Molecular Therapy
Nucleic Acids
Original Article

Amount of Escape Estimation Based on Bayesian and MCMC Approaches for RNA Interference

Tian Liu,[1] Yongzhen Pei,[1,2] Changguo Li,[3] and Ming Ye[1,4]

[1]School of Computer Science and Technology, Tiangong University, Tianjin 300387, China; [2]School of Mathematical Sciences, Tiangong University, Tianjin 300387, China; [3]Department of Basic Science, Army Military Transportation University, Tianjin 300387, China; [4]Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

The amount of short interfering RNA (siRNA) escaping from the endosome has a significant impact on the efficiency of RNAi. In general, the initial injected amount of siRNAs during the experiment is known, and also the amount of siRNAs after the experiment can be revealed by the level of mRNA measured. However, it is impossible to measure the amount of siRNAs that escape from the endosome and really take part in the chemical reaction of RNAi by detecting the biological organism and its tissues. Inspired by the bottleneck effect in the virus, we introduce the Bayesian approach to infer the amount of escape based on a single type and multiple types of siRNA, respectively. With the consideration of the large calculation quantity of the accurate posterior distribution and the unavailable analytic expression of the likelihood function, our article proposes to take samples by the improved Markov chain Monte Carlo (MCMC) method. The article takes the silencing gene of the synthesis of chitin and the interfering multiple target oncogene as numerical examples to show that our improved MCMC method has higher operation efficiency compared to the Bayesian approach. Our research models siRNA endosome escape using statistical methods for the first time. It perhaps provides a theoretical basis to decrease the cost of a biotic experiment for the future and the standardized statistical approaches for the amount of escape estimation.

INTRODUCTION

RNAi refers to a highly conserved biological process that recognizes double-stranded RNA (dsRNA) in the cell to induce the specific degradation of homologous mRNA during evolution.[1] Endogenously expressed long dsRNA is first cleaved into short interfering RNA (siRNA) by the enzyme, such as Dicer, that is the component of a gene-silencing mechanism, and then the short RNA molecules are exploited as guides to target homologous RNA species.[2,3] The specific suppression of gene expression possibly actualizes through injecting or feeding with dsRNA. The introduction of siRNA into insect cells and silencing of target genes expression offer a new potential tool for the biological pest control method.[4] For example, the RNAi pathway could be applied to reduce the breeding of lepidopteran and coleopteran insect pests via restraining the planta expression,[5] and Mao et al.[6] provide a strategy to impair larval tolerance of gossypol by interfering a cotton bollworm RNA. As a highly efficient technology, RNAi has also developed rapidly in the field of infectious disease and tumor gene therapy,[7,8] and it can cure humans with various diseases that traditional drugs cannot, such as chronic hepatitis B virus.[9] In addition, individualized treatment schemes can be designed according to different conditions of patients.

The significant barrier for efficient siRNA uptake lies in the plasma membrane. In spite of the small size of siRNA molecules, they are still prevented from crossing biological membranes because of their negative charge and hydrophilicity. The procedure of the intracellular transportation of siRNAs begins with early endosomal vesicles. Subsequently, with the fusion of these early endosomes and sorting endosomes, siRNAs are transferred to the late endosomes. Only a small part of siRNAs could escape from the endosomes, and another part with the endosomal contents is removed to the lysosomes. The lysosomes that contain various nucleases acidify the endosomal content, and the siRNAs are degraded in turn. Figure 1 provides a schematic diagram that describes the process of the uptake and intracellular trafficking of a targeted siRNA. So, in order to avert lysosomal degradation, siRNAs have to escape from the endosomes and get into the cytosol, where they will associate with the RNAi mechanism.[10] Besides, it has been found that some of the generated siRNAs are not directly derived from the cleavage of dsRNA but rather, from a chain reaction of RNA polymerase. With the allowance of a single strand of siRNA as a primer and the target mRNA as a template, this reaction amplifies the target mRNA under the action of RNA-mediated RNA polymerase (RdRP) and generates a new siRNA subpopulation.[11] These, in turn, would continue to react to the target mRNA and degrade it.[12] This cyclical amplification process of RNAi explains the reason why a small amount of dsRNA can induce strong gene-silencing effects.

We find that the process of siRNA delivery resembles the biological effect called bottleneck. The bottleneck describes the phenomenon that the number of individuals in a group is reduced drastically or even extinct due to drastic changes in the environment. When we

Received 8 March 2019; accepted 8 October 2019;
https://doi.org/10.1016/j.omtn.2019.10.010.

Correspondence: Yongzhen Pei, School of Mathematical Sciences, Tiangong University, Binshui West Road, Tianjin 300387, China.
E-mail: yongzhenpei@163.com

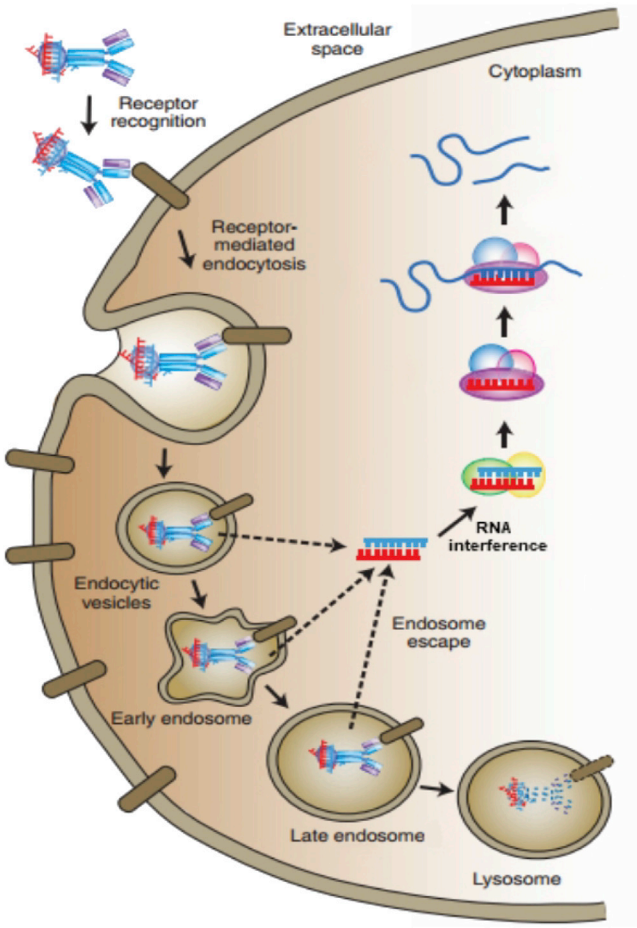

Figure 1. The Process of Escape of siRNA
Uptake and intracellular trafficking of a targeted siRNA delivery vehicle.[10]

inject a certain amount of siRNA into a pest, only a small fraction of the siRNA can across the plasma membrane and participate in the RNAi, and the remaining siRNAs will be degraded. The lower amount of escaping siRNA (commonly known as bottleneck size) will lead to a form of a new population by the amplification process.[13] Accurate quantification of the amount of escape for RNAi is vital for several reasons. First, the estimation of the amount of siRNAs escaping from the endosome helps us to research the biological mechanism of endosomal escape more definitively. Second, the knowledge of the amount of siRNAs of escape in RNAi processes is important to design rationally the strategies that optimize the amount of siRNA to interfere with the target RNA. Finally, the amount of escape impacts the levels of the types that can escape from the endosome into the cytosol when we inject multiple types of siRNA and thereby, impact the effect of interference.

