Markovian approaches to modeling intracellular reaction processes with molecular memory

Jiajun Zhang[a,b] and Tianshou Zhou[a,b,1]

[a]Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, 510275 Guangzhou, P. R. China; and [b]School of Mathematics, Sun Yat-sen University, 510275 Guangzhou, P. R. China

Edited by Hong Qian, University of Washington, Seattle, WA, and accepted by Editorial Board Member Curtis G. Callan Jr. October 11, 2019 (received for review August 13, 2019)

Many cellular processes are governed by stochastic reaction events. These events do not necessarily occur in single steps of individual molecules, and, conversely, each birth or death of a macromolecule (e.g., protein) could involve several small reaction steps, creating a memory between individual events and thus leading to nonmarkovian reaction kinetics. Characterizing this kinetics is challenging. Here, we develop a systematic approach for a general reaction network with arbitrary intrinsic waiting-time distributions, which includes the stationary generalized chemical-master equation (sgCME), the stationary generalized Fokker–Planck equation, and the generalized linear-noise approximation. The first formulation converts a nonmarkovian issue into a markovian one by introducing effective transition rates (that explicitly decode the effect of molecular memory) for the reactions in an equivalent reaction network with the same substrates but without molecular memory. Nonmarkovian features of the reaction kinetics can be revealed by solving the sgCME. The latter 2 formulations can be used in the fast evaluation of fluctuations. These formulations can have broad applications, and, in particular, they may help us discover new biological knowledge underlying memory effects. When they are applied to generalized stochastic models of gene-expression regulation, we find that molecular memory is in effect equivalent to a feedback and can induce bimodality, fine-tune the expression noise, and induce switch.

chemical-master equation | biochemical-reaction system | nonmarkovian reaction kinetics | gene-expression noise | molecular memory

Quantitative understanding of the dynamics of single living cells is one of the main goals of modern molecular-systems biology. Traditionally, modeling of intracellular biochemical processes is based on the markovian hypothesis, i.e., the stochastic motion of the reactants is uninfluenced by previous states, only by the current state. This memoryless property implies that markovian reaction kinetics can be described by poissonian processes with constant rates, which are characterized by exponential waiting-time distributions (1, 2). The mathematical tractability of markovian reaction processes enables great simplifications in problem formulation, leading to important successes in the description of many intracellular processes (1–4). These studies revealed that the characteristic parameters in the waiting-time distributions, as well as the complex properties of the intracellular process, can have a strong impact on the reaction kinetics.

However, intracellular reaction processes are not necessarily markovian but may be nonmarkovian. First, as a general rule, the dynamics of a given reactant resulting from its interactions with the environment cannot be described as a markovian process since this interaction can create "molecular memory" characterized by nonexponential waiting-time distributions. Second, the reduction of multistep reactions into a single step one may lead to a nonmarkovian process (5). Third, synthesis of a macromolecule would involve several small single- or multimolecular reaction steps (referring to Fig. 1*A*), creating a memory between individual events (6–9). This is evidenced by the fact that inactive phases of the promoter involving the prolactin gene in a mammalian cell are differently distributed, showing strong memory (10). More generally, the complex control process of gene expression, which would involve several repressors, transcription factors (TFs), and mediators as well as chromatin remodeling or changes in supercoiling, can generate nonexponential time intervals between transcription windows. Indeed, molecular memory, which can result in nonmarkovian (or nonpoissonian) reaction kinetics, has been confirmed by the increasing availability of time-resolved data on different kinds of interactions (10–16).

The extensive existence of molecular memory raises important yet unsolved questions: e.g., how does this memory affect nonmarkovian reaction kinetics? In what way and how accurately do life forms achieve the order required to develop and sustain their lives from the disordered reaction events? These questions were partially addressed in a seminal paper (6) wherein Pedraza and Paulsson analyzed a queuing model of stochastic gene expression with molecular memory, which was further generalized and analyzed (7, 17, 18). Recently, Park et al. (19) analyzed another class of queuing models of stochastic gene expression with molecular memory and presented the chemical fluctuation theorem, which provides an accurate relationship between the environment-coupled chemical dynamics of gene expression and gene-expression variability. These analyzed models are based on queuing theory, which is useful in treating simple single-molecular reaction networks such as birth–death processes, but it seems difficult to extend

Significance

Modeling intracellular processes has long relied on the markovian assumption. However, as soon as a reactant interacts with its environment, molecular memory definitely exists and its effects cannot be neglected. Since the Markov theory cannot translate directly to modeling and analysis of nonmarkovian processes, this leads to many significant challenges. We develop a formulation, namely the stationary generalized chemical-master equation, to model intracellular processes with molecular memory. This formulation converts a nonmarkovian question to a markovian one while keeping the stationary probabilistic behavior unchanged. Both a stationary generalized Fokker–Planck equation and a generalized linear noise approximation are further developed for the fast evaluation of fluctuations. These formulations can have broad applications and may help us discover new biological knowledge.

Author contributions: J.Z. and T.Z. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. H.Q. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: mcszhtsh@mail.sysu.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1913926116/-/DCSupplemental.

First published November 4, 2019.

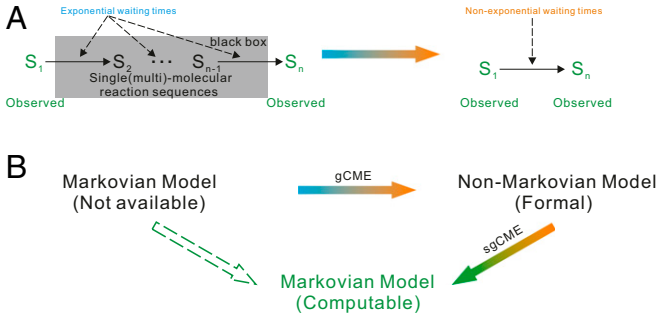

Fig. 1. A framework for analysis of stochastic reaction processes with molecular memory. (*A*) A possibly involved multistep multimolecular reaction process with exponential waiting times is mapped onto a single-step reaction process with nonexponential waiting times, where the initial and final states are observable or measurable, but the intermediate states, which are consequences of single-molecular or multimolecular reactions, would be unknown or have not been unspecified. Note that $S_i \to S_{i+1}$ does not represent a reaction but represents the transformational relation between states S_i and S_{i+1}, which possibly involves single- or multimolecular reactions. (*B*) Our theory shows that a nonmarkovian problem can be converted into a markovian one via an sgCME.

it to complex bi- or multimolecular reactions, which are a characteristic of many intracellular reaction networks. On the other hand, the continuous time random walk (CTRW) framework provides a different modeling way for stochastic reaction processes and has been extensively used in modeling and analysis of stochastic processes (20–22). Despite this, we still lack a general theory for modeling and analysis of general reaction networks with intrinsic waiting-time distributions. Since the sources of nonmarkovianity and network topology can jointly influence the behavior of the whole biochemical system, the reliable tools and mathematical machinery of the Markov theory do not translate directly to modeling and analysis of nonmarkovian reaction processes, leading to many significant challenges. It is needed to develop new methods to characterize nonmarkovian reaction kinetics.

In recent years, the interest in nonmarkovian processes within scientific communities has been blossoming, from the viewpoints of experimental observations (10–16), mathematical modeling (5–7, 17–30), and numerical simulation (30–38). However, because of difficulties in treating nonmarkovian reaction kinetics, only a few exact master equations are known in the literature, most of which are rather formal, or concerning specific systems or stochastic processes (6, 7, 17–20, 39, 40). The aim of this paper is to fill this gap by introducing useful formulations for a general biochemical-reaction system with molecular memory. The basic idea is that we first construct a topology-equivalent, memoryless reaction network by introducing an effective transition rate (ETR) for each reaction in the original nonmarkovian system and then establish an effective master equation for the equivalent reaction network, whose stationary equation is termed as a stationary generalized chemical-master equation (sgCME). The sgCME exactly captures the stationary probabilistic behavior of the original nonmarkovian system. In other words, the original nonmarkovian question is converted into a markovian one (Fig. 1*B*), greatly simplifying analysis of nonmarkovian processes. Based on the sgCME, we further develop another 2 useful techniques for studying stochastic reaction processes on networks: stationary generalized Fokker–Planck equation (sgFPE) and generalized linear noise approximation (gLNA), each enabling fast yet effective evaluation of fluctuations, including the effect of molecular memory.

We demonstrate the power of the above 3 techniques by analyzing 4 examples: a generalized model of constitutive gene expression, a generalized model of gene self-regulation, a generalized ON-OFF model, and a generalized model of genetic toggle switch, each considering molecular memory. These examples are chosen so that analytical or numerical approximations of solutions to the sgCMEs can easily be obtained, and these approximations are then used to clarify the origins of nonmarkovianity and fluctuations in each case and to trace the effects of different parameters on the stochastic properties of the systems. The results reveal the importance of the effect of molecular memory on reaction kinetics.

