L}^{-1}\{\dots\}(T)$ is the inverse Laplace transform of the expression in the curly brackets, evaluated at T.

In the rest of this section, we compute nascent RNA distribution for a range of initiation processes. We first consider three initiation processes for which we compute nascent RNA distributions analytically. These are the Poisson process, the telegraph process and the fully irreversible process with arbitrary number of steps with equal rates. We then consider a three-state initiation process proposed in Ref. [21] that accounts for RNAP recruitment. Finally, as a full demonstration of our framework, we consider a ten-state initiation process based on a canonical model of eukaryotic transcription initiation that includes the on and off switching of the promoter, the binding and unbinding of six general transcription factors (IID, IIA, IIB, IIF, IIE and IIH) and RNAP, the unwinding of the double-stranded DNA, and the promoter proximal pausing of RNAP in metazoans. For the last two processes, we compute the nascent RNA distribution numerically.

B. The Poisson process

The Poisson process describes a constitutive promoter with one gene state ($M = 1$). The reaction scheme for this process is given by

$$S_1 \xrightarrow{k_1} S_1 + N. \tag{24}$$

Since there is only one state, $K = 1$. The pdf $f(t)$ is an exponential,

$$f(t) = k_1 e^{-k_1 t}, \quad f^*(s) = \frac{k_1}{s+k_1}, \tag{25}$$

and the mean is given by $\mu = 1/k_1$. Inserting Eq. (25) into Eqs. (19) and (20) and inverting the Laplace transform we get

$$P(N=n) = \frac{(k_1 T)^n}{n!}\mathrm{e}^{-k_1 T}, \tag{26}$$

which is the Poisson distribution.

The Poisson process is special because the forward recurrence time τ has the same exponential distribution as $f(t)$, $f_{t_0}(\tau) = k_1 e^{-k_1\tau}$, i.e., it does not depend on t_0. This is the memoryless property of the exponential distribution. For $t < T$, $P(N(t)=n)$ is a convolution of f and $K_{n-1} - K_n$ evaluated at t, which yields

$$P(N(t)=n) = \frac{(k_1 t)^n}{n!}e^{-k_1 t}, \quad t < T, \tag{27}$$

where we have used the fact that $K_n(t)$ is an Erlang distribution with the shape parameter n and the rate parameter k_1,

$$K_n(t) = 1 - \sum_{m=0}^{n-1}\frac{(k_1 t)^m}{m!}e^{-k_1 t}. \tag{28}$$

For $t > T$, $P(N(t)=n)$ is a convolution of f_{t-T} and $K_{n-1} - K_n$ evaluated at T, which gives

$$P(N(t)=n) = \frac{(k_1 T)^n}{n!}e^{-k_1 T}, \quad t > T. \tag{29}$$

Since this distribution is the same as the stationary distribution, we conclude that the steady state is reached immediately after the first round of transcription elongation, i.e., at time T.

C. The telegraph process

The telegraph process describes a bursty promoter that switches between two gene states ($M = 2$), an inactive state S_1 and an active state S_2. The reaction scheme for this process is given by

$$S_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} S_2 \xrightarrow{k_2} S_2 + N. \tag{30}$$

After initiation, the process remains in the active state ($K = 2$), see Fig. 1(b). The Laplace transform of $f(t)$ has been computed in Appendix B and reads

$$f^*(s) = \frac{k_2(s+k_{12})}{s^2 + (k_{21}+k_{12}+k_2)s + k_{12}k_2}. \tag{31}$$

The mean initiation time μ is given by

$$\mu = -\left.\frac{df^*}{ds}\right|_{s=0} = \frac{k_{21}+k_{12}}{k_{12}k_2}. \tag{32}$$

To find $P(N=n)$, we write $f^*(s)$ as

$$f^*(s) = \frac{k_2(s+k_{12})}{(s+\lambda_1)(s+\lambda_2)}, \tag{33}$$

where $-\lambda_1$ and $-\lambda_2$ are roots of the quadratic equation

$$(s+\lambda_1)(s+\lambda_2) = s^2 + s(k_{21}+k_{12}+k_2) + k_{12}k_2. \tag{34}$$

Inserting Eq. (33) into Eqs. (19) and (20) we get for $n = 0$ and $n \geqslant 1$, respectively,

$$P^*(0,s) = \frac{s+\lambda_1+\lambda_2-1/\mu}{(s+\lambda_1)(s+\lambda_2)}, \tag{35}$$

$$P^*(n,s) = \frac{(s+k_{21}+k_{12})^2[k_2(s+k_{12})]^{n-1}}{\mu[(s+\lambda_1)(s+\lambda_2)]^{n+1}}. \tag{36}$$

The inverse Laplace transform of $P^*(n,s)$ is

$$P(N=0) = \frac{\lambda_2 - 1/\mu}{\lambda_2-\lambda_1}e^{-\lambda_1 T} + \frac{\lambda_1 - 1/\mu}{\lambda_1-\lambda_2}e^{-\lambda_2 T}, \tag{37}$$

$$P(N=n) = \sum_{i=0}^{n}\frac{T^{n-i}(A_{i,n}e^{-\lambda_1 T} + B_{i,n}e^{-\lambda_2 T})}{(n-i)!}, \tag{38}$$

where $A_{i,n}$ and $B_{i,n}$ are computed in Appendix B. The probability generating function $G(z)$ is given by Eq. (B17) in Appendix B. The expression for $G(z)$ is the same as in Supplemental Material of Ref. [28], Eq. (1.15). There it was shown that the nascent RNA distribution, obtained from $G(z)$,

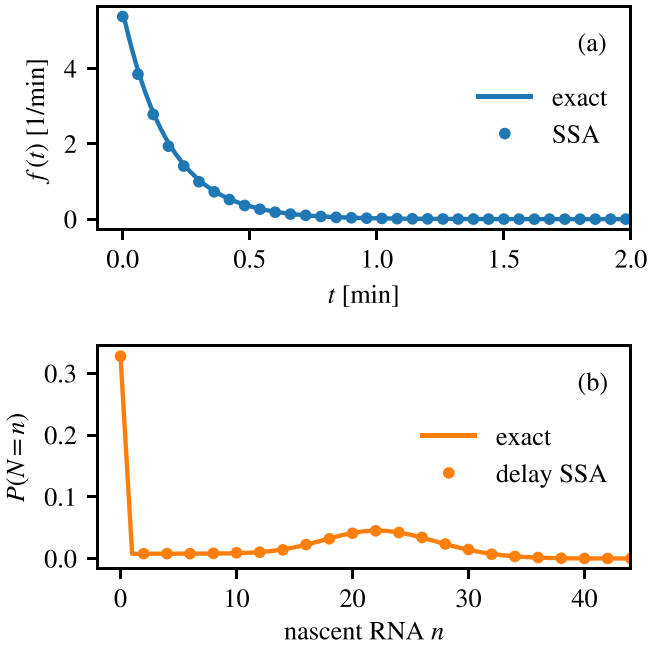

FIG. 2. (a) Probability density function $f(t)$ of the waiting time between successive nascent RNA production events for the initiation processes described by Eq. (30) and depicted in Fig. 1(b). The solid line is the theoretical prediction from Eq. (3), and the points are from stochastic simulations (SSA). (b) The respective probability distribution of the nascent RNA, obtained by inverting Eqs. (19) and (20), and compared to stochastic simulations (delay SSA) performed using `DelaySSAToolkit.jl` package in Julia [29]. Model parameters are: $k_{12} = 0.06\ \text{min}^{-1}$, $k_{21} = 0.042\ \text{min}^{-1}$, $k_2 = 5.52\ \text{min}^{-1}$ and $T = 4.167$ min.

matches the one obtained in Ref. [6]. The predicted distributions $f(t)$ and $P(N = n)$ are in excellent agreement with ones obtained by stochastic simulations, see Fig. 2.