Bottleneck has been extensively researched by many articles that mostly focus on the qualitative analysis of transmission bottleneck sizes,[14] and Abel et al.[15] provide a biologically motivated introduction to bottlenecks. Sobel et al.[16] use the deep-sequencing data to construct the likelihood expression of transmission bottleneck on the basis of the beta-binomial sampling method. Inspired by the above opinions with bottleneck, new ideas aiming at gauging the escaping amounts of siRNA for a single type and multiple types are suggested, respectively. After the observed data are simulated by the Gillespie algorithm, the probability distributions of escaping amounts of siRNA are estimated by means of two algorithms, consisting of the Bayesian approach and the nearest neighbor method.[17] However, both algorithms are inefficient in the course of actual implementation, because the multiple invoking and running of the Gillespie algorithm take much time. So we provide an alternative approach to sample the escaping amounts of siRNA based on the Markov chain Monte Carlo (MCMC) method and take the means of samples as the estimation of escaping amounts to improve the speed of the computer. Finally, comparisons indicate that the estimations inferred by both Bayesian and MCMC methods approximate the true value.

RESULTS

Silence Gene Controlling the Synthesis of Chitin

The oriental migratory locust is a crucial pest in agriculture.[18] Recently, the locust plague has broken out more frequently and severely in China.[19] As we know, the growth and development of locust strictly depend on the biosynthesis and degradation of chitin, which is absent in plants and vertebrates. So, chitin metabolism represents an attractive target for developing safe and effective insecticides.[20]

RNAi can be used to silence genes that control the synthesis of chitin, sequentially leading to the death of locusts. After siRNAs are injected into the locust, they are governed by stochastic processes, including amplification, degradation, immigration, and emigration, which are dominated by a parameter set $\boldsymbol{\theta} = \{\alpha, \lambda, \mu, \sigma\}$. Let $S(t)$ be the amount of the current siRNAs. Then, four stochastic processes are modeled by four biochemical reactions as follows:

$$\begin{array}{ll} S \rightarrow 2S & (a) \\ S \rightarrow \varnothing & (b) \\ \varnothing \rightarrow S & (c) \\ S \rightarrow \varnothing & (d) \end{array} . \qquad \text{(Equation 1)}$$

Next, the biological significance of the construction and parameters in Equation 1 are presented.

- α is the amplification rate of siRNAs that have escaped. Equation 1a means that given the current amount $S(t)$, a unit of new siRNA is generated in the time interval $(t, t + dt)$ with probability $\alpha S(t)dt$.
- λ is the degradation rate of siRNA due to the endocytosis. Equation 1b, represents that a unit of siRNA is degraded by lysosomes with probability $\lambda S(t)dt$ in the time interval $(t, t + dt)$ for given the current states $S(t)$.
- μ is the immigration rate of a new siRNA molecule. Equation 1c reveals that a unit of siRNA immigrates in our system from the neighboring cells with probability μdt in the time interval $(t, t + dt)$.
- σ is the emigration rate of siRNA. Equation 1d shows that siRNA will decrease one unit with the emigration of siRNA into the

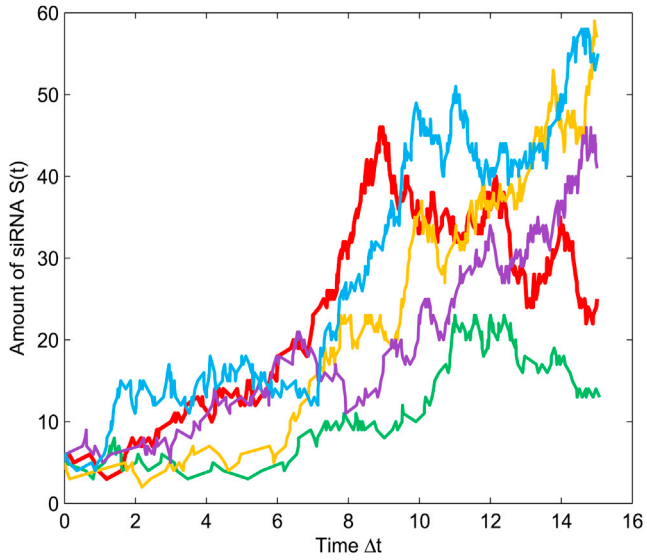

Figure 2. Time Evolutions of the siRNA

The simulations for the dynamic of the siRNA by the Gillespie algorithm[21] are illustrated and the lines with five different colors represent five simulations.

neighboring cells in the time interval $(t, t + dt)$ with the probability $\sigma S(t)dt$ for the given current state $S(t)$.

Take the parameter values $\alpha = 0.6, \lambda = 0.3, \mu = 0.6, \sigma = 0.23$, for example, when the initial value is given by $S(0) = 5$, simulations for the dynamic of the siRNA by the Gillespie algorithm are illustrated in Figure 2. So, the value at $\Delta t = 12$h could be recorded as our observation data s_2 being the amount of siRNA after amplification.

Next, the above observation data s_2 are employed to estimate the amount of escape s_1 or its posterior distribution $p(s_1 \mid s_2)$ and meanwhile, demonstrate the efficacy of Algorithm 1 and Algorithm 2 for the single type of siRNA.

1. Given the target amount of escape $s_1^* \in \{1, 3, 5, 7, 70, 140, 700\}$.
2. Get the data $\{(s_2)_1, (s_2)_2, \ldots, (s_2)_{101}\} \overset{i.i.d}{\sim} Gillespie(s_1^*, \Delta t, \boldsymbol{\theta})$.
3. Make s_2^* be the median of $\{(s_2)_j\}_{j=1,\ldots,101}$.
4. Acquire $p(s_1 \mid s_2^*)$ by Algorithm 1 and the mean s_1 of samples by Algorithm 2, respectively.
5. Compare $p(s_1 \mid s_2^*)$ and the mean with target s_1^*, respectively.

For targets $s_1^* = 7, s_1^* = 70$, and $s_1^* = 700$, we obtained the posterior distributions $p(s_1 \mid s_2)$ of the escaping amount by Algorithm 1 in Figures 3A–3C. Furthermore, we take their modes 9, 67, and 687 as the estimations of the escaping amount, respectively. For the same targets, the samples of the escaping amount are displayed in Figures 4A–4C by Algorithm 2, and their means are estimated as 5, 78, and 687 after burn-in. Obviously, the two kinds of estimations fit the targets very well. This indicates that the two algorithms are efficient.

