Stochastic Kinetic Modeling of Vesicular Stomatitis Virus Intracellular Growth

Sebastian C. Hensel James B. Rawlings
John Yin[*]
Department of Chemical and Biological Engineering
University of Wisconsin-Madison
Madison, WI 53706-1607

February 6, 2009

Abstract

By building kinetic models of biological networks one may advance the development of new modeling approaches while gaining insights into the biology. We focus here on building a stochastic kinetic model for the intracellular growth of vesicular stomatitis virus (VSV), a well-studied virus that encodes five genes. The essential network of VSV reactions creates challenges to stochastic simulation owing to (i) delayed reactions associated with transcription and genome replication, (ii) production of large numbers of intermediate proteins by translation, and (iii) the presence of highly reactive intermediates that rapidly fluctuate in their intracellular levels. We address these issues by developing a hybrid implementation of the model that combines a delayed stochastic simulation algorithm (DSSA) with Langevin equations to simulate the reactions that produce species in high numbers. Further, we employ a quasi-steady state approximation (QSSA) to overcome the computational burden of small time steps caused by highly reactive species. The simulation is able to capture experimentally observed patterns of viral gene expression. Moreover, the simulation suggests that early levels of a low-abundance species, VSV *L* mRNA, play a key

[*] `yin@engr.wisc.edu`

role in determining the production level of VSV genomes, transcripts, and proteins within an infected cell. Ultimately, these results suggest that stochastic gene expression contribute to the distribution of virus progeny yields from infected cells.

1 Introduction

1.1 Stochastic Background

Virus infections are noisy. When a virus encounters a susceptible cell, it binds to receptor molecules on the cell surface, initiating events that enable entry of the virus into the cell and release of its genome, which triggers reactions that ultimately pirate the biosynthetic resources of the cell to produce virus progeny. These processes often involve small numbers: a single virus particle, a handful of receptor molecules, or a single virus genome. The stochastic or noisy behavior of reactions initiated by small numbers of reactants can be especially accentuated in the case of viruses, where the functions encoded by the virus genome often amplify intermediates through processes that can be described by autocatalytic or positive feedback loops. Simulations of noisy gene expression in model viruses support the notion that the noise associated with small numbers of viral intermediates can significantly impact the behavior and productivity of virus infections [2, 32, 29]. Moreover, Delbrück's classical experiments on single cells infected by single virus particles showed how infected cells could produce virus yields that span a broad range of one to two orders of magnitude [8]. Delbrück further speculated that the source of the large variations in yield might easily be accounted for by fluctuations in autocatalytic reactions underlying the virus growth [8, 7].

To better understand mechanistically how noisy reactions may impact the distribution of virus productivity one may develop stochastic kinetic models of virus intracellular growth. In general, one would expect intrinsic fluctuations to impact the dynamics during the earliest stages of infection when levels of viral species are low. These effects may result in extinction of virus species from infected host cells or contribute to broad distributions in virus progeny production. For low numbers of molecules, a continuous or smooth description of the system is not strictly valid because the numbers of molecules are small integer values and reactions cause integer jumps in these values. These systems are typically modeled as discrete jump Markov processes. The stochastic simulation algorithm (SSA), also known as the Gillespie Algorithm [11], is an exact simulation method for these Markov processes. The delayed stochastic simulation algorithm (DSSA) is an extended version of the SSA. It accounts for delays that are involved in various reactions, including gene transcription and replication reactions, as well as protein translation reactions [5, 6]. Other methods directly analyze the master equation and approximation methods based on the Fokker-Planck or Langevin equations [12, 13]. Several hybrid models, that are based on the separation of time scales between fast and slow reactions, have been proposed to reduce the computational effort of the full stochastic models [16, 21, 23, 26, 15, 9, 24, 14, 25]. Partitioning based purely on fast and slow reactions, however, does not produce an efficient simulation because it cannot handle the rapid switching of low concentration species that takes place in the VSV virus infection model.

The challenge of the current work is to advance a stochastic model of a virus infection, allowing for initially low numbers of virus molecules at the initiation of infection. While some species may be rapidly amplified, others may stay at low numbers, and be produced and consumed with high reaction rates, or take part in delayed reactions. Such features cannot, in general, be handled by a simple simulation algorithm. Instead, the model has to be implemented using a strategy that adapts the methods to the ever-changing conditions of the simulation. Here we focus on advancing a stochastic simulation strategy for the intracellular growth of vesicular stomatitis virus (VSV), a relatively well-characterized virus that carries an RNA genome. As a foundation, we build on a deterministic model of VSV growth that accounts for the production, interactions and decay of essential VSV molecular species [19].

1.2 Elements of VSV Biology

Vesicular stomatitis virus (VSV) is a widely studied member of the Rhabdoviridae family, which includes the rabies virus. It carries a single 11 kilobase negative-sense single-stranded RNA genome that encodes five genes, and the molecular processes that define its reproduction within infected cells have been the subject of extensive study [22]. The five genes encode five proteins: the nucleocapsid protein N, the phosphoprotein P, the matrix protein M, the glycoprotein G, and the large polymerase protein L. The latter-most protein is used to transcribe the genome into its messages (mRNAs). These viral messages are then translated into their proteins using ribosomes and other resources of the host cell. Every protein contributes essential functions for the generation of viable virus progeny particles. The N protein encapsidates the genome and stabilizes it, while at the same time it is responsible for the switch between transcription and replication. Once the (−)RNA genome is fully encapsidated by N protein, it serves as a template for the polymerase to synthesize the anti-genome ((+)RNA)). The (+)RNA strand must also be encapsidated by N proteins in order to serve as a template for synthesis of the (−)RNA genome. The P protein also plays a role in the encapsidation process of the genome as well as in the transcription and replication reactions. The G protein forms spikes at the membrane of the host cell, which are then incorporated into the viral membrane as progeny viruses bud from the cell surface. The G protein spikes on the outside of the virus particles enable the virus to bind to other susceptible cells and initiate new infections. The M protein is part of the inner membrane of the virus, and it serves to shut down the host and the viral translation. Although the P, M and G proteins have distinct roles in the virus infection cycle, the current work focuses on early stages of infection, and for simplicity we neglect their contributions and processes of viral binding, particle entry and genome release into the cell cytoplasm. A single VSV particle consists of one (−)RNA genome strand, and about 1258 N, 466 P, 1826 M, 1205 G and 50 L proteins, components that serve to define initial intracellular conditions for our simulations.

An overview of the VSV reaction network is shown in Fig. 1. The (−)RNA and (+)RNA genomes include the naked and the partially encapsidated strands, but not the fully encapsidated strands. (I) shows the encapsidation reaction of the (−)RNA genome. The (+)RNA encapsidation follows the same procedure. In (II) we can see how the replication of the encapsidated (−)RNA genome consumes one L protein at initiation and how the (+)RNA template is formed. The L protein is then released with the completion of the (+)RNA strand synthesis. Synthesis of the (−)RNA strand follows a symmetric mechanism, employing the encapsidated (+)RNA genome as template. (III) shows how the mRNAs are formed from the naked or partially encapsidated (−)RNA genome. The L protein can be released at each gene junction, which leads to different ratios of mRNAs. The genome is shown for a recombinant form of VSV that encodes an additional protein, green fluorescent protein(GFP); it has been found experimentally that carrying GFP in this position has minimal effect on virus yields (unpublished result) while serving as a useful experimental marker to identify and isolate single infected cells.

2 Reactions

Our model includes transcription of the genome to produce viral mRNAs, translation of these message RNA into their corresponding proteins, and replication reactions to synthesize the full length genomic and anti-genomic RNA strands. The transcription reactions are treated as delay reactions, while the translation reactions are non-delayed. Further, the model includes a chain reaction to encapsidate both (+)RNA and (−)RNA genomes, and replication reactions that use the fully encapsidated genome (+)RNA and (−)RNA as templates.

Every transcription reaction involves a different delay. They are initiated at one time but the products appear much later, with time delays ranging from several minutes to more than one hour. These reactions are modeled with a delayed stochastic simulation algorithm (DSSA). Every transcription reaction consumes one polymerase (L protein) at initiation and releases it upon completion. The translation reactions are much faster than the transcription reactions and are therefore modeled with the simpler stochastic simulation algorithm (SSA). The replication reactions consist of non-delayed chain reactions to encapsidate the (+) and (−)RNA genome and a delayed replication reaction that consumes the polymerase at initiation and releases it upon completion of the process. In the following, the reactions that involve a delay are marked with a ($*$). All reactions are modeled as irreversible with rate expressions based on mass action kinetics.