D. Fully irreversible process with arbitrary number of steps with equal rates

This process describes a constitutive promoter with multiple rate-limiting steps with equal rates. The reaction scheme for this process is given by

$$S_1 \xrightarrow{\lambda} S_2 \xrightarrow{\lambda} \ldots \xrightarrow{\lambda} S_M \xrightarrow{\lambda} S_1 + N, \tag{39}$$

where M is the number of rate-limiting steps and all reactions have the same rate λ. We have assumed that the process returns to state S_1 after each round of initiation. If the process returns to some other state S_K, then the results in this section pertain but with M replaced by $M - K + 1$. The waiting time distribution between successive nascent RNA production events is an Erlang distribution with a shape parameter M and a rate parameter λ, whose pdf reads

$$f(t) = \frac{\lambda^M t^{M-1}}{(M-1)!} e^{-\lambda t}, \quad f^*(s) = \frac{\lambda^M}{(s+\lambda)^M}. \tag{40}$$

This distribution is ubiquitous in cellular biology and naturally arises from Markov multistep processes [30]. The mean initiation time μ is given by

$$\mu = -\left.\frac{df^*}{ds}\right|_{s=0} = \frac{M}{\lambda}. \tag{41}$$

Inserting $f^*(s)$ into Eqs. (19) and (20) we get for $n = 0$ and $n \geqslant 1$, respectively,

$$P^*(0,s) = \frac{(Ms-\lambda)(s+\lambda)^M + \lambda^{M+1}}{Ms^2(s+\lambda)^M}, \tag{42}$$

$$P^*(n,s) = \frac{\lambda^{M(n-1)+1}[(s+\lambda)^M - \lambda^M]^2}{Ms^2(s+\lambda)^{M(n+1)}}. \tag{43}$$

The partial fraction decomposition and the inverse Laplace transform of $P^*(n,s)$ were computed in Appendix B. The final result is

$$P(N=n) = \frac{e^{-\lambda T}}{M}\left\{\sum_{j=0}^{N_+-1}(N_+ - j)\frac{(\lambda T)^j}{j!} - 2\sum_{j=0}^{N_0-1}(N_0 - j)\frac{(\lambda T)^j}{j!} + \sum_{j=0}^{N_--1}(N_- - j)\frac{(\lambda T)^j}{j!}\right\}, \tag{44}$$

where $N_+ = M(n+1)$, $N_0 = Mn$, $N_- = M(n-1)$, and we have used a convention according to which a sum in which the upper bound is lower than the lower bound is equal to zero.

E. Three-state process that accounts for RNAP recruitment

This process describes a bursty promoter that accounts for binding and unbinding of transcription factors and RNAP [21]. The reaction scheme for this process is given by

$$S_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} S_2 \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} S_3 \xrightarrow{k_{31}} S_1, \; S_3 \xrightarrow{k_3} S_2 + N, \tag{45}$$

see also Fig. 1(c). Here, S_1 is an inactive gene state, S_2 is a gene state in which the gene is bound to transcription factors, and S_3 is a gene state in which RNAP is bound and paused. During the initiation process, transcription factors may unbind from the DNA ($S_2 \to S_1$). The bound RNAP may also unbind from the DNA, either alone ($S_3 \to S_2$) or together with the transcription factors ($S_3 \to S_1$). Once the RNAP is released into productive elongation, the gene state returns to state S_2, waiting for the next RNAP.

The Laplace transform of the pdf $f(t)$ has been computed in Appendix B and reads

$$f^*(s) = \frac{k_3 k_{23}(s+k_{12})}{s^3 + bs^2 + cs + k_{12}k_{23}k_3}, \tag{46}$$

where b and c are given by

$$b = k_{12} + k_{21} + k_{23} + k_{32} + k_{31} + k_3, \tag{47}$$

$$c = (k_{12}+k_{21})(k_{31}+k_{32}+k_3) + k_{23}(k_{12}+k_{31}+k_3). \tag{48}$$

The mean initiation time μ can be written as

$$\mu = -\left.\frac{df^*}{ds}\right|_{s=0} = \frac{(k_{12}+k_{21})(k_{31}+k_{32}+k_3)}{k_{12}k_{23}k_3} + \frac{k_{23}(k_{12}+k_{31})}{k_{12}k_{23}k_3}. \tag{49}$$

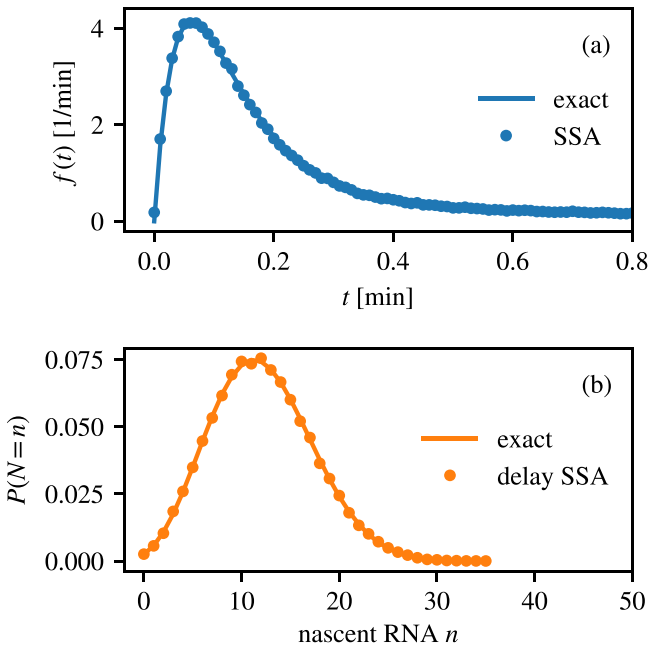

FIG. 3. (a) Probability density function $f(t)$ of the waiting time between successive nascent RNA production events for the initiation processes described by Eq. (45) and depicted in Fig. 1(c). The solid line is the theoretical prediction from Eq. (3), and the points are from stochastic simulations (SSA). (b) The respective probability distribution of the nascent RNA, obtained by inverting Eqs. (19) and (20), and compared to stochastic simulations (delay SSA) performed using `DelaySSAToolkit.jl` package in Julia [29]. Model parameters are: $k_{12} = 1.92\ \text{min}^{-1}$, $k_{21} = 1.92\ \text{min}^{-1}$, $k_{23} = 9.6\ \text{min}^{-1}$, $k_{32} = 0.96\ \text{min}^{-1}$, $k_{31} = 1.92\ \text{min}^{-1}$, $k_3 = 19.2\ \text{min}^{-1}$ and $T = 4.167$ min.

From here we can get $P(N = n)$ by inserting $f^*(s)$ into Eqs. (19) and (20) and taking the inverse Laplace transform. It is possible, albeit tedious, to do this analytically. Instead, we do this numerically—the results are shown in Fig. 3 and are in excellent agreement with the results of stochastic simulations.