Methods

A Theoretical Framework for Reaction Processes on a Network. Before presenting our sgCME, let us give the representation of a general reaction network in terms of intrinsic waiting-time distributions. In order to cast the dynamics of N different species that participate in L different reactions (numbered by i with $1 \le i \le L$) into a CTRW framework, we first define a state space. Denote by X_j the chemical species and by n_j the corresponding particle number, where $1 \le j \le N$. Let $\boldsymbol{n} = (n_1, \ldots n_N)^{\mathrm{T}}$ represent the state vector of particle numbers, where T denotes transpose. Denote by r_{ij} and s_{ij} the stoichiometric coefficients for the ith reaction, i.e., r_{ij} and s_{ij} represent the loss and gain in particle number n_j during the ith reaction, respectively. On the other hand, a single event of reaction i is characterized by the intrinsic waiting time τ_i, whose probability density function (PDF), denoted by $\psi_i(t;\boldsymbol{n})$, depends, in general, on the system state $\boldsymbol{n}$. To that end, the above reaction network can be represented by

$$\sum_{j=1}^{N} r_{ij} X_j \xrightarrow{\psi_i(t;\boldsymbol{n})} \sum_{j=1}^{N} s_{ij} X_j,\ i = 1,2,\cdots,L. \qquad [1]$$

Note that $\nu_{ji} = s_{ij} - r_{ij}$ represents the net change of X_i in the ith reaction. Function $\psi_i(t;\boldsymbol{n})$ is called the intrinsic waiting-time distribution for reaction i (20, 39, 40). We emphasize that such a waiting-time distribution is an extension of exponential waiting-time distribution in the markovian models but is different from waiting-time distributions in queuing theory (21). In addition, intrinsic waiting-time distributions are suitable to the description of any reactions, including bimolecular or multimolecular reactions, so they are also extensions of delay distributions introduced in refs. 5 and 28, wherein the authors proposed a methodology to represent chains of single-molecular reactions by simpler, reduced models.

If all of the reaction events happen in a markovian (or memoryless) manner, then $\psi_i(t;\boldsymbol{n})$ reduces to $\psi_i(t;\boldsymbol{n}) = a_i(\boldsymbol{n})e^{-a_i(\boldsymbol{n})t}$ for all $t \ge 0$ (hereafter, we always assume $t \ge 0$), where $a_i(\boldsymbol{n})$ is the reaction-propensity function for reaction i. This expression is natural since exponential waiting-time distributions, which have a clear biophysical foundation (41), are a main characteristic of markovian reaction processes. In contrast to the times until the next molecular events happen in single-step processes (elementary reactions), the waiting times between reaction events are not exponentially distributed in general. If the process goes through a series of (identical) exponential steps, the waiting times will be gamma-distributed (42). If we consider an intrinsic waiting-time distribution of the gamma type: $\psi_i(t;\boldsymbol{n}) = ((\lambda_i(\boldsymbol{n}))^{L_i}/\Gamma(k_i))t^{L_i-1}e^{-\lambda_i(\boldsymbol{n})t}$, where L_i may be understood as the number of small yet possibly unspecified reaction steps, then $L_i = 1$ corresponds to an exponential waiting-time distribution, whereas $L_i \ne 1$ (called memory index) to a nonexponential waiting-time distribution with the noise intensity (defined as the ratio of the variance over the squared mean) being $1/L_i$. Note that for a reaction network, as long as there is a reaction such that the corresponding memory index is not equal to 1, the whole reaction kinetics is nonmarkovian. In this paper, we will only consider intrinsic waiting-time distributions of $L_i = 1$ (exponential) and $L_i \ne 1$ (nonexponential).

sgCME.

Effective transition rates. For convenience, let $\Psi_i(t;\boldsymbol{n})$ be the cumulative distribution function of $\psi_i(t;\boldsymbol{n})$, i.e., $\Psi_i(t;\boldsymbol{n}) = \int_0^t \psi_i(t';\boldsymbol{n})dt'$. If we define $\varphi_i(t;\boldsymbol{n}) = \psi_i(t;\boldsymbol{n})\prod_{j\ne i}[1-\Psi_j(t;\boldsymbol{n})]$, then $\varphi_i(n;t)dt$ represents the probability that the ith reaction happens and the reaction waiting time is in interval $[t, t+dt]$. Note that the conservative condition $\sum_{i=1}^{L}\int_0^\infty \varphi_i(t';\boldsymbol{n})dt' = 1$ always holds.

Now, we introduce a memory function for reaction i, denoted by $M_i(t;\boldsymbol{n})$, which is defined through the Lapalace transform: $\tilde{M}_i(s;\boldsymbol{n}) = s\tilde{\varphi}_i(s;\boldsymbol{n})/[1-\sum_{j=1}^{L}\tilde{\varphi}_j(s;\boldsymbol{n})]$, where $1 \le i \le L$, and the above bar represents the Laplace transform of a function, e.g., $\tilde{\varphi}_i(s;\boldsymbol{n}) = \int_0^\infty e^{-st}\varphi_i(t;\boldsymbol{n})dt$. To help the reader understand function $M_i(t;\boldsymbol{n})$, let us consider waiting-time distributions of special forms: $\psi_i(t;\boldsymbol{n}) = a_i(\boldsymbol{n})e^{-a_i(\boldsymbol{n})t}$, with $a_i(\boldsymbol{n})$ being the reaction-propensity function for reaction i $(1 \le i \le L)$, implying that all of the reaction events

happen in markovian manners. Then, by the expression of $\tilde{M}_i(s;\boldsymbol{n})$ in combination with the inverse of the Laplace transform, we can show $M_i(t;\boldsymbol{n})=a_i(\boldsymbol{n})$ for all $t\geq 0$. In this case, the memory functions reduce to the common reaction-propensity functions.

Interestingly, we find that the limit $\lim_{s\to 0}\tilde{M}_i(s;\boldsymbol{n})$ exists, given waiting-time distributions for a reaction network. Moreover, if this limit is denoted by $K_i(\boldsymbol{n})$, then it can be analytically expressed as (see *SI Appendix* for details)

$$K_i(\boldsymbol{n})=\frac{\int_0^{+\infty}\psi_i(t;\boldsymbol{n})\prod_{j\neq i}[1-\Psi_j(t;\boldsymbol{n})]dt}{\int_0^{+\infty}\prod_{j=1}^{L}[1-\Psi_j(t;\boldsymbol{n})]dt}, \qquad [2]$$

which is termed as the ETR for reaction i, where $i=1,\cdots,L$. If there exists some subscript i_0 such that $\psi_{i_0}(t;\boldsymbol{n})=a_{i_0}(\boldsymbol{n})e^{-a_{i_0}(\boldsymbol{n})t}$ for all $t\geq 0$, i.e., if the waiting time of the i_0th reaction follows an exponential distribution, function $K_{i_0}(\boldsymbol{n})$ reduces to the common reaction-propensity function, i.e., $K_{i_0}(\boldsymbol{n})=a_{i_0}(\boldsymbol{n})$ (see *SI Appendix* for details). If waiting-time distributions for all of the reactions are exponential, the nonmarkovian reaction network reduces to a markovian reaction one. Therefore, ETRs are extensions of reaction-propensity functions. However, different from function $a_i(\boldsymbol{n})$ that is a polynomial of $\boldsymbol{n}$, function $K_i(\boldsymbol{n})$ is in general a rational function of $\boldsymbol{n}$. For example, consider a generalized birth–death process: $\varnothing \xrightarrow{\psi_1(t;n)} X$ and $X \xrightarrow{\psi_2(t;n)} \varnothing$, where n represents the number of X molecules. If $\psi_2(t;n)=n\lambda_2 e^{-n\lambda_2 t}$ and $\psi_1(t;n)=((\lambda_1)^{L_1}/\ \Gamma(k_1))t^{L_1-1}e^{-\lambda_1 t}$ with L_1 being a positive integer, we can show $K_2(n)=n\lambda_2$ and $K_1(n)=(\lambda_1)^{L_1}(n\lambda_2)/((\lambda_1+n\lambda_2)^{L_1}-(\lambda_1)^{L_1})$, which is apparently a rational function of n if $L_1>1$. *SI Appendix* also gives the explicit expressions of ETRs in the case of environmental perturbations or external noise that can lead to stochastic reaction delays.

