MODELING OF CELL CYCLE CHECKPOINTS WITH APPLICATIONS TO THE ANALYSIS OF INTERMITOTIC TIME DATA

---

by

Zachary Jones

A Dissertation

Presented to the Faculty of the Department of Computational Sciences

Middle Tennessee State University

December 2016

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in Computational Sciences

Dissertation Committee:

Dr. Rachel Leander, Chair

Dr. Zachariah Sinkala

Dr. Joshua Phillips

Dr. John Wallin

This dissertation is dedicated to my wonderful mother whose undying love and support helped shape me into the person I am today. I could not be more proud of her display of strength and determination over the last few years while fighting her sickness.

ACKNOWLEDGMENTS

I would like to express my appreciation to Dr. Rachel Leander for her guidance during the last few years of my study. Her expertise, patience, and guidance have helped push me over the edge to completing this work. Thank you for the countless hours meeting with me to make this possible. I would also like to thank Dr. Zachariah Sinkala for his guidance and fatherly presence, especially during my first few years as a graduate student. You helped me adjust to a new environment and pushed me to do my best work. I am also indebted to my other committee members, Dr. Joshua Phillips and Dr. John Wallin. Thank you both for your constructive criticism and suggestions over the last several months.

I am grateful for the financial support from the Computational Sciences program, the Department of Mathematical Sciences, and the Graduate School. All three allowed me to venture outside of MTSU to conferences, summer workshops, and other opportunities which allowed me to grow as a young scientist. Finally, thank you to all the incredible educators I've encountered over the last two decades. Collectively you all taught me how to think critically, write effectively, and speak confidently - tools that I have used to the best of my ability over the last six years as a graduate student.

ABSTRACT

The mammalian cell can be thought of as an information processing unit, much like a computer. External stimuli initiate very complex networks of interacting proteins. When activated, these different signaling networks result in varying cell fate decisions such as cell growth, division, differentiation, and cell death. These signaling networks are subject to randomness, so that cell fate decisions are variable. For example, in a population of homogeneous cells, the time between two successive mitotic events (cell divisons), or intermitotic time (IMT), is subject to considerable, seemingly random, variation. To determine the potential sources of variability in IMT, we first use an existing model of the Rb-E2F network, which controls cell cycle entry at the restriction point. A network perturbation analysis is performed on the model to determine which part or parts of the network contribute to temporal variability in cell cycle entry. Network perturbation reveals that regulation of the *Rb* node contributes to variability in IMT. This analysis is complemented by the development and application of numerical methods for the analysis of IMT data. Specifically, we apply multi-part stochastic models to the study of IMT data, develop and test procedures for performing maximum likelihood estimation of model parameters, explain how the random variables associated with the model can be linked to intracellular protein concentrations, and analyze IMT variability within our model framework. Model selection theory is then used to determine which model performs best from a set of candidate models. Our results show that the cell cycle is best conceptualized as a two-part stochastic process. Collectively, the presented approaches provide a greater understanding of how mammalian cells process information and the noise sources involved in temporal variability in cell cycle entry.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Information theory is a branch of mathematics focused on quantifying information, its storage, and communication. Claude Shannon's 1948 paper, "A Mathematical Theory of Communication," is considered the birth of information theory. In it, Shannon detailed a general communication system consisting of several parts: an information source, a transmitter, a channel, a receiver, a destination, and a noise source [10]. The system is initiated with a message produced by the information source which is then transmitted into a signal to be moved over a channel. The signal moves through the channel until it meets the receiver where the signal is then reconstructed back into the message to be delivered to its destination. Throughout this process, a noise source acts on the signal as it travels through the system, resulting in variability in what final message is delivered to the destination [27]. In fact, "the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point [10]." This problem, Shannon explained, is one of disorder and uncertainty.

Information theory has many applications in areas of science where disorder and uncertainty in the flow of information have challenged scientists and researchers for decades. Cell biology is certainly one of those fields where randomness plays an important role in cell fate decisions [23]. We can think of the mammalian cell very much like a general communication system that Shannon detailed. Cells respond to an external stimuli (message) by secreting proteins (signal) which travel through the extracellular medium and are received by external receptors (first receiver) located on the cell surface. Through the receptor, the signal propagates to the cell's interior where it is processed by complex protein networks (receivers). This message is eventually delivered to the nucleus (destination). External stimuli initiate many dif-

ferent signaling pathways that result in varying cell fate decisions such as cell growth, division, differentiation, and cell death [11].

The cell cycle is typically divided into four phases: G1, S, G2, and M (Figure 1). Mitosis (M) is the process by which a cell splits into two daughter cells. The M phase is followed by the first growth (G1) phase. During the first part of G1, cells grow, respond to mitotic signaling, and decide if they are ready to traverse the restriction point (R-point) into the DNA synthesis (or S) phase. During this phase, the cell's genetic material doubles. If the cell doesn't receive mitotic signaling during early G1, it will revert back to a resting (G0) phase until it is met with external signals. After S phase, the cell enters a second growth (G2) phase. During G2, the cell checks itself for DNA damage, and it also ensures that the necessary proteins needed for M phase are present [22].

