A Mathematical Model of RNA3 Recruitment in the Replication Cycle of Brome Mosaic Virus

Tori Huffman, Kathryn Link, John Nardini, Laura Poag,
Kevin Flores, and H.T. Banks
Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC
and
Bernat Blasco, Jennifer Jungfleisch, and Juana Diez
Molecular Virology Group
Department of Experimental and Health Sciences
Univesitat Pompeu Fabra, PRBB
08003 Barcelona, Spain

October 22, 2013

Abstract

Positive-strand RNA viruses, such as the brome mosaic virus (BMV) and hepatitis C virus, utilize a replication cycle which involves the recruitment of RNA genomes from the cellular translation machinery to the viral replication complexes. Here, we coupled mathematical modeling with a statistical inverse problem methodology to better understand this crucial recruitment process. We developed a discrete-delay differential equation model that describes the production of BMV protein 1a and BMV RNA3, and the effect of protein 1a on RNA3 recruitment. We validated our model with experimental data generated in duplicate from a yeast strain that was engineered to express protein 1a and RNA3 under the control of inducible promoters. We used a statistical model comparison technique to test which biological assumptions in our model were correct. Our results suggest that protein 1a expression is governed by a nonlinear phenomenon and that a time delay is important for modeling RNA3 recruitment. We also performed an uncertainty analysis of two experimental designs and found that we could improve our data collection procedure in future experiments to increase the confidence in our parameter estimates.

Key Words: Positive-strand RNA virus, brome mosaic virus, *Saccharomyces cerevisiae*, inverse problem, uncertainty analysis, time delay.

1 Introduction

A virus is an obligate intracellular parasite forced by its limited coding capacity to employ the machinery of its host to replicate and disperse its genome. This relationship between viruses and their hosts has been studied extensively in an attempt to further understand the complex molecular processes involved in viral replication. Of particular interest are positive-strand RNA viruses [(+)RNA], accounting for over one third of all virus genera. They contain a large number of serious plant, animal, and human pathogens, including the hepatitis C virus (HCV), which is a major cause of liver disease and is estimated to have chronically infected 130-170 million people worldwide [12].

Research has revealed a number of common fundamental features in the replication processes of all (+)RNA viruses. Unlike other viral groups, they do not encapsulate viral polymerases required for replication. Thus, upon gaining entry into the cytoplasm, their genomic RNA acts as mRNA to be directly translated in order to produce the necessary viral replication factors. These replication factors then specifically recognize viral RNA and recruit it from translation into replication complexes where it serves as a template for replication. These two functions are mutually exclusive since the 5' to 3' movement of the ribosomes directly conflicts with the 3' to 5' polymerase copying [11, 16]. Therefore, a key step in the replication cycle of any (+)RNA virus is the exit of the genomic RNA from translation to replication, a process that must be highly regulated in order to allow both sufficient translation and replication.

Due to the complexity of viral infections in eukaryotes, numerous biological systems have been created to study infection in the simpler and better understood yeast *Saccharomyces cerevisiae* [10]. One such system for studying fundamental aspects of (+)RNA virus biology is the replication of the plant brome mosaic virus (BMV) [1]. The BMV genome is comprised of three genomic RNAs with 5' caps and, differing from cellular mRNAs, 3' tRNA-like ends that are aminoacylated in vivo by host enzymes [15]. RNA1 and RNA2 encode essential viral RNA replication factors protein 1a and 2a respectively [15]. RNA3 encodes a cell-to-cell movement protein and, through the production of a subgenomic RNA, the capsid protein. Both are required for systemic infection in BMV's natural host but not for replication [15]. The replication factor protein 1a plays a key role in replication. It directs itself and 2a polymerase to the endoplasmic reticulum (ER), where it induces the formation of membrane-enveloped spherules that line the replication complex [5, 18]. Moreover, it acts independently of 2a polymerase through specific sequences in BMV RNA to recruit it out of cellular translation machinery and into replication complexes within the ER [13]. Here, 2a polymerase, with the help of protein 1a, initiates BMV RNA synthesis to produce viral RNA progeny.

To date, a mathematical model has not been developed to study BMV RNA replication. Such a mathematical model could be used to test hypotheses regarding molecular features and mechanisms underlying crucial steps in (+)RNA virus replication, and to formulate new assumptions about similarities in the replication cycles of other (+)RNA viruses, such as HCV. Here, we develop a mathematical model of the recruitment phase of BMV RNA replication for this purpose. We validated and refined our mathematical model using inverse problem methods applied to protein and RNA data collected from yeast cells expressing protein 1a from non-viral mRNA and BMV RNA3. Without RNA2 expression, this biological system allows recruitment of RNA3 to the replication complex built by protein 1a but not synthesis of the viral progeny, thus isolating the recruitment phase of BMV RNA replication from the synthesis phase.

2 Mathematical Model description

Our mathematical model describes the dynamics of protein 1a and RNA3 production and recruitment. In our model, the total amount of RNA3 is divided into RNA3 that has not been recruited to a replication complex and recruited RNA3. Thus, the model accounts for the amount of protein 1a, unrecruited RNA3, and recruited RNA3; these quantities are denoted by $x(t)$, $y(t)$, and $z(t)$ in equations (1)-(3), respectively, with the corresponding compartmental model depicted in Figure 1.

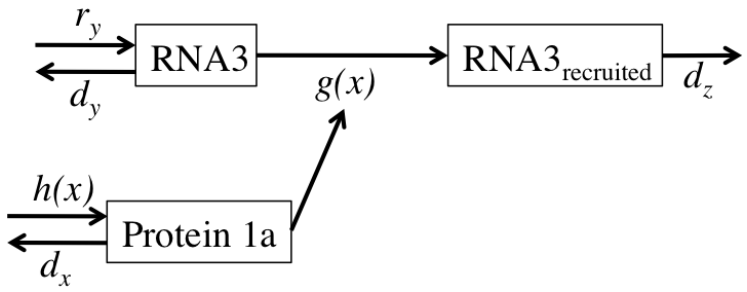

Figure 1: Model diagram for equations (1) - (3).

The model equations are given by

$$\frac{dx}{dt} = h(x) - d_x x \tag{1}$$

$$\frac{dy}{dt} = r_y - d_y y - g(x)y \tag{2}$$

$$\frac{dz}{dt} = g(x)y - d_z z \tag{3}$$

with initial conditions

$$(x(0), y(0), z(0)) = (0, y_0, 0). \tag{4}$$

The parameters r_y and d_i, $i \in \{x, y, z\}$, are production and degradation rates, respectively. For $y(t)$, d_y represents a combination of degradation and translation. The positive initial condition, y_0, for $y(t)$ and zero initial conditions for $x(t)$ and $z(t)$ in (4) are described in Section 3.1 below. The function $h(x)$ describes the production of protein 1a. We compared the performance of $h(x) = r_x$ to $h(x) = \frac{r_x}{1+Ae^{-x}}$ in fitting our experimental data in order to test whether protein 1a production is governed by linear or non-linear growth, respectively. The function $g(x)$ describes the interaction of protein 1a with RNA3 leading to recruitment of RNA3 to a replication complex. Our primary goal was to determine whether the delayed effect of protein 1a induction on RNA3 recruitment is more accurately described by a time delay or a threshold in the interaction between RNA3 and protein 1a (see Figure 2). To evaluate these different mechanisms, we tested whether a mass action function, $g(x) = mx$, a threshold function, $g(x) = \frac{mx^H}{1+Bx^H}$, or a mass action function with a time delay, $g(x) = mx(t-\tau)$, more accurately fit our experimental data.