Effective CME. For the above (nonmarkovian) reaction network, if we let $P(\boldsymbol{n};t)$ be the probability that the system is in state $\boldsymbol{n}$ at time t, then according to the CTRW theory (20, 39, 40), we can derive the following CME expressed in Laplace transforms

$$s\tilde{P}(\boldsymbol{n};s)=P(\boldsymbol{n};0)+\sum_{i=1}^{L}\tilde{M}_i(s;\boldsymbol{n}-\boldsymbol{\nu}_i)\tilde{P}(\boldsymbol{n}-\boldsymbol{\nu}_i;s)-\sum_{i=1}^{L}\tilde{M}_i(s;\boldsymbol{n})\tilde{P}(\boldsymbol{n};s). \qquad [3]$$

By the final value theorem (42), we know that if the limit $\lim_{t\to\infty}f(t)$ exists, 2 limits $\lim_{t\to\infty}f(t)$ and $\lim_{s\to 0}s\tilde{f}(s)$ are equal, i.e., $\lim_{t\to\infty}f(t)=\lim_{s\to 0}s\tilde{f}(s)$. Multiplying s on both sides of Eq. 3, letting $s\to 0$ and using $K_i(\boldsymbol{n})=\lim_{s\to 0}\tilde{M}_i(s;\boldsymbol{n})$, we can obtain the following stationary equation (*SI Appendix*):

$$\sum_{i=1}^{L}[K_i(\boldsymbol{n}-\boldsymbol{\nu}_i)P(\boldsymbol{n}-\boldsymbol{\nu}_i)-K_i(\boldsymbol{n})P(\boldsymbol{n})]=0, \qquad [4]$$

which is called the sgCME, where $P(\boldsymbol{n})$ is the stationary distribution corresponding to $P(\boldsymbol{n};t)$ and is assumed to exist (numerical simulations verifies this, referring to Fig. 2*C*). Eq. 4 with Eq. 2 is one of the main results of this paper.

On the other hand, we can use $K_i(\boldsymbol{n})$ to construct a reaction network. This is done by taking $K_i(\boldsymbol{n})$ as the reaction-propensity function for the ith reaction in an topologically equivalent reaction network with the same substrates but without molecular memory. For this reaction network, we can establish its CME. In fact, if we let $Q(\boldsymbol{n};t)$ be the probability that the system is in state $\boldsymbol{n}$ at time t, the corresponding CME takes the form

$$\frac{\partial Q(\boldsymbol{n};t)}{\partial t}=\sum_{i=1}^{L}[K_i(\boldsymbol{n}-\boldsymbol{\nu}_i)Q(\boldsymbol{n}-\boldsymbol{\nu}_i;t)-K_i(\boldsymbol{n})Q(\boldsymbol{n};t)], \qquad [5]$$

which is called an effective CME for the original nonmarkovian reaction system, where "effective" will be interpreted below.

Notably, the stationary equation corresponding to Eq. 5 is exactly the same as Eq. 4 except for notation, implying that the stationary probabilistic behavior of the original nonmarkovian reaction system is exactly the same as that of the constructed markovian reaction network. In this sense, the original nonmarkovian problem is converted to a markovian one. However, there would exist differences in dynamic probability behavior between the 2 reaction networks, referring to Fig. 2*C*, which also verifies that the stationary distribution indeed exists even in the presence of molecular memory. Despite this difference, Eq. 5 provides a way for studying complex nonmarkovian reaction kinetics.

Generalized Linear Noise Approximation. In analysis of reaction networks, an extensively used technique is the linear noise approximation (LNA) (1, 2). Here, we derive a generalized LNA (gLNA) for the above general reaction system with arbitrary waiting-time distributions.

First, the rate equations corresponding to the constructed-above markovian reaction network read

$$\frac{dx_k}{dt}=\sum_{i=1}^{L}\nu_{ik}K_i(\boldsymbol{x}),\ k=1,2,\cdots,N, \qquad [6]$$

where $\boldsymbol{x}=(x_1,\cdots,x_N)^{\mathrm{T}}$, x_k represents the concentration of reactive species X_k, i.e., $x_k=\lim_{\Omega\to\infty}n_k/\Omega$ with Ω being the volume of the system, and $K_i(\boldsymbol{x})$ is given by Eq. 2 if $\boldsymbol{n}$ is replaced with $\boldsymbol{x}$. Let $\mathbf{S}=(\nu_{ij})$ be the stoichiometric matrix and $\boldsymbol{K}(\boldsymbol{x})=(K_1(\boldsymbol{x}),\cdots,K_L(\boldsymbol{x}))^{\mathrm{T}}$ be a column vector of ETRs. Then, the algebraic equation

$$\mathbf{S}\boldsymbol{K}(\boldsymbol{x})=\mathbf{0} \qquad [7]$$

determines the steady state of the deterministic system described by Eq. 6, which is denoted by $\boldsymbol{x}_s$.

Then, we adopt the Ω-expansion method (1, 2) to derive an algebraic equation for covariance matrix. Write $\boldsymbol{n}=\Omega\boldsymbol{x}+\Omega^{1/2}\boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of random variables. Under this transform, the original $P(\boldsymbol{n};t)$ becomes another

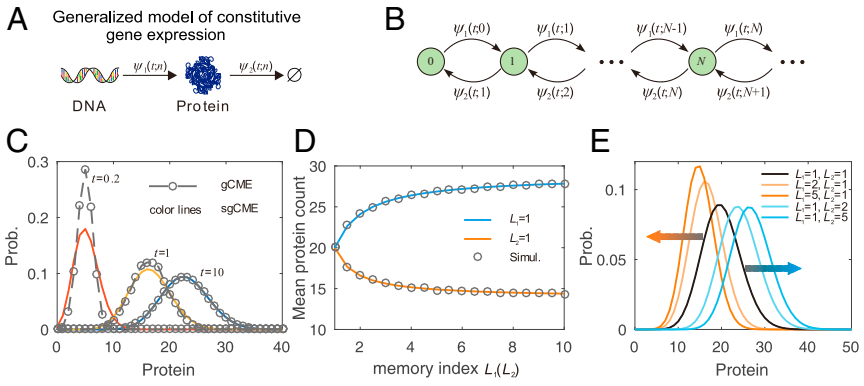

Fig. 2. Analysis of a generalized model of constitutive gene expression without self-regulation, i.e.,$\psi_1(t;n)$ is independent of n. (*A*) Schematic representation of the model. (*B*) Transitions between states. (*C*) Protein distributions at several different time points, where curves with empty circles correspond to the original nonmarkovian model, whereas colored curves correspond to the constructed markovian model. (*D*) Influences of memory index L_1 (L_2) on mean protein levels, where empty circles represent the results obtained by the modified "first-reaction" Gillespie algorithm (41), whereas the lines represent the results obtained by theoretical predictions. (*E*) Stationary protein distributions for different values of L_1 and L_2. Parameter values are set as: $L_1=4$, $\lambda_1=120$, and $L_2=1$, $\lambda_2=1$ (*C*) and $\lambda_1=20L_1$, $\lambda_2=L_2$ (*D* and *E*).

$\Pi(\boldsymbol{z};t)$. According to *SI Appendix*, we can derive the following Lyapunov matrix equation:

$$\mathrm{A_S}\boldsymbol{\Sigma_S}+\boldsymbol{\Sigma_S}\mathrm{A_S^T}+\mathbf{D_S}=\mathbf{0}, \quad [8]$$

where $\mathrm{A_S}=(A_{ij})$ is a matrix with elements $A_{ij}=\sum_{k=1}^{L}\nu_{ki}\frac{\partial K_k(\boldsymbol{x}_s)}{\partial x_j}=\frac{\partial[\mathbf{S}_i\boldsymbol{K}(\boldsymbol{x}_s)]}{\partial x_j}$, $\mathbf{D_s}=\mathbf{BB}^{\mathrm{T}}$ is a noise matrix with elements $[\mathbf{BB}^{\mathrm{T}}]_{ij}=\sum_{k=1}^{L}\nu_{ki}\nu_{kj}K_k(\boldsymbol{x}_s)=[\mathbf{S}^{\mathrm{T}}\mathrm{diag}(K_1(\boldsymbol{x}_s),\cdots,K_L(\boldsymbol{x}_s))\mathbf{S}]_{ij}$, and $\sum_s=\langle\langle(\boldsymbol{x}-\boldsymbol{x}_s)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}_s)\rangle\rangle$ is a covariance matrix to be unknown. Note that the diagonal elements of $\boldsymbol{\Sigma_S}$ represent the variances of the random variables, and the vector of the mean concentrations of the reactive species is approximately given by $\langle\boldsymbol{x}\rangle\approx\boldsymbol{x}_s$.

We point out that the above analysis framework is convenient for both clarifying the origins of nonmarkovianity (including fluctuations) and tracing the effects of different parameters on the stochastic properties of the underlying systems. *SI Appendix* provides details for examples of analysis by the gLNA. We point out that a slow-scale LNA (ssLNA) has been developed, which describes a class of nonmarkovian systems stemming from timescale separation (43). Specifically, starting with a markovian system composed of fast and slow species, the authors of that paper derived the LNA for the nonmarkovian system, which describes only the slow species (the observables).

sgFPE. As an effective approximation of the CME in some situations, the Fokker–Planck equation (FPE) has extensively been used (1, 2, 44, 45), mainly because the latter is more easily analyzed and can often provide more intuitive understanding of a biochemical system than the former. However, the FPE has not been established in the presence of molecular memory. Here, we derive a sgFPE for a general nonmarkovian reaction network with arbitrary waiting-time distributions.

