Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Systems Biology

Inferring delays in partially observed gene regulatory networks

Hyukpyo Hong [1, 2], Mark Jayson Cortez [3], Yu-Yu Cheng [4], Hang Joon Kim [5], Boseung Choi [2,6,*], Krešimir Josić [7,8,*] and Jae Kyoung Kim [1,2,*]

[1] Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, 34141, Korea,
[2] Biomedical Mathematics Group, Institute for Basic Science, Daejeon, 34126, Korea,
[3] Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, College, Laguna, 4031, Philippines,
[4] Department of Biochemistry, University of Wisconsin - Madison, Madison, WI, 53706, USA,
[5] Division of Statistics and Data Science, University of Cincinnati, Cincinnati, OH, 45221, USA,
[6] Division of Big Data Science, Korea University Sejong Campus, Sejong, 30019, Korea,
[7] Department of Mathematics, University of Houston, Houston, TX, 77204, USA,
[8] Department of Biology and Biochemistry, University of Houston, Houston, TX, 77204, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Cell function is regulated by gene regulatory networks (GRNs) defined by protein-mediated interaction between constituent genes. Despite advances in experimental techniques, we can still measure only a fraction of the processes that govern GRN dynamics. To infer the properties of GRNs using partial observation, unobserved sequential processes can be replaced with distributed time delays, yielding non-Markovian models. Inference methods based on the resulting model suffer from the curse of dimensionality.
Results: We develop a simulation-based Bayesian MCMC method for the efficient and accurate inference of GRN parameters when only some of their products are observed. We illustrate our approach using a two-step activation model: An activation signal leads to the accumulation of an unobserved regulatory protein, which triggers the expression of observed fluorescent proteins. With prior information about observed fluorescent protein synthesis, our method successfully infers the dynamics of the unobserved regulatory protein. We can estimate the delay and kinetic parameters characterizing target regulation including transcription, translation, and target searching of an unobserved protein from experimental measurements of the products of its target gene. Our method is scalable and can be used to analyze non-Markovian models with hidden components.
Availability: Accompanying code in R is available at https://github.com/Mathbiomed/SimMCMC.
Contact: jaekkim@kaist.ac.kr or kresimir.josic@gmail.com or cbskust@korea.ac.kr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in microscopy allow us to observe the dynamics of gene regulatory networks (GRNs) in unprecedented detail. Novel statistical techniques have helped interpret the resulting wealth of data. However, even the best experimental methods can provide observations of only a fraction of the components constituting a GRN, and thus offer only partial information about the dynamics of gene circuits. Statistical methods thus need to take into account the effect of unobserved processes to correctly interpret the data, and accurately characterize genetic circuits and their dynamics.

Recently, inference methods have been proposed based on the assumption that the unobserved processes are sequential, and thus can be modeled by introducing a delay (Jiang *et al.*, 2021; Heron *et al.*, 2007; Calderazzo *et al.*, 2018; Choi *et al.*, 2020; Cortez *et al.*, 2022; Barrio *et al.*, 2013; Leier *et al.*, 2014; Gomez *et al.*, 2016; Kim *et al.*, 2022). The resulting models are non-Markovian, as system dynamics depends not only on the present, but also past states. This model drastically reduces the number of parameters and reactions that need to be inferred. However, this approach can only account for effects of sequential processes such as transcription and translation, which are modeled as delays in interactions between the genes in the network (Fig. 1). Thus, it is unclear how to analyze data and perform inference when the products of some genes in the circuit are unobserved. For example, in a network of two genes *x* and *y*, where protein X regulates the expression of gene *y*, present methods can be applied when counts of both proteins X and Y are known. Often it is impossible to observe the products of both genes concurrently. Is it possible to characterize the dynamics of X, if only Y is observed?

Currently available Bayesian Markov chain Monte Carlo (MCMC) methods for inference of non-Markovian systems are not well suited to address this question. While these methods are applicable even to systems with high intrinsic noise (i.e., low copy numbers of molecules) (Choi *et al.*, 2020) and cell-to-cell heterogeneity (Cortez *et al.*, 2022), they rely on the assumption that all proteins in a GRN are observed. When some protein counts are unobserved, extending such methods directly requires that we characterize reactions involving the unobserved proteins at each measurement time. The high dimensionality of the resulting system makes inference challenging, if not impossible.

Here, we present a simulation-based Bayesian method for the inference of kinetic and delay parameters of a GRN when only the products of some of the genes in the network are observed. The approach is applicable generally even if only the most downstream genes, i.e., the final outputs, of the network are observed. We illustrate the method using a two-step activation model, where an initial signal activates gene *x* whose product protein X is unobserved. Protein X triggers the expression of gene *y* and the production of an observed protein Y. By performing an identifiability analysis, we characterize what information about protein Y dynamics is needed to obtain accurate and precise information about the unobserved protein X. We find that the dilution rate and synthesis time delay of Y need to be known in order to infer the kinetic and time delay parameters characterizing the dynamics of X. We apply this approach to a plasmid-borne two-step activation circuit in *Escherichia coli* (*E. coli*) where unobserved AraC protein triggers the expression of yellow fluorescent protein (YFP). When the dilution rate and the time delay for the synthesis of the observed protein YFP are separately measured, we are able to infer the time delay for target regulation by the unobserved AraC protein (i.e., the delay due to transcription, translation, and target searching). This finding can play a critical role in synthetic circuit design because the AraC protein, whose kinetics has yet to be fully characterized, is a widely used transcriptional activator in synthetic biology. Our study also illustrates how information from unobserved proteins can be inferred from the dynamics of observed proteins in GRNs. Our approach is scalable and provides a tool for characterizing non-Markovian systems from partial observations.

2 Materials and methods

2.1 Derivation of a likelihood function of kinetic and delay parameters

We first derive a likelihood function to construct a Bayesian inference method for estimating kinetic and delay parameters (Choi *et al.*, 2017, 2020; Hong *et al.*, 2022; Cortez *et al.*, 2022). Consider a biochemical reaction network with u species $Z_1, \ldots, Z_u$ and v reactions $R_1, \ldots, R_v$. Reaction R_k can be represented as

$$R_k : p_{k1} Z_1 + \cdots + p_{ku} Z_u \rightarrow q_{k1} Z_1 + \cdots + q_{ku} Z_u \qquad (1)$$

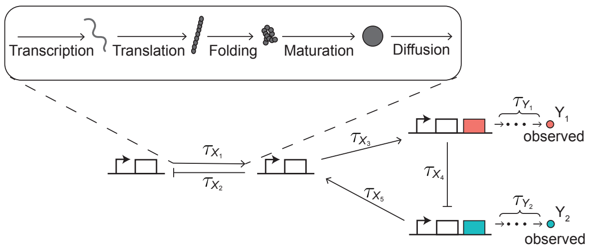

Fig. 1. Gene regulatory networks consist of genes whose interactions can be described with distributed time delays. In such networks, we can often observe the product (i.e., protein) of only some genes by using fluorescence microscopy (Y_1 and Y_2). Protein synthesis consists of multiple steps (transcription, translation, folding, maturation), and its duration can be described using a distributed time delays (τ_{Y_i}). Further delays, τ_{X_i}, can also be used to describe interactions between unobserved and observed genes. Such delays take into account protein synthesis, 3D diffusion inside a cell, and sliding along a strand of DNA to find a promoter region.

where p_{kj} and q_{kj} are the stoichiometric coefficients of species Z_j. For each reaction R_k, the reaction initiation rate, $h_k(z(t), \theta_k)$, is a function of the current state $z(t) = (z_1(t), \ldots, z_u(t))$ and the associated kinetic parameters θ_k, where $z_j(t)$ is the number of species Z_j at time t. We assume that each reaction takes a random time to complete. Therefore, after a reaction is initiated, the system state changes only after a random delay. This delay follows a distribution η_k fully determined by a vector of parameters, Δ_k. For example, in a GRN, when the synthesis of a transcriptional activator protein is initiated, a functional protein is produced after a sequence of steps including transcription, translation, and maturation (Golding *et al.*, 2005; Kærn *et al.*, 2005). Each of these steps takes time, and only after all steps are completed can the functional protein diffuse to its target binding site (Cheng *et al.*, 2017; Elf *et al.*, 2007; Hammar *et al.*, 2012). Thus, the number of functional activator proteins increases only when all intermediate steps are completed. If a reaction is not delayed, then the associated delay distribution is the Dirac delta measure at time 0.