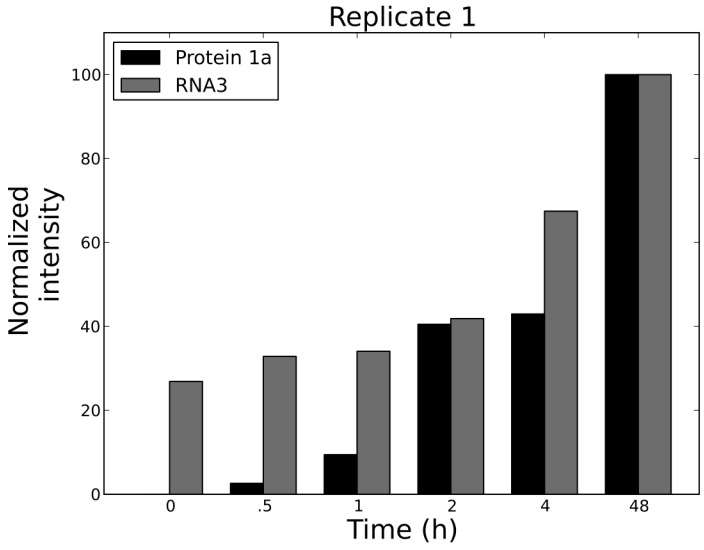

Figure 2: Data for protein 1a and RNA3 for replicate 1 (see Data and Methods for details). RNA3 is initially at steady state in our experimental setup. Once protein 1a is induced at time t = 0 hours, protein 1a increases and RNA3 increases due to the stabilizing effect of recruitment by protein 1a. However, there appears to be a delayed response between the increase in protein 1a and the increase in RNA3 for the first few hours of the experiment. Between time t = .5 hours and t = 2 hours, the fold-increase in protein 1a (15.52-fold) is much higher than RNA3 (1.27-fold). In contrast, the fold-increase in protein 1a (2.47-fold) and RNA3 (2.39-fold) is approximately the same between time t = 2 hours and t = 48 hours.

3 Data and Methods

3.1 Data sets

RNA3 and protein 1a expression were measured from yeast cells (YPH500 WT) containing two plasmids: pB3VG1-URA, expressing RNA3 under the control of a Copper promoter and pB1YT3H, expressing protein 1a under the control of a Galactose promoter. Yeast wild-type cells were transformed with the plasmids in a medium containing copper (0.5mM) and raffinose (2%). RNA3 is transcribed but protein 1a is not expressed under these conditions. Cells were allowed to grow until an optical density (OD) of approximately 0.5. At this OD, the cell growth rate is constant and RNA3 reaches a steady state where the production rate is equivalent to the degradation rate. Thus, the initial condition for RNA3 is assumed to be at a positive steady state, whereas protein 1a and recruited RNA3 are initially absent from the system. At time t = 0, Galactose (2%) was added to the medium, inducing the expression of protein 1a. Samples were collected from two biological replicates at 0 minutes, 30 minutes, 1 hour, 2 hours, 4 hours, and 48 hours post-induction.

Protein and RNA expression were quantified by western and northern blotting, respectively. Protein 1a levels were quantified using the Odyssey infrared imaging system, which measures signal intensity in K Counts/mm^2. RNA3 levels were quantified using a Phosphoimager, which measures

the intensity of photon emissions (in arbitrary units) released from the storage phosphor screen during scanning. PGK and 18S RNA were used as loading controls for protein 1a and RNA3, respectively. The expression levels (intensity) of protein 1a and RNA3 were normalized using the formula $100 * (\frac{N_i}{N_{max}})/(\frac{C_i}{C_{max}})$, where N_i and C_i are the intensities at the i-th time point, N_{max} and C_{max} are the maxima of those intensities over all time points, N is either protein 1a or RNA3, and C is either PGK or 18S. Thus, the protein 1a and RNA3 observables (see Section 3.2) are non-dimensional. Consequently, the parameters $\{A, B\}$ are also non-dimensional, and the parameters $\{r_x, d_x, r_y, d_y, m, d_z\}$ are all rates with units hour^{-1}.

3.2 Inverse problem methodology

For each mathematical model, we estimated parameters from our data using the ordinary least squares (OLS) framework with a constant error model. The observation operators were either $f(t, q) = x(t, q)$ for the protein 1a data or $f(t, q) = y(t, q) + z(t, q)$ for the RNA data. Forward simulations were run using *ode*45 or *dde*23 in Matlab; cases where we used either solver are stated within each section below. Initial conditions were fixed using the initial data point from either protein 1a or RNA3 for each replicate. Parameters were estimated separately for each replicate using the Simbiology 2012a package from Matlab for $\{r_x, d_x, A\}$ and either Simbiology or the *lsqnonlin* function for $\{r_y, d_y, m, d_z, \tau, B\}$. We note that the number of time points in our data was less than the total number of parameters in the system (1)-(3). In order to numerically implement a nonlinear regression in Matlab, the number of estimated parameters needs to be less than or equal to the number of data points. To satisfy this requirement, we first used the protein 1a data to estimate the parameters in equation (1), i.e., $\{r_x, d_x, A\}$. We then used these parameter estimates as fixed values to estimate parameters in the following sets for equations (2) and (3), where our choice of parameter set depended on the form of $g(x)$: $\{r_y, d_y, m, d_z\}$, $\{r_y, d_y, m, d_z, B\}$, or $\{r_y, d_y, m, d_z, \tau\}$. We justified this approach based on two observations. First, we are primarily interested in the mechanism of RNA3 recruitment and not on the production of protein 1a. Second, the measured concentration of protein 1a is not affected by RNA3.