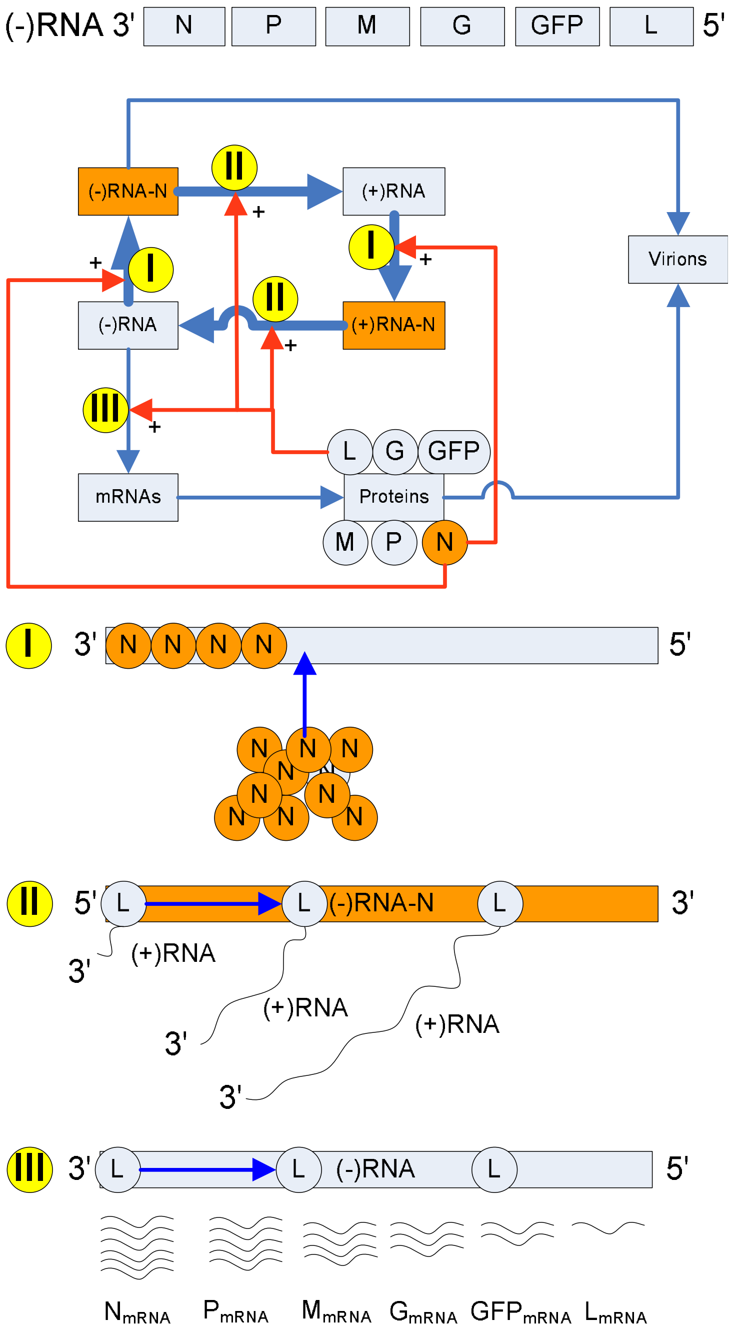

Figure 1: Schematic of vesicular stomatitis virus (VSV) genome, reaction network, and essential reactions.

2.1 Transcription

The transcription reactions transcribe all genes of the genome into their mRNAs, which serve as templates for the viral proteins.

$$
\begin{aligned}
(-)RNA + L + S_1 &\xrightarrow{k_{1(1)}} (-)RNA + L + N_{mRNA} \quad (*) && \text{(1a)}\\
(-)RNA + L + S_1 &\xrightarrow{k_{1(2)}} (-)RNA + L + N_{mRNA} + P_{mRNA} \quad (*) && \text{(1b)}\\
(-)RNA + L + S_1 &\xrightarrow{k_{1(3)}} (-)RNA + L + N_{mRNA} + P_{mRNA} + \\
&\qquad M_{mRNA} \quad (*) && \text{(1c)}\\
(-)RNA + L + S_1 &\xrightarrow{k_{1(4)}} (-)RNA + L + N_{mRNA} + P_{mRNA} + \\
&\qquad M_{mRNA} + G_{mRNA} \quad (*) && \text{(1d)}\\
(-)RNA + L + S_1 &\xrightarrow{k_{1(5)}} (-)RNA + L + N_{mRNA} + P_{mRNA} + \\
&\qquad M_{mRNA} + G_{mRNA} + GFP_{mRNA} \quad (*) && \text{(1e)}\\
(-)RNA + L + S_1 &\xrightarrow{k_{1(6)}} (-)RNA + L + N_{mRNA} + P_{mRNA} + \\
&\qquad M_{mRNA} + G_{mRNA} + GFP_{mRNA} + \\
&\qquad L_{mRNA} \quad (*) && \text{(1f)}
\end{aligned}
$$

The $(-)$RNA in this equation stands for all $(-)$RNA genomes that are not fully encapsidated. The model parameters and delays are given in Table 1. The six transcription reactions involve different delays. The delays have been calculated by dividing the nucleotide length of the genes by the L polymerase elongation rate [17]. While L is released with the last mRNA, messages more closely positioned to the 3' transcription initiation site of the genome are released earlier in time. The initiation rates are calculated by multiplying the different attenuation factors Φ, [4, 17, 22], with the transcription initiation rate k_1^*, [33]. The L protein does not always read through the whole genome, but stops at intergenic regions. The attenuation factors reflect the probability of the L protein transcribing until it reaches the intergenic regions after each gene, where it may fall off. The nucleic acids S_1 are assumed to be in abundance. The transcription reactions have the following rate expressions.

$$r_{1(1)} = k_{1(1)}(-)RNA \cdot L \tag{2a}$$
$$r_{1(2)} = k_{1(2)}(-)RNA \cdot L \tag{2b}$$
$$r_{1(3)} = k_{1(3)}(-)RNA \cdot L \tag{2c}$$
$$r_{1(4)} = k_{1(4)}(-)RNA \cdot L \tag{2d}$$
$$r_{1(5)} = k_{1(5)}(-)RNA \cdot L \tag{2e}$$
$$r_{1(6)} = k_{1(6)}(-)RNA \cdot L \tag{2f}$$

2.2 Translation

The translation reactions produce proteins by translating the messages ($mRNAs$) using the host cell ribosomes:

$$N_{mRNA} + S_2 \xrightarrow{k_2} N_{mRNA} + N \tag{3a}$$
$$P_{mRNA} + S_2 \xrightarrow{k_2} P_{mRNA} + P \tag{3b}$$
$$M_{mRNA} + S_2 \xrightarrow{k_2} M_{mRNA} + M \tag{3c}$$
$$G_{mRNA} + S_2 \xrightarrow{k_2} G_{mRNA} + G \tag{3d}$$
$$GFP_{mRNA} + S_2 \xrightarrow{k_2} GFP_{mRNA} + GFP \tag{3e}$$
$$L_{mRNA} + S_2 \xrightarrow{k_2} L_{mRNA} + L \tag{3f}$$

The parameters for the translation reactions are given in Table 2. The translation rate constant k_2 is the same for all translation reactions as the ribosomal elongation rate is the same for all mRNAs. It is calculated by dividing the translational elongation rate by the ribosome footprint [28] [19]. The ribosome footprint is a parameter that has been estimated in our previous deterministic model [19] by fitting it to four independent sets of data from the literature and our own experimental data. The ribosomes ($S2$) are assumed to be unlimited host resources. The delays associated with translation, which are much smaller than the transcriptional delays, are neglected. The

Table 1: Model parameters and delays of the VSV transcription reactions.

Parameter	Symbol	Φ [4, 17, 22]	Value
Initiation rate constant [33]	k_1^*		0.0461 s^{-1}
Reaction (1a) rate constant	$k_{1(1)}$	1.0	0.0461 s^{-1}
Reaction (1b) rate constant	$k_{1(2)}$	.75	0.0346 s^{-1}
Reaction (1c) rate constant	$k_{1(3)}$	.5625	0.0259 s^{-1}
Reaction (1d) rate constant	$k_{1(4)}$	.422	0.0195 s^{-1}
Reaction (1e) rate constant	$k_{1(5)}$	.422	0.0195 s^{-1}
Reaction (1f) rate constant	$k_{1(6)}$	.0633	0.0029 s^{-1}
Polymerase elongation rate [17]	$k_{e,p}$		3.7 nt/s
Length of N gene [22]	l_N		$1,333$ nt
Length of P gene [22]	l_P		822 nt
Length of M gene [22]	l_M		838 nt
Length of G gene [22]	l_G		$1,672$ nt
Length of GFP gene [22]	l_{GFP}		720 nt
Length of L gene [22]	l_L		$6,380$ nt
Reaction (1a) delay	τ_{t1}		600.27 s
Reaction (1b) delay	τ_{t2}		1062.4 s
Reaction (1c) delay	τ_{t3}		1528.9 s
Reaction (1d) delay	τ_{t4}		2220.8 s
Reaction (1e) delay	τ_{t5}		2655.4 s
Reaction (1f) delay	τ_{t6}		4619.7 s

Table 2: Model parameters of the VSV translation reactions.