F. Ten-state process of eukaryotic transcription initiation

As our final example, we consider a ten-state initiation process based on a canonical model of eukaryotic transcription initiation [18,22], see Fig. 1(d) and Fig. 4. The reaction scheme for this process is given by

$$S_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} S_2 \text{ (gene activation/deactivation)}, \tag{50a}$$

$$S_2 \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} S_3 \text{ (TFIID binding/unbinding)}, \tag{50b}$$

$$S_3 \underset{k_{43}}{\overset{k_{34}}{\rightleftharpoons}} S_4 \text{ (TFIIA binding/unbinding)}, \tag{50c}$$

$$S_4 \underset{k_{54}}{\overset{k_{45}}{\rightleftharpoons}} S_5 \text{ (TFIIB binding/unbinding)}, \tag{50d}$$

$$S_5 \underset{k_{65}}{\overset{k_{56}}{\rightleftharpoons}} S_6 \text{ (TFIIF} \cdot \text{RNAP binding/unbinding)}, \tag{50e}$$

$$S_6 \underset{k_{76}}{\overset{k_{67}}{\rightleftharpoons}} S_7 \text{ (TFIIE binding/unbinding)}, \tag{50f}$$

$$S_7 \underset{k_{87}}{\overset{k_{78}}{\rightleftharpoons}} S_8 \text{ (TFIIH binding/unbinding)}, \tag{50g}$$

$$S_8 \underset{k_{98}}{\overset{k_{89}}{\rightleftharpoons}} S_9 \text{ (open/closed PIC)} \tag{50h}$$

$$S_9 \xrightarrow{k_{910}} S_{10}, \text{ (promoter proximal pausing)} \tag{50i}$$

$$S_{10} \xrightarrow{k_{10}} S_K + N \text{ (release into elongation)}. \tag{50j}$$

S_1 and S_2 are the off and on states of the promoter, respectively. General transcription factors and RNA polymerase bind the promoter in the following order: TFIID (S_3), TFIIA (S_4), TFIIB (S_5), TFIIF and RNAP (S_6), TFIIE (S_7) and TFIIH, resulting in the closed preinitiation complex (PIC, S_8). The TFIIH unwinds the promoter DNA, creating an open PIC (S_9) that begins the elongation. In metazoans, the elongating RNA polymerase pauses shortly after the initiation (S_{10}) [23,24]. The RNAP is eventually released into productive elongation,

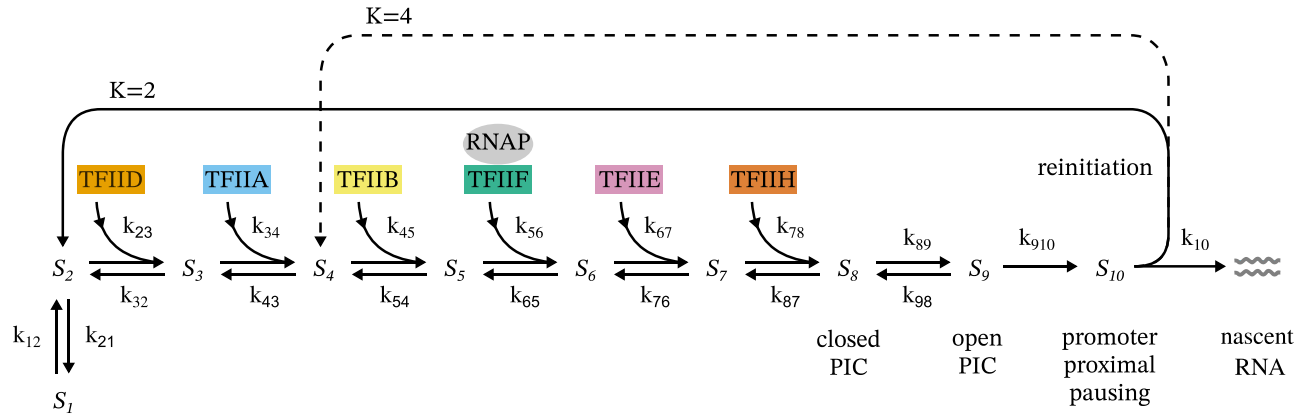

FIG. 4. A ten-state initiation process based on a canonical model of eukaryotic transcription initiation [18,22], which includes the on and off switching of the promoter, the binding and unbinding of six general transcription factors (IID, IIA, IIB, IIF, IIE and IIH) and RNAP, the unwinding of the double-stranded DNA, and the promoter proximal pausing of RNAP in metazoans [23,24]. We consider two scenarios for reinitiation: in one, the gene returns to the on state S_2 ($K = 2$), and in the other, the gene returns to the state S_4 ($K = 4$) with transcription factors IID and IIA bound to the promoter [31,32].

TABLE I. Parameter values for the ten-state process of eukaryotic transcription initiation sourced from the literature. See the main text for their full description.

Parameter	Description	Value	Reference
k_{12}	*PLEC* on rate	0.04 min^{-1}	[1]
k_{21}	*PLEC* off rate	0.53 min^{-1}	[1]
k_{23}	TFIID binding	$1/1.1$ min^{-1}	[33]
k_{32}	TFIID unbinding	$1/130$ min^{-1}	[33]
k_{34}	TFIIA binding	$1/20$ s^{-1}	[34]
k_{43}	TFIIA unbinding	$1/8$ min^{-1}	[35]
k_{45}	TFIIB binding	$1/3.2$ s^{-1}	[34]
k_{54}	TFIIB unbinding	$1/1.5$ s^{-1}	[34]
k_{56}	TFIIF & RNAP binding	2.3×10^{-3} s^{-1}	[36]
k_{65}	TFIIF & RNAP unbinding	3×10^{-3} s^{-1}	[36]
k_{67}	TFIIE binding	1.7×10^{-2} s^{-1}	[37]
k_{76}	TFIIE unbinding	5×10^{-2} s^{-1}	[37]
k_{78}	TFIIH binding	5×10^{-2} s^{-1}	[37]
k_{87}	TFIIH unbinding	$1/5$ min^{-1}	[37]
k_{89}	closed to open PIC	1.9×10^{-3} s^{-1}	[38]
k_{98}	open to closed PIC	1.1×10^{-4} s^{-1}	[38]
k_{910}	release into elongation	0.17 s^{-1}	[38]
k_{10}	promoter proximal unpausing	$1/5$ min^{-1}	[39]
T	elongation time	14 min	[40]

clearing the promoter for the next round of initiation. We assumed two scenarios for the reinitiation: in one, the gene returns to the on state S_2 ($K = 2$), and in the other, transcription factors IID and IIA remain bound to the promoter [31,32]—this scenario corresponds to $K = 4$ (dashed line in Fig. 4).

Kinetic parameters describing the progression from the on state (S_2) to the production of nascent RNA were sourced from the literature [1,33–40] and were assumed to be representative of eukaryotic genes (Table I). The values of k_{12} and k_{21} were matched to the on and off rates inferred from the transcription kinetics of *PLEC* gene promoter in mouse fibroblast cells (gene id 18810) [1]. For this promoter, the mean time it took to initiate from the on state, assuming no return to the off state, was reported to be 5.2 min [1]. The values of k_{54}, k_{56}, k_{89}, and k_{10} were adjusted to match this value to the mean initiation time obtained from $f(t)$ in Eq. (A11) after setting $k_{21} = 0$ and $K = 2$, yielding $k_{54} = 1/30$ s^{-1}, $k_{56} = 6.9 \times 10^{-2}$ s^{-1}, $k_{89} = 5.7 \times 10^{-2}$ s^{-1}, and $k_{10} = 2$ min^{-1}. The elongation time T was obtained by dividing the gene length of *PLEC* gene (60.404 kb) by the RNA polymerase II speed of 4.3 kb/min [40].