3.3 Uncertainty quantification: Asymptotic theory

We calculated standard errors and confidence intervals [3, 4] in order to quantify the uncertainty in estimating each element of the parameter estimate $\hat{q}^n$ for a given model with scalar observation $f(t, q)$. To compute these values, we must first define a few other terms. Recall that the statistical model in the OLS case is of the form

$$Y_j = f(t_j, q_0) + \mathcal{E}_j,\ j = 1, 2, \ldots, n, \tag{5}$$

where $f(t_j, q_0)$ (either $x(t_j, q_0)$ or $y(t_j, q_0) + z(t_j, q_0)$) is the model observation with the hypothesized "true" parameter vector q_0, and the error terms $\mathcal{E}_j$ are independent and identically distributed (*i.i.d.*) random variables with mean $E[\mathcal{E}_j] = 0$ and constant variance var($\mathcal{E}_j$)=σ_0^2. Then the observations $\vec{Y} = \{Y_j\}$ are also *i.i.d.* with mean $E[Y_j] = f(t_j, q_0)$ and variance var(Y_j)=σ_0^2.

The n x p sensitivity matrix, where n is the number of data points and p is the number of parameters, is given by the partial derivatives of the model with respect to each parameter:

$$\chi_{jk}(q) = \frac{\partial f(t_j, q)}{\partial q_k}. \tag{6}$$

Given the data $\{y_j\}_{j=1}^n$ and the resulting parameter estimate $\hat{q}^n$, the variance σ_0^2 can be approximated by

$$\sigma_0^2 \approx \hat{\sigma^2} = \frac{1}{n-p}\sum_{j=1}^{n}[y_j - f(t_j, \hat{q}^n)]^2. \tag{7}$$

With these values, we can calculate the following approximation:

$$\Sigma_0^n \approx \hat{\Sigma}^n(\hat{q}^n) = \hat{\sigma}^2[\chi^{nT}(\hat{q}^n)\chi^n(\hat{q}^n)]^{-1}, \tag{8}$$

where $\chi^n = \sum_{j=1}^n \chi_{jk}$. This matrix is used to compute the standard errors for each element $(k = 1, 2, \ldots, p)$ of $\hat{q}^n$, given by

$$SE_k(\hat{q}^n) = \sqrt{\hat{\Sigma}_{kk}^n(\hat{q}^n)}. \tag{9}$$

The 100(1-α)% confidence intervals can be computed based on the confidence level parameters associated with the parameter estimators $q^n = q^n(\vec{Y})$

$$Prob\{q_k^n - t_{1-\alpha/2}SE_k(q^n) < q_{0k} < q_k^n + t_{1-\alpha/2}SE_k(q^n)\} = 1 - \alpha, \tag{10}$$

where α is chosen to be small (e.g., α=0.05 for 95% confidence intervals) and $t_{1-\alpha/2}$ is determined by $Prob(T \geq t_{1-\alpha/2}) = \alpha/2$, where $T \sim t^v$ for $v = n - p$ degrees of freedom. The corresponding 95% confidence intervals are then given by

$$[\hat{q}_k^n - t_{1-\alpha/2}SE_k(\hat{q}^n), \hat{q}_k^n + t_{1-\alpha/2}SE_k(\hat{q}^n)]. \tag{11}$$

Asymptotic standard errors were calculated using the Simbiology 2012a package in Matlab for Tables 2, 5, and 7.

3.4 Uncertainty quantification: Bootstrapping

Rather than using asymptotic theory to compute the standard errors and confidence intervals in parameter estimation, one can alternatively use the *bootstrapping* technique, as described in [3, 6, 8, 9, 14]. As was previously stated, we implemented the bootstrapping technique for the RNA3 delay differential equation model. To do this, an initial parameter vector estimate, $\hat{q}^n$, must first be estimated using OLS techniques. The next step is to calculate the standardized residuals, $\overline{r}_j$, of these estimates:

$$\overline{r}_j = \sqrt{\frac{n}{n-p}}(y_j - f(t_j, \hat{q}^n)),\ j = 1, ...n, \tag{12}$$

where n is again the number of data points and p is the number of parameters. We then create bootstrap sample points by sampling residuals $\overline{r}_j^m$ with replacement from $\{\overline{r}_j\}_{j=1}^n$ and adding them to the model solution:

$$y_j^m = f(t_j, \hat{q}^n) + \overline{r}_j^m,\ j = 1, ...n,\ m = 1, ..., M. \tag{13}$$

After creating $M = 1000$ simulated bootstrap data sets in this fashion, this technique is completed by conducting M inverse problems to fit the model to each of these simulated data sets and

storing the parameter estimates $\hat{q}^m$ in a matrix, Q_{BOOT}. With these values, the mean, variance, and standard errors for the parameters can be calculated using the following formulas given in [3]:

$$\hat{q}_{BOOT} = \frac{1}{M}\sum_{m=1}^{M}\hat{q}^m,$$
$$Var(\hat{q}_{BOOT}) = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{q}^m - \hat{q}_{BOOT})(\hat{q}^m - \hat{q}_{BOOT})^T,$$
$$SE_k(\hat{q}_{BOOT}) = \sqrt{Var(\hat{q}_{BOOT})_{kk}}.$$

The confidence intervals are then calculated by using equation (11).

3.5 Model comparison testing

In order to compare the effectiveness of various model components, we used a statistical model comparison test [2, 4] to test the null hypothesis, H_0, that a certain parameter is not needed to describe the system. If we can reject the null hypothesis, then we determine that the parameter in question is needed to accurately describe the system. The parameter vector q belongs to the parameter set $\mathcal{Q}$, and the restricted parameter set $\mathcal{Q}_H \subset \mathcal{Q}$ is defined for each model comparison test by fixing the parameter in question. For example, in Section 4.1, $\mathcal{Q}_H = \{r_x, d_x, 0\}$ and $\mathcal{Q} = \{r_x, d_x, A\}$, and in Section 4.3, $\mathcal{Q}_H = \{r_y, d_y, m, d_z, 0\}$ and $\mathcal{Q} = \{r_y, d_y, m, d_z, \tau\}$. Given data $(\vec{t}, \vec{y}) = (\{t_j\}, \{y_j\})$, with n data points, one defines the OLS cost to be $J_n(\vec{y}, q) = \frac{1}{n}\Sigma_{i=1}^{n}[y_j - f(t_j, q)]^2$. Then the realizations of the OLS estimators over the sets $\mathcal{Q}$ and $\mathcal{Q}_H$ are given by:

$$\hat{q}^n = \arg\min_{q\in\mathcal{Q}} J_n(\vec{y}, q) \qquad \text{and} \qquad \hat{q}^n_H = \arg\min_{q\in\mathcal{Q}_H} J_n(\vec{y}, q). \tag{14}$$

After obtaining the two values above, we calculate the following test statistics:

$$T_n(\vec{y}) = n(J_n(\vec{y}, \hat{q}^n_H) - J_n(\vec{y}, \hat{q}^n)) \qquad \text{and} \qquad U_n(\vec{y}) = \frac{T_n(\vec{y})}{J_n(\vec{y}, \hat{q}^n)}. \tag{15}$$

Note that $J_n(\vec{y}, \hat{q}^n_H)$ is greater than or equal to $J_n(\vec{y}, \hat{q}^n)$, so $T_n(\vec{y})$ and $U_n(\vec{y})$ are non-negative values. One can argue [2] that $U_n(\vec{Y})$ is asymptotic to a χ^2 distribution with $r = 1$ degrees of freedom which we use with parameters of interest (ξ, α), where α is the significance level, and ξ is the threshold corresponding to α in the $\chi^2(r)$ table. The degrees of freedom, r, is 1 in this case, since we are only eliminating one variable for each test. Once we calculate the test statistic $U_n(\vec{y})$, we find the corresponding α^*, which is the *P-value*. If this P-value $\alpha^* < \alpha$, or if the test statistic $U_n(\vec{y}) > \xi$, then we *reject* H_0 *as false* with confidence $(1 - \alpha^*)100\%$. Otherwise, we *do not reject* H_0 *as true*.