Parameter	Symbol	Value
Ribosome elongation rate [28]	$k_{e,r}$	18 nt/s
Ribosome foot print [19]	s_{rib}	238.5 nt
Translation rate constant	k_2	0.0755 s^{-1}

translation reactions have the following rate expressions.

$$r_{2(1)} = k_2 N_{mRNA} \tag{4a}$$
$$r_{2(2)} = k_2 P_{mRNA} \tag{4b}$$
$$r_{2(3)} = k_2 M_{mRNA} \tag{4c}$$
$$r_{2(4)} = k_2 G_{mRNA} \tag{4d}$$
$$r_{2(5)} = k_2 GFP_{mRNA} \tag{4e}$$
$$r_{2(6)} = k_2 L_{mRNA} \tag{4f}$$

2.3 Replication

The replication consists of four different types of reactions. In order to synthesize a copy of the $(-)$strand RNA genome, we need all four reactions to happen in the following order. The $(-)$strand RNA has to be encapsidated with 1258 N proteins before it can serve as a template for (+)strand RNAs via polymerase-mediated replication. The (+)strand RNA also has to be encapsidated by 1258 N proteins in order to serve as a template to synthesize a naked $(-)$strand RNA. The full set of encapsidation reactions, modeled as a set of chain reactions, follows:

$$
\begin{aligned}
(-)RNA + N &\xrightarrow{k_3} (-)RNA_1 \\
(-)RNA_1 + N &\xrightarrow{k_3} (-)RNA_2 \\
\dots &\xrightarrow{k_3} \dots \\
(-)RNA_{1257} + N &\xrightarrow{k_3} (-)RNA_{1258} \\
(-)RNA_{1258} + L &\xrightarrow{k_3} (-)RNA_{1258} + (+)RNA + L \quad (*) \\
(+)RNA + N &\xrightarrow{k_3} (+)RNA_1 \\
(+)RNA_1 + N &\xrightarrow{k_3} (+)RNA_2 \\
\dots &\xrightarrow{k_3} \dots \\
(+)RNA_{1257} + N &\xrightarrow{k_3} (+)RNA_{1258} \\
(+)RNA_{1258} + L &\xrightarrow{50 \cdot k_3} (+)RNA_{1258} + (-)RNA + L \quad (*)
\end{aligned} \tag{5}
$$

An exact stochastic model would include the simulation of all chain reactions and all species that take part in it. Full simulation of this model is computationally expensive owing to the memory needed to store all the states and track their changes. Fig. 2 shows one such simulation run that includes all chain reactions described in Equation 5. $(-)$RNA genomes includes all naked and partially encapsidated $(-)$RNA strands, but not the fully encapsidated $(-)$RNA strands. The first drop (I) in the N protein level is the first full encapsidation reaction described in (I) in Fig. 1. The delay before the first N mRNA occurs is the time required to produce transcripts from the genome, which has been explained in (III) in Fig. 1. The time between (I) and (III) is the time it takes to synthesize the whole genome, which has been described in (II) in Fig. 1. It can be seen that the N protein is highly reactive and begins to fluctuate intensively right after (II), when the first $(+)$RNA genomes are replicated, which consume the N protein.

The simulation was stopped at 100,000 iterations. The full infection cycle cannot be simulated with the model used for this simulation run, because of the fast switching species in the system. It can be seen that all the production and consumption reactions of the N protein dictate the step size, while its level fluctuates intensively. To reduce the computational cost we implement full model up to the point where N protein begins to switch quickly. Then, all chain reactions are implemented using only one single delayed reaction for the full encapsidation of the genome, where the delay is calculated using

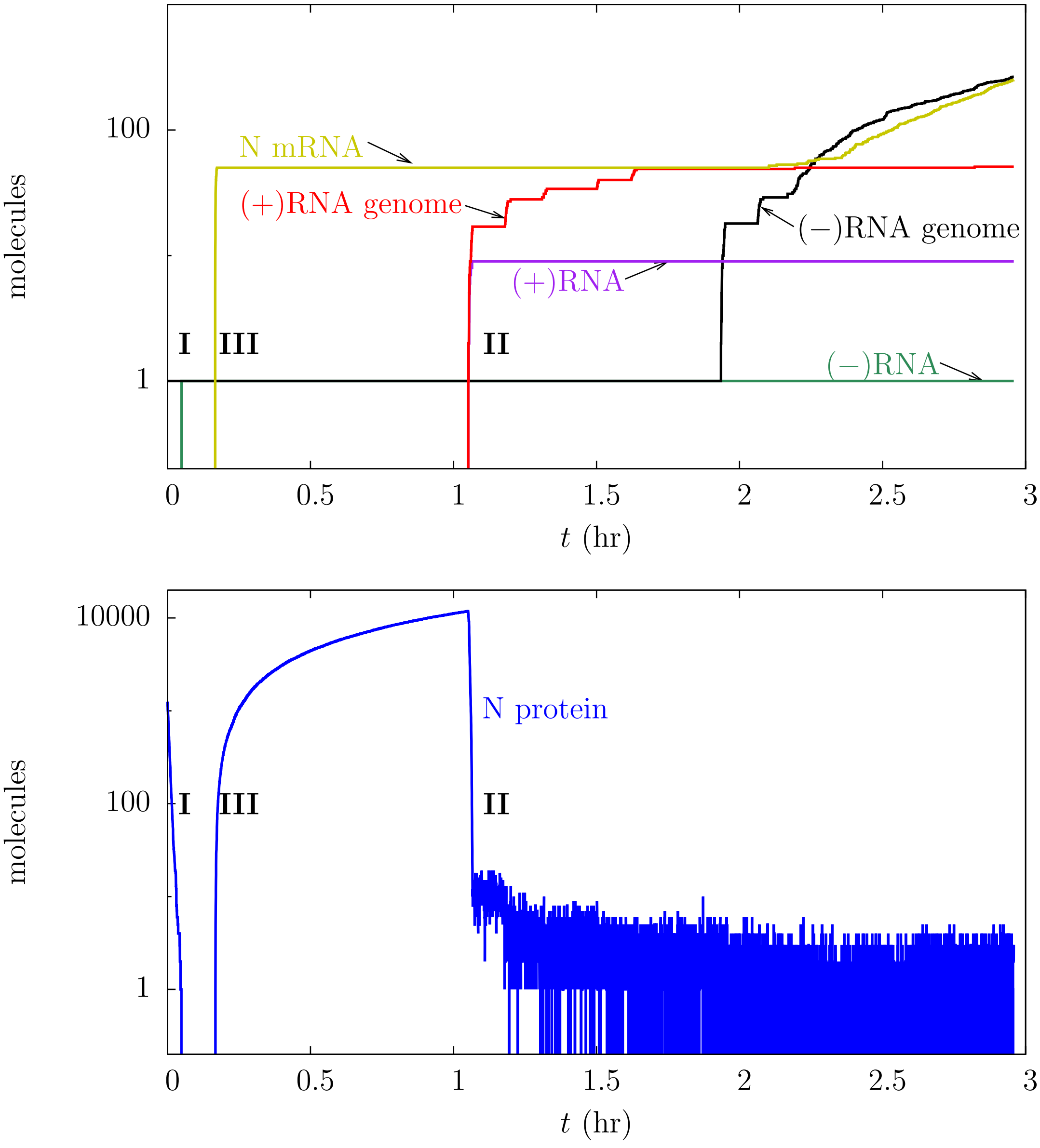

Figure 2: Full simulation of VSV (+)RNA and (−)RNA genome levels and the N mRNA and N protein levels.

the mean of the QSSA distribution of the N protein that we will describe later. The encapsidation reaction of this model has the following form

$$\begin{aligned}
(-)RNA + 1258\,N &\xrightarrow{k_3} (-)RNA_{1258} \quad (*) && \text{(6a)}\\
(-)RNA_{1258} + L &\xrightarrow{k_4} (-)RNA_{1258} + (+)RNA + L \quad (*) && \text{(6b)}\\
(+)RNA + 1258\,N &\xrightarrow{k_3} (+)RNA_{1258} \quad (*) && \text{(6c)}\\
(+)RNA_{1258} + L &\xrightarrow{k_5} (+)RNA_{1258} + (-)RNA + L \quad (*) && \text{(6d)}
\end{aligned}$$