The Laplace transform of $f(t)$ was computed analytically for this process, because the matrix $\mathbf{L}$ in Eq. (4) is tridiagonal, for which the inverse is known explicitly [41]. The derivation of $f^*(s)$ is presented in Appendix A. The plot of $f(t)$ obtained by inverting $f^*(s)$ numerically is presented in Figs. 5(a) and 5(b) for two reinitiation scenarios $K = 2$ and $K = 4$, respectively. The corresponding nascent RNA distribution $P(N = n)$ is presented in Figs. 5(c) and 5(d). Both $f(t)$ and $P(N = n)$ are in excellent agreement with the results from stochastic simulations.

Next, we performed sensitivity analysis in order to understand how the kinetic rates of the model affect the nascent RNA distribution. In particular, we computed the local sensitivity coefficient s_{ij} defined as

$$s_{ij} = \frac{k_{ij}}{\text{FF}_N} \frac{\partial \text{FF}_N}{\partial k_{ij}}, \tag{51}$$

where FF_N is the Fano factor of the nascent RNA distribution,

$$\text{FF}_N = \frac{\text{Var}[N]}{\langle N \rangle}. \tag{52}$$

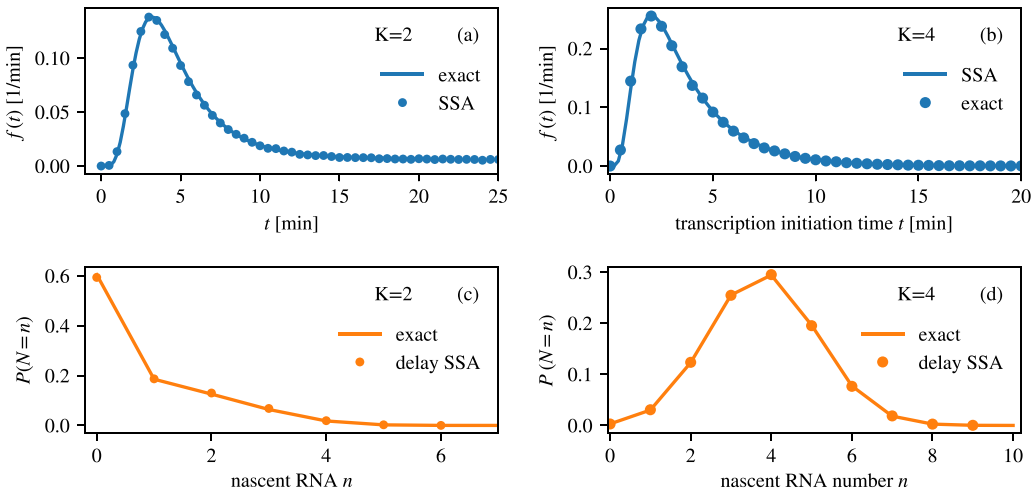

FIG. 5. (a) Probability density function $f(t)$ of the waiting time between successive nascent RNA production events for the initiation processes described by Eqs. (50a)–(50j) and depicted in Fig. 4. The solid line is the theoretical prediction from Eq. (3), and the points are from stochastic simulations (SSA). (b) The respective probability distribution of the nascent RNA, obtained by inverting Eqs. (19) and (20), and compared to stochastic simulations (delay SSA) performed using `DelaySSAToolkit.jl` package in Julia [29].

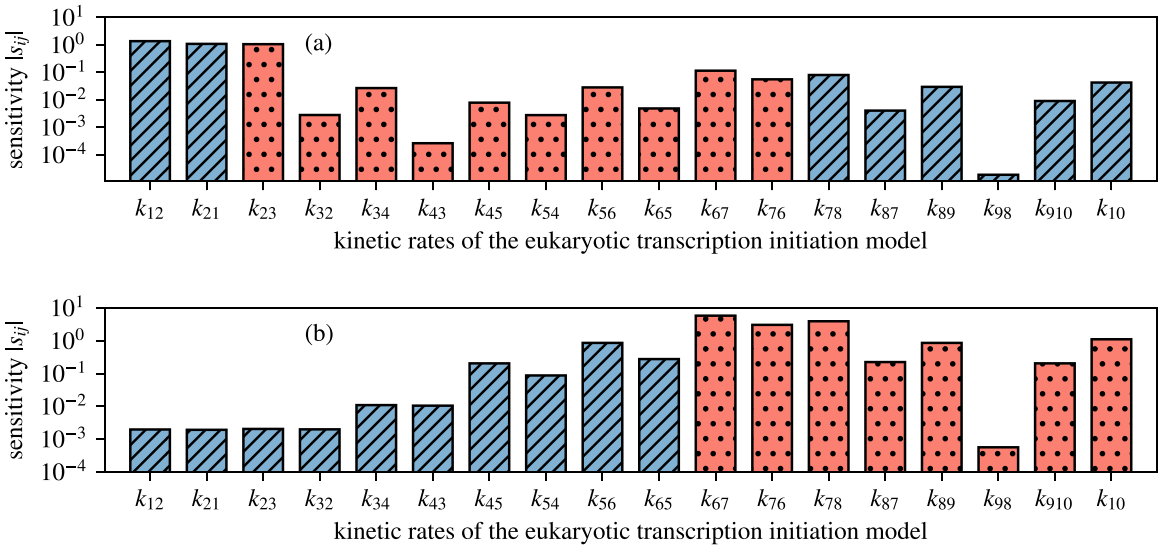

FIG. 6. Absolute value of the local sensitivity coefficient $|s_{ij}|$ defined in Eq. (51). Rates are labeled according to Eqs. (50a)–(50j) and Fig. 4. Blue bars with diagonal lines have $s_{ij} < 0$, whereas red bars with dots have $s_{ij} > 0$. (a) $K = 2$. (b) $K = 4$.

A value x of s_{ij} means that a 1% change in the value of k_{ij} causes a x% change in the value of FF_N. To find s_{ij}, we write Eq. (51) as

$$s_{ij} = \frac{k_{ij}}{\mathrm{FF}_n}\left[\frac{D_2}{\langle N\rangle} - D_1\left(\frac{\mathrm{Var}[N]}{\langle N\rangle^2} + 2\right)\right], \tag{53}$$

$$D_1 = \frac{\partial\langle N\rangle}{\partial k_{ij}}, \quad D_2 = \frac{\partial\langle N^2\rangle}{\partial k_{ij}}. \tag{54}$$

The advantage of our analytical solution over stochastic simulations is that D_1 and D_2 can be computed directly from Eqs. (23a) and (23b), respectively. The results for s_{ij} for all 18 kinetic rates are presented in Fig. 6. When reinitiation occurs from the on state ($K = 2$), we find that FF_N is most sensitive to changes in the on rate, the off rate, and the rate of TFIID binding [Fig. 6(a)], which indeed have been identified previously as rate-limiting steps in transcription initiation [21,42]. However, we get markedly different values of s_{ij} if we assume that TFIID and TFIIA remain bound at the promoter until reinitiation [Fig. 6(b)]. We note that s_{ij} measures local sensitivity, which means that its value will likely change for a different choice of parameters.