First, although Eq. 3 holds for the discrete variables, it also holds for the corresponding continuous variables. Second, Taylor-expanding the CME in the case of continuous variables to the second-order term yields

$$\begin{aligned}s\tilde{P}(\boldsymbol{x},s)-\tilde{P}(\boldsymbol{x},0)\approx&-\sum_{i=1}^{L}\sum_{k=1}^{N}\nu_{ik}\frac{\partial}{\partial x_k}\left[\tilde{M}_i(s;\boldsymbol{x})\tilde{P}(\boldsymbol{x},s)\right]\\&+\frac{1}{2}\sum_{i=1}^{L}\sum_{k,l=1}^{N}\nu_{ik}\nu_{il}\frac{\partial^2}{\partial x_k\partial x_l}\left[\tilde{M}_i(s;\boldsymbol{x})\tilde{P}(\boldsymbol{x},s)\right]\end{aligned}. \quad [9]$$

Multiplying s on both sides of Eq. 9, letting $s\to 0$ and making use of the facts: $K_i(\boldsymbol{n})=\lim_{s\to 0}\tilde{M}_i(s;\boldsymbol{n})$ and $P(\boldsymbol{x})=\lim_{s\to 0}s\tilde{P}(\boldsymbol{x},s)$, we can arrive at the following sgFPE:

$$-\sum_{k=1}^{N}\frac{\partial}{\partial x_k}\left[\sum_{i=1}^{L}\nu_{ik}K_i(\boldsymbol{x})P(\boldsymbol{x})\right]+\frac{1}{2}\sum_{k,l=1}^{N}\frac{\partial^2}{\partial x_k\partial x_l}\left[\sum_{i=1}^{L}\nu_{ik}\nu_{il}K_i(\boldsymbol{x})P(\boldsymbol{x})\right]=0. \quad [10]$$

In the next section, we will use Eq. 10 to analyze generalized stochastic models of gene expression and obtain some interesting results on the effect of molecular memory.

In the following section, we will apply the above general theory to 4 gene models: a generalized model of constitutive gene expression, a generalized model of gene self-regulation, a generalized ON-OFF model, and a generalized model of genetic toggle switch, where by "generalized," we mean that each model considers molecular memory or nonmarkovianity. Then, we discover biological knowledge, e.g., molecular memory is in effect equivalent to a feedback and can induce bimodality, fine-tune the expression noise, and induce switch. (The related data will be available from the corresponding author upon request.)

Results

The Effect of Molecular Memory Is Equivalent to the Introduction of a Feedback. Understanding how a gene is turned on at a mechanistic level has been one of the big challenges in molecular biology and has received extensive attention over decades. Identifying the actual sequence of events during gene expression and establishing the method of recruitment have turned out to be a surprisingly difficult task (46). Here, we introduce a generalized model to mimic complex biochemical processes underlying gene expression, referring to Fig. 2*A*, where the proteins are assumed to be produced instantaneously after messenger RNAs (mRNAs) are produced. Fig. 2*B* is a schematic representation of transitions between protein states with time. Let $\psi_1(t;n)$ and $\psi_2(t;n)$ be waiting-time distributions for protein synthesis and degradation, respectively, where n represents the number of protein molecules.

First, consider the case without regulation but with molecular memory. Two waiting-time distributions are set as $\psi_1(t;n)=\lambda_1{}^{L_1}/\Gamma(L_1)t^{L_1-1}e^{-\lambda_1 t}$ and $\psi_2(t;n)=(n\lambda_2)^{L_2}t^{L_2-1}e^{-n\lambda_2 t}$, where λ_1 and λ_2 are positive constants (which may be understood as the mean synthesis and degradation rates, respectively). Before presenting analytical results, we perform numerical calculation with results shown in Fig. 2, where Fig. 2*C* demonstrates that the stationary protein distribution indeed exists even in the presence of molecular memory. Hereafter, we will vary memory index (L_1 or L_2) while keeping the constant average time between successive reactions by scaling parameter λ_1 (λ_2) appropriately with L_1 (L_2), i.e., keeping the ratio L_i/λ_i fixed.

As pointed out above, the stationary probabilistic behavior of the original nonmarkovian reaction system is exactly the same as that of the constructed markovian reaction network, but there would exist differences in dynamic probability behavior between the 2 networks. Fig. 2*C* shows that 2 dynamic distributions are different at the initial stage, but this difference gradually reduces and finally disappears with time. Fig. 2*E* demonstrates how molecular memory (i.e., $L_1>1$ or $L_2>1$) affects the stationary protein distributions, whereas Fig. 2*D* shows that L_1 always decreases the mean protein number but L_2 always increases this number. These numerical results imply that the effect of molecular memory is equivalent to the introduction of a feedback. For this, we give an intuitive interpretation. First, note that in simulation, we keep the average waiting times between successive reactions constant by scaling $\lambda_1(\lambda_2)$ appropriately with $L_1(L_2)$, i.e., we keep ratios $L_1/\lambda_1\equiv 1/\lambda_1^*$ and $L_2/\lambda_2=1/\lambda_1^*$ constant. This implies that the average of waiting times remains unchanged, but their variances decrease with increasing $L_1(L_2)$. Second, the waiting-time distribution will collapse onto a Dirac delta function due to zero variance if $L_1(L_2)$ tends to infinity. Third, note that the reaction event that actually occurs is the one whose waiting time is minimum. Therefore, for a fixed L_2, if L_1 increases, the variability in birth waiting times decreases. As a result, the probability of birth events decreases and hence the effective reaction rate decreases. Similarly, if the protein-decay rate decreases with L_2, the mean protein number will increase. Finally, we emphasize that such a memory-induced feedback stems from fluctuations in waiting times rather than changes in their means.

In order to obtain analytical results, we consider the case of $L=L_1>1$ and $L_2=1$. In this case, 2 ERTs are given by $K_1(n)=n\lambda_2\lambda_1{}^L/[(\lambda_1+\lambda_2 n)^L-\lambda_1{}^L]$ with $K_1(0)=\lambda_1/L$ and $K_2(n)=n\lambda_2$. Note that function $f(x)=x\lambda_2\tilde{\lambda}_1{}^L/[(\tilde{\lambda}_1+x)^L-\tilde{\lambda}_1{}^L]$ with $\tilde{\lambda}=\lambda_1/\lambda_2$ has the following properties: $f(0)=0$ and the derivative $f'(x)$ is less than zero, i.e., $f'(x)<0$ for all $x\geq 0$. Therefore, the effect of molecular memory is equivalent to the introduction of a negative feedback. In addition, we can show that the stationary protein distribution is given by $P(n)=(\tilde{\lambda}_1{}^L)^n/[n!(a_1)_n\cdots(a_{L-1})_n\rho]$ with $\rho=[{}_1F_L(1,1,a_1,\cdots,a_{L-1};\tilde{\lambda}_1{}^L)]^{-1}$, where symbol ${}_1F_L(b,c_1,\cdots,c_L;z)$ is a confluent hypergeometric function (47), and symbol $(c)_n$ is defined as $(c)_n=c(c+1)\cdots(c+n-1)$. Constants $a_1,\cdots,a_{L-1}$ are determined by comparing the coefficients for the same power of x in the equality of $\tilde{\lambda}^{L-1}+x\tilde{\lambda}^{L-1}+\cdots+x^{L-1}=(x+a_1)\cdots(x+a_{L-1})$, where $a_1,\cdots,a_{L_1-1}$ are assumed to be real (the case of complex roots can be similarly analyzed). This form of the distribution is similar to that of the stationary mRNA distribution in a stochastic gene model with a DNA loop (48).

Molecular Memory Can Induce Bimodality. In the markovian case, gene self-regulating systems have been extensively studied, and some analytical results have been obtained (49–52). However, gene self-regulating processes are in general nonmarkovian as

pointed out in the introduction, raising the question of how nonmarkovianity impacts gene-product distributions. To address this question, let us consider the following gene model: $\mathrm{ON}\xrightarrow{\psi_1(t;n)}\mathrm{ON}+\mathrm{protein}$, $\mathrm{protein}\xrightarrow{\psi_2(t;n)}\varnothing$, where n represents the number of protein molecules, and $\psi_1(t;n)$ and $\psi_2(t;n)$ are the intrinsic waiting-time distributions for protein synthesis and degradation, which take the forms $\psi_1(t;n)=(\lambda_1(n))^L/\Gamma(L)t^{L-1}e^{-\lambda_1(n)t}$ and $\psi_2(t;n)=n\lambda_2 e^{-n\lambda_2 t}$, where $\lambda_1(n)=[\mu_0+\mu_1(n/K)^H]\ /[1+(n/K)^H]$ is a regulation function of Hill type, H is the Hill coefficient, and μ_0, μ_1 (representing the feedback strength), K, and λ_2 are positive parameters. This setting enables us to have a description in terms of one promoter state rather than 2 switching states but also hides the fact that proteins produced by the gene in the ON state bind to the gene and take it back to the OFF state. If promoter switching between ON and OFF states is very quick and if the protein–DNA binding rate is not much larger than the unbinding rate, the use of an effective Hill-type function is reasonable (53).