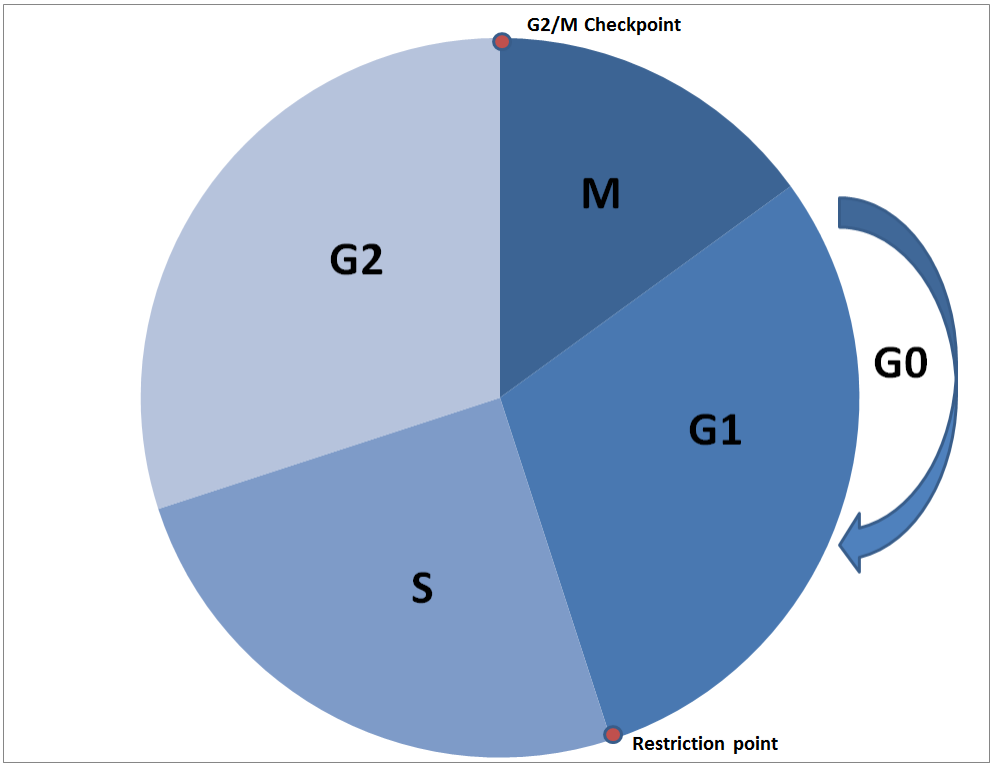


Figure 1: The mammalian cell cycle. The cell progresses orderly through G1, S, G2, and M phases. If growth factors are not present, the cell reverts back to the resting G0 phase.

The cell cycle is tightly controlled to ensure that each phase proceeds normally before entering the next phase. Cell cycle 'checkpoints' check if the conditions for progression are met at critical points during the cycle [16]. Two well studied checkpoints are the R-point and the G2/M checkpoint. These checkpoints are controlled by protein networks, which have been studied extensively over the last several decades. The R-point is a critical time point in the G1 phase that controls whether the cell enters the cell cycle or not [7, 4]. The R-point separates the G1 phase into two parts: G1-pm and G1-ps [4]. G1-pm is the post-mitotic interval and G1-ps is the pre-synthesis interval. During G1-pm, if cells are deprived of growth signals, the cell will revert back

to G0 phase until it is re-stimulated back into G1, whereas cells in G1-ps complete the cell cycle in a growth factor independent fashion [4]. The G2/M checkpoint allows the cells to repair any damaged DNA that may have occurred during S phase before proceeding to mitosis [12].

Despite these checkpoints, the time it takes for a cell to divide is subject to considerable, seemingly random, variation. Experimental data shows that across multiple cell lines and under various conditions, the time between two successive mitotic events (cell divisions), or intermitotic time (IMT), is heterogeneous. This leads us to ask the questions: Why do isogenic cells (same genotype) process information differently, and what is the noise source that is driving the variability in division time for cells? This work attempts to answer these questions in two ways. First, an existing model of the Rb-E2F network, which controls the R-point, is used to identify potential sources of temporal variability in cell cycle entry, and secondly, we develop and apply numerical methods for the analysis of single-cell IMT data with multi-part stochastic models.

2. SOURCES OF VARIABILITY IN THE Rb-E2F NETWORK

2.1. Introduction

The R-point is regulated by the Rb-E2F network. Retinoblastoma (Rb) is a negative regulator of cell cycle progression [32, 31]. The E2F family of proteins is important for regulating the transcription of many promitotic genes and includes the promitotic transcription factors E2F1-3 [31]. Post mitosis and in quiescent cells, hypophosphorylated Rb and E2F are bound in complex. Rb-binding renders E2F inactive. Moreover, the Rb-E2F complex is proposed to actively inhibit the transcription of some promitotic genes, including that of cyclin E [32]. In response to mitogenic signaling, traversal of the R-point occurs in the following manner. First, the transcription factor Myc is activated [9, 31]. Active Myc positively regulates the transcription of cyclin-dependent kinases (CDKs) and E2F [6, 21, 15, 34]. The cyclin/CDK complexes act to phosphorylate Rb [13]. Rb phosphorylation neutralizes inhibitory Rb complexes and activates the E2F transcription factors [31, 32].