4 Results

We developed the mathematical model in equations (1)-(3) in order to test several biological hypotheses about protein 1a production and the recruitment of RNA3 by protein 1a. We tested these hypotheses using inverse problem methodology, data from two biological replicates, and statistical

model comparison tests to evaluate different forms for $h(x)$ and $g(x)$ in the model (1)-(3). Below, we comment on the consistency of our findings between the replicate data sets. We quantified the uncertainty in our parameter estimates and propose a new experimental design to reduce this uncertainty.

4.1 Comparison of a Non-linear vs. Linear model for protein 1a production

We tested whether protein 1a production could be described using a non-linear equation by comparing two models represented in equation (1) with $h(x) = \frac{r_x}{1+Ae^{-x}}$. When $A = 0$, equation (1) reduces to a standard linear model of protein production with a constant expression rate, r_x, and an exponential degradation rate, d_x. We refer to this case as the "Linear model". When $A > 0$, the constant expression rate becomes a threshold function which tends to a maximum r_x as $x \to \infty$. We refer to this case as the "Non-linear model". We note that for these models the forward simulations were run using *ode*45, the inverse problems were solved using Simbiology, and the standard errors were also computed using Simbiology.

We found that the Non-linear model resulted in a lower OLS cost than the Linear model for each replicate data set (Table 1). We note that this result is expected, since the Non-linear model is a one parameter extension of the Linear model. We used a statistical model comparison technique (see Section 3.5) in order to test whether the lower OLS cost for the Non-linear model was significant, i.e., if $A > 0$ or $A = 0$ for each replicate data set. We found that $A > 0$ for replicate 2 and that A can be taken to be zero for replicate 1.

Replicate	Linear P1a model OLSC	Non-linear P1a model OLSC	Model comparison P-value
1	40.651	38.439	0.55679
2	52.095	7.4835	$2.22 * 10^{-9}$

Table 1: Ordinary least squares costs (OLSC) and model comparison P-values for the protein 1a models. The Linear P1a model is the case where $A = 0$ and the Non-linear P1a model is the case where $A > 0$ in equation (1) with $h(x) = \frac{r_x}{1+Ae^{-x}}$. We used a statistical model comparison technique to test whether the OLSC was significantly lower for the Non-linear P1a model. The resulting model comparison P-value was significant in the second replicate, indicating that the parameter A is important for describing protein 1a production in this data set.

These results suggest that using a non-linear function to describe protein 1a production may be a correct assumption, since it improved the OLS cost in both replicates and provided a statistically significantly lower OLS cost for the second replicate. These results reflect the difference in the dynamics between replicate 1 and replicate 2. The data from replicate 2 displays an inflection point, whereas the replicate 1 data do not (Figure 3).

Taking both replicates into consideration, the Non-linear model more accurately fit the experimental data than the Linear model, since it is able to cover a wider range of biological dynamics. The parameter estimates, their standard errors, and 95% confidence intervals are presented in Table 2. The second replicate shows much lower standard errors and narrower 95% confidence intervals than the first replicate. The high standard error for the parameter A resulted in a negative lower bound for the 95% confidence interval for replicate 1. This result may reflect the high P-value from the model comparison test for replicate 1 which showed that the parameter A could be taken equal to zero (Table 1).

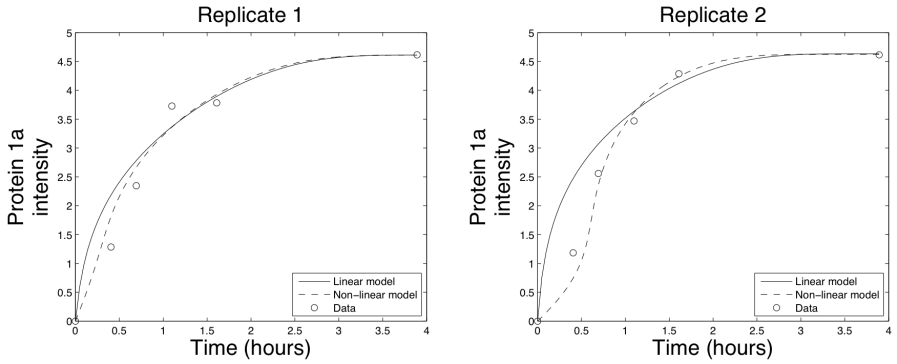

Figure 3: The protein 1a model solution (equation (1)) is plotted together with the data for two replicates. Both the abscissa and ordinate axes are plotted with the transformation $log(w+1)$.

Replicate	Parameter	Estimate	Standard Error	95% CI
1	r_x	18.245	5.419	(7.62376,28.86624)
1	d_x	0.18343	0.058	(0.06975,0.29711)
1	A	3.7577	8.907	(-13.70002,21.21542)
2	r_x	34.691	4.166	(26.52564,42.85636)
2	d_x	0.34448	.0466	(0.253144,0.435816)
2	A	22.706	6.222	(10.5148,34.8972)

Table 2: Parameter estimates, standard errors, 95% confidence intervals (CI) for two replicate data sets using the non-linear model for protein 1a (equation (1)).

4.2 Comparison of a mass action vs. threshold model for RNA3 recruitment

We tested whether RNA3 recruitment could be described using a threshold equation by comparing two models represented in equation (2) with $g(x) = \frac{mx^H}{1+Bx^H}$. When the Hill coefficient $H = 1$ and the threshold parameter $B = 0$, the function $g(x)$ represents the mass action interaction between protein 1a and RNA3. To estimate parameters in the RNA3 model, i.e., equations (2) and (3), we first fixed the parameters in (1) for each replicate using the values in Table 2. We then estimated the parameters $\{r_y, d_y, m, d_z, B\}$ by fixing $H = 1, 2, ..., 10$. We note that for these models the forward simulations were run using *ode*45 and the inverse problems were solved using *lsqnonlin* with parameter bounds $\{[0, 100], [0, 10], [0, 8], [0, 1], [0, 2]\}$ for $\{r_y, d_y, m, d_z, B\}$, respectively. We found that for $H = 1$ the parameter B was close to zero (see Table 3), suggesting that $B = 0$ in this case. When $H > 1$ we found that the OLS costs were greater in each replicate than for $H = 1$. Taken together, these findings suggest that the form of $g(x)$ with $B = 0$ and $H = 1$ is the most accurate model for these data sets, i.e., a threshold form for $g(x)$ is not an accurate assumption.