Similar to the case of no regulation, 2 effective transition rates, $K_1(n)$ and $K_2(n)$, can also be given analytically. Although the exact stationary protein distribution cannot be analytically given, the solution to Eq. 10, i.e., the stationary distribution of continuous variables, can be approximately expressed as (see *SI Appendix* for details)

$$P(x)=\mathcal{N}\frac{[\lambda_1(x)+\lambda_2 x]^L-[\lambda_1(x)]^L}{[\lambda_1(x)+\lambda_2 x]^L\lambda_2 x}\exp\left(-2x+\int_0^x 4\left[\frac{\lambda_1(x')}{\lambda_1(x')+\lambda_2 x'}\right]^L dx'\right), \quad [11]$$

where $\mathcal{N}$ is a normalization factor.

Numerical results are demonstrated in Fig. 3, where Fig. 3 *A* and *C* shows how the number of the most probable protein molecules obtained by a statistical method depends on memory index L, whereas Fig. 3 *B* and *D* demonstrates that the stationary distribution predicted by Eq. 11 can well approximate that obtained by solving the sgCME. Note that $L=1$ corresponds to the markovian case (red dashed lines), whereas $L<1$ or $L>1$ to the nonmarkovian case. In Fig. 3 *A* and *C*, the shadowed areas represent that bimodality exists. Also note that the number of the most probable protein molecules shown in Fig. 3 *A* and *C* is in agreement with that predicted by a deterministic system (*SI Appendix*). From Fig. 3 *B* and *D*, we observe that bimodal protein distribution exists only for moderately large L with $L>1$ or only for moderately small L with $L<1$. Since memory index L determines the strength of molecular memory, Fig. 3 implies that the molecular memory only with moderate strengths can induce bimodality. Here, we also give an intuitive explanation for the demonstrated numerical results. First, the occurrence of bimodality needs appropriate nonlinearity. Second, if L is far away from 1, then the system's nonlinearity enhances. However, values of L must be appropriately chosen to generate bimodal protein distributions since only the appropriate nonlinearity can lead to bimodality.

Molecular Memory Can Fine-Tune the Gene Expression Noise. Transcription is a key step in gene expression. Biochemical processes associated with transcription often involve a variety of TFs, which regulate the promoter kinetics. For bacterial cells, promoters can exist in a surprisingly large number of regulatory states, e.g., the PRM promoter of phage lambda in *Escherichia coli* is regulated by 2 different TFs binding to 2 sets of 3 operators that can be brought together by looping out the intervening DNA, and, as a result, the number of regulatory states of the PRM promoter is up to 128 (54). In contrast, eukaryotic promoters are more complex, involving nucleosomes competing with or being removed by TFs (55). In addition to the conventional regulation by TFs, the eukaryotic promoters can also be epigenetically regulated via histone modifications (56–58), and such regulation may lead to very complex promoter structures (59). Given this complexity, we introduce intrinsic waiting-time distributions to model promoter kinetics. Specifically, assume that the switch times from OFF to ON and vice versa follow gamma distributions given, respectively, by $\psi_{\mathrm{on}}(t;m)=\frac{\lambda_{\mathrm{on}}{}^{L_{\mathrm{on}}}}{\Gamma(L_{\mathrm{on}})}t^{L_{\mathrm{on}}-1}e^{-\lambda_{\mathrm{on}}t}$ and

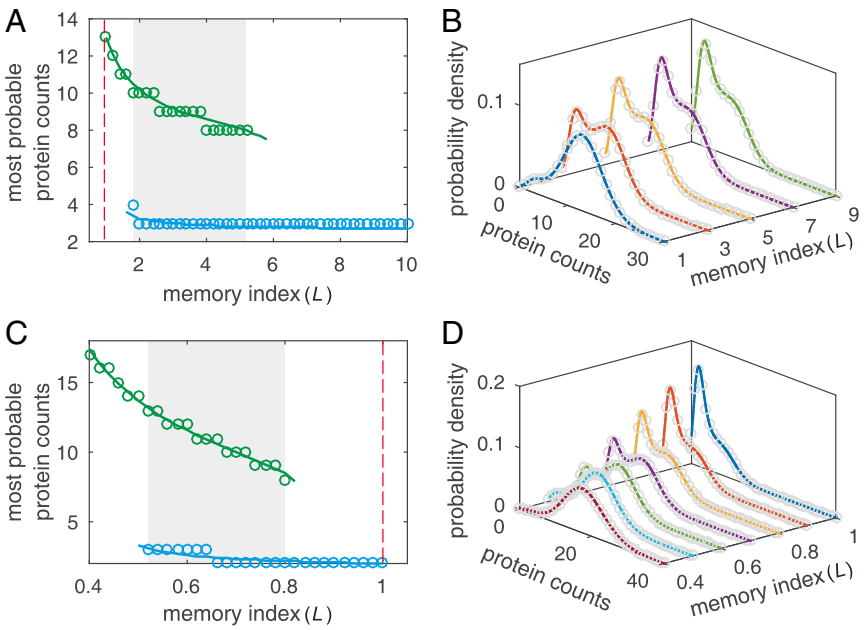

Fig. 3. Molecular memory can induce bimodal protein expression in the presence of feedback: the gene model depicted in Fig. 2*A*, where $\psi_1(t;n)$ depends on n. Empty circles represent the results obtained by the sgCME, whereas the solid lines represent the results predicted by Eq. 11. *A* and *B* correspond to the case of $L\geq 1$, where *A* demonstrates the dependence of most probable protein numbers on memory index L, whereas *B* shows stationary protein distributions. *C* and *D* correspond to the case of $L\leq 1$, where *C* demonstrates the dependence of most probable protein numbers on L, whereas *D* shows stationary protein distributions. Parameter values are set as: $\mu_0=4L$, $\mu_1=14L$, $K=6$, $n=4$, $\lambda_2=1$(*A* and *B*); $\mu_0=2.5L$, $\mu_1=10L$, $K=6$, $n=4$, $\lambda_2=1$ (*C* and *D*).

$\psi_{\text{off}}(t;m)=\frac{\lambda_{\text{off}}{}^{L_{\text{off}}}}{\Gamma(L_{\text{off}})}\,t^{L_{\text{off}}-1}e^{-\lambda_{\text{off}}t}$, where m represents the number of mRNA molecules. In addition, assume that the waiting-time distributions for transcription and degradation are given by $\psi_{\text{g}}(t;m)=\frac{\mu^{L_{\text{g}}}}{\Gamma(L_{\text{g}})}\,t^{L_{\text{g}}-1}e^{-\mu t}$ and $\psi_{\text{deg}}(t;m)=\frac{(m\lambda_{\text{deg}})^{L_{\text{deg}}}}{\Gamma(L_{\text{deg}})}\,t^{L_{\text{deg}}-1}e^{-m\lambda_{\text{deg}}t}$. The corresponding gene model is schematically depicted in Fig. 4*A*.

Let $K_i(m)$ $(i=1,2,3,4)$ be ETRs for transition from OFF to ON, for transition from ON to OFF, for transcription, and for degradation, respectively. According to Eq. 2, we can obtain the analytical expressions of the ETRs, which are given in *SI Appendix*. Moreover, the analytical expression of $K_1(m)$ with $L_{\text{on}}>1$ or $K_2(m)$ with $L_{\text{off}}>1$ can imply that the effect of molecular memory is equivalent to the introduction of a feedback, as interpreted above. Let $P_0(m)$ and $P_1(m)$ be the probabilities that mRNA has m molecules at states OFF and ON, respectively. Then, the corresponding sgCME takes the following form:

$$\begin{aligned}&-K_1(m)P_0(m)+K_2(m)P_1(m)+K_4(m+1)P_0(m+1)\\&\quad-K_4(m)P_0(m)=0\\&K_1(m)P_0(m)-K_2(m)P_1(m)+K_4(m+1)P_1(m+1)-K_4(m)P_1(m)\\&\quad+K_3(m-1)P_1(m-1)-K_3(m)P_1(m)=0.\end{aligned}\qquad[12]$$

In general, Eq. 12 has no analytical solution but can be solved numerically (see *SI Appendix* for details). Fig. 4 shows numerical results, where we vary memory index (L) but always keep the constant average time between successive reactions by scaling the characteristic parameter (e.g., α) appropriately with L.

From this figure, we observe that the mRNA mean is monotonically decreasing in memory index L_{on} (Fig. 4*B*) or in memory index L_{g} (Fig. 4*D*) if the other memory indices are set as 1, but monotonically increasing in memory index L_{off} (Fig. 4*C*) or in memory index L_{deg} (Fig. 4*E*) if the other memory indices are set as 1. Fig. 4*B* shows that memory index L_{on} amplifies the mRNA noise, whereas Fig. 4 *C*–*E* demonstrates that the other 3 memory indices can reduce the mRNA noise. In a word, Fig. 4 indicates that molecular memory plays an unneglectable role in gene expression.