Because the Rb-E2F network is frequently deregulated in cancer [32], it has been the subject of considerable research. The network has been shown to function as a bistable switch [15]. That is, E2F is activated in an all-or-none fashion as growth signaling increases, and once active, it remains active when growth signaling is diminished. Moreover, in [14], the Rb-E2F network was distilled into a robust, minimal model that reproduces the resettable bistability that Yao et al observed in their experiments (Figure 2). Specifically, a simple model consisting of only three variables (nodes): MD, RP, and EE, was proposed. Each of these nodes represent a collection of proteins that possess similar function in the full Rb-E2F network. The MD node represents Myc and its associated proteins (e.g. Ras, p15, p16, cyclin D/cdk4-6), the RP node represents Rb and other tumor suppressor proteins (e.g. p130, p107,

p27), and the EE node represents E2F and its associated proteins (e.g. E2F1-3, cyclin E/cdk2, cdc25A). These nodes are connected via regulatory links (1, 3, 5, 6, 7) which either activate or inhibit the other nodes (Figure 2). Link 1 represents the activation of Myc in response to mitogenic signaling, link 3 represents inactivation (phosphorylation) of Rb by Myc induced cyclin D/CDK4-6 complex, link 5 represents inactivation (phosphorylation) of Rb by cyclin E/CDK2 complex, link 6 represents the inactivation of E2F when Rb and E2F are bound in complex, and link 7 represents activation of E2F by Myc. These links are modeled using Michaelis-Menten kinetics.

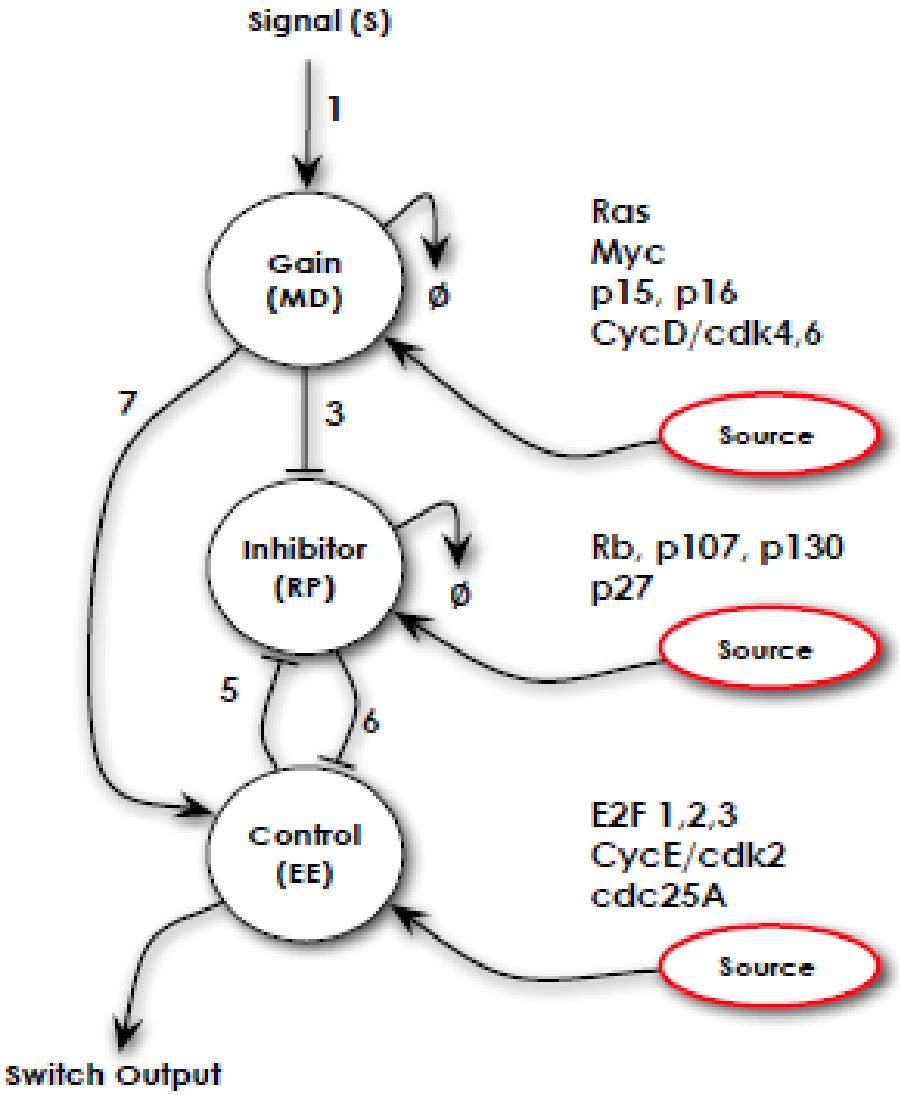


Figure 2: Rb-E2F network with source terms, adapted from [14]

In this work, we seek to identify the sources of IMT variability within the Rb-E2F network. Specifically, we use the robust minimal model to investigate how the time that a cell spends in G1 phase before traversing the R-point varies with the level of

proteins in the Rb-E2F pathway.