4.3 Comparison of a Delay vs. Non-delay model for RNA3 recruitment

We tested whether RNA3 recruitment could more accurately be described using a discrete time delay by comparing two models represented in equations (2) - (3) with $g(x) = mx(t-\tau)$. We refer

Replicate	OLSC	r_y	d_y	m	d_z	B	H
1	1.865	71.694	2.1907	0.046953	0.54328	4.44×10^{-14}	1
2	6.4498	49.118	4.7462	0.090914	0.33433	4.44×10^{-14}	1
1	7.2991	88.411	3.1	0.033291	0.20353	0.051429	2
2	15.447	44.535	4.8707	0.02859	0.17706	0.012453	2
1	7.4622	71.418	2.642	0.33116	0.19539	0.45694	5
2	30.441	49.104	6.0935	0.22848	0.14623	0.099123	5
1	7.4973	74.594	2.7883	0.31623	0.1951	0.43802	10
2	32.518	50.17	6.0106	1.2352	0.14448	0.55289	10

Table 3: OLS costs and parameter estimates for threshold function with $H = 1, 2, 5, 10$. Hill coefficients with $H > 1$ resulted in higher OLS costs than the case for $H = 1$. The most accurate fit when $H = 1$ resulted in $B \approx 0$.

to the cases where $\tau > 0$ and $\tau = 0$ as the "Delay model" and "Non-delay model", respectively. To estimate parameters in the RNA3 model, i.e., equations (2) and (3), we first fixed the parameters for (1) for each replicate using the parameter estimates in Table 2. We note that the forward simulations for the Non-delay models were run using *ode*45, the forward simulations for the the Delay models were run using *dde*23, and the inverse problems were solved using *lsqnonlin* with parameter bounds $\{[0, 100], [0, 10], [0, 8], [0, 1], [0, 3]\}$ for $\{r_y, d_y, m, d_z, \tau\}$, respectively. Similar to the protein 1a model calculations, extending the model from $\tau = 0$ to $\tau > 0$ lowers the OLS cost for each of the two replicate data sets. However, in contrast to the protein 1a model analysis, we found that this lower OLS cost was statistically significant in both replicates (Table 4 and Figure 4). That is, we used a statistical model comparison technique to test whether the OLSC was significantly lower for the Delay model. The resulting model comparison P-value was significant for both replicates, indicating that the time delay (τ) is important for describing RNA3 recruitment in these data sets.

Replicate	RNA3 Non-delay model OLSC	RNA3 Delay model OLSC	Model comparison P-value
1	1.8653	0.35928	5.302×10^{-7}
2	6.4499	0.66937	6.099×10^{-13}

Table 4: Ordinary least squares costs (OLSC) and model comparison P-values for the RNA3 models. The Non-delay model is the case where $\tau = 0$ and the Delay model is the case where $\tau > 0$ for equations (2) and (3) with $g(x) = mx(t - \tau)$.

The range of parameter estimates was consistent between the two replicates (Table 5) despite the difference in qualitative behavior in the second replicate, i.e., the initial decrease in RNA3 intensity (Figure 4). We postulate that this initial decrease in RNA3 intensity may have been due to a fluctuation RNA3 expression around time $t = 0$. Despite the possible heterogeneity in experimental conditions, the estimate for the time delay in our model was highly consistent between both replicates. Importantly, our parameter estimates agreed with the biological observation that recruited RNA3 is more stable than un-recruited RNA3, i.e., $d_z < d_y$. This finding is significant because this inequality was not imposed in any way by our inverse problem methodology.

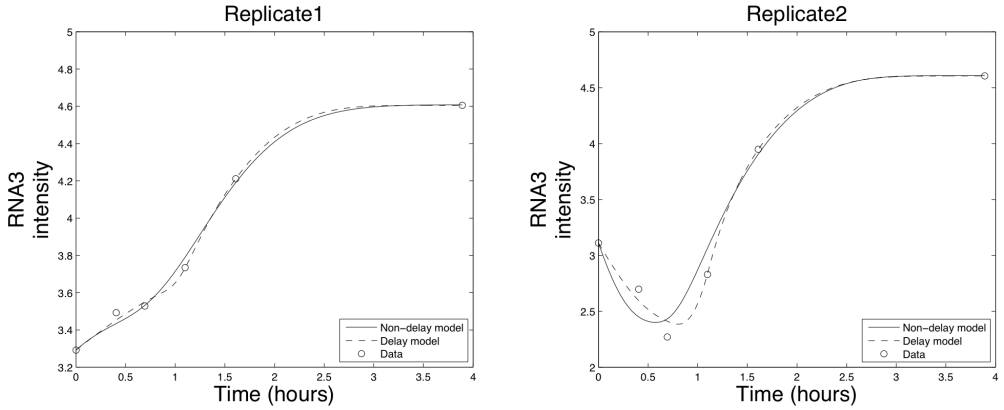

Figure 4: The RNA3 model solution (equations (2) and (3)) is plotted together with the data for two replicates. The abscissa and ordinate axes are plotted with the transformation $log(w+1)$ and $log(w)$, respectively.

4.4 Uncertainty quantification for the Delay model

We quantified the uncertainty in our parameter estimates for the Delay model by calculating standard errors. Asymptotic standard errors were calculated while simultaneously calculating parameter estimates using Simbiology. We assigned a time-dependent function for $x(t-\tau)$ in Simbiology by first solving for $x(t)$ using fixed values for $\{r_x, d_x, A\}$ listed in Table 2. We computed the asymptotic estimates of the standard errors and found the standard errors to be unreasonably high for all estimated parameters (Table 5). Since all but one of these standard errors were larger than the parameter estimates themselves, we did not compute 95% confidence intervals. We next computed the standard errors using bootstrapping. We note that for bootstrapping we ran forward simulations using *dde*23 and used *lsqnonlin* to solve each inverse problem with the same parameter bounds stated above.

Replicate	Parameter	Estimate	SE
1	r_y	31.641	53.249
1	d_y	0.7562	1.7005
1	d_z	0.3139	0.5047
1	m	0.5557	0.8791
1	τ	1.2374	1.2224
2	r_y	27.057	47.016
2	d_y	2.6264	3.3181
2	d_z	0.2681	.5763
2	m	4.7993	26.707
2	τ	1.3717	0.2275

Table 5: Parameter estimates and asymptotic standard errors for the RNA3 Delay model (equations (2) and (3) with a positive time delay τ).