Molecular Memory Can Induce Switch. Recall that a toggle-switch network (Fig. 5*A*) can model the cross-repression between the determinants of different cellular states, which can result in a definite choice between 2 outcomes (60–62). Conventional models of genetic toggle switch consider exponential waiting-time distributions. However, the expression of a gene in general involves a multistep process. Indeed, transcriptional repressor monomer (A or B) binds first to dimers and then to specific DNA sequences near the promoter, repressing the production of transcriptional repressor monomer (B or A). This multistep process can lead to nonexponential waiting times, creating a memory between individual reaction events. Here, we consider a generalized model of genetic toggle switch, which is schematically shown in Fig. 5*A* with 4 reactions listed in Fig. 5*B*,

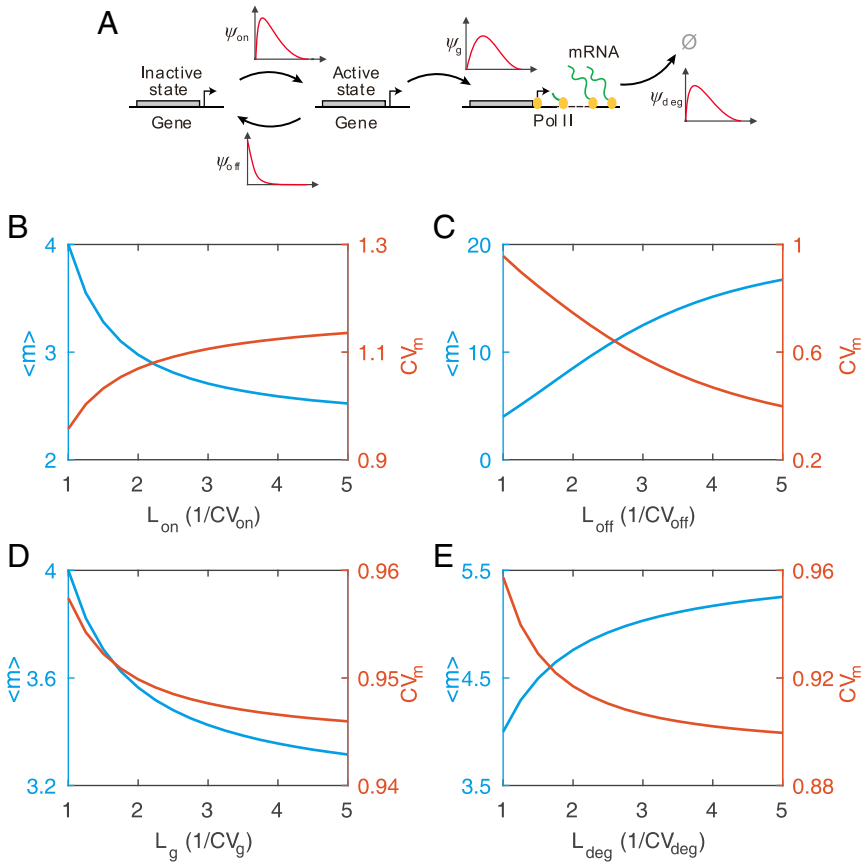

Fig. 4. Effect of molecular memory on gene expression. (*A*) Schematic representation of a model of stochastic transcription. (*B*) Dependence of the mean mRNA and the noise intensity on L_{on}, where $L_{\text{off}}=L_{\text{g}}=L_{\text{deg}}=1$. (*C*) Dependence of the mean mRNA and the noise intensity on L_{off}, where $L_{\text{on}}=L_{\text{g}}=L_{\text{deg}}=1$. (*D*) Dependence of the mean mRNA and the noise intensity on L_{g}, where $L_{\text{on}}=L_{\text{off}}=L_{\text{deg}}=1$. (*E*) Dependence of the mean mRNA and the noise intensity on L_{deg}, where $L_{\text{on}}=L_{\text{off}}=L_{\text{g}}=1$. The parameter values are set as $\lambda_{\text{on}}=L_{\text{on}}$, $\lambda_{\text{off}}=4L_{\text{off}}$, $\lambda_{\text{g}}=20L_{\text{g}}$ and $\lambda_{\text{deg}}=L_{\text{deg}}$. This setting implies the average time between successive reactions is kept fixed by scaling λ_i appropriately with L_i.

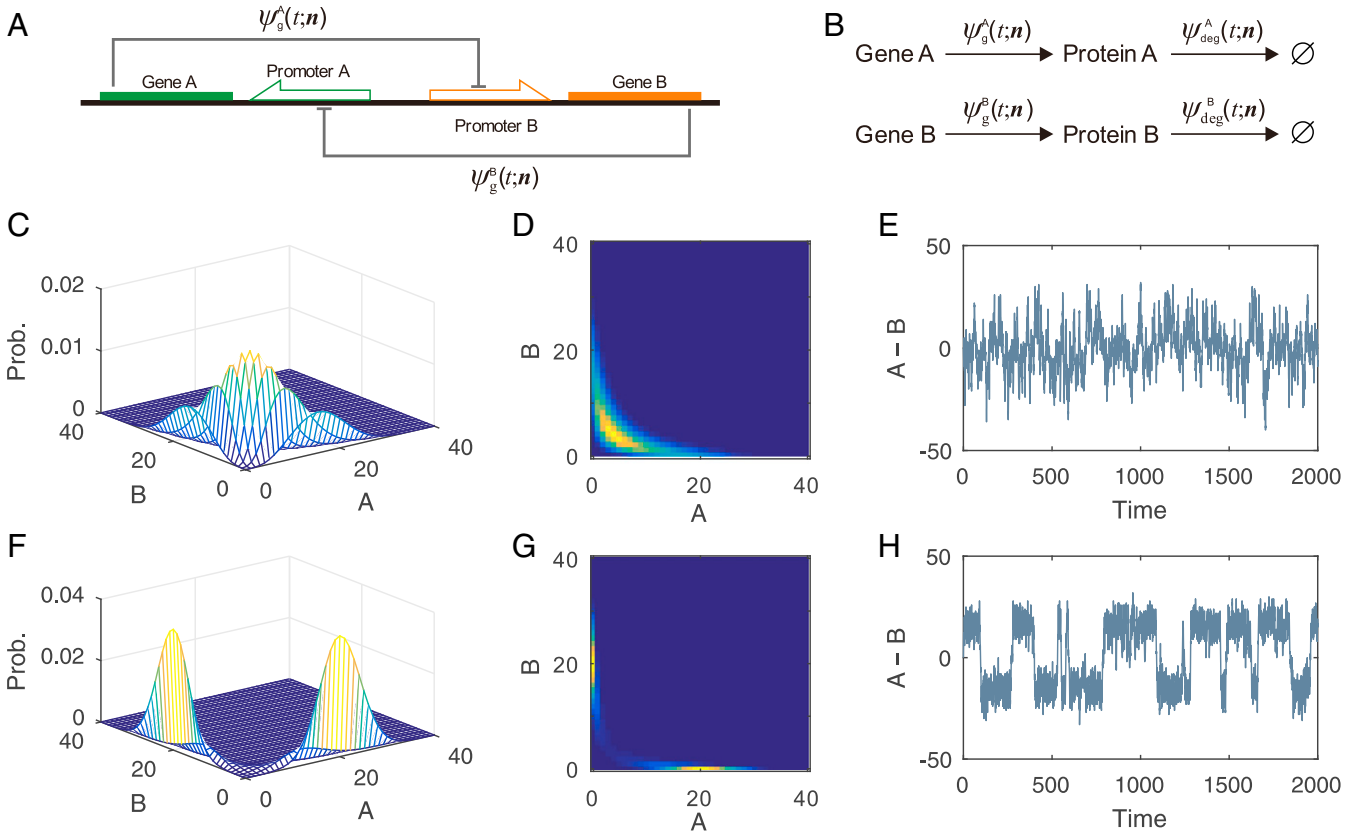

Fig. 5. (*A*) Schematic representation of a genetic toggle-switch model with molecular memory, where 2 genes are repressed by each other. (*B*) Four reactions corresponding to *A*, where waiting times for synthesis and degradation of each protein follow distributions. Default parameter values are taken as: $\beta_A=\beta_B=1$, $H_A=H_B=1$, $\lambda^A_{deg}=\lambda^B_{deg}=1$. (*C* and *F*) Joint distributions of proteins A and B, obtained by a numerical algorithm (*SI Appendix*). (*D* and *G*) Heat maps in the plane of protein A and B. (*E* and *H*) Time series of the difference between the levels of protein A and B, obtained by sgGA. *C–E* correspond to exponential waiting times, where parameter values are set as $L^A_g=L^B_g=1$, $\alpha_A=\alpha_B=20$. *F–H* correspond to nonexponential waiting times, where parameter values are set as $L^A_g=L^B_g=2$, $\alpha_A=\alpha_B=40$. Prob., probability.