Schlicht and Winkler, 2008 have proven that a reaction *completion* propensity, $f_k(t, \mathbf{z_c}, \theta_k, \Delta_k)$, describes the effective reaction rate of R_k at time t. This propensity depends on $\mathbf{z_c}$, the *complete* trajectory of all species from time 0 to the maximum measurement time T. The reaction completion propensity is a function of the reaction initiation propensity, h, and the delay distribution, η, and is given by:

$$f_k(t, \mathbf{z_c}, \theta_k, \Delta_k) = \int_0^t h_k(z(t-s), \theta_k)\, d\eta_k(s). \qquad (2)$$

Intuitively, the completion propensity is an average of past reaction initiation propensities weighted by the probability that they have occurred a given time in the past.

We can define the likelihood of the kinetic and delay parameters, θ and $\mathbf{\Delta}$, respectively, for the given trajectory $\mathbf{z_c}$ (Gupta and Rawlings, 2014):

$$L(\mathbf{z_c} | \theta, \mathbf{\Delta}) = \left[\prod_{j=1}^{M} f_{k_j}\left(t_j, \mathbf{z_c}, \theta_{k_j}, \Delta_{k_j}\right) \right] \times \exp\left[-\Lambda_0(T, \mathbf{z_c}, \theta, \mathbf{\Delta})\right]. \qquad (3)$$

Here the pair (t_j, k_j) for $j = 1, \ldots, M$ denotes the completion time and type of a reaction that completes within the time interval $(0, T]$. In addition,

$$\Lambda_0(t, \mathbf{z_c}, \theta, \mathbf{\Delta}) = \sum_{k=1}^{v} \Lambda_k(t, \mathbf{z_c}, \theta_k, \Delta_k),$$
$$\Lambda_k(t, \mathbf{z_c}, \theta_k, \Delta_k) = \int_0^t f_k(\hat{t}, \mathbf{z_c}, \theta_k, \Delta_k)\, d\hat{t}. \qquad (4)$$

This is analogous to the likelihood provided by Boys *et al.* (2008) for a system without delays.

The trajectories of biochemical species can be measured experimentally only at discrete time points. Complete reaction histories are thus unknown. If we measure the trajectories at discrete time points $t = 0, \ldots, T$, and denote these measurements by $\mathbf{z}$, then according to Choi *et al.* (2020); Cortez *et al.* (2022), an approximate likelihood function, $\hat{L}$, is given by

$$\hat{L}\left(\mathbf{z}|\mathbf{r},\theta,\mathbf{\Delta}\right) = \left[\prod_{i=1}^{T}\prod_{k=1}^{v}\frac{\hat{f}_k(i,\mathbf{z},\theta_k,\Delta_k)^{r_{ki}}}{r_{ki}!}\right] \times \exp\left(-\hat{\Lambda}_0\left(T,\mathbf{z},\theta,\mathbf{\Delta}\right)\right) \times \chi(\mathbf{z}|\mathbf{r}), \tag{5}$$

where $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_v)$ is a vector of reaction counts completing within each of the time intervals, i.e., $\mathbf{r}_k = (r_{k1}, \ldots, r_{kT})$ and r_{ki} is the count of reaction of type k that completes within the time interval $(i-1, i]$. Note that $\chi(\mathbf{z}|\mathbf{r})$ is the indicator function that is one if the trajectory of species matches the reaction counts and zero otherwise (see Supplementary Methods for details). In Eq. (5), $\hat{f}_k$ is an approximate reaction completion propensity computed by linearly interpolating the exact completion propensity (Eq. (2)) using:

$$\hat{f}_k\left(i,\mathbf{z},\theta_k,\Delta_k\right) = \sum_{m=0}^{i-1}\int_{m}^{m+1}\int_{t-1}^{t}\left[\left(s+1-t\right)h_k\left(z\left(i-m\right),\theta_k\right) + \left(t-s\right)h_k\left(z\left(i-m+1\right),\theta_k\right)\right]d\eta_k\left(s\right)dt, \tag{6}$$

and $\hat{\Lambda}_0$ is defined analogously to Λ_0 with $\hat{f}_k$ replacing f_k in Eq. (4).

Since the approximate likelihood given in Eq. (5) takes into account the number of reactions completed between discrete time points, it corresponds to the τ-leaping approach (Gillespie, 2001). The exact likelihood in Eq. (3) corresponds to the exact stochastic simulation algorithm (SSA) for a system with delays (Cai, 2007).

2.2 Derivation of a likelihood function given noisy observations of a subset of species

In GRNs, we cannot measure the activity of all components directly. However, we can measure the activity of fluorescent reporter proteins (Fig. 1). These measurements are often contaminated by observational noise, which we assume is characterized by the vector of noise parameters σ. We derive a likelihood function for these noisy observations, assuming that only some protein counts are observed.

Let $\mathbf{x}$ and $\mathbf{y}$ be the trajectories of the unobserved and observed species, respectively, in a biochemical reaction network with delays so that $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. We let $\mathbf{y}_{\text{obs}}$ be the vector of noisy observations of species $\mathbf{y}$ at discrete time points $t = 0, \ldots, T$. We assume that $\mathbf{y}_{\text{obs}}$ is obtained by adding i.i.d. noises from a Gaussian distribution to each observation in $\mathbf{y}$. The joint likelihood function of the unknown kinetic and delay parameters and reaction counts is then given by

$$\hat{L}(\mathbf{y}_{\text{obs}}|\mathbf{r},\theta,\mathbf{\Delta},\sigma) = \sum_{\mathbf{x},\mathbf{y}} L(\mathbf{y}_{\text{obs}}|\mathbf{x},\mathbf{y},\sigma) \times \hat{L}\left(\mathbf{x},\mathbf{y}|\mathbf{r},\theta,\mathbf{\Delta}\right) \tag{7}$$

$$= L(\mathbf{y}_{\text{obs}}|\bar{\mathbf{x}}(\mathbf{r}),\bar{\mathbf{y}}(\mathbf{r}),\sigma) \times \hat{L}\left(\bar{\mathbf{x}}(\mathbf{r}),\bar{\mathbf{y}}(\mathbf{r})|\mathbf{r},\theta,\mathbf{\Delta}\right) \tag{8}$$

$$= L(\mathbf{y}_{\text{obs}}|\bar{\mathbf{y}}(\mathbf{r}),\sigma) \times \hat{L}\left(\bar{\mathbf{x}}(\mathbf{r}),\bar{\mathbf{y}}(\mathbf{r})|\mathbf{r},\theta,\mathbf{\Delta}\right). \tag{9}$$