Bootstrapping results were grouped by parameter and the estimates were plotted in histograms. Since some of the distributions were non-normal, there were instances where the usual computation of standard errors and confidence intervals could not be used. Details explaining the assumptions and algorithm used to calculate these quantities are found in [6, 7, 8, p. 285 - 287]. Alternatively, 95% confidence intervals were computed by eliminating the first and last 2.5% of the parameter distribution. For example, given our 1000 bootstrap sample parameter estimates, we ordered each parameter vector and selected the 26th and the 974th value to be the lower and upper bound of the confidence interval, respectively. This process was repeated for both replicates. The resulting histograms and 95% confidence intervals for the two replicates are given in Figure 5 and Table 6, respectively.

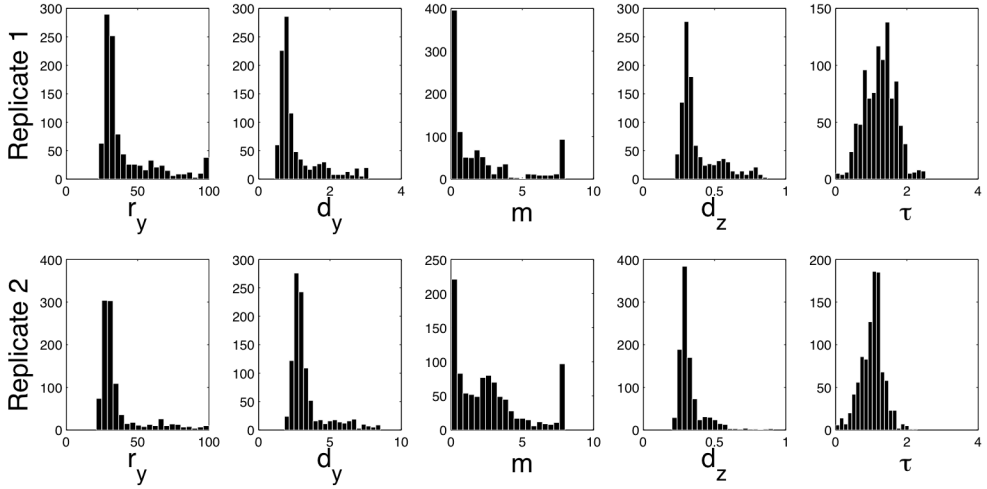

Figure 5: Boostrapping distributions for $\{r_y, d_y, m, d_z, \tau\}$ for replicates 1 and 2.

Replicate	Parameter	95% CI
1	r_y	$(24.7190, 99.9618)$
1	d_y	$(0.5523, 2.8175)$
1	m	$(0.0578, 8)$
1	d_z	$(0.2480, 0.7795)$
1	τ	$(0.4649, 2.0107)$
2	r_y	$(23.9962, 86.3599)$
2	d_y	$(2.1491, 7.3002)$
2	m	$(0.0961, 8)$
2	d_z	$(0.2369, 0.5660)$
2	τ	$(0.3217, 1.6503)$

Table 6: Confidence intervals for bootstrapping estimates from replicates 1 and 2.

For replicate 1, we note that the parameter estimates are not normally distributed, nor symmetric. For example, the distribution of parameter estimates for m accumulate along the bounds of

optimization, and a number of the r_y estimates accumulate at 100, which was the upper bound of optimization. There are a number of reasons that the distributions may not be normally distributed, including the possibility that $M = 1000$ were not enough bootstrap samples (since the Central Limit Theorem is based on the assumption that $M \longrightarrow \infty$), or that we used constrained minimization. An exception to this non-normality is the τ distribution, which appears to be approximately normal. Similar results were found for replicate 2.

4.5 Proposed experimental design for future data collections

We showed, using both asymptotic theory and bootstrapping, that the standard errors for our parameter estimates were unreasonably high for both of our replicates. Here, we propose an experimental design that is likely to result in significantly lower standard errors. Our current experimental design induces RNA3 and allows it to reach steady state prior to protein 1a induction and prior to collecting either RNA3 or protein 1a data. Under the assumption that RNA3 is in steady state at time $t = 0$, i.e. the time when protein 1a is induced, we could estimate the ratio $\frac{r_y}{d_y}$ by $y(0)$. We propose that we could also estimate d_y by removing the copper and galactose inducers at $t = 48$ hrs and then collecting RNA3 data. When there is no inducer present, $r_y = 0$ and the model for $y(t)$ would then be $\frac{dy}{dt} = -d_y y$. Once d_y is estimated, we could then estimate r_y using the formula $r_y = d_y y(0)$. Then, once we have estimates for $\{r_y, d_y\}$, we could proceed to estimate $\{m, \tau, d_z\}$ using the post-protein 1a induction data (e.g., see Figure 6). We exemplified the theoretical effect of this design on the standard errors by fixing $\{r_y, d_y\}$, re-estimating $\{m, \tau, d_z\}$, and then calculating the standard errors and 95% confidence intervals for these parameter estimates. Asymptotic standard errors were calculated and bootstrapping was run using the same methodology as stated in Section 4.4.

We first computed the asymptotic standard errors and found that the new design could theoretically reduce the range of the confidence intervals by several orders of magnitude (Table 7). We found that both m and τ had larger confidence intervals in replicate 2, indicating that the qualitatively different behavior of this replicate may have strongly influenced the uncertainty in parameter estimation. Thus, we recommend to repeat the replicate if an initial decrease in RNA3 intensity is observed.

Replicate	Parameter	Estimate	SE	95% CI
1	m	0.2343	0.0067	(0.221168,0.247432)
1	d_z	0.3474	0.0031	(0.341324,0.353476)
1	τ	1.0358	0.0014	(1.033056,1.038544)
2	m	1.5665	0.3060	(0.96624,2.16576)
2	d_z	0.2731	0.0446	(0.185684,0.360516)
2	τ	1.1497	0.0056	(1.138724,1.160676)

Table 7: Parameter estimates, standard errors, and 95% confidence intervals for the delay model of RNA3 recruitment (equations (2) and (3) with a positive time delay τ) and $\{r_y, d_y\}$ fixed using hypothesized data.