Interfere Multiple Target Oncogene

Related studies have found that the cancerization of normal cells is the consequence of interaction of multiple genes. However, conventional therapies, which are only targeted toward a single gene mostly, cannot completely inhibit the growth of tumors. It is obvious that RNAi technology can be utilized to silence gene. Yin et al.[22] suggested that injecting multiple types of siRNA can specifically interfere with multiple target oncogenes simultaneously and thereby inhibit the growth and proliferation of cancer cells synergistically.

Consequently, for multiple types, a hypothesis is given that we inject seven types of siRNA $v_0^{[1]}, v_0^{[2]}, \ldots, v_0^{[7]}$ for gene therapy. Then, the observation data $\boldsymbol{v}_2$ could be simulated by the Gillespie algorithm, as previously mentioned. Algorithm 3 and Algorithm 4 are applied to estimate the amount of escaping siRNAs and verify the efficacy of these two methods by the following steps.

1. Given the initial injected amount, $\boldsymbol{v}_0 = \langle 600^{[1]}, 600^{[2]}, \ldots, 600^{[7]} \rangle$.
2. Given the target amount of escape, $s_1^* \in \{1, 3, 5, 7, 70, 140\}$.
3. Generate a mode $\boldsymbol{v}_1^*$ using the multivariate hypergeometric distribution related to random samples of size s_1^* from $\boldsymbol{v}_0$.
4. Get the data $\{(\boldsymbol{v}_2)_1, (\boldsymbol{v}_2)_2, \ldots, (\boldsymbol{v}_2)_{101}\} \overset{i.i.d}{\sim} Gillespie(\boldsymbol{v}_1^*, \Delta t, \boldsymbol{\theta})$.
5. Then, make $\boldsymbol{v}_2^*$ be the median of $\{(\boldsymbol{v}_2)_j\}_{j=1,\ldots,101}$.

Algorithm 1 Estimation of Probability Distributions $p(s_1 \mid s_2)$

Input: the amount of siRNAs after amplification s_2, time interval Δt, and the parameter set $\boldsymbol{\theta}$.

Output: the probability $p(s_1 \mid s_2)$ when $s_1 = 1, \ldots, s_{max} (s_{max} \leq s_2)$.

1. **For** $s_1 = 1$ to s_{max}, **do**
2. Simulate $\{(s_2)_1, (s_2)_2, \ldots, (s_2)_{100}\}$ from s_1 by the Gillespie algorithm
3. Get $\widehat{p}(s_2 \mid s_1)$ from $\{(s_2)_j\}_{j=1,\ldots,100}$ by the nearest neighbor method
4. Set $prob = \widehat{p}(s_2 \mid s_1)$.
5. Set $[\widehat{p}(s_1 = 1|s_2), \ldots, \widehat{p}(s_1 = s_{max} \mid s_2)] = \frac{prob}{sum(prob)}$.
6. **Return** $[\widehat{p}(s_1 = 1|s_2), \ldots, \widehat{p}(s_1 = s_{max} \mid s_2)]$.

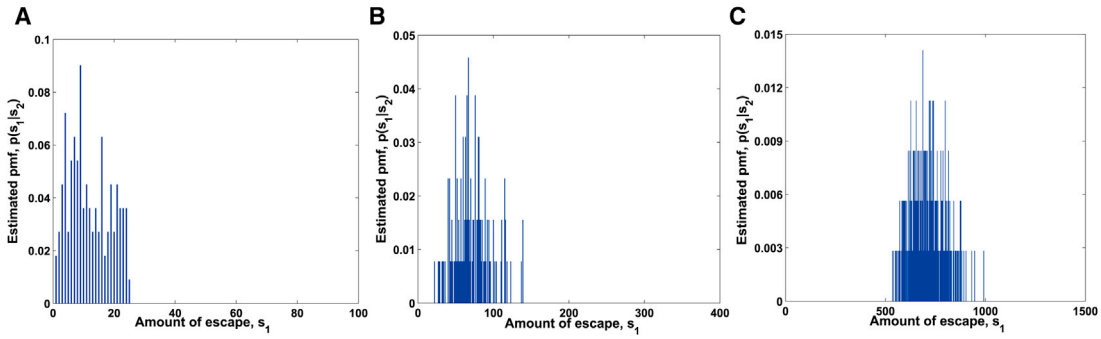

Figure 3. Posterior Distributions p(s_1|s_2) Estimated by Bayesian Inference
Posterior distributions $p(s_1 \mid s_2)$ of amount of escape estimated using Algorithm 1 for (A) $s_1^* = 7$, (B) $s_1^* = 70$, and (C) $s_1^* = 700$.

6. Acquire $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ by Algorithm 3 and the mean s_1 of samples by Algorithm 4, respectively.
7. Compare $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ and the mean with target s_1^*, respectively.

Estimated posterior distributions $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ by Algorithm 3 are shown in Figures 5A–5C for targets $s_1^* = 7, s_1^* = 70$, and $s_1^* = 140$. The modes, as the estimations of the escaping amount, are 9, 65, and 135, respectively. For the same targets, we perform 10,000 samples by Algorithm 4 and report in Figures 6A–6C. After burn-in, we get the estimations 9, 69, and 133 by calculating their means. It can be seen that our predicted results approximate accurately to real ones.

DISCUSSION

The amount of siRNAs escaping from the endosome is one of the important essentials dominating the efficiency of RNAi, but it is intractable to be observed and calculated in experiments. In this paper, two methods are proposed to estimate the amount of escape in terms of the knowledge of the dynamics during amplification from the amount after the reaction and the amount of injection. One is to estimate the posterior distribution of escaping the amount according to the Bayesian approach; the other one is to get the samples of the escaping amount by the MCMC method and to use the mean of samples as an estimate. For the traditional Bayesian