The sum over $\mathbf{y}$ in Eq. (7) can be reduced to a single term in Eq. (8) because the vector of reaction counts, $\mathbf{r}$, uniquely determines the trajectory in the indicator function in Eq. (5), $\chi(\mathbf{x},\mathbf{y}|\mathbf{r})$. Thus, the other terms in the sum vanish, and we denote the trajectories of unobserved and observed species without noise matching the vector of reaction counts by $\bar{\mathbf{x}}(\mathbf{r})$ and $\bar{\mathbf{y}}(\mathbf{r})$, respectively. Furthermore, Eq. (8) simplifies to Eq. (9) as the noisy observations depend only on the trajectory of the observed species, $\mathbf{y}(\mathbf{r})$, and are thus conditionally independent of $\mathbf{x}(\mathbf{r})$. On the right-hand side of Eq. (9), the first factor is the likelihood function of the noisy observation of the given trajectory at discrete time points, $\mathbf{y}_{\text{obs}}$, while the second factor is the approximate likelihood function given in Eq. (5). Based on this approximate joint likelihood, we developed a Bayesian MCMC algorithm for a delayed reaction system with noisy measurements of the observed components.

2.3 Simulation-based MCMC for discrete noisy observation from a gene regulatory network with time delays

Sampling from the conditional posterior distribution of the parameters characterizing the unobserved processes requires protein counts as input. However, because we cannot measure all protein counts directly, we have to generate samples of the unobserved protein counts as well. As we explain below, the random walk approach used previously for this purpose suffers from the curse of dimensionality (Boys *et al.*, 2008; Choi *et al.*, 2020; Cortez *et al.*, 2022). To circumvent this problem, we use stochastic simulations to generate samples of the unobserved protein counts. We describe the general idea behind our approach in this section. We provide details and the example of the two-activation model in the Supplementary Methods.

Using Bayes' theorem, we can obtain the joint posterior distribution of the model parameters, $(\theta, \mathbf{\Delta})$, and the number of reactions, $\mathbf{r}$, given the noisy measurements of the observed species, $\mathbf{y}_{\text{obs}}$, by multiplying the priors and the joint likelihood of the unknowns (Eq. (9)):

$$\pi(\mathbf{r},\theta,\mathbf{\Delta}|\mathbf{y}_{\text{obs}},\sigma) \propto \pi(\mathbf{r})\pi(\theta)\pi(\mathbf{\Delta}) \times \hat{L}(\mathbf{y}_{\text{obs}}|\mathbf{r},\theta,\mathbf{\Delta},\sigma).$$

When we generate samples from this joint posterior distribution, the dimension of the sampling distribution increases with the number of parameters and measurements. To generate samples in high dimensions, we exploit the Gibbs sampling approach and decompose each sampling step in the high dimensional space into separate low dimensional sampling steps. We use the Metropolis-Hastings (MH) -within-Gibbs sampler method to sample from each conditional posterior.

Although we divided each high dimensional sampling step into iterative low dimensional sampling steps, there are other practical challenges to implementing the block updating method we have used previously (Boys *et al.*, 2008; Choi *et al.*, 2020; Cortez *et al.*, 2022). First, because the block updating method is based on the MH algorithm, we need to tune too many hyperparameters (i.e., the variances of proposal distributions). Second, even if we tuned all proposal distributions, the MCMC algorithm converges slowly because all reaction counts are updated independently, while the reaction counts on subsequent time intervals are strongly correlated. Thus, using the random walk chain for each reaction count can significantly reduce the acceptance probability of proposed samples, leading to slow convergence of the proposed MCMC algorithm.

To address these problems, we utilize an algorithm that generates proposal reaction counts based on simulations of the biochemical reaction network in Eq. (1) (Wilkinson, 2018). Simulating the model directly obviates the need for parameter tuning and captures correlations between the reaction counts on subsequent time intervals. For simulations we chose τ-leaping (Gillespie, 2001) because it corresponds to the approximate likelihood in Eq. (9) and it is computationally more efficient than the exact delayed SSA (Cai, 2007).

The resulting MCMC procedure can be described as follows (Fig. 2), with the superscript (j) denoting samples at the j-th MCMC iteration step:

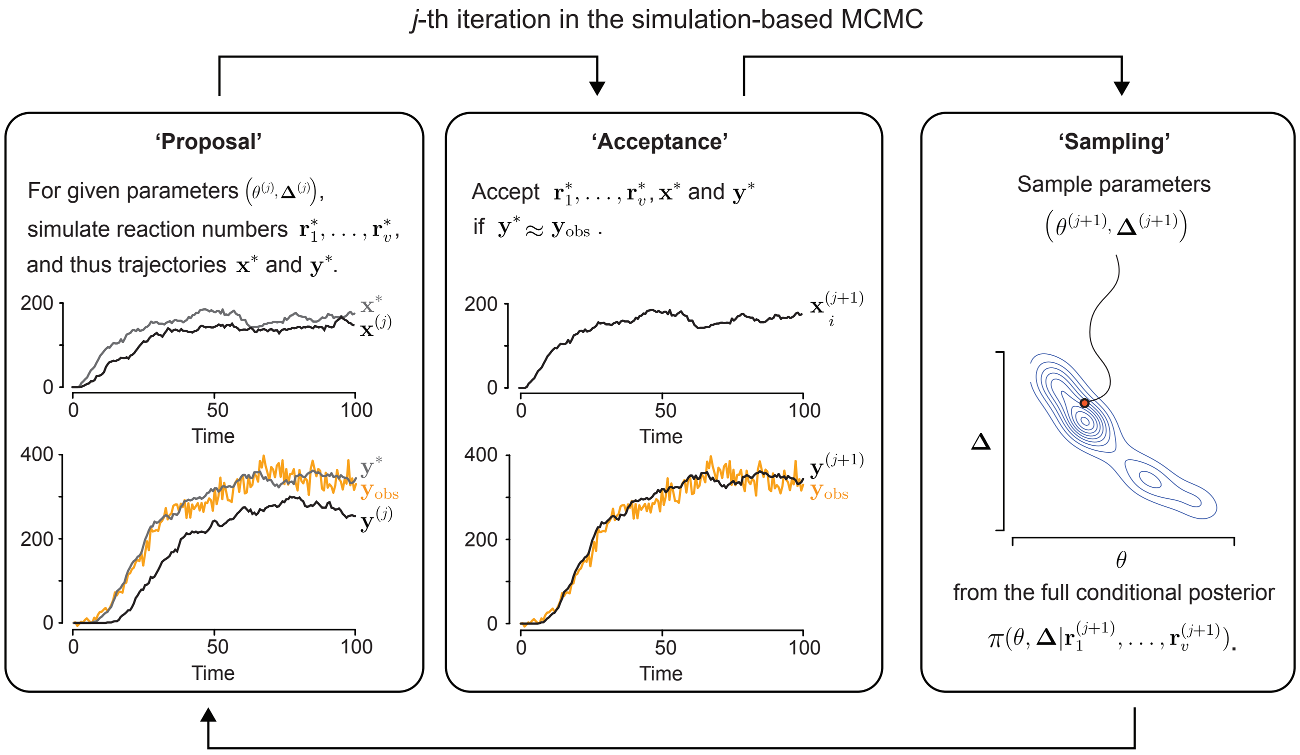
j-th iteration in the simulation-based MCMC
'Proposal'
For given parameters $\left(\theta^{(j)}, \mathbf{\Delta}^{(j)}\right)$,
simulate reaction numbers $\mathbf{r}_1^*, \ldots, \mathbf{r}_v^*$,
and thus trajectories $\mathbf{x}^*$ and $\mathbf{y}^*$.
200
0
$\mathbf{x}^*$
$\mathbf{x}^{(j)}$
0
50
100
Time
400
200
0
$\mathbf{y}^*$
$\mathbf{y}_{\text{obs}}$
$\mathbf{y}^{(j)}$
0
50
100
Time
'Acceptance'
Accept $\mathbf{r}_1^*, \ldots, \mathbf{r}_v^*, \mathbf{x}^*$ and $\mathbf{y}^*$
if $\mathbf{y}^* \approx \mathbf{y}_{\text{obs}}$.
200
0
$\mathbf{x}_i^{(j+1)}$
0
50
100
Time
400
200
0
$\mathbf{y}^{(j+1)}$
$\mathbf{y}_{\text{obs}}$
0
50
100
Time
'Sampling'
Sample parameters
$\left(\theta^{(j+1)}, \mathbf{\Delta}^{(j+1)}\right)$
$\mathbf{\Delta}$
θ
from the full conditional posterior
$\pi(\theta, \mathbf{\Delta}|\mathbf{r}_1^{(j+1)}, \ldots, \mathbf{r}_v^{(j+1)})$.