2.2. Methods

In this work, we adapt the minimal model to include additional source terms for all three nodes (Figure 2). These source terms represent an external perturbation to the Rb-E2F network. The adapted model is a system of ordinary differential equations (ODEs) (Equations 1-3).

$$\frac{d[MD]}{dt} = \frac{1}{\tau_{MD}}\left(\frac{[S]^{n1}}{K_1^{n1}+[S]^{n1}} - [MD]\right) + source \quad (1)$$

$$\frac{d[RP]}{dt} = \frac{1}{\tau_{RP}}\left(\frac{K_3^{n3}}{K_3^{n3}+[MD]^{n3}}\frac{K_5^{n5}}{K_5^{n5}+[EE]^{n5}} - [RP]\right) + source \quad (2)$$

$$\frac{d[EE]}{dt} = \frac{1}{\tau_{EE}}\left(\frac{K_6^{n6}}{K_6^{n6}+[RP]^{n6}}\frac{[MD]^{n7}}{K_7^{n7}+[MD]^{n7}} - [EE]\right) + source \quad (3)$$

The system of ODEs representing the minimal Rb-E2F network contains 17 parameters and 4 regulatory links. Parameter ranges, with the exception of the parameters representing the new source terms, are taken from [14]; K_i, τ_{MD}, τ_{RP}, and τ_{EE} are selected such that E2F would switch to the active state (K_i = [0.9597, 0.5517, 0.1472, 0.1578, 0.2649, 0.8423, 0.2617, 0.8161, 0.2511, 0.9300], τ_{MD} = 7.1297, τ_{RP} = 4.0926, and τ_{EE} = 5.1715). The parameters n^i (Hill exponents) are set to 1 for simplicity. The initial conditions for MD, RP, and EE are taken as 0.01, 11.0, and 0.01 respectively, which corresponds to the quiescent or resting state. The model is implemented into FORTRAN 90. For each of the experiments described below, we run 10^4 simulations (one simulation = one cell) in which the ODE system is solved using a fourth-order Runge-Kutta with 10^4 time steps over the interval [0 700].

To determine the source of heterogeneity in time to R-point passage (R time),

all three source terms are varied. The model is simulated to investigate the effects of three different perturbations: MD up regulation, RP up regulation, and EE up regulation. For each simulation, the source parameter of interest is sampled from a log normal distribution with $\mu = 0.5$ $\sigma = 0.01$ and a normal distribution with $\mu = 0.5$ $\sigma = 0.01$, for comparison. For each choice of the source parameter, we let the system equilibrate over 150 time units with $S = 0.01$ (low signal strength). We then stimulate the system with a pulse of signal $S = 5.0$ for 200 time units and monitor the dynamics of EE. To determine the time step at which EE is switched to the active state, we use the following criterion: $[EE](t) - [EE](0) \geq 0.1$ where $[EE](0)$ is the equilibrium level. Specifically, as in [14], when this condition is satisfied, we say that EE is active and the cell has traversed the R-point. In this way we generate distribution of R times which we plotted in the statistical software package R.

For each distribution and node, we produced plots showing the corresponding distribution of R time (gray). For consistency with our experiments, which include approximately 100 cells, we also generated R time distributions by sampling the total data (black). Using R, we fit each distribution to the exponentially modified Gaussian model (EMG, in red) using the built in Kolmogorov-Smirnov test (Figures 3-5). The EMG model was previously used to fit distributions of intermitotic times [1]. Sensitivity (source versus the rate of change of R time) plots are also given side by side for comparison (Figure 6).