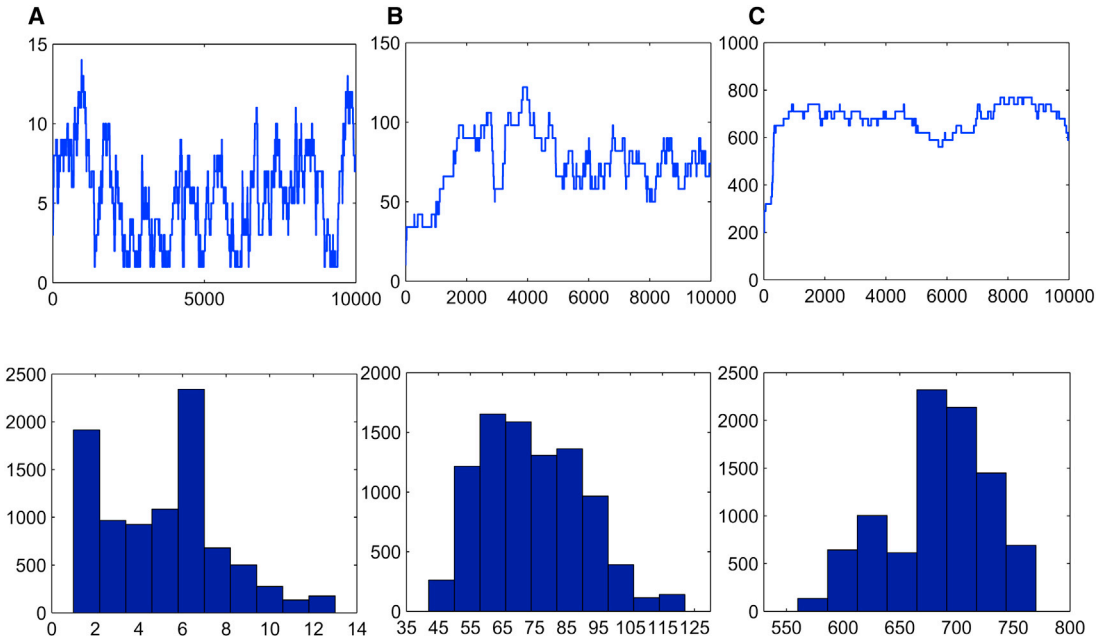

Figure 4. The Results of Sampling for Single Type of siRNA Obtained by MCMC Method
The three panels at the top visualize the sampled data of s_1. For all other panels, the posterior distributions $p(s_1 \mid s_2)$ obtained using Algorithm 2 are delineated. (A) refers to the target $s_1^* = 7$, (B) to the $s_1^* = 70$, and (C) to $s_1^* = 700$.

Algorithm 2 Generating the Samples of s_1

Input: the amount of siRNAs after amplification s_2, time interval Δt, the parameter set $\boldsymbol{\theta}$, initial value $s_1^{(0)}$, number of iterations N, and cycle index $k = 0$.

Output: the sample $s_1^{(0)}, s_1^{(1)}, \ldots, s_1^{(N)}$.

1. Simulate $s_2^{(k)}$ from $s_1^{(k)}$ by the Gillespie algorithm, and calculate $d = \left| s_2^{(k)} - s_2 \right|$
2. **For** $k = 0$ to N, **do**
3. Generate a proposed value s_1' from proposal distribution $q(s_1' \mid s_1^{(k)})$
4. Simulate s_2' from s_1' by the Gillespie algorithm, and calculate $d' = \left| s_2' - s_2 \right|$
5. Sample u from uniform distribution $U(0, 1)$
6. Calculate the acceptance probability α by (Equation 7)
7. **If** $u \leq \alpha(s_1', s_1^{(k)})$, **then**
8. Accept s_1', and set $s_1^{(k+1)} = s_1', d = d'$
9. **else**
10. Reject s_1', and set $s_1^{(k+1)} = s_1^{(k)}, d = d$
11. **Return** $s_1^{(0)}, s_1^{(1)}, \ldots, s_1^{(N)}$

approach, we present the specific algorithms combined with the nearest neighbor method, which is used for the estimation of $p(s_2 \mid s_1)$. For the MCMC method, the acceptance probability of the Metropolis-Hastings (MH) algorithm is controlled by the distance function between the simulation with the observed data. Furthermore, with the contraposition of the single type of siRNAs and multiple types of siRNAs, the algorithms of the estimate of the escaping amount are given, respectively. To inspect the validity of our algorithms, two examples on the silencing gene for the synthesis of chitin and blocking multiple target oncogenes are derived. Our pursuit offers statistical ways to infer the exact amount of siRNAs participating in the actual RNAi reaction. Meanwhile, it perhaps provides a theoretical basis to decrease the cost of the biotic experiment for the future.

Even so, there are still some problems worth exploring further. First, the MCMC method failed to estimate the posterior distribution that could express the uncertainty through the variance of the distributions, although it improves the efficiency. It indicates that a more comprehensive method that takes into account the accuracy of estimation, efficiency, and expression of uncertainty together is required. Besides, the estimation of the bottleneck size is only built on the assumption that the dynamics during amplification are known. When the partial data are missing, how to estimate the amount of escape and the parameters together is the problem for further consideration. In future research, we will try to find the solutions to these problems.

MATERIALS AND METHODS

Single Type of siRNA

In general, we only introduce a single type of siRNA aimed at a specific RNA into the organisms. The processes for which siRNAs escape from the endosome and amplify intracellularly have been described in the first part, and now, we picture them in Figure 7. Define the initial injected amount of siRNAs as s_0, the amount of siRNAs that escape from endosome as s_1, and the amount of siRNAs after amplification as s_2 (Figure 7). Obviously, $s_1 \leq s_2$. Then, on the premise of the amount of siRNAs after amplification, Bayesian inference or MCMC can be applied to estimate the posterior distribution of the escaping amount of siRNAs, as well as their value.

Bayesian Inference

According to the Bayesian framework, the amount of siRNAs escaping from the endosome can be estimated by the posterior probability distributions. Given the observations of the amount after amplification, the distribution is given by

$$p(\text{amount of escape } (s_1) \mid \text{amount after amplification } (s_2)).$$

The merit of the use of the Bayesian approach is that we not only could get the estimates of the most probable amount of escape (in terms of the modes of the distribution), but also, we could be aware of the uncertainty via the variance of the distributions. Then, the posterior probability $p(s_1 \mid s_2)$ is given by

$$p(s_1 \mid s_2) = \frac{p(s_1)p(s_2 \mid s_1)}{\sum_{s_1} p(s_1)p(s_2 \mid s_1)} \propto p(s_1)p(s_2 \mid s_1). \quad \text{(Equation 2)}$$

With the further assumption of the prior $p(s_1)$ to be equally likely, one gets

$$p(s_1 \mid s_2) = \frac{p(s_2 \mid s_1)}{\sum_{s_1} p(s_2 \mid s_1)}. \quad \text{(Equation 3)}$$

Algorithm 3 Estimation of Probability Distributions $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$

Input: the initial injected amount of siRNAs of various types $\boldsymbol{v}_0$, the amount of siRNAs after amplification $\boldsymbol{v}_2$, time interval Δt, and the parameter set $\boldsymbol{\theta}$.

Output: the probability $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ when $s_1 = 1, \ldots, s_{max}$.