where $\psi^A_g(t;\boldsymbol{n}),\psi^A_{deg}(t;\boldsymbol{n})$ and $\psi^B_g(t;\boldsymbol{n}),\psi^B_{deg}(t;\boldsymbol{n})$ are intrinsic waiting-time distributions for the synthesis and degradation of protein A and protein B, respectively, and $\boldsymbol{n}=(n_A,n_B)^T$ with n_A and n_B representing the numbers of protein A and protein B molecules, respectively. Assume that these waiting-time distributions are given by $\psi^A_g(t;\boldsymbol{n})=[\Gamma(L^A_g)]^{-1}[\lambda^A_g(\boldsymbol{n})]^{L^A_g}t^{L^A_g-1}e^{-\lambda^A_g(\boldsymbol{n})t}$ with $\lambda^A_g(\boldsymbol{n})=\alpha_A/(1+\beta_A n_B^{H_A})$, $\psi^A_{deg}(t;\boldsymbol{n})=\lambda^A_{deg}n_Ae^{-\lambda_2 n_A t}$; $\psi^B_g(t;\boldsymbol{n})=[\Gamma(L^B_g)]^{-1}[\lambda^B_g(\boldsymbol{n})]^{L^B_g}t^{L^B_g-1}e^{-\lambda^B_g(\boldsymbol{n})t}$ with $\lambda^B_g(\boldsymbol{n})=\alpha_B/(1+\beta_B n_A^{H_B})$, $\psi^B_{deg}(t;\boldsymbol{n})=\lambda^B_{deg}n_Be^{-\lambda_2 n_B t}$. Note that $L^A_g=1$ and $L^B_g=1$ correspond to the markovian case, whereas $L^A_g>1$ or $L^B_g>1$ corresponds to the nonmarkovian case.

Numerical results are demonstrated in Fig. 5 *C–H*, where Fig. 5 *C–E* corresponds to the case of exponential waiting times, whereas Fig. 5 *F–H* to the case of nonexponential waiting times. We observe that if the waiting times for synthesis of protein A and B follow exponential distributions (i.e., if we set $L^A_g=L^B_g=1$), the steady-state joint distribution of proteins A and B is unimodal, referring to Fig. 5 *C* and D. However, if the waiting times for synthesis of protein A and B follow nonexponential distributions (e.g., if we set $L^A_g=L^B_g=2$), the steady-state joint distribution of proteins A and B is bimodal, referring to Fig. 5 *F* and *G*. To examine the time dependence of the populations of 2 proteins in a single cell, we first perform stochastic simulations with a numerical algorithm (see *SI Appendix* for details) and then calculate the difference between the levels of proteins A and B. Numerical results are shown in Fig. 5 *E* and *H*. Comparing Fig. 5*E* with Fig. 5*H*, we find that 2 switching states occur only in the case of nonexponential waiting times or molecular memory. Thus, we conclude from Fig. 5 that molecular memory can induce bimodal distributions in the toggle-switch model depicted in Fig. 5*A* or Fig. 5*B*.

Discussion

Previous studies of biochemical-reaction processes on networks are mainly based on markovian (i.e., memoryless) hypothesis. However, as soon as a reactant interacts with its environment, the effect of molecular memory cannot be neglected. We have derived an exact sgCME, an sgLNA, and an sgFPE for a general biochemical-reaction network with molecular memory characterized by nonexponential waiting-time distributions. These derived equations allow one to retain analytical and/or numerical tractability, being general in scope, and thus are of a potential applicability in a wide variety of problems that transcend pure physics applications. The derived sgCME is particularly useful in finding stationary distributions in a number of nonmarkovian biochemical systems, as demonstrated in this article. Analysis of stochastic gene expression examples has indicated that the sgCME can help us find new biological knowledge, e.g., the effect of molecular memory is equivalent to the introduction of a feedback, and molecular memory can induce bimodality, although the distribution is not bimodal in the corresponding markovian case. The power of the sgCME can be enhanced by analyzing other examples, such as nonmarkovian random walks

and diffusion on networks (21, 63–65) and nonmarkovian open quantum systems (66).

Our general theory can reproduce some known results for queuing models of biological processes. First, recall that Pedraza and Paulsson (6) analyzed a $GI^X/M/\infty$ (including $G/M/\infty$) model of gene expression with a general queuing waiting-time distribution for the arrival of bursts and an exponential waiting-time distribution for the decay of mRNAs and derived an approximate formula for the mRNA noise. In *SI Appendix*, we have used the above theory to reproduce this formula. Second, environmental perturbations or external noise, which is often inevitable in cellular processes, can be modeled with time delay (20). We have derived the analytical expressions of effective transition rates and further established the corresponding gLNA (see *SI Appendix* for details). Functionally, this gLNA may be analogous to the chemical fluctuation theorem for $G_t/G/\infty$ models of gene expression (19), where subscript "t" represents that the corresponding waiting-time distributions are time-varying.

Our theoretical framework can also be used in the inference of the structure and parameters involved in system modes for a broad class of nonmarkovian biochemical-reaction processes on networks. For example, the structure of gene promoters and their kinetics, which would be complex due to, e.g., TF regulation, can be inferred based on experimental data. In fact, we can first infer the key parameter k in the Erlang waiting-time distribution from experimental data, since it can represent the number of small, difficultly specified reaction steps involved in transitions from ON to OFF states or vice versa, implying that the promoter structure can be determined. Then, we can use the standard method (e.g., the maximum likelihood estimation) to infer the values of other parameters from the experimental data, such as the mean switching rates between ON and OFF states, the mean transcription or translational rate. These inferred kinetic parameters in turn determine promoter kinetics and gene-expression dynamics. Furthermore, the sgCME can be used in the analysis of the corresponding stationary probabilistic behavior. In a word, we expect that our analytical framework will be of use for studying a variety of phenomena in biological and physical sciences and, indeed, in other areas where individual-based models with general waiting-time distributions and/or delayed interactions are relevant.

In the realistic world, "non-Markov is the rule, Markov is the exception," as remarked by N. G. van Kampen (67). A stochastic process (i.e., the biological phenomenon evolving in time) may be or may not be markovian, depending on the variables used to describe it. If all of the variables are observable or measurable, the process is markovian. In general, however, this is impractical and even impossible. Therefore, most of real stochastic processes we observe are nonmarkovian. To model real stochastic processes with some unobservable variables, many different methods of modeling have been proposed, e.g., queuing models (6, 7, 17–19), delay models (5, 28), Langevin equations with color noise (68), and CTRW models (20–22, 26). Correspondingly, some simulation algorithms have also been developed, e.g., those based on general renewal processes (30–33) and the one by introducing some exogenously reaction channels (38). These approaches, despite their own advantages, have finite applications, e.g., queuing models are inconvenient to treating bi- or multimolecular reaction networks.

Finally, we point out that CTRWs used in our theory incorporate the timing of move, where a random walker waits between 2 moves for a duration that independently follows an intrinsic waiting-time distribution. In other words, the move events are generated by a renewal process. On the other hand, CTRWs can be categorized into the 2 classes of active CTRWs and passive CTRWs, depending on whether a random walker actively initializes them as it travels or passively follows states when available (21). In active CTRWs, the interevent time of a state is reinitialized when a random walker lands on it. In passive CTRWs to which queuing models correspond, however, the interevent time of a state is not reset, and the waiting time depends on the last activation time. Usually, active CTRWs generates interevent times from a given PDF, more suitable to the analytical study of random processes, whereas passive CTRWs use interevent times observed in real data, less favorable to model and analyze. Based on the active CTRW framework, we have established a set of theories for a general biochemical network with arbitrary intrinsic waiting-time distributions. Our analysis of generalized birth and death processes based on the passive CTRW framework (*SI Appendix*) has provided a general thinking by establishing the relationship between the active and passive CTRWs.

ACKNOWLEDGMENTS. T.S. was supported by National Natural Science Foundation of China (Grants 11932019, 11775314, and 91530320). J.J. was supported by National Natural Science Foundation of China (Grants 11475273 and 11631005), Science and Technology Program of Guangzhou (Grant 201707010117), and Guangdong Key Research and Development Project 2019B0233002.