We verified the asymptotic results using bootstrapping and again found that our proposed experimental design could decrease the range of the confidence intervals substantially (Table 8). Fixing the parameters $\{r_y, d_y\}$ changed the bootstrapping distributions to be more normally distributed

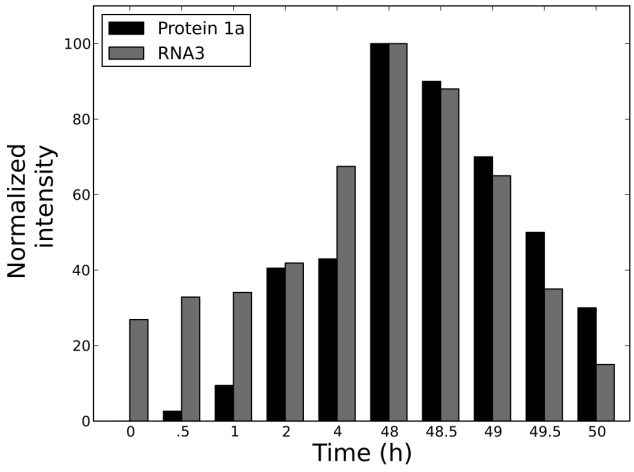

Figure 6: A proposed data collection for RNA3 and protein 1a for future experiments. The data after time $t = 0$ are the same data shown in Figure 2, i.e., $t = 0$ is the time when protein 1a expression is induced. We propose to collect data for RNA3 after removing both the protein 1a and RNA3 inducers after $t = 48$ hrs. An example of additional data collection times are shown as $t = 48.5, 49, 49.5, 50$ hrs. The data at $t > 48$ are artificial and only meant to show a trend towards degradation.

and narrow than the distributions calculated for the parameters $\{r_y, d_y, m, \tau, d_z\}$ (Figure 7). We note that the bootstrapping estimates in Table 8 do not agree with the asymptotic estimates in Table 7. This disagreement reflects the differences in the algorithm for computing parameter estimates and standard errors for each of the methods. The Simbiology 2012a package was used to simultaneously compute the parameter estimates and standard errors for the asymptotic estimates, whereas the bootstrapping estimates used the *lsqnonlin* function with the same initial guesses used in the above bootstrapping efforts. However, both the asymptotic and bootstrapping results showed that the proposed experimental design could decrease the uncertainty in our parameter estimates, regardless of the algorithm used to calculate them.

Replicate	Parameter	Bootstrap estimate	95% CI
1	m	0.2313	(0.22514,0.23958)
1	d_z	0.3459	(0.33737,0.35552)
1	τ	1.0349	(1.0313,1.041)
2	m	3.0589	(2.6531,3.5573)
2	d_z	0.2825	(0.27613,0.28945)
2	τ	1.1458	(1.1432,1.1539)

Table 8: Bootstrapped parameter estimates and 95% confidence intervals for the RNA3 model (equations (2) and (3) with a positive time delay τ) with $\{r_y, d_y\}$ fixed.

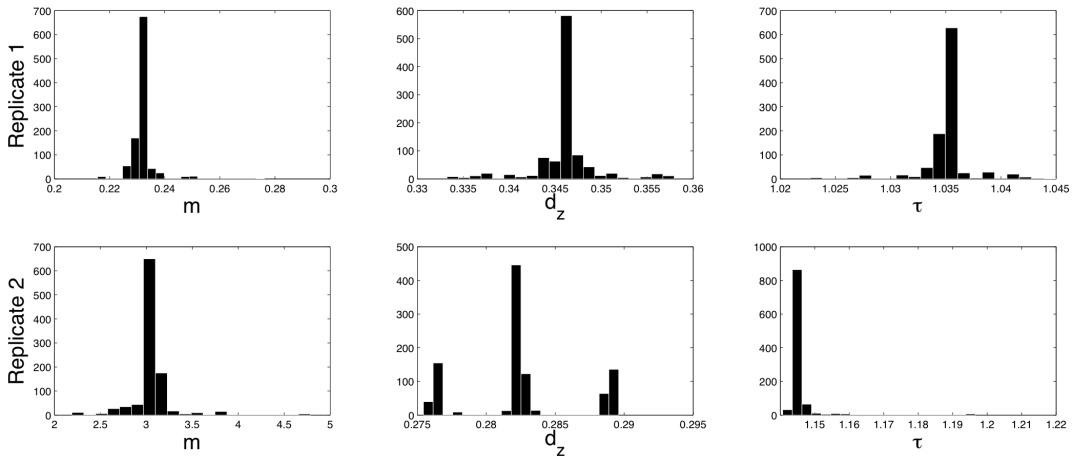

Figure 7: Bootstrapping distributions for $\{m, d_z, \tau\}$ for replicates 1 and 2 with $\{r_y, d_y\}$ fixed.

5 Discussion

Our results indicate that a mathematical model can accurately fit BMV recruitment data and that we may be able to reduce the uncertainty in parameter estimates for the model by collecting RNA3 data after removing the inducers from the experiment. Our model accurately captured the features of two biological replicates, even though the data were qualitatively different between replicates. For example, the RNA3 data for replicate 2 showed a distinct initial decrease before converging toward a steady state (Figure 4). The parameter estimates for r_y and d_y indicate that the RNA3 steady state without protein 1a induction for this replicate was lower than the initial condition. One possible biological explanation for this behavior is that the values at these time points were so low that quantification of these values are not completely accurate since they are very close to background levels. Consequently, in these conditions fluctuations are normally observed.

Our analysis of the protein 1a data suggest that a non-linear model governs the production of protein 1a in one out of the two replicates. This is not completely unexpected since transcription from a promoter does not necessarily need to follow a linear mode. Changes in physiological state of the cells that affect the the uptake of galactose or the adaptation to a new carbon source might explain the observed behavior.

Our main finding was that the delayed interaction between protein 1a and RNA3 is likely due to a time delay rather than a threshold effect in the recruitment process. We note that the nature of this interaction was not understood prior to this study. The high accuracy of the delay model also resulted in a consistent parameter range for the delay in RNA3 recruitment of around 1.3 hours. To the best of our knowledge, this recruitment time has not been measured experimentally. If no experimental technique exists for measuring this recruitment time, then the methodology outlined in the paper, namely collecting longitudinal protein 1a and RNA3 intensity data combined with inverse problem methodology, may currently be the best technique for estimating this recruitment time. Although we showed that a time delay model provided accurate fits to the data, it is still unclear what this time delay may biologically represent. For example, does the 1.3 hours correspond to a rate at which an RNA state transition occurs, intracellular transport time, or the time needed by protein 1a to induce the formation of the membrane-enveloped spherules where RNA3 will be recruited? If it is any of these cases, then are there any other observable intermediate states or interacting molecules, e.g., host factors [17], that remain to be measured? We could explore these questions in future models by adding intermediate transition states between y and z and using model comparison techniques, e.g., Akaike information criteria, to test whether such models are more accurate than a time delay model.