1. **For** $s_1 = 1$ to s_{max}, **do**
2. **For** $k = 1$ to $1,000$, **do**
3. Sample $\boldsymbol{v}_1$ from the multivariate hypergeometric distribution with s_1, $\boldsymbol{v}_0$
4. Calculate $p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0)$ by (Equation 11)
5. Set $a = p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0)$
6. Set $b = 1$
7. **For** $v_1^{[i]} \in \boldsymbol{v}_1$, **do**
8. Simulate $\{(v_2^{[i]})_1, (v_2^{[i]})_2, \ldots, (v_2^{[i]})_{100}\}$ from $v_1^{[i]}$ by the Gillespie algorithm
9. Get $\widehat{p}(v_2^{[i]} \mid v_1^{[i]})$ from $\{(v_2^{[i]})_j\}_{j=1,\ldots,100}$ by the nearest neighbor method
10. Set $p = \widehat{p}(v_2^{[i]} \mid v_1^{[i]})$
11. Set $b = b \times p$ ▷ b is $\widehat{p}(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)$ at last
12. Set $prob = a \times b$
13. Set $numert = sum(prob)$
14. Set $[\widehat{p}(s_1 = 1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2), \ldots, \widehat{p}(s_1 = s_{max} \mid \boldsymbol{v}_0, \boldsymbol{v}_2)] = \frac{numert}{sum(numert)}$
15. Get the modes of $[\widehat{p}(s_1 = 1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2), \ldots, \widehat{p}(s_1 = s_{max} \mid \boldsymbol{v}_0, \boldsymbol{v}_2)]$ as an estimation of s_1
16. **Return** $[\widehat{p}(s_1 = 1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2), \ldots, \widehat{p}(s_1 = s_{max} \mid \boldsymbol{v}_0, \boldsymbol{v}_2)]$

Then, the posterior distribution $p(s_1 \mid s_2)$ can be obtained through estimating all of the probability $p(s_1 \mid s_2)$ for $s_1 = 1, \ldots, s_{max}$, where s_{max} is the maximum of escaping amount s_1. The detailed process is shown as follows.

First, starting from s_1, we perform n simulations using the Gillespie stochastic algorithm,[21] according to a parameter set $\boldsymbol{\theta}$ for the dynamics, and obtain the finite simulating samples of s_2 after time interval Δt:

$$\{(s_2)_1, (s_2)_2, \ldots, (s_2)_n\},$$

from which $p(s_2 \mid s_1, \boldsymbol{\theta})$ is estimated using the nearest neighbor method,[17] which is a classical nonparametric estimation method.

Second, with the substitution of all probabilities $p(s_2 \mid \cdot, \cdot)$ into Equation 3, one gets the estimation of the probability distribution $p(s_1 \mid s_2)$.

In detail, the algorithm for estimating distribution $p(s_1 \mid s_2)$ is given as follows.

Algorithm 1 implies that the Gillespie algorithm runs n times when the loop executes one time. It reveals that Algorithm 1 is time consuming if simulating time n is large. So, in order to improve the running efficiency of program, we adopt the MCMC method to estimate the escaping amount of siRNAs.

MCMC Method

The MCMC method includes Gibbs and MH, which are techniques simulating the random variables by using the Markov chain.[23] In this paper, we choose MH to sample single variable s_1, rather than Gibbs from the target distribution, being the conditional distribution of interest. Here, the target distribution, that is, posterior distribution $p(s_1 \mid s_2)$ in Equation 2, is proportional to the product of prior $p(s_1)$ and likelihood $p(s_2 \mid s_1)$.

From the ideas of MCMC, we need to compute the acceptance probability $\alpha(s_1^{'}, s_1^{(k)})$,

where $s_1^{(k)}$ is k th sample, and $s_1^{'}$ is a proposed value. From the symmetry of proposal distribution, namely $q(s_1^{'} \mid s_1^{(k)}) = q(s_1^{(k)} \mid s_1^{'})$,[24] and equally likely possibility of prior

$p(s_1)$, the acceptance probability can be simplified to

$$\alpha\left(s_1^{'}, s_1^{(k)}\right) = \min\left\{1, \frac{p(s_2 \mid s_1^{'})}{p\left(s_2 \mid s_1^{(k)}\right)}\right\}. \qquad \text{(Equation 4)}$$

Again, because $p(s_2 \mid s_1^{'})$ and $p(s_2 \mid s_1^{(k)})$ in Equation 4 are unknown, next, we pursue a novel approach to compute them. For $p(s_2 \mid s_1^{(k)})$, first of all, we simulate one value $s_2^{(k)}$ from $s_1^{(k)}$ after a certain time Δt by the Gillespie algorithm. Second, we compute the distance

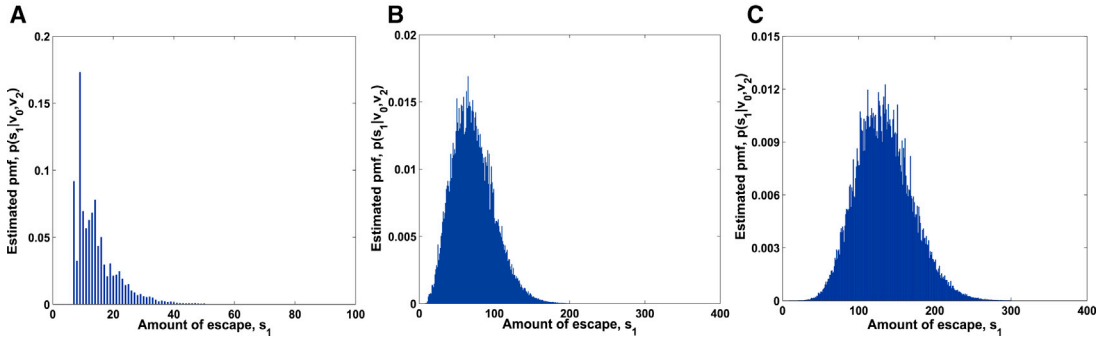

Figure 5. Posterior Distributions $p(s_1|v_0,v_2)$ Estimated by Bayesian Inference

Posterior distributions of amount of escape estimated using Algorithm 3 for (A) $s_1^* = 7$, (B) $s_1^* = 70$, and (C) $s_1^* = 140$.

between the given value s_2 and the simulation $s_2^{(k)}$ denoted by $d = \left| s_2^{(k)} - s_2 \right|$. Finally, the likelihood[25] is calculated by

$$p\left(s_2 \mid s_1^{(k)}\right) = e^{-d}. \quad \text{(Equation 5)}$$

Similarly, another likelihood in Equation 4 is calculated by

$$p\left(s_2 \mid s_1^{'}\right) = e^{-d'}, \quad \text{(Equation 6)}$$

where $d' = \left| s_2^{'} - s_2 \right|$, while $s_2^{'}$ is simulating from $s_1^{'}$ by the same way as $s_2^{(k)}$.

From all of the above, the acceptance probability in Equation 4 is renovated by

$$\alpha\left(s_1^{'}, s_1^{(k)}\right) = min\left\{1, \frac{e^{-d'}}{e^{-d}}\right\}. \quad \text{(Equation 7)}$$

Now, the procedure of sampling s_1 by MCMC methods is listed as follows.

Multiple Types of siRNA

With the consideration of injecting multiple types of siRNAs to affect different target RNAs, the stochastic process of siRNAs is shown in Figure 8. Assume that we inject m types of siRNA for which the initial injected amount consists of $v_0^{[1]}, v_0^{[2]}, \ldots, v_0^{[m]}$, where $v_0^{[i]} \geq 0$ is the amount of i th siRNA. The amount of siRNA is declined to the relatively lower values of $v_1^{[1]}, v_1^{[2]}, \ldots, v_1^{[m]}$ because of endocytosis. After amplification, the composition of siRNA develops into $v_2^{[1]}, v_2^{[2]}, \ldots, v_2^{[m]}$ (Figure 8). With initial injected amount $\boldsymbol{v}_0 = (v_0^{[1]}, v_0^{[2]}, \ldots, v_0^{[m]})$ and the amount $\boldsymbol{v}_2 = (v_2^{[1]}, v_2^{[2]}, \ldots, v_2^{[m]})$ after amplification known, the Bayesian inference and MCMC method can be applied to estimate the posterior distribution $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ and sample s_1 from this posterior distribution, respectively.