The replication parameters and delays are given in Table 3. The initiation of the (+)RNA synthesis reaction and all chain reactions have the same rate constant as the transcription initiation rate constant k_1^*. It has been found that the promoter strength of the fully encapsidated (+)RNA strand is higher than the promoter strength of the negative strand [27, 10]. The (+)RNA strand can be found in a level that is up to 50 times higher than the level of the (−)RNA strand. Therefore, the initiation reaction rate constant for the (−)RNA synthesis reaction was set at a value 50 times higher than the rate constant of the (+)RNA synthesis. The total time needed for the genome synthesis is almost an hour and it is calculated by dividing the total genome nucleotide length by the L polymerase elongation rate [17]. The delay time for the encapsidation reaction varies with the total number of genomes and the mean of the QSSA distribution of the N protein. There are methods to calculate the delay time via a gamma distribution of multiple reaction events if the reaction rate is not changing over time. However, this is not the case for the chain reaction rate in this model, where the changes in the highly reactive protein N and the genome level cause fluctuating reaction rates. The method to calculate the delay and the QSSA distribution of the N protein level is discussed later. The replication reactions have the following rate expressions.

$$\begin{aligned}
r_{3(1)} &= k_3(-)RNA \cdot N && \text{(7a)}\\
r_{3(2)} &= k_4(-)RNA_{1258} \cdot L && \text{(7b)}\\
r_{3(3)} &= k_3(+)RNA \cdot N && \text{(7c)}\\
r_{3(4)} &= k_5(+)RNA_{1258} \cdot L && \text{(7d)}
\end{aligned}$$

Table 3: Model parameters and delays of the VSV replication reactions.

Parameter	Symbol	Value
Length of genome [22]	l_g	$11,765$ nt
Polymerase elongation rate [17]	$k_{e,p}$	3.7 nt/s
Reaction (6a,6b) rate constant [33]	k_3	0.0461 s^{-1}
Reaction (6b) rate constant [33]	k_4	0.0461 s^{-1}
Reaction (6d) rate constant [33]	k_5	2.305 s^{-1}
Reaction (6a) delay	τ_{r1}	varies
Reaction (6b) delay	τ_{r2}	3179.7 s
Reaction (6c) delay	τ_{r3}	varies
Reaction (6d) delay	τ_{r4}	3179.7 s

2.4 Host Factors

Among multiple host factors or characteristics that could influence our reaction network, we consider only the cell size, which we assume, for simplicity, to be constant during the infection cycle. Cell size is given in Table 4. The shape of the virus is assumed to be spherical and we used average diameters of BHK host cells to estimate the volume of the host cells. Reaction rates depending on concentrations that have been derived from experimental data is converted into molar reaction rates using this average host cell volume.

Table 4: Host parameters of the BHK-21 cells.

Parameter	Symbol	Value
Average cell diameter	d	16 μm
Average cell volume	v	2140 $(\mu\mathrm{m})^3$

2.5 Delayed Stochastic Simulation Algorithm (DSSA)

The model is implemented with a delay stochastic simulation algorithm (DSSA) developed by Bratsun et al. [6] and Barrio et al. [5]. The delays have to be treated carefully because all delayed reactions consume species upon initiation and release several products at different times. The DSSA handles delayed and non-delayed reactions, and is an extended version of the original Gillespie Algorithm [11]. The following algorithm uses waiting and delay times, and also delayed reactions that change the state of the system at both initiation and completion. The stoichiometric matrix of the non-delayed reactions and the initiation of delayed reactions is denoted ν_i, and the stoichiometric matrix of the completion of the delayed reactions is denoted ν_d. The stored reaction times are saved in T_d.

1. Set time t equal to zero, the number of species x to x_0 and the first stored completion time t_d to ∞

2. Calculate all m reaction rates, $r_j(x) = k_j a_j(x)$

3. Calculate the total reaction rate, $r_{tot} = \sum_{j=1}^{m} r_j$

4. Generate two random numbers (p_1, p_2) uniformly distributed on (0,1)

5. Calculate the stochastic time step $\Delta t = -\ln(p_1)/r_{tot}$

6. if there is a stored reaction n to finish in $[t, t+\Delta t)$:

    - Discard steps 4–5 and update time $t = \min(T_d) = t_d$
    - Update species number $x = x + \nu_{d(n)}$
    - Repeat steps 2–7 while $t \leq t_{final}$

7. else:

    - Find reaction n, such that $\sum_{j=1}^{n-1} r_j(x) < p_2 r_{tot} \leq \sum_{j=n}^{m} r_j(x)$
    - Update time $t = t + \Delta t$
    - Update species number $x = x + \nu_{i(n)}$
    - If reaction is delayed, store the time $t + \tau$ at which the system should be updated according to the completion of reaction n
    - Repeat steps 2–7 while $t \leq t_{final}$

2.6 The Langevin Equation

Initial runs indicated that some species were produced at levels in the millions of molecules (not shown), significantly reducing the speed of the simulation. To enable faster simulation an approximation was employed to allow for larger time steps, while still accounting for fluctuations. A method that has already been explored in various stochastic models is the use of Langevin equations [30, 18]. The formulation of the chemical Langevin equation has been addressed by Gillespie [13]. The Langevin equation is a good approximation under certain conditions that may change during a simulation run.

$$x(i+1) = x(i) + r\Delta t + \sqrt{r\Delta t}\, R \qquad R \sim N(0,1) \tag{8}$$

The equation above characterizes a first order approximation of a continuous time stochastic process, in which $x(i+1)$ is the next state, $x(i)$ is the initial state, $r\Delta t$ is the first order change in $x(i)$, and $\sqrt{r\Delta t}\, R$ is the standard deviation of that first order change, multiplied with a normally distributed random number R. Therefore, the next state is not only calculated by the mean of the reaction rate, but by a normally distributed probability distribution around that mean. This equation is valid only when the rate r is large and the time step is chosen so that the change in x is small. The number of all species that are influenced by reaction r must therefore be large as well. Then the reaction can be modeled with a Langevin equation and is not updated as part of the DSSA. The simulation algorithm has to be capable of switching between the implementation via the Langevin equations and the DSSA.

2.7 QSSA on the N Protein

The N protein is a highly reactive species in the system. It switches among small integer values so the fast time scale of N protein fluctuations cannot be treated by partitioning the system into fast and slow reactions. Model reduction methods for highly reactive species have recently been derived by [20]. We next present how to apply that reduction method to handle the N protein. The N protein is produced by its mRNA with a high rate, while the N protein is consumed by both positive and negative RNA encapsidation processes. All rates for the chain reactions are calculated by the same reaction rate constant k_3. The N_{mRNA} and the genomes are present in large amounts, while N is highly reactive. Considering the N_{mRNA} and the genome levels to be constant at some values over some interval of interest, the master equation for N can be written as follows:

$$\begin{aligned}\frac{dP(N,t)}{dt} &= -k_2 N_{mRNA} P_N + k_2 N_{mRNA} P_{N-1} \\ &-k_3 \sum_{i=0}^{1257}((-)RNA_i + (+)RNA_i) N P_N \\ &+k_3 \sum_{i=0}^{1257}((-)RNA_i + (+)RNA_i)(N+1) P_{N+1} \qquad (9)\end{aligned}$$

in which P_N is shorthand for $P(N,t)$. If the production and consumption rates are high such that N equilibrates to its steady-state condition on a fast time scale compared to the evolution of level mRNA and the genomes, the steady-state probability density of N can be found by setting $dP(N,t)/dt = 0$. Using

$$\begin{aligned} r_1 &= k_2 N_{mRNA} \\ r_2 &= k_3 \sum_{i=0}^{1257}((-)RNA_i + (+)RNA_i) \qquad (10)\end{aligned}$$

the following equation can be derived:

$$0 = -r_1 P_N + r_1 P_{N-1} - r_2 N P_N + r_2 (N+1) P_{N+1} \qquad (11)$$

Evaluating this equation for N = 0,1,2,...

$$\begin{aligned} N=0 &\qquad 0 = -r_1 P_0 + r_2 P_1 \\ N=1 &\qquad 0 = -r_1 P_1 + \underbrace{r_1 P_0 - r_2 P_1}_{\text{zero from } N=0} + 2 r_2 P_2 \\ N=2 &\qquad 0 = -r_1 P_2 + \underbrace{r_1 P_1 - 2 r_2 P_2}_{\text{zero from } N=1} + 3 r_2 P_3 \\ \ldots &\qquad \ldots \\ N=n &\qquad 0 = -r_1 P_n + \underbrace{r_1 P_{n-1} - n r_2 P_n}_{\text{zero from } N=n-1} + (n+1) r_2 P_{n+1} \qquad (12)\end{aligned}$$