Fig. 2. An illustration of the simulation-based MCMC method to estimate the posterior distribution of the kinetic (θ) and delay parameters ($\mathbf{\Delta}$) of a GRN with unobserved, X, and observed, Y, components. (**Proposal**) At the j-th MCMC iteration, for given parameter samples, $\theta^{(j)}$ and $\mathbf{\Delta}^{(j)}$, we simulate a stochastic model with time delays and propose candidates for the reaction counts in the model (i.e., $\mathbf{r}_1^*, \ldots, \mathbf{r}_v^*$), which uniquely determine the trajectories for X and Y (i.e., $\mathbf{x}^*$ and $\mathbf{y}^*$). (**Acceptance**) The proposed reaction counts and trajectories are more likely to be accepted if $\mathbf{y}^*$ is closer to the observation $\mathbf{y}_{\text{obs}}$ than the previous sample $\mathbf{y}^{(j)}$ (see text and Materials and methods for details). If the proposed reaction counts and trajectories are not accepted, those from the previous iteration are kept as the current samples. The updated reaction counts and trajectories are referred to as $\mathbf{r}_1^{(j+1)}, \ldots, \mathbf{r}_v^{(j+1)}$, $\mathbf{x}^{(j+1)}$, and $\mathbf{y}^{(j+1)}$. (**Sampling**) For given updated reaction counts, we sample the kinetic and delay parameters from their full conditional posterior distribution $p(\theta, \mathbf{\Delta}|\mathbf{r}_1^{(j+1)}, \ldots, \mathbf{r}_v^{(j+1)})$ using MH-within-Gibbs sampling approach.. These steps are repeated until a convergence criterion is met.

1. Initialize the kinetic and delay parameters $(\theta^{(0)}, \mathbf{\Delta}^{(0)})$ and reaction counts $\mathbf{r}^{(0)}$.
2. By performing the stochastic simulation for given $(\theta^{(j)}, \mathbf{\Delta}^{(j)})$, propose new trajectories $\mathbf{x}^*$ and $\mathbf{y}^*$ and the underlying reaction counts $\mathbf{r}^*$.
3. Accept the proposed reaction counts and corresponding trajectories $(\mathbf{r}^*, \mathbf{x}^*, \mathbf{y}^*)$ based on the MH acceptance probability of $\rho(\mathbf{r}^*, \mathbf{r}^{(j)})$ where

$$\rho(\mathbf{r}^*, \mathbf{r}^{(j)}) = \min\left\{\frac{\pi(\mathbf{r}^*)}{\pi(\mathbf{r}^{(j)})} \times \frac{L(\mathbf{y}_{\text{obs}}|\mathbf{y}^*)}{L(\mathbf{y}_{\text{obs}}|\mathbf{y}^{(j)})}, 1\right\}. \qquad (10)$$

4. Sample the kinetic and delay parameters $(\theta^{(j+1)}, \mathbf{\Delta}^{(j+1)})$ from their full conditional posterior distribution for the given reaction counts $\mathbf{r}^{(j+1)}$ using an MH-within-Gibbs sampling approach.
5. Repeat Steps 2–4 until a convergence criterion is met.

The acceptance probability in Eq. (10) can be viewed as the conventional MH acceptance probability when a sample from a proposal distribution is replaced with a sample generated by a stochastic simulation. We provide the derivation of the acceptance probability, $\rho(\mathbf{r}^*, \mathbf{r}^{(j)})$, and the explicit form of conditional posterior distributions of the kinetic and delay parameters in Step 4, $\pi(\theta, \mathbf{\Delta}|\mathbf{r}^{(j+1)})$, in the Supplementary Methods.

3 Results

3.1 Kinetic and delay parameters can be accurately estimated as long as unidentifiable parameters are known.

We first applied our inference algorithm to synthetic data obtained from simulations of a two-step activation model (Fig. 3a) to test if our method can accurately estimate the kinetic and delay parameters of a GRN. In the generative model, the production of the unobserved protein X is initiated at rate of A_X and the protein activates the production of the observed protein Y. Transcription of Y is initiated at a rate that is modeled by a Michaelis-Menten function, after a regulation delay τ_X. Protein Y takes a random time to mature, and hence each molecule becomes observable after a time delay, τ_Y. Proteins X and Y are diluted at the same rate B due to cell growth. When trying to infer all parameters in this model, we encountered identifiability issues. However, we show that we can obtain accurate and precise parameters estimates when some of the unidentifiable parameters are known.

To test our inference algorithm, we generated 40 time series of the observed protein Y (i.e., $Y(t)$) using a delayed stochastic simulation algorithm (Cai, 2007), and added combined additive and multiplicative noise, sampled from $N(0, Y(t) + 10)$ at each time t, to obtain 40 synthetic measurement trajectories, $Y_{\text{obs}}(t)$, that we subsequently used for inference (Fig. 3b). Trajectories were indistinguishable when both parameters A_X and K_M were scaled by the same factor 2 or 0.5 (Fig. 3c). This degeneracy follows directly from the equation defining the production