Bayesian Inference

For the posterior distribution $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$, we have

$$p\left(s_1 | \boldsymbol{v}_0, \boldsymbol{v}_2\right) = p\left(\bigvee_{\substack{\boldsymbol{v}_1 \\ s.t.sum\,(v_1)=s_1}} \boldsymbol{v}_1 \,\middle|\, \boldsymbol{v}_0, \boldsymbol{v}_2\right) = \sum_{\substack{\boldsymbol{v}_1 \\ s.t.sum\,(v_1)=s_1}} p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2), \quad \text{(Equation 8)}$$

where $sum(\boldsymbol{v}_1) = \sum_i v_1^{[i]}$. Again, from Bayes' theorem, one gets

$$p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2) = \frac{p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1, \boldsymbol{v}_0)}{\sum_{\boldsymbol{v}_1} p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1, \boldsymbol{v}_0)} = \frac{p(\boldsymbol{v}_1 | \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)}{\sum_{\boldsymbol{v}_1} p(\boldsymbol{v}_1 | \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)}. \quad \text{(Equation 9)}$$

Therefore, the incorporation of Equations 8 and 9 yields

$$p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2) = \frac{\sum_{\substack{\boldsymbol{v}_1 \\ s.t.sum(v_1)=s_1}} p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)}{\sum_{\boldsymbol{v}_1} p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)} = \frac{\sum_{s1} \sum_{\substack{\boldsymbol{v}_1 \\ s.t.sum(v_1)=s_1}} p(\boldsymbol{v}_1 | \boldsymbol{v}_0) p(\boldsymbol{v}_2 | \boldsymbol{v}_1)}{\sum_{\boldsymbol{v}_1} p(\boldsymbol{v}_1 | \boldsymbol{v}_0) p(\boldsymbol{v}_2 | \boldsymbol{v}_1)} \quad \text{(Equation 10)}$$

Assume that all types of siRNAs are phenotypically identical and have the same probability of escaping from the endosome. Then, the distribution $v_1^{[1]}, v_1^{[2]}, \ldots, v_1^{[m]}$ of s_1 could be considered as sampling randomly without replacement from the initial injected amount

Algorithm 4 Generating the Samples of s_1

Input: the initial injected amount of siRNAs $\boldsymbol{v}_0$, the amount of siRNAs after amplification $\boldsymbol{v}_2$, time interval Δt, parameter set $\boldsymbol{\theta}$, initial value $s_1^{(0)}$, number of iterations N, and cycle index $k = 0$.

Output: the samples $s_1^{(0)}, s_1^{(1)}, \ldots, s_1^{(N)}$.

1. Sample $\boldsymbol{v}_1^{(k)}$ from the multivariate hypergeometric distribution with $\boldsymbol{v}_0$, $s_1^{(k)}$
2. Simulate $\boldsymbol{v}_2^{(k)}$ from $\boldsymbol{v}_1^{(k)}$ using the Gillespie algorithm, and calculate $d = \|\boldsymbol{v}_2^{(k)} - \boldsymbol{v}_2\|$
3. **For** $k = 0$ to N, **do**
4. Generate a proposed value s_1' from proposal distribution $q(s_1' \mid s_1^{(k)})$
5. Sample $\boldsymbol{v}_1'$ from the multivariate hypergeometric distribution with $\boldsymbol{v}_0$, s_1'
6. Simulate $\boldsymbol{v}_2'$ from $\boldsymbol{v}_1'$ by the Gillespie algorithm, and calculate $d' = \|\boldsymbol{v}_2' - \boldsymbol{v}_2\|$
7. Sample u from uniform distribution $U(0, 1)$
8. Calculate the acceptance probability α by (Equation 17)
9. **If** $u \leq \alpha(s_1', s_1^{(k)})$, **then**
10. Accept s_1', and set $s_1^{(k+1)} = s_1', d = d'$
11. **else**
12. Reject s_1', and set $s_1^{(k+1)} = s_1^{(k)}, d = d$
13. **Return** $s_1^{(0)}, s_1^{(1)}, \ldots, s_1^{(N)}$

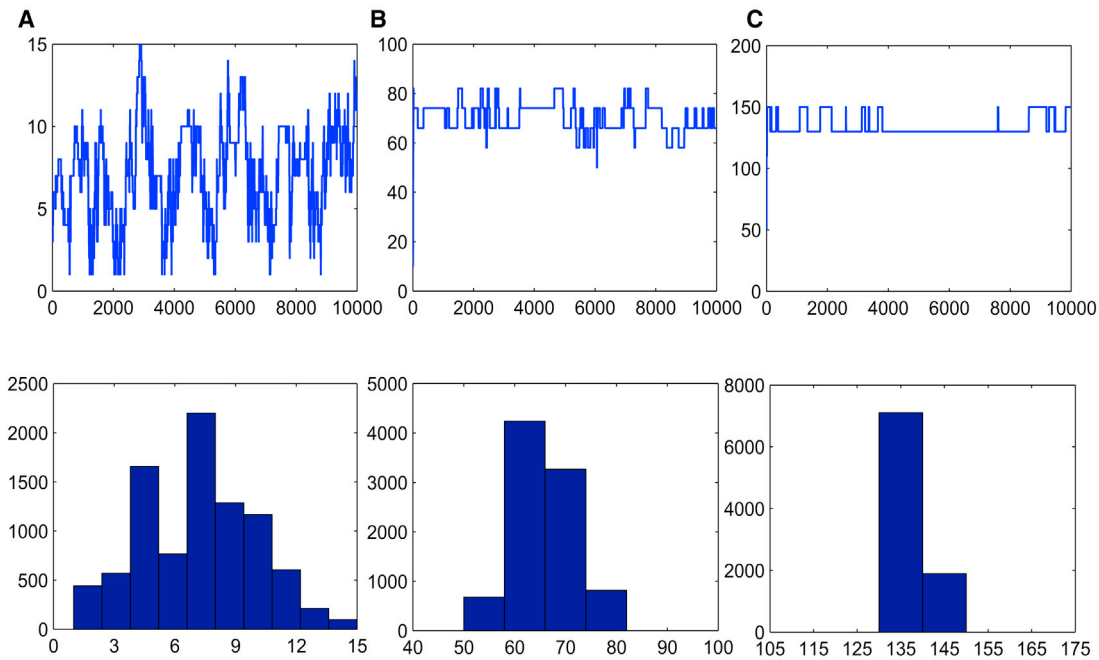

Figure 6. The Results of Sampling for Multiple Types of siRNAs Obtained by MCMC Method

The three panels at the top visualize the sampled data of s_1. For all other panels, the posterior distributions $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ obtained using Algorithm 4 are delineated. (A) refers to the target $s_1^* = 7$, (B) to the $s_1^* = 70$, and (C) to $s_1^* = 140$.