This relation provides the following recursion in terms of P_0.

$$\begin{aligned} P_N &= \frac{1}{N} \alpha P_{N-1} \\ P_N &= \frac{1}{N!} \alpha^N P_0 \qquad \alpha = \frac{r_1}{r_2} = \frac{k_2 N_{mRNA}}{k_3 \sum_{i=0}^{1257}((-)RNA_i + (+)RNA_i)} \qquad (13)\end{aligned}$$

Summing P_N over N gives us:

$$\begin{aligned}\sum_{N=0}^{\infty} P_N &= (1 + \frac{\alpha}{1!} + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \cdots)P_0 \\ 1 &= e^{\alpha} P_0 \\ P_0 &= e^{-\alpha} \qquad (14)\end{aligned}$$

The quasi-steady probability density of N is therefore:

$$P_N = \frac{1}{N!}\alpha^N e^{-\alpha} \qquad (15)$$

2.8 Delayed Replication Reaction

As mentioned earlier, the encapsidation reaction of the genome is modeled as a delayed reaction. The reaction rate constant k_3 for all chain reactions is the same. When the distribution of the N protein stays constant over the amount of time that it takes to encapsidate a whole genome, the average reaction rate for a single chain reaction is calculated using only the mean of the QSSA distribution derived above. All chain reactions follow the same reaction rate and time, and summing over all reaction times, the total time needed to encapsidate the whole genome can be calculated. One should note that the resulting delay is an approximation that does not reflect the full stochasticity of all encapsidation reactions. The approximation allows simulation to larger times, however, by avoiding the firing of all single chain reactions occurring at small time steps. In the full model, all chain reactions can then be modeled as a delayed reaction that is initiated by the first chain reaction with the N protein level drawn from the QSSA probability density in Equation 15. All delayed chain reactions are stored in $T_{d,rep}$. When the distribution of N changes, and therefore the reaction rate and the delay time changes, the remaining time of each stored delayed encapsidation reaction has to be updated accordingly. This method will be described subsequently.

3 Simulation Strategy

The VSV model in this work consists of a variety of delayed and non-delayed reactions that cannot be simulated by only using the DSSA owing to the high computational burden of the fast reactions. Our simulation strategy focuses on combining different methods and approximations to attain a computationally inexpensive and exact stochastic simulation of the model.

3.1 Hybrid Langevin Implementation

Computational efficiency can be gained if reaction systems can be partitioned into subsets of different time scales, or fast and slow reaction subsets [16, 21]. These methods can be applied only when molecule levels are high, but the methods do not work accurately for fast switching states at low levels. In this work, the fast reactions are further divided into two sets: those that influence molecules present at high numbers only and those that influence molecules present at low numbers. The "high" or "low" molecule number fast reactions are approximated via Langevin equations or sampled as stochastic events, respectively. For this model, we separated "high" and "low" molecule number fast reactions by a threshold set to $n_{sw} = 100$, which divides all fast reactions into m_1 "low" molecule number fast reactions and m_2 "high" molecule number fast reactions. The value of n_{sw} was chosen by comparing the solution of the full model with the approximate model using different n_{sw} values and choosing the smallest value for which the approximate model remained in good agreement with the full model. The maximum Langevin step is set to a value that we do not change the molecular numbers more than one percent, $tol = 0.01$, and in no case exceed a maximum stepsize, $\bar{\Delta}_h = 10$ s. When the stochastic step is bigger than the Langevin step, only the continuous Langevin states are updated and the stochastic time step is discarded. This is valid because the reaction rates are exponentially distributed and therefore memoryless. Using this separation, one may define a hybrid Langevin algorithm:

1. Set time t equal to zero, the number of species x to x_0 and define n_{sw}
2. Separate the reactions into "high" and "low" molecule fast reactions by comparing n_{sw} to all non-zero stoichiometric species of each fast reaction
3. Round all non-integer molecule levels that are smaller than n_{sw}
4. Calculate all "low" molecule number fast reaction rates, $r_l = k_l a_l(x)$ and the "high" molecule number fast reaction rates, $r_h = k_h a_h(x)$
5. Calculate the total "low" molecule number reaction rate, $r_{tot} = \sum_{l=1}^{m_1} r_l$

6. Generate two random numbers (p_1, p_2) uniformly distributed on (0,1)

7. Calculate the stochastic time step $\Delta_l t = -\ln(p_1)/r_{tot}$

8. Calculate the maximum Langevin step
   $\Delta_h t = \min(\bar{\Delta}_h, tol \cdot \min(\text{abs}(x_h/r_h)))$

9. if $\Delta_l t \leq \Delta_h t$

   - Set time step $\Delta t = \Delta_l t$ and update $t = t + \Delta t$
   - Find reaction α, such that $\sum_{l=1}^{\alpha-1} r_l < p_2 r_{tot} \leq \sum_{l=\alpha}^{m_1} r_l$
   - Update species number $x = x + \nu_\alpha$

10. else

    - Set time step $\Delta t = \Delta_h t$ and update $t = t + \Delta t$

11. Generate m_2 random numbers $R \sim N(0, 1)$

12. Update the Langevin part $x_h = x_h + r_h \Delta_t + \sqrt{r_h \Delta t} \cdot R$

13. Repeat steps 2–11, while $t < t_{final}$

This algorithm is much faster than the full SSA when the reaction network has a "fast" reaction subset and it still captures fluctuations of reactions with high molecule levels via Langevin equations. The separation of reactions is implemented within the algorithm, not off-line. This gives the system the flexibility to run exact stochastic sampling for fast reactions affecting low molecule numbers and via Langevin equations for fast reactions affecting only high molecule levels. It can also be extended for use with a DSSA. The same algorithm can be used for the initiation reactions of the DSSA, but in addition $\Delta_l t$ and $\Delta_h t$ have to be compared to the first stored completion time of the delayed reactions $t_d = \min(T_d)$. If $t_d \leq t + \Delta_l t$ and $t_d \leq t + \Delta_h t$, the delayed reaction completes. If not, steps 9–11 of the hybrid Langevin algorithm have to be followed; if a delayed reaction is initiated, that information is stored at time $t + \tau$ and the system is updated at the completion of the delayed reaction.

The QSSA can be used for the fast switching state in the system. If a state equilibrates to its steady-state condition on a time scale that is fast compared to the evolution of other species that are involved in reactions with this state, then we find the steady state probability density for this state as mentioned earlier. In the case of the N protein, the simulation switches to the QSSA of N at around $t \approx 1.1$ hr as shown in Figure 2. Instead of modeling all production reactions, the N protein level is calculated by drawing randomly from its probability density at each time step. The approximated N protein level is used to calculate the initiation reaction rate r_{encap} and the delay time τ_{encap} of each encapsidation reaction. When the mean α of the N steady-state probability density is changing, all stored reaction delays $T_{d,rep}$ have to be updated using the new delay time.

$$\begin{aligned} r_{encap}(i) &= k_3 \cdot \alpha(i) \\ \tau_{encap}(i) &= 1258 \cdot \frac{1}{r_{encap}(i)} \\ T_{d,rep}(i) &= \frac{T_{d,rep}(i-1) - \Delta}{\tau_{chain}(i-1)} \cdot \tau_{chain}(i) \end{aligned} \tag{16}$$

Although the QSSA on N allows us to take larger time steps, the L protein also exhibits fast switching at low molecule levels. An approximation for the delayed species L will need to be developed before the simulation can be run to the end of the viral infection.