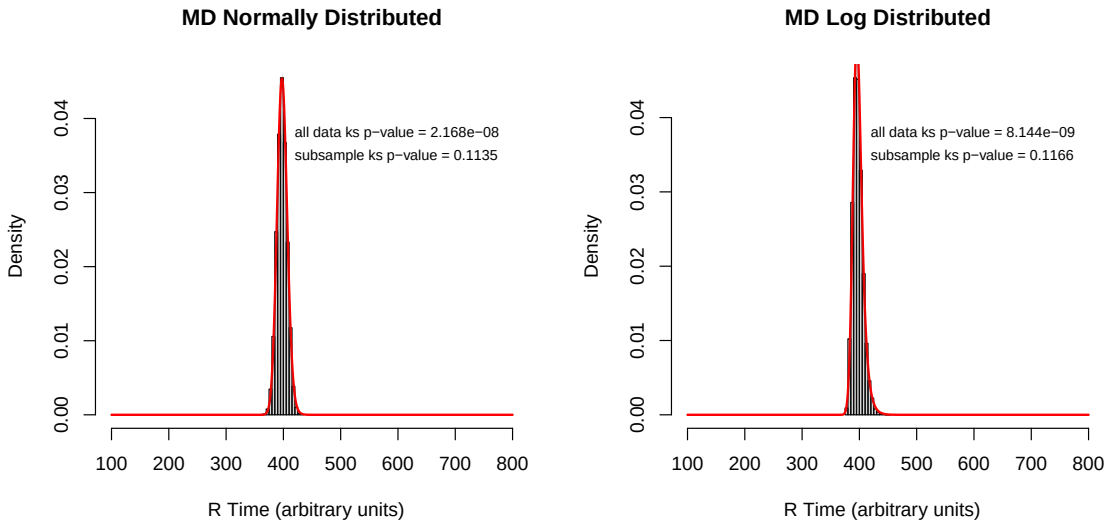


Figure 3: R time under MD regulation

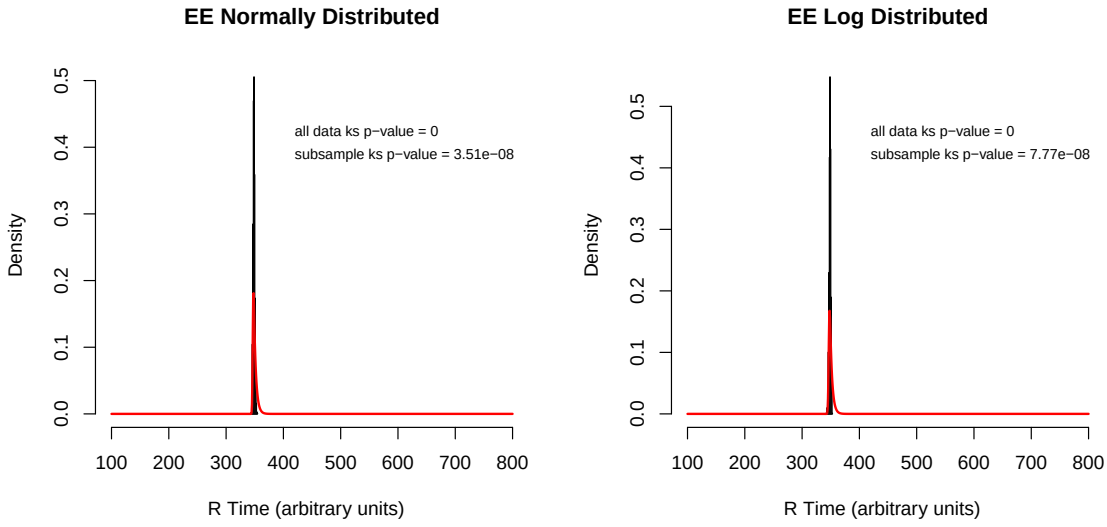


Figure 4: R time under EE regulation

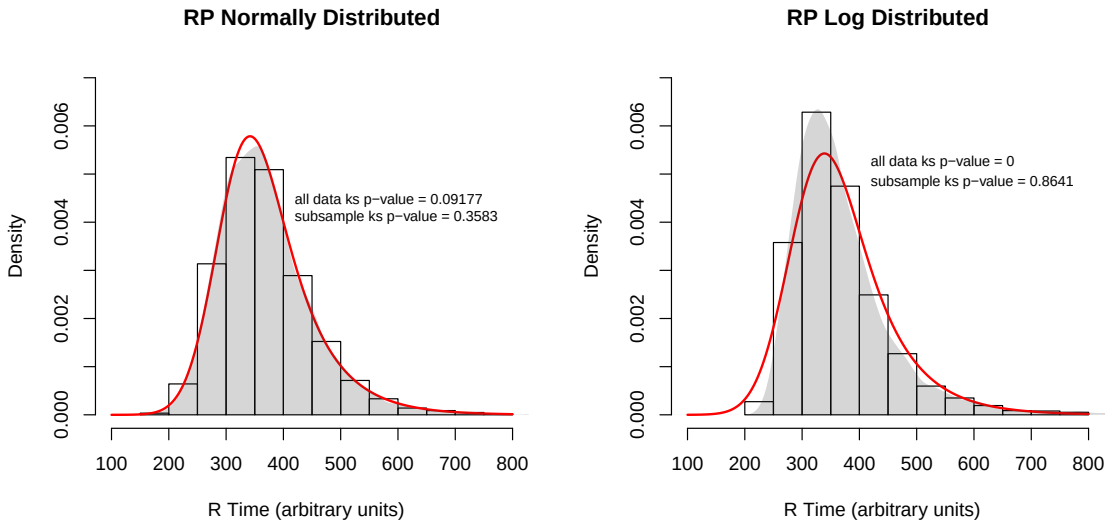


Figure 5: R time under RP regulation

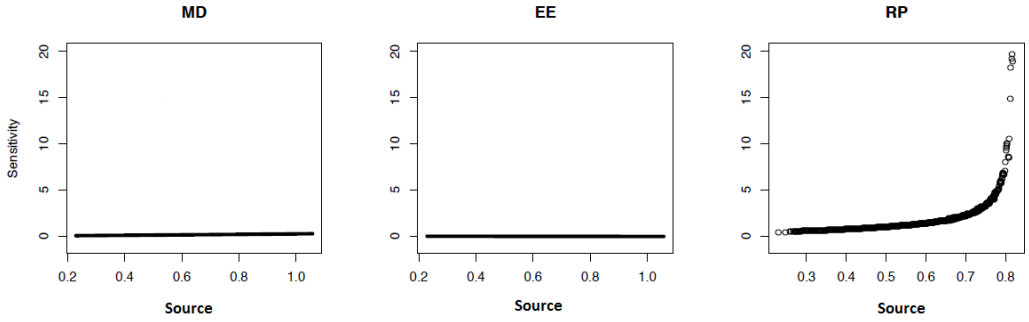


Figure 6: R time sensitivity (source versus the rate of change of R time)

2.3. Results and discussion

Our simulation results show that according to this model, RP is a likely source of heterogeneity in R time, while MD or EE do not significantly alter R time. This indicates that, in the minimal network and under this criterion for R-point passage, the level of RP determines the response to stimulation. Indeed, there is very little

variance in the R-time distributions for regulated MD and EE (Figures 3 and 4). Looking at the network structure, and considering our criterion for activation, we speculate that, in the absence of stimulation, an increase in the level of RP does not significantly alter the steady state level of EE. That is, the RP node may be specifically important for determining the network's response to stimulation. Our simulations also reveal that when the distribution of the source parameter is log normal, the distribution of R time is sharply peaked, and this peak is not captured by the EMG distribution; the same peak is present in our experimental data and is observed in other data sets as well.