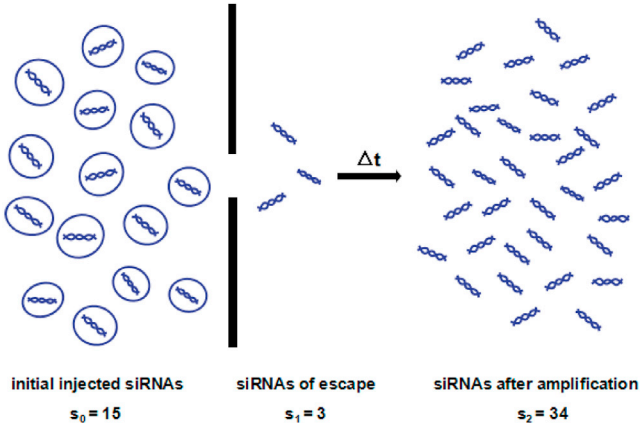

Figure 7. Diagrammatic Representation of the Process that siRNAs Escape and the Amount at Each Stage

Firstly, the siRNAs with an initial injected amount s_0 escape from endosome. And then, the escaping siRNAs s_1 are amplified to s_2 after Δt time.

with distribution $\boldsymbol{v}_0$. So, we can select the amount of escape from a multivariate hypergeometric distribution with $\boldsymbol{v}_0$ and $sum(\boldsymbol{v}_1) = s_1$. The probability of drawing $\boldsymbol{v}_1$ from $\boldsymbol{v}_0$ is given by

$$p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0,\ sum(\boldsymbol{v}_1) = s_1) = \frac{\binom{v_0^{[1]}}{v_1^{[1]}}\binom{v_0^{[2]}}{v_1^{[2]}}..\binom{v_0^{[m]}}{v_1^{[m]}}}{\binom{v_0^{[1]}+v_0^{[2]}+\ldots+v_0^{[m]}}{v_1^{[1]}+v_1^{[2]}+\ldots+v_1^{[m]}}}. \quad \text{(Equation 11)}$$

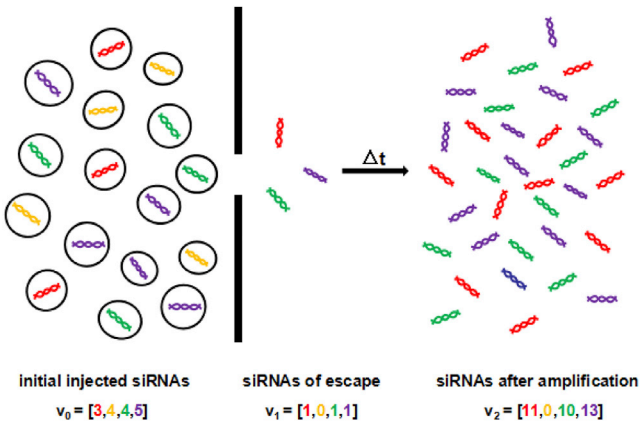

Figure 8. Diagrammatic Representation of the Process in which siRNAs Escape and Their Propensity to Stochastic Variability in Terms of Both the Amount and the Composition of Their Population

Different colors express different types of siRNAs. Initial injected multiple types of siRNA consist of $\mathbf{v_0}$. After endocytosis, their amount decline to $\mathbf{v_1}$. Subsequently , the escaping siRNAs are amplified to $\mathbf{v_2}$ after Δt time.

In reality, components of $\boldsymbol{v}_2$ are simulated by the Gillespie algorithm in view of parameter vector $\boldsymbol{\theta}$. So, for convenience, $p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)$ is denoted by $p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1, \boldsymbol{\theta})$, which is factorized in accordance with the independence between each type of siRNA as follows:

$$p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1, \boldsymbol{\theta}) = \prod_i p\left(v_2^{[i]} \,\middle|\, \boldsymbol{v}_1, \boldsymbol{\theta}\right) = \prod_i p\left(v_2^{[i]} \,\middle|\, v_1^{[i]}, \boldsymbol{\theta}\right). \quad \text{(Equation 12)}$$

Then, $p(v_2^{[i]} \,|\, v_1^{[i]}, \boldsymbol{\theta})$ could be estimated the same way that we estimate $p(s_2 \mid s_1, \boldsymbol{\theta})$, used in Algorithm 1.

The acquisition of $p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0)$ and $p(\boldsymbol{v}_2 \mid \boldsymbol{v}_1)$ that are desired for Equation 10 has been solved in the previous segment, but we should count all of the summands when $\boldsymbol{v}_2$ gets every possible value, such that $\sum_i v_1^{[i]} = s_1$. One key problem is that all possible values of $\boldsymbol{v}_1$ grow superexponentially with s_1 when we give a value of s_1.[26] Now, we face a combinatorial and computational challenge, and so a replaceable approach is required.

To avoid the combinatorial problem, the more probable configuration of $\boldsymbol{v}_1$, such as the modes of $\boldsymbol{v}_1$, could replace the summands that consider all possibilities of $\boldsymbol{v}_1$ in Equation 10. Requena et al.[27] have elaborated an algorithm to solve this question, but now, we provide a simpler sampling method that is to sample points $\boldsymbol{v}_1$ randomly from multivariate hypergeometric distribution $p(\boldsymbol{v}_1 \mid \boldsymbol{v}_0)$ for enough times so that most of these points would be adjacent to the modes. The concrete execution of the sampling procedure is shown in Algorithm 3.

Likewise, as discussed in the context above, there are problems of efficiency with this approach. Therefore, it is tempting to attempt to use the MCMC method.

MCMC Method

Multiple types are also appropriate for the MCMC method. Similar to the single type, our target distribution is posterior distribution $p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2)$ now. From Equation 9, we get

$$p(s_1 \mid \boldsymbol{v}_0, \boldsymbol{v}_2) \propto p(s_1 \mid \boldsymbol{v}_0) p(\boldsymbol{v}_2 \mid s_1). \quad \text{(Equation 13)}$$

In view of the equal possibility of the prior $p(s_1)$ and the previous Equation 13, the acceptance probability about the MH method is given by

$$\alpha\left(s_1^{'}, s_1^{(k)}\right) = min\left\{1, \frac{p(\boldsymbol{v}_2 \mid s_1^{'})}{p\left(\boldsymbol{v}_2 \mid s_1^{(k)}\right)}\right\}. \quad \text{(Equation 14)}$$