4 Simulation Results

The complexity of our model of the VSV reaction network lies in the coupling of the fast switching species and the delayed reactions. We are able to implement our model using Langevin equations and the QSSA assumption for the N protein, which allows us to take larger time steps and simulate further in time. Fig. 3, Fig. 4 and Fig. 5 show the means of the genomes, the viral mRNAs and the viral proteins, respectively, for 1000 simulations. In Fig. 3 we find that the first (+)RNA appears shortly after one hour, which is the time required to encapsidate and replicate from the (−)RNA genomic template. Owing to the delay and the transcriptional attenuation between adjacent genes, our simulation shows a gradual decrease of the mRNA level, as shown in Fig. 4 following $N > P > M > G > GFP > L$. The proteins in Fig. 5 show the same order for the P, M, G and GFP proteins because they are not currently implemented in other reactions. The levels of free N and L proteins rapidly fluctuate because they are rapidly produced and consumed, so these free protein levels deviate from the pattern observed for their mRNA

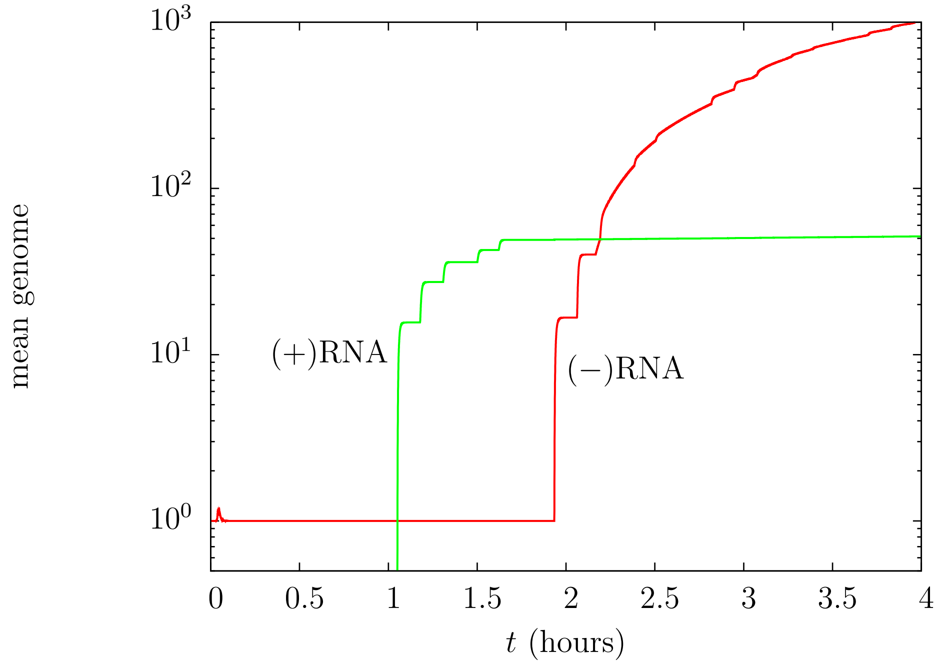

Figure 3: Mean of the genome levels versus time. (−)RNA includes all naked, partially and fully encapsidated (−)RNA strands, and (+)RNA includes all naked, partially and fully encapsidated (+)RNA strands.

expression levels. More simulation runs would be needed to obtain smoother mean values in the levels of free N and L proteins.

An intriguing result is the emergence of clusters in the distribution of (−)RNA genome levels across 1000 simulation runs, as shown in Fig. 6. These clusters arise owing to the sensitivity of (−)RNA genome replication to the availability of polymerase, reflected in L protein, the VSV intermediate that is present at the lowest levels in the model. Here the (−)RNA genome includes all forms of the molecule: (−)RNA strands that are naked as well as partially and fully encapsidated by N protein. We see that the mean level of (−)RNA genome increases over time and that the distribution divides into clusters beginning at about 2.5 hr post infection (HPI). The clusters arise because formation of L mRNA during the first round of transcription from the initial entering (−)RNA genome is a rare event, yet this event is essential for the early production of L protein, which is required for initial VSV gene expression and replication of the (−)RNA genome. To better appreciate the

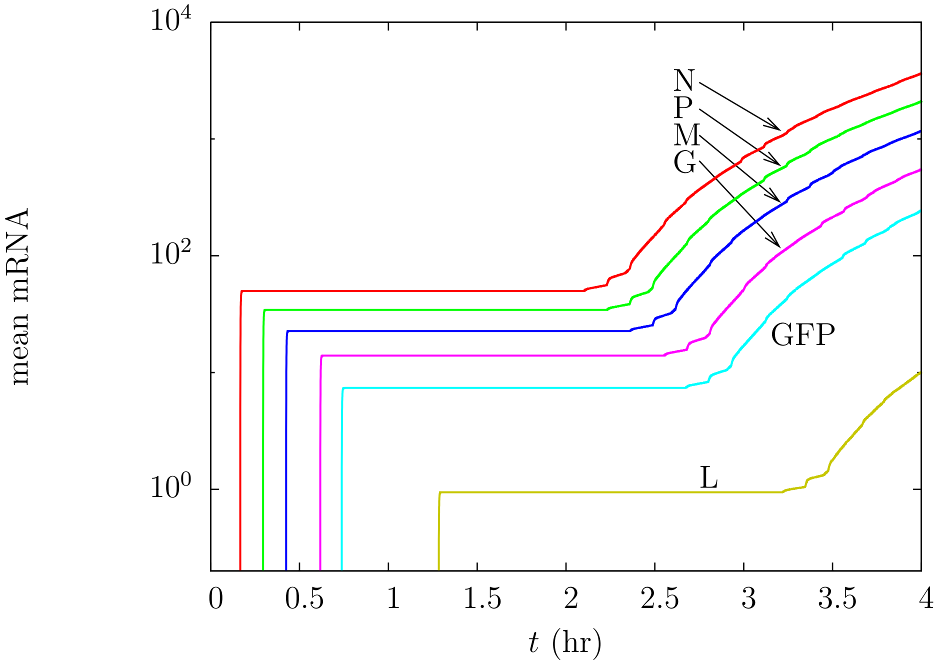

Figure 4: Mean of mRNA levels versus time.

dependence of (−)RNA genome production on the early production of L mRNA one may examine how the level of (−)RNA genomes at 4 hr post-infection depends on the level of L mRNA at 1.5 hr post-infection, as shown in Fig. 7. Here the relatively rare infected cells that have produced five L mRNA molecules by 1.5 hr produce about a factor of two more (−)RNA genomes at 4 hr than the relatively common infected cells that have produced one L mRNA molecule by 1.5 hr post-infection.

The distribution of GFP levels at different times, shown in Fig. 8 share similarities with the distribution of (−)RNA genomes. Both exhibit a separation of populations across their distributions. However, the separation emerges at a later time in the case of GFP protein, reflecting the delayed impact of (−)RNA genome production on the subsequent production of virus-encoded mRNAs and protein.

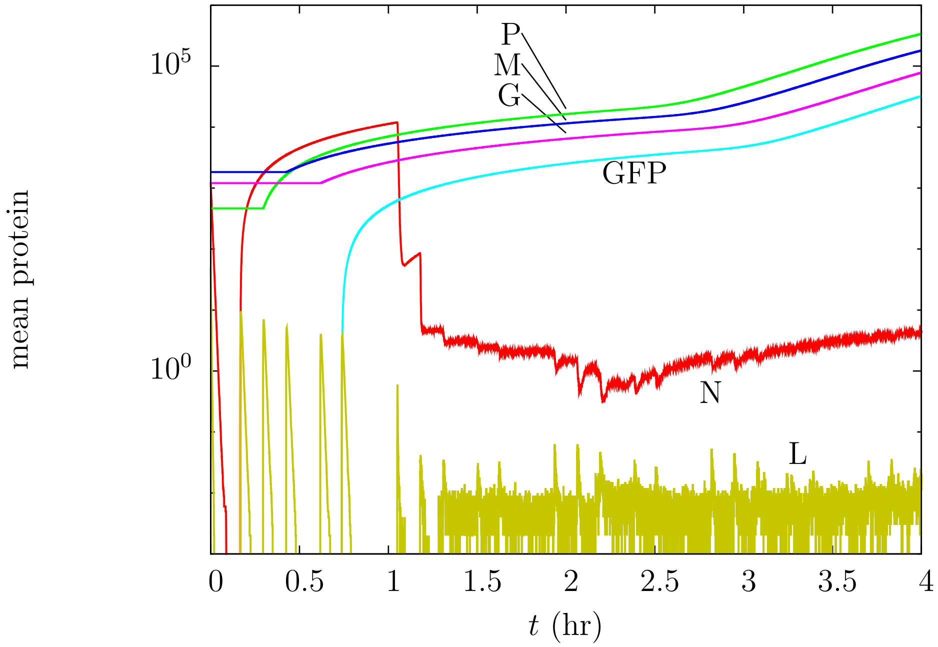

Figure 5: Mean of protein levels versus time.