While this network perturbation analysis allows us to suggest that regulation of the RP node is the main driver in temporal variability in cell cycle entry, this work does not provide methods for the analysis of IMT data nor does it explain the distinctive shape of IMT distributions. The next sections discuss a top-down stochastic model of the cell cycle to complement this previously described bottom up approach, which can aid in the description of cell division at the population level.

3. APPLYING MULTI-PART STOCHASTIC MODELS TO THE ANALYSIS OF SINGLE-CELL IMT DATA

3.1. Introduction

The time it takes a cell to divide, or intermitotic time (IMT), is subject to considerable, seemingly random, variation. Although the sources and implications of this variability are topics of considerable research [33, 1, 23, 20, 3], the form of the IMT distribution is yet to be determined. Research into the form of the IMT distribution is of interest because it has the potential to provide new insights into the biological mechanisms that govern cellular proliferation. Moreover, it can aid in the description of proliferation at the population level because estimates of population level parameters are sensitive to the form of the underlying distribution [5].

In order to explain the distribution of intermitotic times in terms of biological mechanisms, we have developed a model of a cell cycle "part" as a simple stochastic process. From this model, it follows that the time spent in a single part of the cell cycle follows an inverse Gaussian distribution, and that the total IMT is a convolution of one or more inverse Gaussian distributions. In [28] we determined that a model of the cell cycle as a two-part stochastic process is superior to several alternative models, including an EMG model, at describing individual data from MCF cells under diverse experimental conditions. Because the cell cycle is complex it is reasonable to ask if a more complex three-part model of the cell cycle would provide an even better description of single-cell data. In this manuscript, we develop a numerical method for fitting two- and three-part stochastic models to single-cell data and compare the descriptive abilities of the two- and three-part models. Our results suggest that the cell cycle is best represented as a two-part stochastic process, in which one part is short and highly variable in length. We propose that this part of the cell cycle may

be identified with G1 presynthesis, and show how the abstract random variable of the stochastic model can be linked to an underlying molecular network.

3.2. Data collection and processing

Data were collected at the Quaranta Lab at the Vanderbilt Ingram Cancer Center from benign mammary epithelial cells MCF10A (hereafter simply MCF), and MCF10A-AT1 cells (hereafter simply AT1). The AT1 cells are derived from MCF cells and have been engineered to express a constitutively active form of the protein kinase Ras, which is relevant to the G1/S transition checkpoint in that it makes cells commit to cell cycle progression more frequently. All cell lines were engineered to express a histone H2B/monomeric red fluorescent protein (H2B-mRFP) fusion protein using lentivirus-mediated transduction [30]. The fluorescent single-cell clones were compared to the parental population using traditional proliferation measurements in order to ensure that the engineered cells are representative of the parent cells in terms of proliferation rates. Extended temporally-resolved automated microscopy was used to image cells proliferating in culture. A temperature- and CO_2 controlled, automated, spinning-disk confocal microscope, the BD Pathway 855 (BD Biosciences, Rockville, MD) was used for cell imaging. In order to determine the time between mitotic events, images were acquired every 6-12 minutes.

The nuclei in the images were enumerated using the freely available ImageJ program (http://rsb.info.nih.gon/ij/). Manual image analysis was used to detect cell divisions as in [29, 30]. Individual cells were tracked through a series of images, and the number of frames between two successive mitoses was used to determine the IMT. Each mitosis was associated with a generation number, an absolute time, and an IMT. The MCF cells were also imaged under a variety of conditions, including treatment with erlotinib, which inhibits signaling through the epidermal growth factor receptor,

treatment with cycloheximide (CHX) which blocks protein synthesis by inhibiting protein translation, and a control, dimethyl sulfoxide (DMSO).

The data from the MCF and AT1 cell lines were processed as follows. Files (.csv) containing the birth times and intermitotic times were uploaded into MATLAB. Prior to data fitting, the partial correlation coefficients (MATLAB, Spearman) between birth-time, generation, and IMT were calculated. There is not a significant correlation between IMT and generation in any of the data. The correlation between IMT and birth time in the data for the untreated MCF cell data was $R = -.1033$, $p = .18125$, which suggests that during the course of the experiment, cell growth was approximately steady. In the AT1 cell data, birth time and IMT are significantly correlated ($R = .30963$, $p = 4.1218 \times 10^{-10}$). This correlation is likely the result of crowding. In order to minimize bias due to the end of experiment and crowding, the data was segregated so that only the IMTs of cells that were (i) born before the last time at which the birth-time is not significantly correlated with IMT as measured by the Spearman correlation coefficient ($p = .01$) and, (ii) according to their birth-time, had at least a 98% probability of dividing for the MCF cell line and a 99% probability of dividing for the AT1 cell line, were included in subsequent analysis. The data for treated MCF cells came to us preprocessed.