The purpose of our current work was to create a biological model and corresponding mathematical model of RNA recruitment within the BMV replication cycle. Thus, we collected data from a yeast system expressing protein 1a and RNA3 alone. In order to more fully understand the dynamics of RNA replication, we would need to collect data from a yeast system that also expresses either RNA2 or protein 2a, since protein 2a is necessary for replication once RNA has reached the replication complex. Such data could be used in an iterative modeling effort similar to this one, in which we can establish an accurate mathematical model of the RNA replication system as depicted in Figure 8. Solid grey lines are processes involved in virion assembly and encapsulation, whereas all other lines are involved in RNA replication. In the BMV-yeast system, RNA1, RNA2, and RNA3 can be expressed to mimic the introduction of these RNAs into a cellular system by a virus. Each RNA can either be translated into its respective protein product, or transported to an RNA replication center. The corresponding RNAs located in replication centers are denoted by

the "rep" subscript. Protein 1a is needed for the recruitment of RNAs to replication centers and protein 2a is needed for RNA replication. RNA3 expresses, through a subgenomic RNA, the capsid protein that is essential for encapsidation of the virion. The capsid protein and all viral RNAs are used to assemble a complete virion. Our current modeling effort incorporates only a portion of RNA recruitment within the full viral replication cycle, i.e., protein 1a, RNA3, and RNA3$_{rep}$. The next system we propose to investigate includes RNA recruitment and replication processes, i.e., protein 1a, protein 2a, RNA3, and RNA3$_{rep}$. We propose that this system would be the next logical scenario to analyze, since it incorporates more variables than our current model and fewer variables than a full model of the BMV replication cycle.

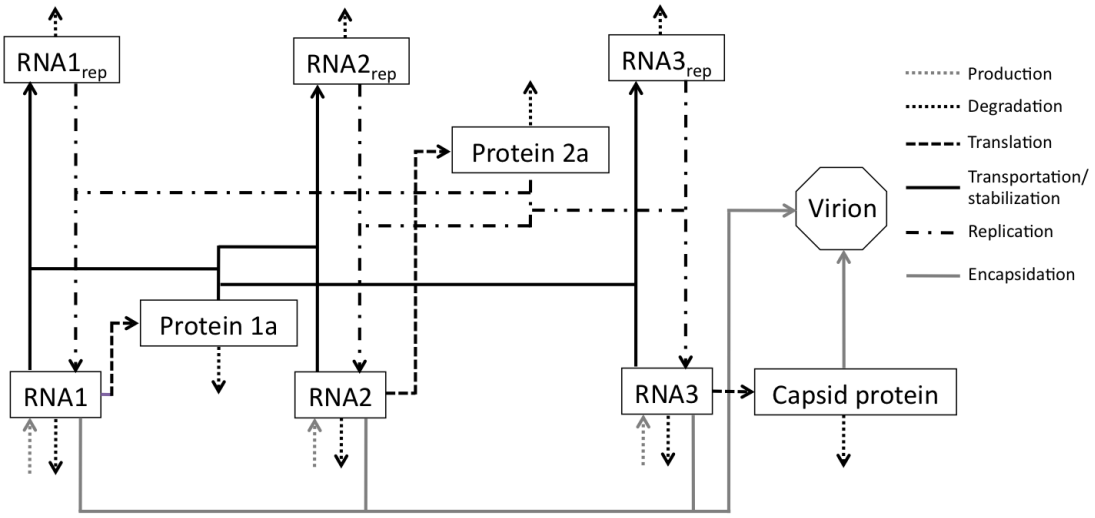

Figure 8: A model diagram of the BMV replication cycle.

6 Acknowledgements

This research was supported in part by grant number NIAID R01AI071915-09 from the National Institute of Allergy and Infectious Diseases, in part by the Undergraduate Biomathematics grant number NSF DBI-1129214 from the National Science Foundation and in part by a grant from the Spanish Ministerio de Ciencia e Innovaciòn (BFU2010-20083). The authors are grateful to Dr. W. Clay Thompson for helpful suggestions in the earlier stages of their investigations.

References

[1] I. Alves-Rodriques, R. P. Galao, A. Meyerhans and J. Diez, *Saccharomyces cerevisiae*: a useful model host to study fundamental biology of viral replication, *Virus Research*, **120** (2006), 49–56.

[2] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, *Journal of Mathematical Biology*, **28** (1990), 501–527.

[3] H.T. Banks, K. Holm and D. Robbins, Standard error computations for uncertainty quantification in inverse problems: asymptotic theory vs. bootstrapping, CRSC Technical Report CRSC-TR09-13, NCSU, May 2010; *Mathematical and Computer Modeling*, **52** (2010), 1610–1625.

[4] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, New York (2009).

[5] J.A. den Boon and P. Ahlquist, Organelle-like membrane compartmentalization of positive-strand RNA virus replication factories, *Annual Review of Microbiology*, **64** (2010), 241–256.

[6] R.J. Carrol and D. Ruppert, *Transformation and Weighting in Regression,* Chapman & Hall, New York (1988).

[7] R.J. Carroll, C.F.J. Wu and D. Ruppert, The effect of estimating weights in weighted least squares, *Journal of the American Statistical Association*, **83** (1988), 1045–1054.

[8] M. Davidian, *Nonlinear Models for Univariate and Multivariate Response,* ST 762 Lecture Notes, Chapters 9 and 11, 2007; http://www4.stat.ncsu.edu/∼davidian/course.html

[9] M.Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data,* Chapman & Hall, London (1998).

[10] R. Galao, N. Scheller, I. Alves-Rodrigues, T. Breinig, A. Meyerhans and J. Diez, Saccharomyces cerevisiae: a versatile eukaryotic system in virology. *Microbial Cell Factories*, **6** (2007), 32.

[11] A. Gamarnik and R. Andino, Switch from translation to RNA replication in a positive-strand RNA virus. *Genes & Development*, **12** (1998), pp. 2293–2304.

[12] L. Gravitz, Introduction: A smouldering public-health crisis, *Nature*, **474** (2011), pp. S2–S4.

[13] M. Janda and P. Ahlquist, Brome mosaic virus RNA replication protein 1a dramatically increases in vivo stability but not translation of viral genomic RNA3, *Proceedings of the National Academy of Sciences*, **95** (1998), 2227–2232.

[14] N. Matloff, R. Rose and R. Tai, A comparison of two methods for estimating optimal weights in regression analysis, *Journal of Statistical Computation and Simulation*, **19** (1984), 265–274

[15] A. Noueiry and P. Ahlquist, Brome mosaic virus RNA replication: revealing the role of the host in RNA virus replication, *Annual Review of Phytopathology*, **41** (2003), 77–98.

[16] N. Scheller and J. Diez, RNA viruses hijack the mRNA decay machinery to multiply, *Cell Cycle*, **8** (2009), 4013–4014.

[17] N. Scheller, L. B. Mina, R. P. Galao, A. Chari, M. Gimenez-Barcons, A. Noueiry, U. Fischer, A. Meyerhans and J. Diez, Translation and replication of hepatitis C virus genomic RNA depends on ancient cellular proteins that control mRNA fates, *Proceedings of the National Academy of Sciences*, **106** (2009), 13517–13722.

[18] M. Schwartz, J. Chen, M. Janda, M. Sullivan, J.A. den Boon and P. Ahlquist, A positive-strand RNA virus replication complex parallels form and function of retrovirus capsids, *Molecular Cell*, **9** (2002), 505–514.