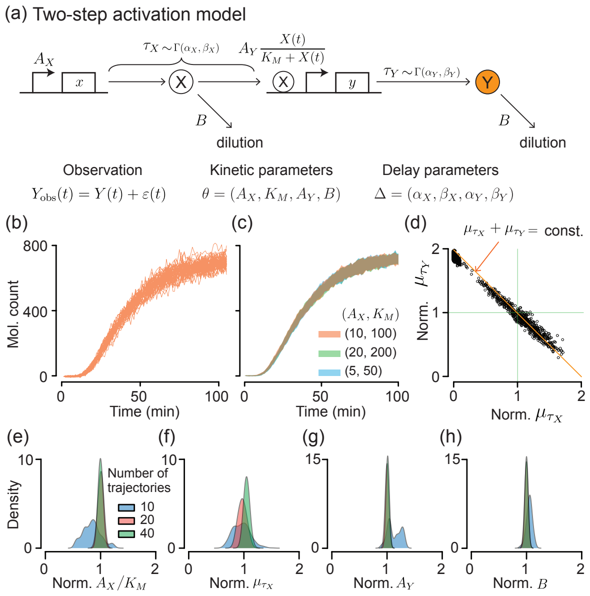

Fig. 3. Estimation of the kinetic and delay parameters using multiple trajectories that reach steady state is accurate and precise when unidentifiable parameters, K_M and τ_Y, are fixed. (**a**) Two-step activation model diagram. The unobserved protein, X, is produced at rate A_X, and activates the production of a downstream protein, Y, after a random delay τ_X. Activation is modeled by a Michaelis-Menten function. After transcription is initiated, protein Y takes a random time, τ_Y, to mature. Both delays, τ_X and τ_Y follow Gamma distributions with distinct parameters. Proteins X and Y are diluted at the same rate, B, due to cell growth. We assume that only protein Y is observed, and the protein count is recorded at discrete times. These measurements are corrupted by a combination of additive and multiplicative observational noise, $\varepsilon(t)$, which follows the normal distribution $N(0, Y(t) + \sigma_e)$. (**b**) 40 noisy trajectories of measurements, $Y_{\text{obs}}(t)$, at time $0, 1, \ldots, 100$ (min) of the two-step activation model in Fig. 2 with kinetic parameters $A_X = 10$ min^{-1}, $A_Y = 60$ min^{-1}, $K_M = 100$, $B = 0.05$ min^{-1}, delays $\tau_X \sim \Gamma(18/5, 3/5)$, $\tau_Y \sim \Gamma(18/5, 3/5)$, and observational noise $\sim N(0, Y(t) + 10)$. (**c**) The simulated measurements of $Y_{\text{obs}}(t)$ are indistinguishable when varying A_X and K_M while keeping their ratio, A_X/K_M, constant. Here $A_X/K_M = 0.1$. The upper and lower boundaries of the shaded region correspond to the mean±SD for the 40 trajectories obtained for each parameter set. (**d**) When estimating parameters using the trajectories in (b), we found that the means of the two delays, μ_{τ_X} and μ_{τ_Y}, were not individually identifiable, but their sum could be accurately estimated. (**e–h**) To resolve this identifiability issue, we assumed that the distribution of τ_Y could be estimated separately, and the distribution was thus fixed in the estimation process. We could then accurately estimate A_X/K_M, μ_{τ_X}, A_Y, and B. These estimates became more accurate and precise as we increased the number of trajectories used for inference. Here, sample values were divided by true parameter values to obtain posterior distributions of the normalized parameters.

of Y, $A_Y X(t)/(K_M + X(t))$: Because the level of $X(t)$ is proportional to A_X, the production term is proportional to $A_Y A_X/(K_M + A_X)$, which can be rewritten as $A_Y A_X/K_M/(1 + A_X/K_M)$. Thus, the production of Y is mainly governed by the ratio A_X/K_M rather than the individual parameters A_X and K_M. This degeneracy leads to the unidentifiability of A_X and K_M. We therefore fixed K_M to an arbitrary value, 100, and estimated the ratio A_X/K_M instead of A_X in the following.

When we estimated the remaining parameters, the posterior samples of the mean time delay needed for gene *x* to regulate gene *y*, τ_X, and the mean synthesis delay of protein Y, τ_Y, were strongly correlated, indicating that their sum could be accurately estimated, but each could not be estimated individually (Fig. 3d). Hence, measurements of Y contain information only about the sum of the two mean time delays. Indeed, simulated trajectories of the model with $\tau_X \sim \Gamma(3.6, 0.6)$ and $\tau_Y \sim \Gamma(3.6, 0.6)$, both with means equal to 6 mins, are indistinguishable from simulations with $\tau_X \sim \Gamma(1.2, 0.6)$ and $\tau_Y \sim \Gamma(6.0, 0.6)$, with means of 2 and 10 mins, respectively (Supplementary Fig. **S1**). This is consistent with the fact that the total delay in a cascade equals the sum of individual delays in the deterministic case (Glass *et al.*, 2021).

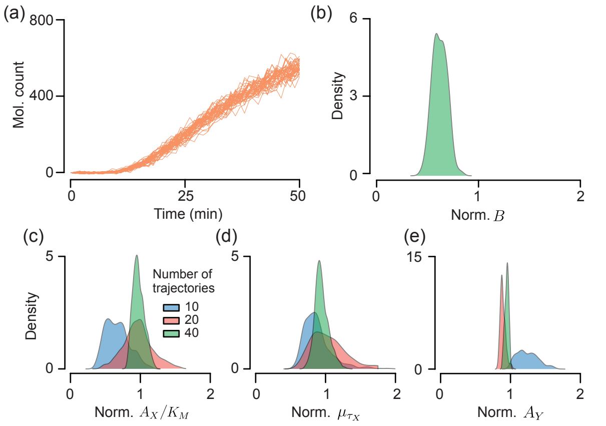

Fig. 4. Estimation of the kinetic and delay parameters using multiple short trajectories that did not reach steady state becomes accurate when the decay rate, B, is assumed known. (**a**) 40 trajectories of noisy measurements at times $0, 1, \ldots, 50$ (min) of the two-step activation model shown in Fig. 2 using the same parameter values as in Fig. 3a. Here, the observation window is too short for the simulated trajectories to reach steady state, unlike those in Fig. 3a. (**b**) We also assumed that K_M and the distribution of τ_Y are known, and found that the dilution rate, B, is underestimated using our algorithm. This indicates that the decay rate, B, can be accurately estimated only when measurements are taken until the system reaches steady state. (**c–e**) To account for this bias, we assume that the decay rate, B, is known, and fixed it at its true value in our inference algorithm. This allowed us to obtain the accurate and precise estimates of the ratio A_X/K_M, μ_{τ_X}, and A_Y, which improved with the number of measurement trajectories. Sample values were divided by true parameter values to obtain posterior distributions of the normalized parameters.

The expression delay of Y, τ_Y, can be directly estimated using an independent experiment with a genetic circuit containing only gene *Y*. We therefore assumed that the distribution of τ_Y is known to resolve the identifiability issue with τ_X and τ_Y. As a result, we obtained accurate and precise estimates of all remaining parameters: A_X/K_M, A_Y, τ_X, and B (Fig. 3e–h). These estimates became more precise when we increased the number of measured trajectories. Precision also increased with measurement frequency (Supplementary Fig. **S2**).

Thus tests with a simple circuit and synthetic data suggest that it is in general impossible to infer all parameters in biochemical reaction networks when some of the components are not observed. Some of the unidentifiable parameters need to be fixed to accurately estimate the remaining parameters in the two-step activation model. For synthetic gene circuits, some of these parameters could be estimated in separate experiments, while only combinations of other parameters may be inferable. This identifiability analysis was possible because we used a Bayesian approach and estimated the joint posterior of the parameter (Hines *et al.*, 2014).

3.2 With short time series, to estimate kinetic and delay parameters, the decay rate needs to be known

In time-lapse microscopy experiments cells can enter stationary phase (Kolter *et al.*, 1993), or the experiment may last too long before the population reaches an equilibrium distribution. To test whether the kinetic and delay parameters can be accurately estimated from data obtained over shorter time intervals we applied our algorithm to synthetic data corresponding to the first half of the experiment we described in the previous sections (Fig. 4a).