3.3. Models and parameterization

We considered four potential stochastic models. These models describe the cell cycle as a one, two or three part process where the duration of each part of the cell cycle is either constant or variable. In the later case, it is controlled by a random variable y, which evolves according to the Itô stochastic differential equation

$$dy = \mu dt + \sigma dWt; \quad y(0) = y_0$$

, where $t = 0$ corresponds to the beginning of the part and and $y(t) = 1$ corresponds to exit form the part, i.e. checkpoint passage. It follows that the exit time from a part of the cell cycle which is variable in length follows an inverse Gaussian distribution [24]. The first and simplest model describes the cell cycle as a one-part stochastic process. Our second and third models divide the cell cycle into two parts. In the second model the IMT distribution is an inverse Gaussian distribution with undetermined origin (i.e. one part of the cell cycle is of constant duration). In the third model IMT is distributed as a convolution of two inverse Gaussian distributions (i.e. both parts of the cell cycle are stochastic in length). Our fourth model divides the cell cycle into three stochastic parts. It implies that the IMT distribution is a three-fold convolution of inverse Gaussian distributions.

All of the models were fit to the data using Matlab's built-in maximum likelihood estimator (mle). This function accepts custom probability density functions and parameter ranges. We used the following form for the inverse Gaussian distribution [24]

$$p(t) = \frac{1}{\sigma\sqrt{2\pi t^3}} \exp \frac{-(\mu t - 1)^2}{2\sigma^2 t} \tag{4}$$

In this form, the parameters μ and σ correspond to the drift and diffusion parameters of the underlying stochastic process.

There are some challenges to fitting the convolution models to our data. First, the likelihood function may have multiple local maxima. In particular, in some simulations the output of the maximum likelihood routine varies with the initial guess. Furthermore, as the distribution of time spent in a part of the cell cycle becomes very concentrated (i.e. approaches a point-mass distribution), it becomes increasingly dif-

ficult to get an accurate numerical approximation of the likelihood of the data. This difficulty is compounded by the fact that for certain initial guesses and data sets the maximum likelihood routine is converges to such solutions. These solutions may correspond to solutions of the sub models in which one or more parts of the cell cycle is deterministic in length. This second problem can lead to long run times, inaccurate estimates of the likelihood of the data and/or prevent convergence. We have improved the runtime and avoided the computational challenges associated with concentrated distributions by treating highly concentrated distributions as point-mass distributions. Specifically, if the standard deviation in the time spent in a part of the cell cycle is (i) less than 0.0025 hours while the mean of the time spent in a part of the cycle is 25 times greater than this standard deviation, or (ii) the mean time spent in a part of the cell cycle is less than .01 hours then that part of the cell cycle is treated as deterministic in length. In this case, the length of the distribution is taken as the mode the corresponding inverse Gaussian distribution [24]. This method of approximation is related to the accuracy of our data as follows: Because the data in generated through experiments in which cells are observed every 6 minutes (i.e. every .1 hours), it is impossible to distinguish between events that are less than six minutes apart. Hence it is reasonable to treat distributions in which a large majority of cells exit a part of the cell cycle within a six minute interval, as deterministic in length. With this in mind, consider the following reparameterization of the inverse Gaussian distribution

$$p(t) = \left(\frac{\lambda}{2\pi t^3}\right)^{1/2} \exp\left(-\frac{\lambda(\nu - t)^2}{2\nu^2 t}\right). \tag{5}$$

Where $\nu = \frac{1}{\mu}$, and $\lambda = \frac{1}{\sigma^2}$ [24]. In terms of these parameters, the mean is ν and the

variance is $\frac{\nu^3}{\lambda}$. Let $var = \frac{\nu^3}{\lambda}$ and suppose that condition (i) above holds. Then for $\nu - 2\sqrt{var} \leq t \leq \nu + 2\sqrt{var}$,

$$p(t) \geq \frac{\lambda}{2\pi\nu^3}\frac{\nu^3}{(\nu+2\sqrt{var})^3}\Bigg)^{1/2} \exp\left(-\frac{1}{var}\frac{\nu}{(\nu-2\sqrt{var})}\frac{(\nu-t)^2}{2}\right) \quad (6)$$

$$\geq \left(\frac{\nu-2\sqrt{var}}{\nu+2\sqrt{var}}\right)^3 N\left(\nu, var\left(\frac{\nu-2\sqrt{var}}{\nu}\right)\right), \quad (7)$$

where N denotes the normal distribution. Since $var\frac{\nu-2\sqrt{var}}{\nu} < var$, it follows from the properties of the normal distribution that for $X \sim N\left(\nu, var\left(\frac{\nu-2\sqrt{var}}{\nu}\right)\right)$ more that 95% of observations of X fall within 2 standard deviations of the mean. Since condition (i) also ensures that $\nu > 25\sqrt{var}$, more than 90% of observations will fall within $2\sqrt{var} < .005$ hr $= 3$ min from the mean [17]. That is, within the accuracy of our data a large majority of cells will exit the corresponding part of the cell cycle at the same time. On the other hand, if condition (ii) holds then it must be that 90% of cells spend fewer than 6 minutes in the corresponding part of the cell cycle. Indeed, if not then

$$\int_{.1}^{\infty} p(t)dt > .1,$$

and thus

$$\nu = \int_0^{\infty} tp(t)dt \geq \int_{.1}^{\infty} tp(t)dt \geq .1\int_{.1}^{\infty} p(t)dt > .01,$$

a contradiction.