IV. COMPARISON WITH EXPERIMENTAL DATA FROM LIVE CELL IMAGING

In Ref. [25], transcription kinetics of a target gene were followed in live *Escherichia coli* cells, one transcription event at a time. This was possible by a method that tags mRNA *in vivo* with MS2-GFP proteins [43,44]. In the experiment, the target gene was controlled by the *tetA* promoter, which was induced (turned on) by anhydrotetracycline (aTc) at a concentration of 15 ng/ml at two temperatures, 24°C and 37°C. The waiting time distribution between successive mature mRNA production events was measured experimentally and fitted to a hypoexponential distribution with three rates denoted by r_1, r_2, and r_3 [Figs. 7(a) and 7(b)]. The Laplace transform of the pdf of this distribution is given by

$$f^*_{\mathrm{fit}}(s) = \frac{r_1 r_2 r_3}{(s + r_1)(s + r_2)(s + r_3)}. \tag{55}$$

To apply our theory to these data, we assumed that the measured intervals between the productions of successive mature RNA molecules were not significantly affected by elongation. According to Ref. [25], this assumption is supported by the fact that the mean duration of the intervals between successive mature mRNA production events was larger than 600 s, whereas the elongation time took only tens of seconds. Sequence-specific transcriptional pauses, if existed, were ruled out as being too short to significantly affect the measured distributions (they are in the range between 10–100 s). Premature termination was also excluded as it would result in a multimodal distribution that was not observed in the experiment.

The fraction of cells with a given number $M(t)$ of mRNA was measured at $t = t_1 = 1$ h after the induction by aTc. Since MS2 tags protected mRNA from degradation, we assumed that the number of mRNA $M(t_1)$ was equal to the number of transcription initiation events from the time of full induction t_0 until $t_1 - T$,

$$M(t_1) = N_I(t_1 - t_0), \tag{56}$$

where T is the elongation time. The latter was only tens of seconds, hence we approximated $t_1 - T$ by t_1. The time of full induction by aTC was estimated to be $t_0 = 20$ min [25]. From Eqs. (9) and (56) it follows that the Laplace transform of $P(M(t) = n)$ is given by

$$\mathcal{L}\{P(M(t) = n)\} = \int_0^\infty dt\ P(M(t) = n)\mathrm{e}^{-st} = \frac{[1 - f^*_{\mathrm{fit}}(s)][f^*_{\mathrm{fit}}(s)]^n}{s}. \tag{57}$$

The distribution of $M(t_1)$ was obtained by computing the inverse of Eq. (57) evaluated at $t_1 - t_0 = 40$ min. The results for 24°C and 37°C are shown in Figs. 7(c) and 7(d), respectively.

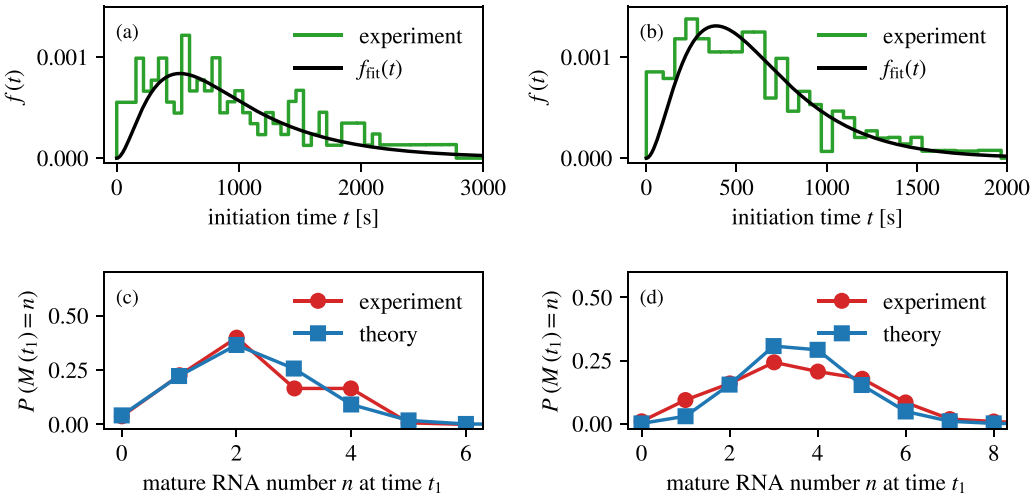

FIG. 7. (a) Probability density function $f(t)$ of the waiting times between successive mature RNA production events measured in Ref. [25] at 24°. Black line is a fit to the hypoexponential distribution obtained by inverting $f^*_{\rm fit}(s)$ in Eq. (55) with $r_1 = 1/620$ s^{-1}, $r_2 = 1/240$ s^{-1}, and $r_3 = 1/115$ s^{-1}. (b) Same as in (a) but at 37°C with $r_1 = 1/245$ s^{-1}, $r_2 = 1/254$ s^{-1} and $r_3 = 1/109$ s^{-1}. (c) Probability distribution $P(M(t_1) = n)$ of the fraction of cells with a given number of mRNA molecules measured 1 h following induction by aTc at 24°C, compared to the theoretical prediction obtained by computing the inverse of Eq. (57) evaluated at $t_1 = 40$ min. (d) Same as in (c) but at 37°C.

Both distributions are in a very good agreement with the experimental data, given that no fitting parameters were used other than those of $f_{\rm fit}(t)$. This suggests that the knowledge of the waiting time distribution was sufficient to correctly predict the distribution of the accumulated mature mRNA.

V. CONCLUSIONS

We have presented a general framework that allows us to predict the distribution of nascent RNA from the distribution of waiting times between successive nascent RNA production events, assuming the time of elongation and termination to be fixed. The significance of our solution is that it applies to any initiation mechanism that is modelled by first-order reactions with time-independent rates. The theory also allows for a model-free prediction of the nascent RNA distribution, provided the waiting time distribution is measured experimentally. We have tested our theory on the number of accumulated nascent RNA in live cells, whereas the theory is yet to be tested on the number of nascent RNA that are actively engaged in transcription under steady state conditions.

A major advantage of our theory is that it provides an exact steady-state solution of the nascent RNA distribution for a wide range of initiation mechanisms. This means that given data on the number of nascent RNA in each cell in a population, we can use the theory to compute the exact likelihood of observing this data and from the maximization of the latter we can estimate the optimal parameter values for a given initiation mechanism. More importantly, by doing this for a number of plausible initiation mechanisms, we can use the Bayesian information criterion to select the best mechanism that explains the data.

Important directions for future work include assessing the importance of stochastic events in transcription elongation, such as RNAP collisions, ubiquitous pausing, and premature termination, which have not been accounted for in this framework.

The code and data to reproduce Figs. 2, 3, and 5 are available on GitHub [45].

ACKNOWLEDGMENTS

This work was supported by a research project grant from the Leverhulme Trust (No. RPG-2020-327).

APPENDIX A: WAITING TIME DISTRIBUTIONS FOR THE INITIATION PROCESSES DEPICTED IN FIG. 1

1. The telegraph process

This process has two gene states ($M = 2$). After initiation, the process returns to the state S_2 ($K = 2$). The conditional probability $P_i(t)$ that the gene is in state S_i at time t, given that it was in state S_2 at time 0 and that no nascent RNA production has yet occurred, satisfies the following master equation:

$$\frac{dP_1}{dt} = k_{21}P_2 - k_{12}P_1,$$
$$\frac{dP_2}{dt} = k_{12}P_1 - (k_2 + k_{21})P_2. \tag{A1}$$

with the initial condition $P_i(0) = \delta_{i,2}$. By taking the Laplace transform of both sides, we get

$$sP_1^* = k_{21}P_2^* - k_{12}P_1^*,$$
$$sP_2^* - 1 = k_{12}P_1^* - (k_2 + k_{21})P_2^*, \tag{A2}$$

and from there the Laplace transform of $f(t)$ in Eq. (31).