5 Discussion

Recent efforts toward understanding the behavior of stochastic reaction networks have focused on reducing the complexity and computational burden in simulating their behavior, especially in cases involving delayed reactions, reactions that produce high levels of essential intermediates, or highly reactive species. However, no single approach enables efficient stochastic simulation of a core reaction network for the growth of VSV, an experimentally well-studied virus. Processes of viral transcription and genome replication involve significant delays, so it is appropriate to employ a delayed stochastic simulation algorithm (DSSA). Fast and productive reactions, such as protein translation, generate high numbers of protein molecules; here, the use of Langevin equations can enable simulation of their trajectories while minimizing their computational burden. In addition, simulation of highly reactive species, such as the rapid formation and depletion of viral N protein by translation and genome encapsidation, respectively, motivate the development of a quasi-steady state approximation (QSSA) to avoid explicit and costly simulation

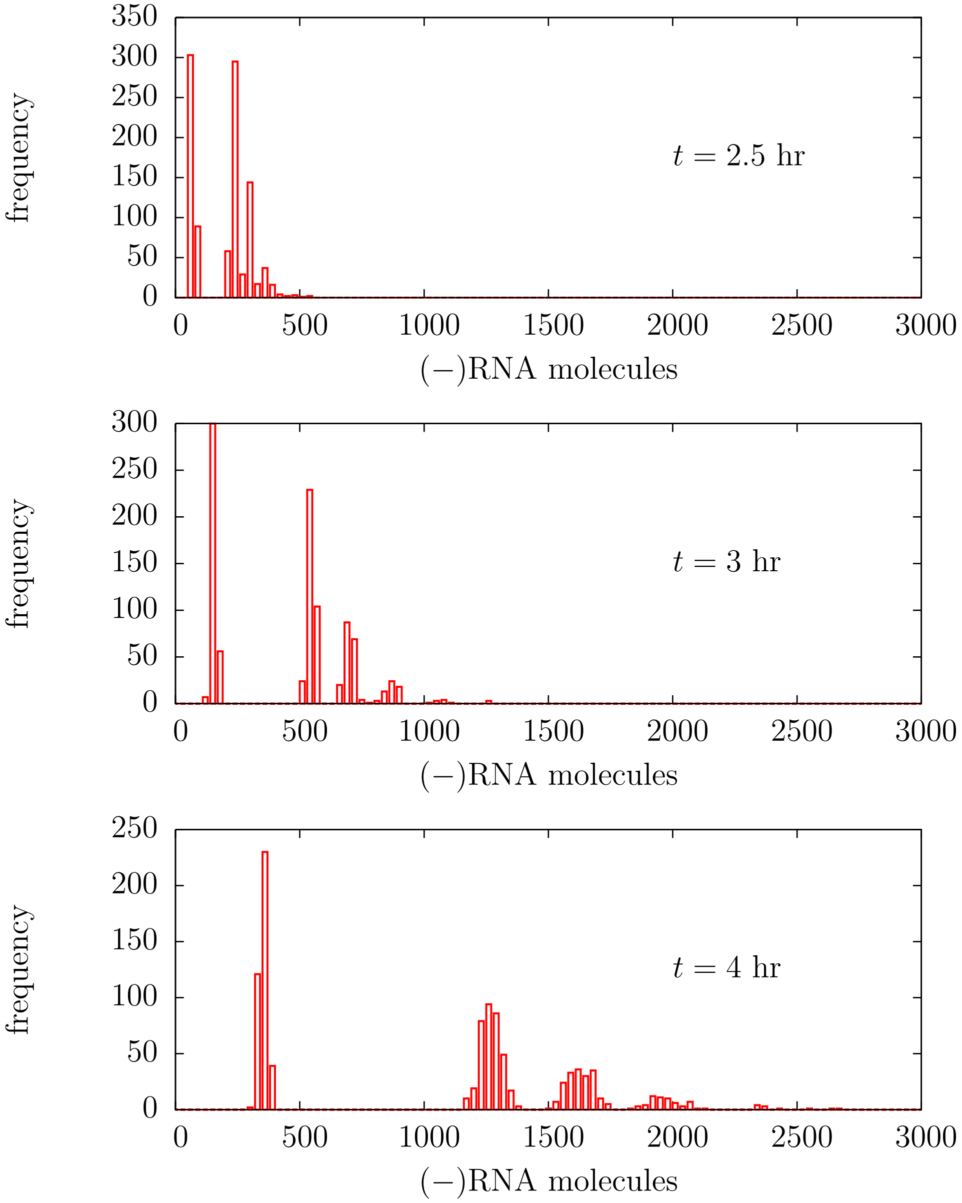

Figure 6: Distribution of the (−)RNA at $t = 2.5, 3$, and 4 hr for 1000 simulations.

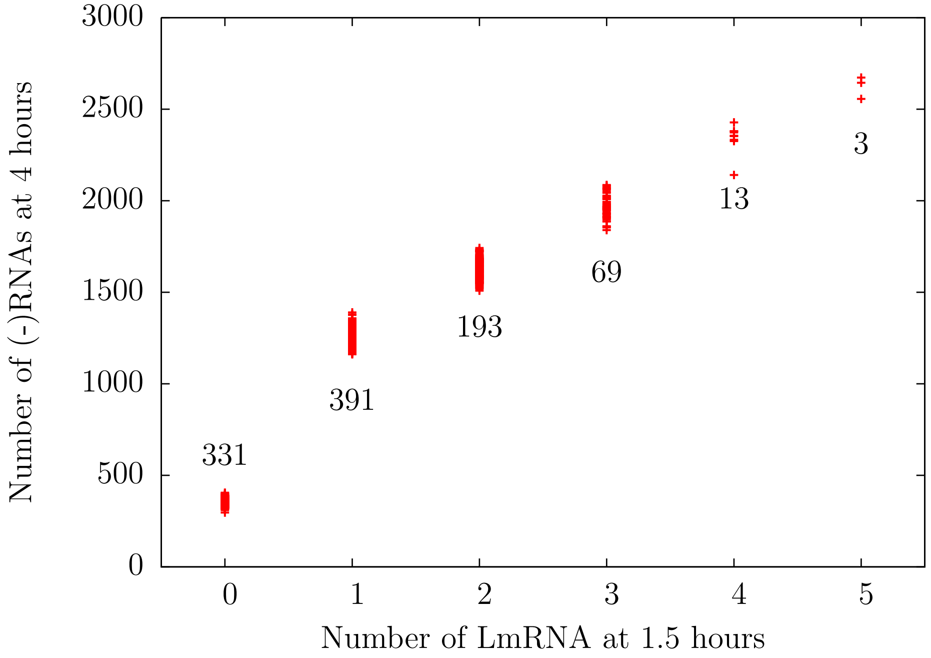

Figure 7: The ($-$)RNA genome level at 4 hr versus the L mRNA level at 1.5 hr. Numbers specify how many simulations expressed the corresponding level of L mRNA at 1.5 hr.

of every reaction.

By combining these approaches we were able to simulate essential early processes of VSV intracellular infections. These include transcription and translation of viral mRNA and proteins, respectively, regulation of transcription by intergenic attenuation, genomic encapsidation by the nucleocapsid (N) protein, and genome replication by the viral polymerase. Our simulations are consistent with experimentally observed patterns of viral mRNA expression. Specifically, the sequential expression of viral genes in the simulation, with mRNA appearing first for gene N, followed by P, M, G, and L, agree with the observed sequence inferred from in vitro transcription studies of VSV [3, 1]. Moreover, the gradient in the level of gene expression, with mRNA for gene N at the highest level, followed by P, M, G, and L, agree with the gradient in the level of expression observed in VSV infected cells [31]. More interesting, our simulations suggest that early levels of a low-abundance species, L mRNA, play an important role in determining the production of

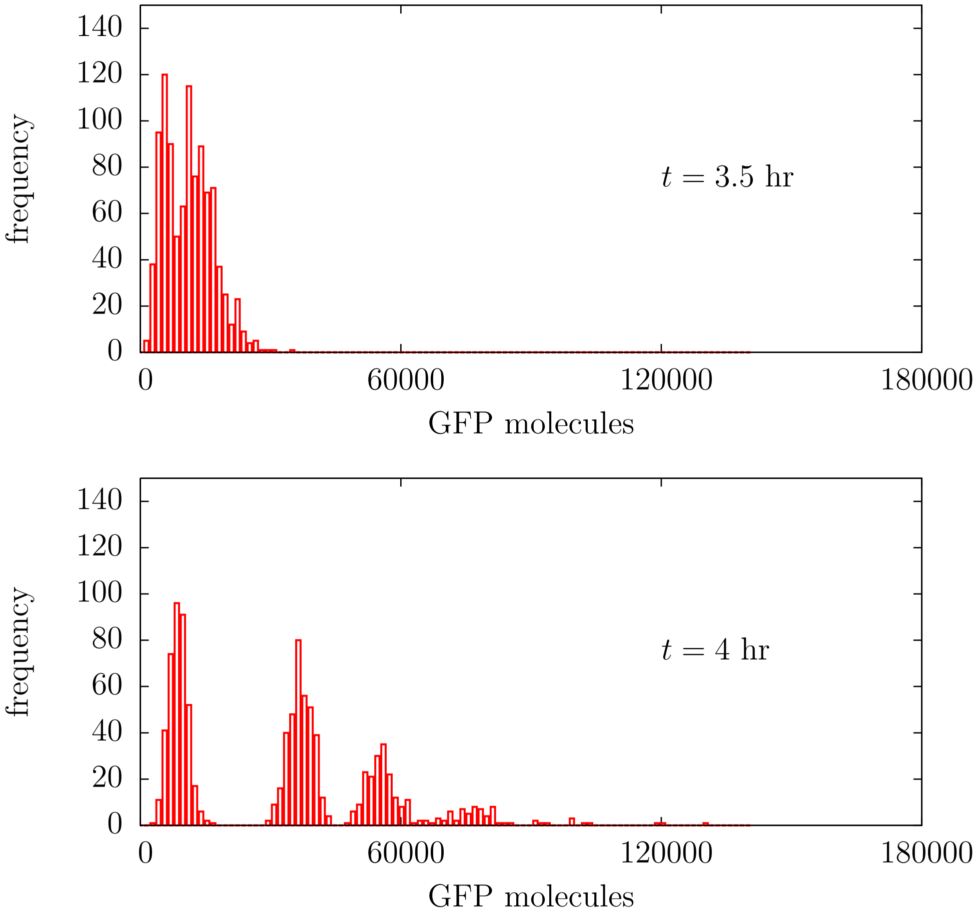

Figure 8: Distribution of the GFP at $t = 3.5$ and 4 hr for 1000 simulations.