Under the conditions leading to accurate and precise estimation in the previous section (i.e., assuming that K_M and the distribution of τ_Y are known), we observed that the posterior mean of the decay rate, B, was considerably lower than the true parameter value (Fig. 4b). This underestimate occurred because the decay rate does not strongly affect the dynamics of the observed and unobserved proteins in the transient regime before their counts reach steady state values, since the terms $BX(t)$ and $BY(t)$ are small until protein counts increase. Thus, B needs to be measured separately to estimate the other parameters accurately. If the decay mainly occurs via growth-induced dilution, B can be estimated by measuring single-cell growth trajectories obtained with time-lapse microscope. By fixing the decay rate to its true value, we obtained accurate estimates for the other parameters $A_X/K_M, \mu_{\tau_X}$, and A_Y, and these became more precise as we increased the number of measurement trajectories (Fig. 4c–e).

We therefore found that the dilution rate needs to be estimated separately to obtain accurate and precise estimates of the remaining kinetic and delay parameters in the two-step activation model (Fig. 3a) when observed trajectories did not reach a steady state. We expect that similar identifiability issues will persist in more complex models.

3.3 Estimation of time delays in unobservable transcriptional regulation

We applied our method to data obtained from a two-step activation circuit in *E. coli* (Fig. 5a). The time-lapse fluorescence images of the cell populations were obtained previously (Cheng *et al.*, 2017). The two-step activation circuit consists of two genes; one encodes the unobserved transcriptional activator AraC and the second encodes the observed YFP, corresponding to X and Y in the two-step activation model, respectively (Fig. 3a). Once IPTG and arabinose are added at time $t = 0$, AraC is activated. After maturation of the expressed AraC, the mature protein searches a downstream target binding site and initiates the synthesis of YFP. The single-cell fluorescence signal from matured YFP was measured over 50 min (Fig. 5b). From the measured fluorescence intensity for two independent experiments with 23 and 25 cells, we obtained molecular counts by multiplying a previously calculated conversion rate (Fig. 5c) (Choi *et al.*, 2020). Because these trajectories did not reach a steady state, we first estimated the decay rate, B, and the time delay for the synthesis of YFP, τ_Y, to be able to accurately estimate the remaining kinetic and delay parameters from data (Fig. 3d and 4b).

The time delay for the synthesis of YFP can be separately estimated using a 'reporter-only' circuit which consists of a P_{BAD} promoter that drives the expression of YFP without the need for accumulation of AraC. Previously, we used data from experiments with such a circuit to obtain the estimate $\tau_Y \sim \Gamma(6.89, 0.89)$ (Choi *et al.*, 2020). We used this estimate in the following. For the decay rate, we can use the dilution rate because dilution is the main driver of decay since YFP is stable and is not enzymatically degraded (Andersen *et al.*, 1998). The dilution rate can be directly estimated from cell area tracked with a time-lapse microscope (Fig. 5d left) (Megerle *et al.*, 2008; Taheri-Araghi *et al.*, 2015). We fit an exponential function to each single-cell growth trajectory to estimate individual dilution rates (Fig. 5d right). We used the average of these estimates, $\hat{B} = 0.022s^{-1}$, as the dilution rate for the follwing estimates.

To estimate the remaining parameters we applied our MCMC algorithm to the measurements of the observed fluorescent protein and obtained estimates of the posterior distributions for the parameters $A_X/K_M, \mu_{\tau_X}$, and A_Y. The posterior means of the kinetic parameter A_X/K_M and A_Y in Experiment 1 were higher than those in Experiment 2 (Fig. 5e). This difference was due to a ∼20% higher intensity in Experiment 1 compared to Experiment 2 (Fig. 5c). On the other hand, the posterior means of the expected time delay for the transcriptional regulation of AraC, μ_{τ_X}, are similar between the two experiments: 7.50±0.21 min and 7.87±0.34 min (Fig. 5e).

The estimated time delay for the transcriptional regulation is the sum of delays corresponding to AraC synthesis, diffusion, and binding-site search. Interestingly, the binding-site search time of transcriptional factors in prokaryotic cells is usually less than a few minutes even for chromosomal genes whose copy number is just one or two (Hammar *et al.*, 2012; Elf *et al.*, 2007). The two-step activation circuit is plasmid-borne, and thus has copy number in the dozens. This indicates that the binding-site search time is much shorter than those for chromosomal genes (i.e., less than a minute), so the estimated time delay (∼7.5 min.) mostly comes from the synthesis, including transcription, translation, folding, and dimerization. This time delay has not been characterized previously. This finding thus provides a better understanding of the kinetics of AraC protein, a widely used transcriptional activator in synthetic biology.

4 Conclusion

We have developed a simulation-based Bayesian MCMC method for the inference of kinetic and delay parameters from noisy measurement of a GRN with unobserved components. We applied the method to a two-step activation model where an unobserved species X regulates the synthesis of an observed species Y.

Using synthetic data, we have shown that certain parameters are not identifiable. However, accurate estimates of some of these parameters can be obtained if the *observable* system components can be characterized in separate experiments. Specifically, we found that the production rate of X, A_X, and the Michaelis-Menten constant for regulating gene Y, K_M, cannot be estimated independently. However, the ratio between these two parameters (i.e., A_X/K_M) is identifiable from the observed trajectories (Fig. 3c). We also found that the sum of two delays, the regulatory delay, τ_X, and synthesis delay, τ_Y, was identifiable, but the delays were not identifiable individually (Fig. 3d). Often τ_Y or K_M can be estimated from separate experiments. In that case the regulatory delay, τ_X, and production rate, A_X, are identifiable with data from the full circuit. Furthermore, if the measured trajectories do not reach steady state, the decay rate, B, is also not identifiable (Fig. 4b). For proteins that are not actively degraded this last identifiability issue can be resolved by estimating the decay rate directly from cell size measurements.

We applied our method to two independent sets of experimental time-lapse fluorescence data from a two-step activation circuit in which AraC protein activates the synthesis of YFP. We fixed the decay rate to the dilution rate estimated from the observed cell area time series (Fig. 5d). Furthermore, we also fixed the delay for the synthesis of YFP to the value estimated with YFP-reporter only circuit in our previous work (Choi *et al.*, 2020). This allowed us to estimate the production rates and regulation delay parameters of unobserved AraC protein using only observations of YFP. The estimated production rates A_X/K_M and A_Y were higher in the second experiment (Fig. 5e). This might be due to the mean trajectory being higher in the second group (Fig. 5c). On the other hand, the estimated regulation delays, τ_X, are similar in the two experiments. This might be because regulation delay is an inherent attribute of AraC and the downstream gene *sfyfp* unlike the fluorescence level, which can be sensitive to camera settings. We hypothesize that the estimated time delay of transcriptional regulation of AraC (∼ 7.5 min.) mostly comes from the synthesis, not binding-site search. This finding can play an important role in building synthetic circuits as the AraC protein, whose kinetics are yet to be fully characterized, is a widely used transcriptional activator in synthetic biology (Romano *et al.*, 2021; Moon *et al.*, 2012).