2. Three-state process that accounts for RNAP recruitment

This process has three gene states ($M = 3$). After initiation, the process returns to the state S_2 ($K = 2$). The conditional probability $P_i(t)$ that the gene is in state S_i at time t, given that it was in state S_2 at time 0 and that no nascent RNA production

has yet occurred, satisfies the following master equation:

$$\begin{aligned}\frac{dP_1}{dt} &= k_{31}P_3 + k_{21}P_2 - k_{12}P_1,\\ \frac{dP_2}{dt} &= k_{12}P_1 + k_{32}P_3 - (k_{21}+k_{23})P_2,\\ \frac{dP_3}{dt} &= k_{23}P_2 - (k_{31}+k_{32}+k_3)P_3,\end{aligned} \tag{A3}$$

with the initial condition $P_i(0) = \delta_{i,2}$. By taking the Laplace transform of both sides, we get

$$\begin{aligned}sP_1^* &= k_{31}P_3^* + k_{21}P_2^* - k_{12}P_1^*,\\ sP_2^* - 1 &= k_{12}P_1^* + k_{32}P_3^* - (k_{21}+k_{23})P_2^*,\\ sP_3^* &= k_{23}P_2^* - (k_{31}+k_{32}+k_3)P_3^*,\end{aligned} \tag{A4}$$

and from there the Laplace transform of $f(t)$ in Eq. (46).

3. A stepwise process that accounts for the binding of general transcription factors and RNAP

This process has M gene states. After initiation, the process returns to the state S_K. The conditional probability $P_i(t)$ that the gene is in state S_i at time t, given that it was in state S_K at time 0 and that no nascent RNA production has yet occurred, satisfies the following master equation:

$$\begin{aligned}\frac{dP_1}{dt} &= k_{21}P_2 - k_{12}P_1,\\ \frac{dP_2}{dt} &= k_{12}P_1 + k_{32}P_3 - (k_{21}+k_{23})P_2,\\ &\vdots\\ \frac{dP_M}{dt} &= k_{M-1M}P_{M-1} - (k_{MM-1}+k_M)P_M,\end{aligned} \tag{A5}$$

with the initial condition $P_i(0) = \delta_{i,K}$. We can write Eq. (A5) compactly as

$$\frac{dP}{dt} = \mathbf{A}P, \quad P(t) = \begin{bmatrix} P_1(t)\\ \vdots\\ P_M(t)\end{bmatrix}, \tag{A6}$$

where $\mathbf{A}$ is a $M \times M$ tridiagonal matrix whose nonzero elements are

$$[\mathbf{A}]_{i,i+1} = k_{i+1,i}, \quad i = 1,\ldots,M-1, \tag{A7}$$

$$[\mathbf{A}]_{i,i} = -e_i, \quad i = 1,\ldots,M, \tag{A8}$$

$$[\mathbf{A}]_{i+1,i} = k_{i,i+1}, \quad i = 1,\ldots,M-1. \tag{A9}$$

Here, $e_1 = k_{12}$, $e_i = k_{ii-1} + k_{ii+1}$ for $i = 2,\ldots,M-1$, and $e_M = k_{MM-1} + k_M$. The solution to Eq. (A6) is

$$P(t) = e^{\mathbf{A}t}P(0), \tag{A10}$$

where $e^{\mathbf{A}t}$ is the matrix exponential. From here, we get the initiation time distribution as

$$f(t) = k_M P_M(t) = k_M(e^{\mathbf{A}t})_{MK}, \tag{A11}$$

and from there, the Laplace transform of $f(t)$,

$$f^*(s) = k_M(s\mathbf{I} - \mathbf{A})^{-1}_{MK}. \tag{A12}$$

Because $\mathbf{A}$ is a tridiagonal matrix, the inverse $(s\mathbf{I}-\mathbf{A})^{-1}$ can be found explicitly [41]. Following [41], we define the following recurrence relations:

$$z_0 = 1, \quad z_1 = s + e_1, \tag{A13a}$$

$$z_i = (s+e_i)z_{i-1} - k_{ii-1}k_{i-1i}z_{i-2}, \tag{A13b}$$

$$y_{M+1} = 1, \quad y_M = s + e_M, \tag{A13c}$$

$$y_j = (s+e_j)y_{j+1} - k_{j+1j}k_{jj+1}y_{j+2}. \tag{A13d}$$

for $i = 2,\ldots M$ and $j = M-1,\ldots,1$. The matrix elements $(s\mathbf{I}-\mathbf{A})^{-1}_{M1}$ and $(s\mathbf{I}-\mathbf{A})^{-1}_{MM}$ can be expressed as, respectively,

$$(s\mathbf{I}-\mathbf{A})^{-1}_{M1} = \frac{\prod_{n=1}^{M-1} k_{nn+1}}{y_2\left(s + e_1 - k_{21}k_{12}\frac{y_3}{y_2}\right)}, \tag{A14}$$

$$(s\mathbf{I}-\mathbf{A})^{-1}_{MM} = \frac{1}{s + e_M - k_{MM-1}k_{M-1M}\frac{z_{M-2}}{z_{M-1}}}, \tag{A15}$$

whereas $(s\mathbf{I}-\mathbf{A})^{-1}_{MK1}$ for $K = 2,\ldots,M-1$ reads

$$(s\mathbf{I}-\mathbf{A})^{-1}_{MK} = \frac{\prod_{n=1}^{M-K} k_{nn+1}}{y_{K+1}\left(s + e_K - k_{KK-1}k_{K-1K}\frac{z_{K-2}}{z_{K-1}} - k_{K+1K}k_{KK+1}\frac{y_{K+2}}{y_{K+1}}\right)}, \quad K = 2,\ldots,M-1. \tag{A16}$$

APPENDIX B: INVERSE LAPLACE TRANSFORM OF $P^*(n,s)$ FOR SELECTED INITIATION PROCESSES

1. The telegraph process

For this process, the Laplace transform $P^*(n,s)$ for $n = 0$ and $n \geqslant 1$ are given by, respectively,

$$P^*(0,s) = \frac{s + \lambda_1 + \lambda_2 - 1/\mu}{(s+\lambda_1)(s+\lambda_2)}, \tag{B1}$$

$$P^*(n,s) = \frac{(s+k_{21}+k_{12})^2[k_2(s+k_{12})]^{n-1}}{\mu[(s+\lambda_1)(s+\lambda_2)]^{n+1}} \tag{B2}$$

By decomposing $P^*(0,s)$ into partial fractions, we get

$$\begin{aligned}P^*(0,s) &= \frac{\mu s + \mu\lambda_1 + \mu\lambda_2 - 1}{\mu(s+\lambda_1)(s+\lambda_2)}\\ &= \frac{\lambda_2 - 1/\mu}{(\lambda_2-\lambda_1)(s+\lambda_1)} + \frac{\lambda_1 - 1/\mu}{(\lambda_1-\lambda_2)(s+\lambda_2)}.\end{aligned} \tag{B3}$$

This can now be inverted using the following identity:

$$\mathcal{L}^{-1}\left\{\frac{1}{(s+\alpha)^n}\right\}(T) = \frac{T^{n-1}}{(n-1)!}e^{-\alpha T}, \tag{B4}$$

from which we get Eq. (37).