(−)RNA genomes, viral transcripts, and ultimately viral proteins within an infected cell.

Ultimately, stochastic gene expression may contribute to the distribution of VSV yields from infected cells. Our recent experimental study of VSV production from single infected cells revealed levels of virus production that spanned 1000-fold, a range that could not be solely accounted for by genetic or environmental factors [34]. Ongoing studies will aim to quantify potential roles for stochastic processes in virus growth.

Simulating the distributions of virus yields will require that the current simulation be extended to account for a full infection cycle. Challenges await.

For example, like free N protein, levels of free L protein fluctuate rapidly, and L protein is also involved in the delayed transcription and genome replication reactions. Further analysis will need to be done to calculate a probability density for L protein, as we did for N protein. This analysis is not straightforward because of the effect of the delayed reactions initiated in the past that influence the present level of L protein. When this distribution can be calculated, the system could be updated via delayed Langevin equations for the species at high molecule levels that are involved in the delayed reactions. Further challenges await. For example, translation reactions have been implemented here without delays because their delays are considered to be short relative to delays associated with transcription reactions, and there is uncertainty in spacing between active ribosomes on a transcript. If one assumes an abundance of host ribosomes, then the ribosomal spacing can define the rate of the protein production and delays can be neglected. However, many viruses divert, inhibit, and shut down host translation resources, so one may need to relax assumptions that host ribosomes are plentiful, and delays may at some stage be appropriate to incorporate into our descriptions of protein synthesis.

Finally, taking a broader perspective, all viruses must make mRNA, synthesize proteins, and construct multi-component assemblies. We have shown here how stochastic simulation of such processes can be treated in the special case of VSV. Given the many common features that diverse virus share, we anticipate that the approaches developed here may find useful applications in simulating the intracellular growth of diverse viruses.

Acknowledgment

We are grateful for financial support from the NIH through a Phased Innovation Award (R21AI071197), the NSF through a Focused Research Group grant (DMS-0553687), and industrial members of the Texas-Wisconsin Modeling and Control Consortium. David Anderson, Gheorghe Craciun, Tom Kurtz, and Rishi Srivastava contributed helpful discussions. All simulations were performed using Octave (http://www.octave.org), which is freely distributed under the terms of the GNU General Public License.

References

[1] G. Abraham and A. K. Banerjee. Sequential transcription of the genes of vesicular stomatitis virus. *Proc. Natl. Acad. Sci. USA*, 73(5):1504–1508, 1976.

[2] A. Arkin, J. Ross, and H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149(4):1633–1648, August 1998.

[3] L. A. Ball and C. N. White. Order of transcription of genes of vesicular stomatitis virus. *Proc. Natl. Acad. Sci. USA*, 73(2):442–446, 1976.

[4] J. N. Barr, S. P. J. Whelan, and G. W. Wertz. Transcriptional control of the RNA-dependent RNA polymerase of vesicular stomatitis virus. *Biochim. Biophys. Acta, Gene Struct. Expression*, 1577(2):337–353, September 13 2002.

[5] M. Barrio, K. Burrage, A. Leier, and T. Tian. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Comput. Biol.*, 2(9):e117, September 2006.

[6] D. Bratsun, D. Volfson, L. S. Tsimring, and J. Hasty. Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. USA*, 102(41):14593–14598, October 2005.

[7] M. Delbrück. Statistical fluctuations in autocatalytic reactions. *J. Chem. Phys.*, 8:120–124, 1940.

[8] M. Delbrück. The burst size distribution in the growth of bacterial viruses (bacteriophages). *J. Bact.*, 50:131–135, 1945.

[9] W. E, D. Liu, and E. Vanden-Eijnden. Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *J. Chem. Phys.*, 123:194107, 2005.

[10] E. B. Flanagan, L. A. Ball, and G. W. Wertz. Moving the glycoprotein gene of vesicular stomatitis virus to promoter-proximal positions accelerates and enhances the protective immune response. *J. Virol.*, 74(17): 7895–7902, September 2000.

[11] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[12] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425, 1992.

[13] D. T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113 (1):297–306, 2000.

[14] J. Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *J. Chem. Phys.*, 122(18):184102, May 2005.

[15] M. Griffith, T. Courtney, J. Peccoud, and W. Sanders. Dynamic partitioning for hybrid simulation of the bistable HIV-1 transactivation network. *Bioinformatics*, 22(22):2782–2789, 2006.

[16] E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.*, 117(15):6959–6969, October 2002.

[17] L. E. Iverson and J. K. Rose. Localized attenuation and discontinuous synthesis during vesicular stomatitis virus transcription. *Cell*, 23(2): 477–484, February 1981.

[18] J. A. M. Janssen. The elimination of fast variables in complex chemical reacions. II. Mesoscopic level (reducible case). *J. Stat. Phys.*, 57(1/2): 171–185, 1989.

[19] K. Lim, T. Lang, V. Lam, and J. Yin. Model-based design of growth-attenuated viruses. *PLoS Comput. Biol.*, 2(9):e116, September 2006.

[20] E. A. Mastny, E. L. Haseltine, and J. B. Rawlings. Two classes of quasi-steady-state model reductions for stochastic kinetics. *J. Chem. Phys.*, 127(9):094106, September 2007.

[21] C. V. Rao and A. P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J. Chem. Phys.*, 118(11):4999–5010, March 2003.

[22] J. Rose and M. Whitt. Rhabdoviridae: The viruses and their replication. In D. Knipe and P. Howley, editors, *Fields Virology*, volume 1, pages 1221–1244. Lippincot Williams & Wilkins, Philadelphia, 4th edition, 2001.

[23] H. Salis and Y. Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *J. Chem. Phys.*, 122(5):054103, February 2005.

[24] H. Salis and Y. Kaznessis. An equation-free probabilistic steadystate approximation: Dynamic application to the stochastic simulation of biochemical reaction networks. *J. Chem. Phys.*, 123:214106, 2005.

[25] A. Samant and D. G. Vlachos. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: A multiscale Monte Carlo algorithm. *J. Chem. Phys.*, 123:144114, 2005.

[26] A. Samant, B. Ogunnaike, and D. Vlachos. A hybrid multiscale Monte Carlo algorithm (HyMSMC) to cope with disparity in time scales and species populations in intracellular networks. *BMC Bioinf.*, 8(1):175, May 2007.

[27] C. C. Simonsen, S. Batt-Humphries, and D. Summers. RNA synthesis of vesicular stomatitis virus-infected cells: in vivo regulation of replication. *J. Virol.*, 31(1):124–132, July 1979.

[28] A. Spirin. *Ribosome structure and protein biosysthesis.* Benjamin/Cummings Publication Company, 1986.

[29] R. Srivastava, L. You, J. Summers, and J. Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *J. Theor. Biol.*, 218: 309–321, 2002.

[30] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry.* Elsevier Science Publishers, Amsterdam, The Netherlands, 2nd edition, 1992.

[31] L. P. Villarreal, M. Breindl, and J. J. Holland. Determination of molar ratios of vesicular stomatitis virus induced RNA species in BHK21 cells. *Biochemistry*, 15(8):1663–1667, April 1976.

[32] L. S. Weinberger, J. C. Burnett, J. E. Toettcher, A. P. Arkin, and D. V. Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, 122(2): 169–182, July 2005.

[33] M. Werner. Kinetic and thermodynamic characterization of the interaction between Q beta-replicase and template RNA molecules. *Biochemistry*, 30(24):5832–5838, 1991.

[34] Y. Zhu, A. Yongky, and J. Yin. Growth of RNA virus in single cells reaveals a broad fitness distribution. To appear in *Virology*, e-publication ahead of print, December 2008.