1. C. Gardiner, *Stochastic Methods-A Handbook for the Natural and Social Sciences* (Springer, New York, NY, 2009).
2. N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, The Netherlands, 2007).
3. E. Pardoux, *Markov Processes and Applications: Algorithms, Networks, Genome and Finance* (Wiley & Sons, New York, NY, 2008).
4. H. Andersson, T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis* (Springer, New York, 2000).
5. M. Barrio, A. Leier, T. T. Marquez-Lago, Reduction of chemical reaction networks through delay distributions. *J. Chem. Phys.* **138**, 104114 (2013).
6. J. M. Pedraza, J. Paulsson, Effects of molecular memory and bursting on fluctuations in gene expression. *Science* **319**, 339–343 (2008).
7. T. Jia, R. V. Kulkarni, Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys. Rev. Lett.* **106**, 058102 (2011).
8. K. Nishinari, Y. Okada, A. Schadschneider, D. Chowdhury, Intracellular transport of single-headed molecular motors KIF1A. *Phys. Rev. Lett.* **95**, 118101 (2005).
9. A. Basu, D. Chowdhury, Traffic of interacting ribosomes: Effects of single-machine mechanochemistry on protein synthesis. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **75**, 021902 (2007).
10. C. V. Harper *et al.*, Dynamic analysis of stochastic transcription cycles. *PLoS Biol.* **9**, e1000607 (2011).
11. M. Salathé *et al.*, A high-resolution human contact network for infectious disease transmission. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22020–22025 (2010).
12. A. Corral, Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys. Rev. Lett.* **92**, 108501 (2004).
13. P. S. Stumpf *et al.*, Stem cell differentiation as a non-Markov stochastic process. *Cell Syst.* **5**, 268–282.e7 (2017).
14. D. M. Suter *et al.*, Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
15. T. Guérin, O. Bénichou, R. Voituriez, Non-Markovian polymer reaction kinetics. *Nat. Chem.* **4**, 568–573 (2012).
16. A. L. Barabási, The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
17. A. Schwabe, K. N. Rybakova, F. J. Bruggeman, Transcription stochasticity of complex gene regulation models. *Biophys. J.* **103**, 1152–1161 (2012).
18. N. Kumar, A. Singh, R. V. Kulkarni, Transcriptional bursting in gene expression: Analytical results for general stochastic models. *PLoS Comput. Biol.* **11**, e1004292 (2015).
19. S. J. Park *et al.*, The Chemical Fluctuation Theorem governing gene expression. *Nat. Commun.* **9**, 297 (2018).
20. T. Aquino, M. Dentz, Chemical continuous time random walks. *Phys. Rev. Lett.* **119**, 230601 (2017).
21. N. Masuda, M. A. Porter, R. Lambiotte, Random walks and diffusion on networks. *Phys. Rep.* **716**, 1–58 (2017).
22. R. Kutner, J. Masoliver, The continuous time random walk, still trendy: Fifty-year history, state of art, and outlook. *Eur. Phys. J. B* **90**, 50 (2017). (A collection of papers in this special issue).
23. A. J. Black, A. J. McKane, A. Nunes, A. Parisi, Stochastic fluctuations in the susceptible-infective-recovered model with distributed infectious periods. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **80**, 021922 (2009).
24. P. Van Mieghem, R. van de Bovenkamp, Non-Markovian infection spread dramatically alters the susceptible-infected-susceptible epidemic threshold in networks. *Phys. Rev. Lett.* **110**, 108701 (2013).
25. M. Starnini, J. P. Gleeson, M. Boguñá, Equivalence between non-Markovian and Markovian dynamics in epidemic spreading processes. *Phys. Rev. Lett.* **118**, 128301 (2017).
26. H. H. Jo, J. I. Perotti, K. Kaski, J. Kertész, Analytically solvable model of spreading dynamics with non-Poissonian processes. *Phys. Rev. X* **4**, 011041 (2014).

27. I. Z. Kiss, G. Röst, Z. Vizi, Generalization of pairwise models to non-Markovian epidemics on networks. *Phys. Rev. Lett.* **115**, 078701 (2015).
28. A. Leier, T. T. Marquez-Lago, Delay chemical master equation: Direct and closed-form solutions. *Proc. Math. Phys. Eng. Sci.* **471**, 20150049 (2015).
29. T. Brett, T. Galla, Stochastic processes with distributed delays: Chemical Langevin equation and linear-noise approximation. *Phys. Rev. Lett.* **110**, 250601 (2013).
30. D. T. Gillespie, Monte Carlo simulation of random walks with residence time dependent transition probability rates. *J. Comput. Phys.* **28**, 395–407 (1978).
31. M. Boguñá, L. F. Lafuerza, R. Toral, M. Á. Serrano, Simulating non-Markovian stochastic processes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **90**, 042108 (2014).
32. C. L. Vestergaard, M. Génois, Temporal Gillespie algorithm: Fast simulation of contagion processes on time-varying networks. *PLoS Comput. Biol.* **11**, e1004579 (2015).
33. N. Masuda, L. E. C. Rocha, A Gillespie algorithm for non-Markovian stochastic processes. *SIAM Rev.* **60**, 95–115 (2018).
34. D. Bratsun, D. Volfson, L. S. Tsimring, J. Hasty, Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14593–14598 (2005).
35. M. Barrio, K. Burrage, A. Leier, T. Tian, Oscillatory regulation of Hes1: Discrete stochastic delay modelling and simulation. *PLoS Comput. Biol.* **2**, e117 (2006).
36. D. F. Anderson, A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *J. Chem. Phys.* **127**, 214107 (2007).
37. X. Cai, Exact stochastic simulation of coupled chemical reactions with delays. *J. Chem. Phys.* **126**, 124108 (2007).
38. M. Voliotis, P. Thomas, R. Grima, C. G. Bowsher, Stochastic simulation of biomolecular networks in dynamic environments. *PLoS Comput. Biol.* **12**, e1004923 (2016).
39. V. M. Kenkre, E. W. Montroll, M. F. Shlesin, Generalized master equations for continuous-time random walks. *J. Stat. Phys.* **9**, 45–50 (1973).
40. U. Landman, E. W. Montroll, M. F. Shlesinger, Random walks and generalized master equations with internal degrees of freedom. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 430–433 (1977).
41. D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
42. W. Feller, "Special densities. Randomization" in *An Introduction to Probability Theory and its Applications* (John Wiley & Sons, New York, NY, 2008), vol. 2, p. 47.
43. P. Thomas, A. V. Straube, R. Grima, The slow-scale linear noise approximation: An accurate, reduced stochastic description of biochemical networks under timescale separation conditions. *BMC Syst. Biol.* **6**, 39 (2012).
44. D. Schnoerr *et al*., Approximation and inference methods for stochastic biochemical kinetics-a tutorial review. *J. Phys. A Math. Theor.* **50**, 093001 (2017).
45. C. N. Angstmann *et al*., Generalized continuous time random walks, master equations, and fractional Fokker-Planck equations. *SIAM J. Appl. Math.* **75**, 1445–1468 (2015).
46. M. R. Green, Eukaryotic transcription activation: Right on target. *Mol. Cell* **18**, 399–402 (2005).
47. N. N. Lebedev, "Hypergfome functions" in *Special Functions and Their Applications* (Dover, New York, NY, 1972), pp. 238–280.
48. J. Zhang, T. Zhou, Promoter-mediated transcriptional dynamics. *Biophys. J.* **106**, 479–488 (2014).
49. H. Ge, H. Qian, X. S. Xie, Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. *Phys. Rev. Lett.* **114**, 078101 (2015).
50. R. Grima, D. R. Schmidt, T. J. Newman, Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).
51. N. Friedman, L. Cai, X. S. Xie, Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302 (2006).
52. Z. Cao, R. Grima, Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 3305 (2018).
53. J. Holehouse, R. Grima, Revisiting the reduction of stochastic models of genetic feedback loops with fast promoter switching. *Biophys. J.* **117**, 1311–1330 (2019).
54. J. M. G. Vilar, L. Saiz, CplexA: A mathematica package to study macromolecular-assembly control of gene expression. *Bioinformatics* **26**, 2060–2061 (2010).
55. G. Hornung *et al*., Noise-mean relationship in mutated promoters. *Genome Res.* **22**, 2409–2417 (2012).
56. A. Halme, S. Bumgarner, C. Styles, G. R. Fink, Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast. *Cell* **116**, 405–415 (2004).
57. L. M. Octavio, K. Gedeon, N. Maheshri, Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression. *PLoS Genet.* **5**, e1000673 (2009).
58. L. Weinberger *et al*., Expression noise and acetylation profiles distinguish HDAC functions. *Mol. Cell* **47**, 193–202 (2012).
59. D. A. Stavreva, L. Varticovski, G. L. Hager, Complex dynamics of transcription regulation. *Biochim. Biophys. Acta* **1819**, 657–666 (2012).
60. T. S. Gardner, C. R. Cantor, J. J. Collins, Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).
61. A. Lipshtat, A. Loinger, N. Q. Balaban, O. Biham, Genetic toggle switch without cooperative binding. *Phys. Rev. Lett.* **96**, 188101 (2006).
62. T. Biancalani, M. Assaf, Genetic toggle switch in the absence of cooperative binding: Exact results. *Phys. Rev. Lett.* **115**, 208101 (2015).
63. I. Scholtes *et al*., Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nat. Commun.* **5**, 5024 (2014).
64. J. P. Gleeson, K. P. O'Sullivan, R. A. Banos, Y. Moreno, Effects of network structure, competition and memory time on social spreading phenomena. *Phys. Rev. X* **6**, 021019 (2016).
65. J. C. Delvenne, R. Lambiotte, L. E. C. Rocha, Diffusion on networked systems is a question of time or structure. *Nat. Commun.* **6**, 7366 (2015).
66. I. D. Vega, D. Alonso, Dynamics of non-Markovian open quantum systems. *Rev. Mod. Phys.* **89**, 015001 (2017).
67. N. G. van Kampen, Remarks on non-Markov processes. *Braz. J. Phys.* **28**, 90 (1998).
68. J. Łuczka, Non-markovian stochastic processes: Colored noise. *Chaos* **15**, 26107 (2005).