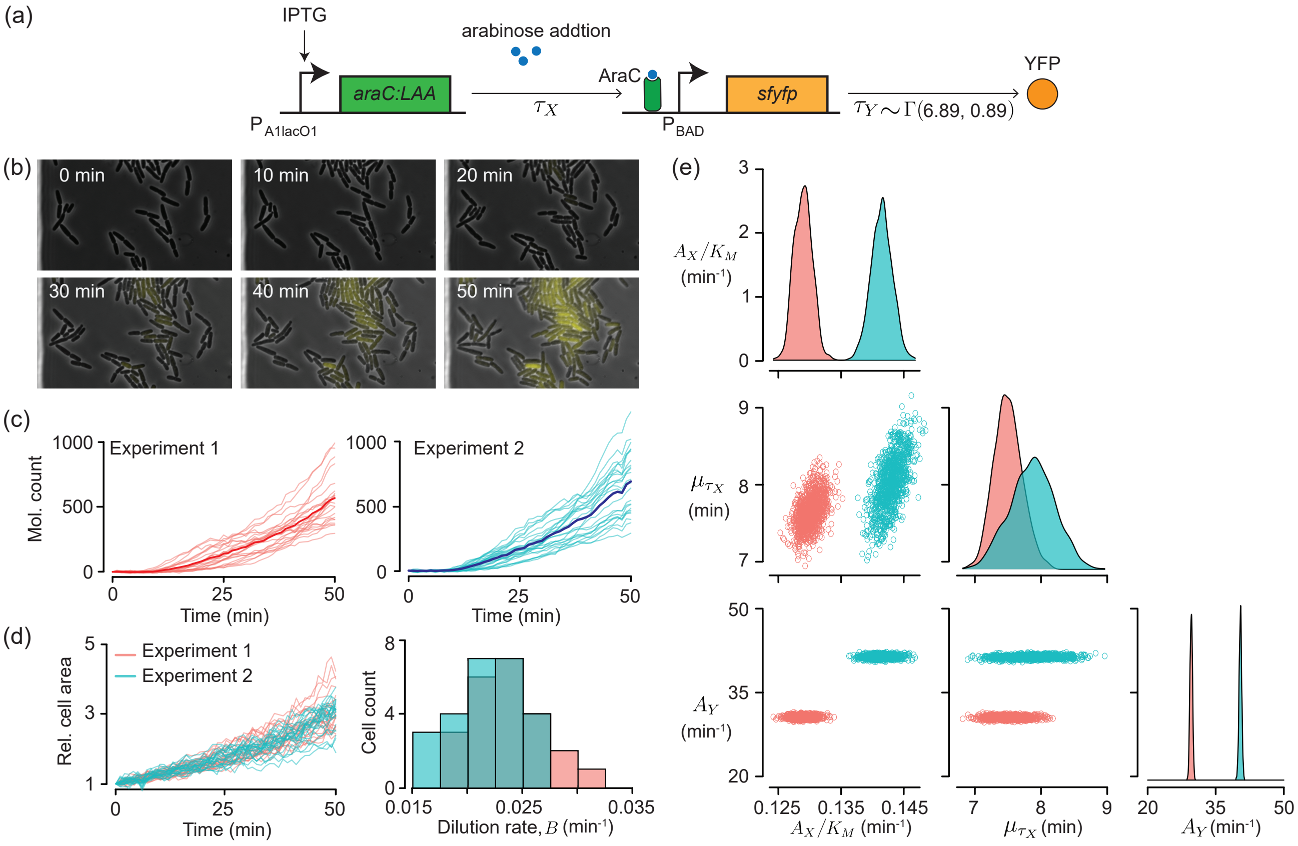

Fig. 5. Our inference method provided similar estimates of the mean regulation delay from two independent experiments with a two-step activation circuit in *E. coli*. (**a**) The two-step activation circuit in *E. coli*. Once IPTG and arabinose are added to the growth media, the expression of *araC* is induced and it activates the synthesis of YFP after the time delay of the regulation, τ_X. The synthesis of YPF also involves a time delay, $\tau_Y \sim \Gamma(6.89, 0.89)$, which we estimated previously using a reporter-only circuit. In this circuit, only the fluorescence level of YFP is measured while the level of AraC is not measurable. (**b**) Time-lapse images of YFP expression from the two-step activation circuit monitored using fluorescence microscopy (Cheng *et al.*, 2017). The fluorescent cells were observed after induction with 0.2mM IPTG and 2% (w/v) arabinose at time 0. (**c**) Molecular counts were obtained by dividing the fluorescence level of each cell by a conversion constant, calculated in our previous paper (Choi *et al.*, 2020). The numbers of cell trajectories are 23 and 25, respectively. The thick lines represent the mean trajectories. (**d**) The area of each cell from two independent experiments was measured by tracking the lineage of each cell. When a mother cell divided into two daughter cells, the area of the mother cell was added to the areas of its daughter cells. Additionally, the area was normalized by its initial value, obtaining the relative cell area of each cell (left). The relative areas were used to estimate the dilution rate, B, by fitting an exponential function to the relative area trajectories (right). (**e**) Applying our inference method to the cell trajectories, we obtained the posterior samples of the parameters A_X/K_M, μ_{τ_X}, and A_Y. To avoid bias and identifiability issues in estimation, we fixed the dilution rate B to its average value, 0.022, as it was directly estimated from the observed cell areas (d) and the time delay of the synthesis of YFP τ_Y to $\Gamma(6.89, 0.89)$ as it was previously estimated (Choi *et al.*, 2020). We obtained the estimates (mean±SD) of A_X/K_M, 12.90±0.14 min^{-1} and 14.15±0.17 min^{-1}, and the estimates of A_Y, 30.64±0.27 min^{-1} and 41.45±0.28 min^{-1}. These estimates are higher in the second experiment because the mean trajectory is higher in the second experiment (c). On the other hand, the estimates of the mean delay, 7.50±0.21 min and 7.87±0.34 min, were similar in both experiments.

For example, transcriptional regulators are being used in constructing cascaded genetic logic gates or oscillators where delay can impact output and dynamics (Moon *et al.*, 2012; Mather *et al.*, 2014).

As the number of parameters and unobserved species in a system increases, identifiability issues often worsen (Raue *et al.*, 2009; Hines *et al.*, 2014; Browning *et al.*, 2020). In addition, more complex models typically result in more local maxima in the corresponding posterior distributions. Markov chains can get trapped at a local maximum, leading to inaccurate estimates. This problem could be resolved by adopting advanced sampling methods, such as the multiset samplers (Leman *et al.*, 2009; Kim and MacEachern, 2015) which relies on many agents in one Markov chain and can more easily escape a local maximum.

While our method may appear similar to approximate-Bayesian computation (ABC) approach (Beaumont *et al.*, 2002), the two are qualitatively different. While in our method the proposed reaction counts are accepted based on the MH acceptance probability, in the ABC approach a proposed sample is accepted based on a metric and a threshold, both of which need to be defined appropriately. Our method is related to the ABC-MCMC approach (Marjoram *et al.*, 2003), in which parameters are accepted stochastically. However, ABC-MCMC uses stochastic simulations only for computing the acceptance probability to update a proposed parameter while in our approach a stochastic simulation of the model system is used to obtain proposed reaction counts.

Thus, the key step of the present method is the use of a stochastic simulation algorithm. The acceptance probability of proposed reaction counts based on the MH algorithm can be computed using the likelihood ratio of only observed processes (Eq. 10), independently of the complex dynamical model of the system. This allows for our method to be easily implemented. Furthermore, this way to obtain proposal reaction counts captures strong correlations between individual counts without the need to tune hyperparameters. In contrast, a direct application of the block-update method (Boys *et al.*, 2008; Choi *et al.*, 2020; Cortez *et al.*, 2022) requires the tuning of numerous proposal distributions, leading to small acceptance rates and slow convergence of the MCMC.