In order to go to the acceptance probability, first, we should draw $\boldsymbol{v}_1^{'}$ from the multivariate hypergeometric distribution with s_1 and given $\boldsymbol{v}_0$. Afterward, simulate one vector of $\boldsymbol{v}_2^{'}$ from $\boldsymbol{v}_1^{'}$ after Δt by the Gillespie algorithm, and then, the distance between the given value $\boldsymbol{v}_2$

and the simulation $\boldsymbol{v}_2^{'}$ is recorded as $d' = \|\boldsymbol{v}_2^{'} - \boldsymbol{v}_2\|$. Finally, we give the numerator in Equation 14 as

$$p\left(\boldsymbol{v}_2 \mid s_1^{'}\right) = e^{-d'}. \quad \text{(Equation 15)}$$

Let $\boldsymbol{v}_2^{(k)}$ be simulating from $s_1^{(k)}$, and $d = \|\boldsymbol{v}_2^{(k)} - \boldsymbol{v}_2\|$, the denominator in Equation 14, is computed by

$$p\left(\boldsymbol{v}_2 \mid s_1^{(k)}\right) = e^{-d}. \quad \text{(Equation 16)}$$

Then, we accept $s_1^{'}$ with probability

$$\alpha\left(s_1^{'}, s_1^{(k)}\right) = min\left\{1, \frac{e^{-d'}}{e^{-d}}\right\}. \quad \text{(Equation 17)}$$

The exact process of the MCMC method is described in Algorithm 4.

AUTHOR CONTRIBUTIONS

Y.P. conceived the project and designed the frame of this paper; T.L. and C.L. finished mathematical analyses, performed simulations and wrote the first draft; M.Y. polished, revised the last draft. All authors contributed to the manuscript and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (11471243 and 11971023).

REFERENCES

1. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.
2. Wilson, J.A., and Richardson, C.D. (2003). Induction of RNA interference using short interfering RNA expression vectors in cell culture and animal systems. Curr. Opin. Mol. Ther. *5*, 389–396.
3. Dykxhoorn, D.M., Novina, C.D., and Sharp, P.A. (2003). Killing the messenger: short RNAs that silence gene expression. Nat. Rev. Mol. Cell Biol. *4*, 457–467.
4. Wilson, J.A., and Richardson, C.D. (2005). Hepatitis C virus replicons escape RNA interference induced by a short interfering RNA directed against the NS5b coding region. J. Virol. *79*, 7050–7058.
5. Baum, J.A., Bogaert, T., Clinton, W., Heck, G.R., Feldmann, P., Ilagan, O., Johnson, S., Plaetinck, G., Munyikwa, T., Pleau, M., et al. (2007). Control of coleopteran insect pests through RNA interference. Nat. Biotechnol. *25*, 1322–1326.
6. Mao, Y.B., Cai, W.J., Wang, J.W., Hong, G.J., Tao, X.Y., Wang, L.J., Huang, Y.P., and Chen, X.Y. (2007). Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. Nat. Biotechnol. *25*, 1307–1313.
7. Yang, W.Q., and Zhang, Y. (2012). RNAi-mediated gene silencing in cancer therapy. Expert Opin. Biol. Ther. *12*, 1495–1504.
8. Ma, T., Pei, Y., Li, C., and Zhu, M. (2019). Periodicity and dosage optimization of an RNAi model in eukaryotes cells. BMC Bioinformatics *20*, 340.
9. Wooddell, C.I., Rozema, D.B., Hossbach, M., John, M., Hamilton, H.L., Chu, Q., Hegge, J.O., Klein, J.J., Wakefield, D.H., Oropeza, C.E., et al. (2013). Hepatocyte-targeted RNAi therapeutics for the treatment of chronic hepatitis B virus infection. Mol. Ther. *21*, 973–985.
10. Dominska, M., and Dykxhoorn, D.M. (2010). Breaking down the barriers: siRNA delivery and endosome escape. J. Cell Sci. *123*, 1183–1189.
11. Dougherty, W.G., and Parks, T.D. (1995). Transgenes and gene suppression: telling us something new? Curr. Opin. Cell Biol. *7*, 399–405.
12. Sijen, T., Fleenor, J., Simmer, F., Thijssen, K.L., Parrish, S., Timmons, L., Plasterk, R.H.A., and Fire, A. (2001). On the role of RNA amplification in dsRNA-triggered gene silencing. Cell *107*, 465–476.
13. Dybowski, R., Restif, O., Price, D.J., and Mastroeni, P. (2017). Inferring within-host bottleneck size: A Bayesian approach. J. Theor. Biol. *435*, 218–228.
14. Moncla, L.H., Zhong, G., Nelson, C.W., Dinis, J.M., Mutschler, J., Hughes, A.L., Watanabe, T., Kawaoka, Y., and Friedrich, T.C. (2016). Selective Bottlenecks Shape Evolutionary Pathways Taken during Mammalian Adaptation of a 1918-like Avian Influenza Virus. Cell Host Microbe *19*, 169–180.
15. Abel, S., Abel zur Wiesch, P., Davis, B.M., and Waldor, M.K. (2015). Analysis of bottlenecks in experimental models of infection. PLoS Pathog. *11*, e1004823.
16. Sobel, L.A., Weissman, D., Greenbaum, B., Ghedin, E., and Koelle, K. (2017). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza Virus. J. Virol. *91*, e00171-17.
17. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (Chapman and Hall/CRC).
18. Zhang, J., Liu, X., Zhang, J., Li, D., Sun, Y., Guo, Y., Ma, E., and Zhu, K.Y. (2010). Silencing of two alternative splicing-derived mRNA variants of chitin synthase 1 gene by RNAi is lethal to the oriental migratory locust, Locusta migratoria manilensis (Meyen). Insect Biochem. Mol. Biol. *40*, 824–833.
19. Xia, J.Y. (2002). Analysis on the outbreak of locusta migratoria manilensis and its control strategies. Plant Protection Technology and Extension *22*, 7–10.
20. Cohen, E. (2001). Chitin synthesis and inhibition: a revisit. Pest Manag. Sci. *57*, 946–950.
21. Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. *81*, 2340–2361.
22. Yin, J.Q., Gao, J., Shao, R., Tian, W.N., Wang, J., and Wan, Y. (2003). siRNA agents inhibit oncogene expression and attenuate human tumor cell growth. J. Exp. Ther. Oncol. *3*, 194–204.
23. Gasparini, M. (1996). Markov Chain Monte Carlo in Practice (Chapman and Hall/CRC).
24. Wilkinson, D.J. (2006). Stochastic modelling for systems biology. In Briefings in Bioinformatics, D.J. Wilkinson, ed. (Chapman and Hall/CRC), pp. 204–205.
25. Pandey, A., Mubayi, A., and Medlock, J. (2013). Comparing vector–host and SIR models for dengue transmission. Math. Biosci. *246*, 252–259.
26. Stanley, R.P. (1997). Enumerative Combinatorics (Cambridge University Press).
27. Requena, F., and Ciudad, N.M. (2003). The Maximum Probability 2 × c Contingency Tables and the Maximum Probability Points of the Multivariate Hypergeometric Distribution. Commun. Stat.-Theor. M. *9*, 1737–1752.