For $n \geqslant 1$, the partial fraction decomposition of $P^*(n,s)$ reads

$$P^*(n,s) = \sum_{i=0}^{n} \frac{A_{i,n}}{(s+\lambda_1)^{n+1-i}} + \sum_{i=0}^{n} \frac{B_{i,n}}{(s+\lambda_2)^{n+1-i}}, \quad \text{(B5)}$$

where $A_{i,n}$ and $B_{i,n}$ for $i = 0, \ldots, n$ are given by

$$A_{i,n} = \frac{1}{i!}\frac{d^i}{ds^i}[P^*(n,s)(s+\lambda_1)^{n+1}]\Big|_{s=-\lambda_1}, \quad \text{(B6)}$$

$$B_{i,n} = \frac{1}{i!}\frac{d^i}{ds^i}[P^*(n,s)(s+\lambda_2)^{n+1}]\Big|_{s=-\lambda_2}. \quad \text{(B7)}$$

To find these coefficients, we define an auxiliary function $u(s;\lambda)$,

$$u(s;\lambda) = \frac{(s+k_{12})^{n-1}}{(s+\lambda)^{n+1}}. \quad \text{(B8)}$$

The coefficients $A_{0,n}$ and $A_{1,n}$ read

$$A_{0,n} = (k_{21}+k_{12}-\lambda_1)^2 u(-\lambda_1;\lambda_2), \quad \text{(B9)}$$

$$A_{1,n} = (k_{21}+k_{12}-\lambda_1)^2 u^{(1)}(-\lambda_1;\lambda_2) + 2(k_{21}+k_{12}-\lambda_1)u(-\lambda_1;\lambda_2). \quad \text{(B10)}$$

For $i \geqslant 2$, applying the general Leibniz rule to the product of $(s+k_{12}+k_{21})^2$ and $u(s;\lambda_2)$ leads to

$$A_{i,n} = \frac{1}{i!}(k_{21}+k_{12}-\lambda_1)^2 u^{(i)}(-\lambda_1;\lambda_2) + \frac{2}{(i-1)!}(k_{21}+k_{12}-\lambda_1)u^{(i-1)}(-\lambda_1;\lambda_2) + \frac{1}{(i-2)!}u^{(i-2)}(-\lambda_1;\lambda_2) \quad i = 2,\ldots,n, \quad \text{(B11)}$$

where we have used the notation

$$u^{(i)}(-\lambda_1;\lambda_2) = \frac{d^i}{ds^i}u(s;\lambda_2)\Big|_{s=-\lambda_1}. \quad \text{(B12)}$$

Applying the general Leibniz rule to $u^{(i)}(s;\lambda_2)$, we get

$$u^{(i)}(-\lambda_1,\lambda_2) = (-n-1)_i \frac{(k_{12}-\lambda_1)^{n-1}}{(\lambda_2-\lambda_1)^{n+1+i}} \times {}_2F_1\left(-i, 1-n, -i-n; \frac{\lambda_2-\lambda_1}{k_{12}-\lambda_1}\right), \quad \text{(B13)}$$

where $(x)_i = x(x-1)\ldots(x-i+1)$ is the falling factorial and ${}_2F_1(a,b,c;z)$ is the Gaussian or ordinary hypergeometric function. The coefficient $B_{i,n}$ can be obtained from $A_{i,n}$ by replacing $\lambda_i \leftrightarrow \lambda_2$. This concludes the derivation of the inverse Laplace transform of $P^*(n,s)$ for the telegraph process.

By inserting Eq. (31) into Eq. (22), we get the following expression for the Laplace transform of the probability generating function $G^*(z,s)$,

$$G^*(z,s) = \frac{s+k_{21}+k_{12}-k_2u+u/\mu}{s^2+s(k_{21}+k_{12}-k_2u)-k_{12}k_2u}, \quad \text{(B14)}$$

where $u = z-1$. Next, we introduce $\lambda_1(u)$ and $\lambda_2(u)$ such that

$$s^2+s(k_{21}+k_{12}-k_2u)-k_{12}k_2u = [s+\lambda_1(u)][s+\lambda_2(u)]. \quad \text{(B15)}$$

We now write

$$G^*(u,s) = \frac{s+k_{21}+k_{12}-k_2u+u/\mu}{[s+\lambda_1(u)][s+\lambda_2(u)]}, \quad \text{(B16)}$$

which can be inverted using the partial fraction decomposition. The final result for $G(u)$ is

$$G(u) = \frac{e^{-\lambda_1(u)T}}{2(k_1+k_{21})\sqrt{\Delta(u)}}\{(k_1+k_{21})[\sqrt{\Delta(u)}-(k_1+k_{21})]-(k_1-k_{21})k_2u + (k_1+k_{21})[\sqrt{\Delta(u)}+(k_1+k_{21})]e^{\sqrt{\Delta(u)}T} + (k_1-k_{21})k_2u\, e^{\sqrt{\Delta(u)}T}\}, \quad \text{(B17)}$$

where $\Delta(u) = (k_{21}+k_{12}-k_2u)^2 + 4k_{12}k_2u$.

2. Fully irreversible process with arbitrary number of steps with equal rates

For this process, the inverse Laplace transform $P^*(0,s)$ and $P^*(n,s)$ for $n \geqslant 1$ are given by, respectively,

$$P^*(0,s) = \frac{(Ms-\lambda)(s+\lambda)^M + \lambda^{M+1}}{Ms^2(s+\lambda)^M}, \quad \text{(B18)}$$

$$P^*(n,s) = \frac{\lambda^{M(n-1)+1}[(s+\lambda)^M - \lambda^M]^2}{Ms^2(s+\lambda)^{M(n+1)}}. \quad \text{(B19)}$$

The partial fraction decomposition of $P^*(n,s)$ reads

$$P^*(n,s) = \sum_{i=0}^{M(n+1)-1} \frac{A_{i,n}}{(s+\lambda)^{M(n+1)-i}}, \quad \text{(B20)}$$

where

$$A_{i,n} = \frac{1}{i!}\frac{d^i}{ds^i}[P^*(n,s)(s+\lambda)^{M(n+1)}]\Big|_{s=-\lambda}, \quad \text{(B21)}$$

and $i = 0, \ldots, M(n+1)-1$. For $n = 0$,

$$A_{i,0} = \frac{(i+1)\lambda^{M-1-i}}{M}, \quad i = 0,\ldots,M-1. \quad \text{(B22)}$$

Inserting this back into Eq. (B20) and inverting the Laplace transform gives

$$P(N=0) = \frac{e^{-\lambda T}}{M} \sum_{i=0}^{M-1} (M-i) \frac{(\lambda T)^i}{i!}. \tag{B23}$$

For $n \geqslant 1$, $A_{i,n}$ is given by

$$A_{i,n} = i!\lambda^{2M-2-i}[(i-2M+1)\mathbb{1}\{i \geqslant 2M\} - 2(i-2M+1)\mathbb{1}\{i \geqslant M\} + i + 1], \tag{B24}$$

where $\mathbb{1}\{x \geqslant y\} = 1$ if $x \geqslant y$ and is 0 if $x < y$. Inserting $A_{i,n}$ into the expression for $P^*(n,s)$ and taking the inverse Laplace transform yields the nascent RNA distribution in Eq. (44).

[1] 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**, 472 (2011).

[2] J. Peccoud and B. Ycart, Markovian modeling of gene-product synthesis, Theor. Popul. Biol. **48**, 222 (1995).

[3] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, Genomic encoding of transcriptional burst kinetics, Nature (London) **565**, 251 (2019).

[4] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, Stochastic mRNA synthesis in mammalian cells, PLoS Biol **4**, e309 (2006).

[5] L. Ham, R. D. Brackston, and M. P. H. Stumpf, Extrinsic Noise and Heavy-Tailed Laws in Gene Expression, Phys. Rev. Lett. **124**, 108101 (2020).