We assumed that all cells in a population are identical, and thus share the same parameters, i.e., we ignored the cell-to-cell variability. Even isogenic cell populations can show significant cell-to-cell variability (Kepler and Elston, 2001; Kærn *et al.*, 2005; Raj and van Oudenaarden, 2008; Smith and Grima, 2018), and heterogeneity in a population plays a crucial role in biological processes such as development. Our method can be extended to account for such variability using hierarchical model, a potential avenue for future work.

Our approach is scalable. Because the likelihood function has been derived for a general biochemical reaction network, the simulation-based MCMC can be tailored to a more complex model. The method does require performing stochastic simulations of the biochemical system for each MCMC iteration, resulting in a high computational cost. This cost could be reduced by developing an emulator, which is a fast data generator replacing a slower computational model to avoid the sampling process (Kennedy and O'Hagan, 2001).

A simulation-based MCMC method can be developed for other dynamical models as long as a likelihood function is available. While we used a continuous-time Markov Chain, which efficiently explains a biochemical reaction network with a low copy number of molecules, one can also use a stochastic differential equation (Calderazzo *et al.*, 2018; Ruttor and Opper, 2009) which is accurate when the copy numbers are higher, an agent-based model (Grazzini *et al.*, 2017), or a delay differential equation (Kim *et al.*, 2022). Thus, we expect that our framework can be extended to various stochastic models of non-Markovian GRNs, and thus characterize the dynamics of a variety of systems from partial observations.

Acknowledgements

Funding

This work was supported by Samsung Science and Technology Foundation [SSTF-BA1902-01 to J.K.K.]; Institute for Basic Science [IBS-R029-C3 to J.K.K.]; National Research Foundation of Korea [NRF-2019-Fostering Core Leaders of the Future Basic Science Program/Global Ph.D. Fellowship Program 2019H1A2A1075303 to H.H., NRF-2020R1F1A1A01066082 to B.C.]; National Science Foundation [MCB-1936770 to K.J.]; NIH grant [R01GM144959 to K.J.]; and Taiwan Studying Abroad Scholarship [Y.Y.C.].

Data Availability

All experimental data has been published previously, and is available upon request.

Conflict of Interest: none declared.

References

Andersen, J. B. *et al.* (1998). New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Applied and Environmental Microbiology*, **64**(6), 2240–2246.

Barrio, M. *et al.* (2013). Reduction of chemical reaction networks through delay distributions. *The Journal of Chemical Physics*, **138**(10), 104114.

Beaumont, M. A. *et al.* (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, **162**(4), 2025–2035.

Boys, R. J. *et al.* (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, **18**(2), 125–135.

Browning, A. P. *et al.* (2020). Identifiability analysis for stochastic differential equation models in systems biology. *Journal of The Royal Society Interface*, **17**(173), 20200652.

Cai, X. (2007). Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of Chemical Physics*, **126**(12), 124108.

Calderazzo, S. *et al.* (2018). Filtering and inference for stochastic oscillators with distributed delays. *Bioinformatics*, **35**(8), 1380–1387.

Cheng, Y.-Y. *et al.* (2017). The timing of transcriptional regulation in synthetic gene circuits. *ACS Synthetic Biology*, **6**(11), 1996–2002. PMID: 28841307.

Choi, B. *et al.* (2017). Beyond the michaelis-menten equation: Accurate and efficient estimation of enzyme kinetic parameters. *Scientific Reports*, **7**(1), 17018.

Choi, B. *et al.* (2020). Bayesian inference of distributed time delay in transcriptional and translational regulation. *Bioinformatics*, **36**(2), 586–593.

Cortez, M. J. *et al.* (2022). Hierarchical Bayesian models of transcriptional and translational regulation processes with delays. *Bioinformatics*, **38**(1), 187–195.

Elf, J. *et al.* (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, **316**(5828), 1191–1194.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, **115**(4), 1716–1733.

Glass, D. S. *et al.* (2021). Nonlinear delay differential equations and their application to modeling biological network motifs. *Nature Communications*, **12**(1), 1788.

Golding, I. *et al.* (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**(6), 1025–1036.

Gomez, M. M. *et al.* (2016). The effects of time-varying temperature on delays in genetic networks. *SIAM Journal on Applied Dynamical Systems*, **15**(3), 1734–1752.

Grazzini, J. *et al.* (2017). Bayesian estimation of agent-based models. *Journal of Economic Dynamics and Control*, **77**(C), 26–47.

Gupta, A. and Rawlings, J. B. (2014). Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology. *AIChE Journal*, **60**(4), 1253–1268.

Hammar, P. *et al.* (2012). The <i>lac</i> repressor displays facilitated diffusion in living cells. *Science*, **336**(6088), 1595–1598.

Heron, E. A. *et al.* (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, **23**(19), 2596–2603.

Hines, K. E. *et al.* (2014). Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *Journal of General Physiology*, **143**(3), 401–416.

Hong, H. *et al.* (2022). *Beyond the Michaelis–Menten: Bayesian Inference for Enzyme Kinetic Analysis*, pages 47–64. Springer US, New York, NY.

Jiang, Q. *et al.* (2021). Neural network aided approximation and parameter inference of non-markovian models of gene expression. *Nature Communications*, **12**(1), 2618.

Kærn, M. *et al.* (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, **6**(6), 451–464.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(3), 425–464.

Kepler, T. B. and Elston, T. C. (2001). Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophysical Journal*, **81**(6), 3116–3136.

Kim, D. W. *et al.* (2022). Systematic inference identifies a major source of heterogeneity in cell signaling dynamics: The rate-limiting step number. *Science Advances*, **8**(11), eabl4598.

Kim, H. J. and MacEachern, S. N. (2015). The generalized multiset sampler. *Journal of Computational and Graphical Statistics*, **24**(4), 1134–1154.

Kolter, R. *et al.* (1993). The stationary phase of the bacterial life cycle. *Annual review of microbiology*, **47**, 855–875.

Leier, A. *et al.* (2014). Exact model reduction with delays: closed-form distributions and extensions to fully bi-directional monomolecular reactions. *Journal of The Royal Society Interface*, **11**(95), 20140108.

Leman, S. C. *et al.* (2009). The multiset sampler. *Journal of the American Statistical Association*, **104**(487), 1029–1041.

Marjoram, P. *et al.* (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100**(26), 15324–15328.

Mather, W. *et al.* (2014). Synchronization of degrade-and-fire oscillations via a common activator. *Phys. Rev. Lett.*, **113**, 128102.

Megerle, J. A. *et al.* (2008). Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophysical Journal*, **95**(4), 2103–2115.

Moon, T. S. *et al.* (2012). Genetic programs constructed from layered logic gates in single cells. *Nature*, **491**(7423), 249–253.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, **135**(2), 216–226.

Raue, A. *et al.* (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**(15), 1923–1929.

Romano, E. *et al.* (2021). Engineering arac to make it responsive to light instead of arabinose. *Nature Chemical Biology*, **17**(7), 817–827.

Ruttor, A. and Opper, M. (2009). Efficient statistical inference for stochastic reaction processes. *Phys. Rev. Lett.*, **103**, 230601.

Schlicht, R. and Winkler, G. (2008). A delay stochastic process with applications in molecular biology. *Journal of Mathematical Biology*, **57**(5), 613–648.

Smith, S. and Grima, R. (2018). Single-cell variability in multicellular organisms. *Nature Communications*, **9**(1), 345.

Taheri-Araghi, S. *et al.* (2015). Cell-size control and homeostasis in bacteria. *Current Biology*, **25**(3), 385–391.

Wilkinson, D. J. (2018). *Stochastic modelling for systems biology*. Chapman and Hall/CRC.