[6] H. Xu, S. O. Skinner, A. M. Sokac, and I. Golding, Stochastic Kinetics of Nascent RNA, Phys. Rev. Lett. **117**, 128101 (2016).

[7] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, Single-cell analysis of transcription kinetics across the cell cycle, eLife **5**, e12175 (2016).

[8] G.-J. Hendriks, L. A. Jung, A. J. Larsson, M. Lidschreiber, O. A. Forsman, K. Lidschreiber, P. Cramer, and R. Sandberg, NASC-seq monitors rna synthesis in single cells, Nat. Commun. **10**, 3138 (2019).

[9] Y. Wan, D. G. Anastasakis, J. Rodriguez, M. Palangat, P. Gudla, G. Zaki, M. Tandon, G. Pegoraro, C. C. Chow, M. Hafner *et al.*, Dynamic imaging of nascent rna reveals general principles of transcription dynamics and stochastic splice site selection, Cell **184**, 2878 (2021).

[10] H. Xu, L. A. Sepúlveda, L. Figard, A. M. Sokac, and I. Golding, Combining protein and mRNA quantification to decipher transcriptional regulation, Nat. Methods **12**, 739 (2015).

[11] L. A. Sepúlveda, H. Xu, J. Zhang, M. Wang, and I. Golding, Measurement of gene regulation in individual cells reveals rapid switching between promoter states, Science **351**, 1218 (2016).

[12] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer, Real-time observation of transcription initiation and elongation on an endogenous yeast gene, Science **332**, 475 (2011).

[13] L. F. Lafuerza and R. Toral, Exact solution of a stochastic protein dynamics model with delayed degradation, Phys. Rev. E **84**, 051121 (2011).

[14] Q. Jiang, X. Fu, S. Yan, R. Li, W. Du, Z. Cao, F. Qian, and R. Grima, Neural network aided approximation and parameter inference of non-Markovian models of gene expression, Nat. Commun. **12**, 2618 (2021).

[15] S. Choubey, J. Kondev, and A. Sanchez, Deciphering transcriptional dynamics in vivo by counting nascent RNA molecules, PLoS Comput. Biol. **11**, e1004345 (2015).

[16] S. Choubey, J. Kondev, and A. Sanchez, Distribution of initiation times reveals mechanisms of transcriptional regulation in single cells, Biophys. J. **114**, 2072 (2018).

[17] S. Braichenko, J. Holehouse, and R. Grima, Distinguishing between models of mammalian gene expression: Telegraph-like models versus mechanistic models, J. R. Soc. Interface. **18**, 20210510 (2021).

[18] S. Sainsbury, C. Bernecky, and P. Cramer, Structural basis of transcription initiation by RNA polymerase II, Nat. Rev. Mol. Cell Biol. **16**, 129 (2015).

[19] R. Milo and R. Phillips, *Cell Biology by the Numbers* (Garland Science, New York, 2015), pp. 132–135.

[20] O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj, Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms, Mol. Cell **58**, 339 (2015).

[21] C. R. Bartman, N. Hamagami, C. A. Keller, B. Giardine, R. C. Hardison, G. A. Blobel, and A. Raj, Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation, Mol. Cell **73**, 519 (2019).

[22] R. E. Kingston and M. R. Green, Modeling eukaryotic transcriptional activation, Curr. Biol. **4**, 325 (1994).

[23] K. Adelman and J. T. Lis, Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans, Nat. Rev. Genet. **13**, 720 (2012).

[24] I. Jonkers, H. Kwak, and J. T. Lis, Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons, eLife **3**, e02407 (2014).

[25] A.-B. Muthukrishnan, M. Kandhavelu, J. Lloyd-Price, F. Kudasov, S. Chowdhury, O. Yli-Harja, and A. S. Ribeiro, Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells, Nucleic Acids Res. **40**, 8472 (2012).

[26] D. Cox, *Renewal Theory* (Methuen, London, 1967).

[27] H. Kung and D. Tong, Fast algorithms for partial fraction decomposition, SIAM J. Comput. **6**, 582 (1977).

[28] X. Fu, H. P. Patel, S. Coppola, L. Xu, Z. Cao, T. L. Lenstra, and R. Grima, Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions, eLife **11**, e82493 (2022).

[29] X. Fu, X. Zhou, D. Gu, Z. Cao, and R. Grima, DelaySSAToolkit.jl: Stochastic simulation of reaction systems with time delays in Julia, Bioinformatics **38**, 4243 (2022).

[30] R. Perez-Carrasco, C. Beentjes, and R. Grima, Effects of cell cycle variability on lineage and population measurements of messenger rna abundance, J. R. Soc. Interface. **17**, 20200360 (2020).

[31] L. Zawel, K. P. Kumar, and D. Reinberg, Recycling of the general transcription factors during RNA polymerase II transcription, Genes Dev. **9**, 1479 (1995).

[32] D. Yean and J. D. Gralla, Transcription reinitiation rate: A potential role for TATA box stabilization of the TFIID:TFIIA:DNA complex, Nucleic Acids Res. **27**, 831 (1999).

[33] B. Hoopes, J. LeBlanc, and D. Hawley, Kinetic analysis of yeast TFIID-TATA box complex formation suggests a multistep pathway, J. Biol. Chem. **267**, 11539 (1992).

[34] Z. Zhang, B. P. English, J. B. Grimm, S. A. Kazane, W. Hu, A. Tsai, C. Inouye, C. You, J. Piehler, P. G. Schultz *et al.*, Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis, Genes Dev. **30**, 2106 (2016).

[35] C. A. Weideman, R. C. Netter, L. R. Benjamin, J. J. McAllister, L. A. Schmiedekamp, R. A. Coleman, and B. Pugh, Dynamic interplay of TFIIA, TBP and TATA DNA, J. Mol. Biol. **271**, 61 (1997).

[36] G. A. Rosen, I. Baek, L. J. Friedman, Y. J. Joo, S. Buratowski, and J. Gelles, Dynamics of RNA polymerase II and elongation factor Spt4/5 recruitment during activator-dependent transcription, Proc. Natl. Acad. Sci. USA **117**, 32348 (2020).

[37] I. Baek, L. J. Friedman, J. Gelles, and S. Buratowski, Single-molecule studies reveal branched pathways for activator-dependent assembly of RNA polymerase II pre-initiation complexes, Mol. Cell **81**, 3576 (2021).

[38] L. Friedman and J. Gelles, Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation, Cell **148**, 679 (2012).

[39] M. S. Buckley, H. Kwak, W. R. Zipfel, and J. T. Lis, Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation, Genes Dev. **28**, 14 (2014).

[40] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer, In vivo dynamics of RNA polymerase II transcription, Nat. Struct. Mol. Biol. **14**, 796 (2007).

[41] Y. Huang and W. F. McColl, Analytical inversion of general tridiagonal matrices, J. Phys. A: Math. Gen. **30**, 7919 (1997).

[42] J. Colgan and J. L. Manley, TFIID can be rate limiting in vivo for TATA-containing, but not TATA-lacking, RNA polymerase II promoters, Genes Dev. **6**, 304 (1992).

[43] I. Golding and E. C. Cox, RNA dynamics in live Escherichia coli cells, Proc. Natl. Acad. Sci. USA **101**, 11310 (2004).

[44] T. T. Le, S. Harlepp, C. C. Guet, K. Dittmar, T. Emonet, T. Pan, and P. Cluzel, Real-time RNA profiling within a single bacterium, Proc. Natl. Acad. Sci. USA **102**, 9160 (2005).

[45] https://github.com/jszavits/nascentRNA.