Although this method provides a way of approximating concentrated distributions, it does not guarantee the accuracy of the likelihood estimates, or that the routine will converge to a global maximum. For this reason, we vary the initial parameter guess and also employ an adaptive method that tracks the error in the likelihood estimation. This error comes from the numerical integration involved in computing the

likelihood of the data using the convolution models. Because our code approximates the convolution as a left-hand Riemann sum (conv.m Matlab), the error scales with the step size, and can be estimated by recalculating the likelihood of the data using a smaller step-size. Thus, for each parameter choice, beginning with a step size of .01, the program reduces the step size by 1/2 until two subsequent estimations of the likelihood of the data are within .001 of the absolute value of the last estimate of the likelihood of the data. That is, our criteria depends on the relative error.

As a second means of evaluating the accuracy of our method, we use it to fit synthetic data. Specifically, for each model we used parameter values close to those returned for the Erlotnib data to generate data sets of increasing size ($n = 200$, $n = 2000$, and $n = 20000$). For a fixed data size we generated and fit 5 sets of data, and then computed the average relative error. The results for the inverse Gaussian, inverse Gaussian with undetermined origin, and convolution of two inverse Gaussian distributions are similar; the relative error at $n = 200$ is small, usually on the order of $10^{-1} - 10^{-2}$, and as the data size increases, the error declines modestly so that the average error at $n = 20,000$ is typically on the order of $10^{-2} - 10^{-3}$. These results suggest that the accuracy of the method is limited, and yet it is possible to achieve a reasonable approximation of the distribution parameters with a relatively small data set. Table 1 shows how the average relative error changes for the two-stage convolution model.

Table 1: **Relative Error in Parameters (two-stage model)**

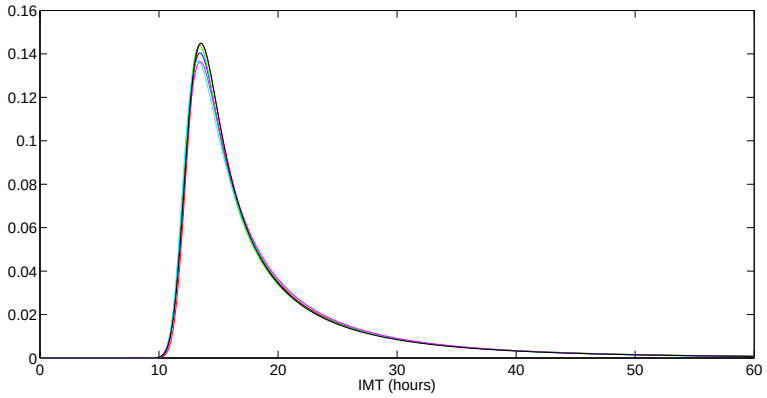
parameter	true value	$(n = 200)$	$(n = 2000)$	$(n = 20000)$
μ_1	.0861	0.0255	0.0100	.0037
σ_1	.0207	0.1445	0.1079	.0271
μ_2	.1292	0.1038	0.0273	.0066
σ_2	.4860	0.0774	0.0546	.0208

Table 2: **Relative Error in Parameters (three-stage model)**

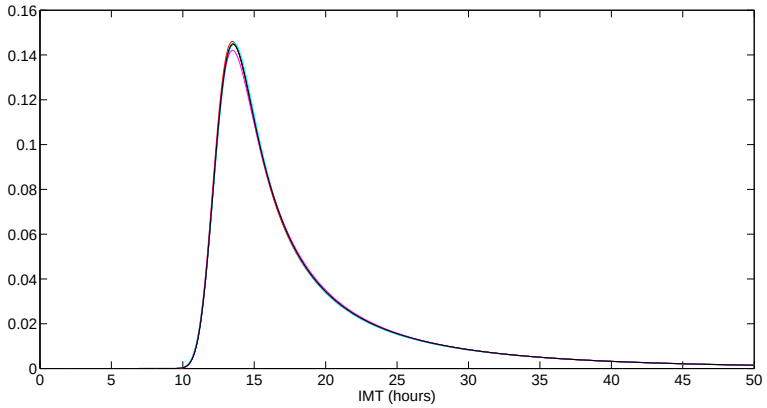
parameter	true value	$(n = 200)$	$(n = 2000)$	$(n = 20000)$
μ_1	.0861	0.0184	0.0193	.0058
σ_1	.0207	0.0535	0.0769	.0295
μ_2	.6437	0.1797	0.2661	.1918
σ_2	2.4344	0.2959	0.2345	.2654
μ_3	.1615	0.0993	0.0994	.0315
σ_3	.6072	0.1157	0.1138	.0393

The results for the convolution of three inverse Gaussian distributions are more complex. In particular, the parameters that correspond to the second part of the cell cycle to not appear to converge to their true value, while those that correspond to the first and third parts of the cell cycle converge similarly as for the two-stage model. However, Figure 7 shows that in four of the five simulations, the graphs of the best-fit probability density functions (pdf) are almost indistinguishable from the

true pdf. This suggests that, for some data sets, it may not be possible to uniquely identify the parameters of the three-stage model.



(a) n=2000



(b) n=20000

Figure 7: Comparison of the pdfs derived through maximum likelihood estimation on synthetic data and the true pdf for $n = 2000$ and $n = 20000$. The black curve corresponds to the true pdf.

Table 3: **Parameters for the three-stage model (n=20000)**

identity	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
black (true)	.0861	.0207	.6437	2.4344	.1615	.6072
blue	0.0860	0.0208	0.7517	3.0055	0.1564	0.5794
green	0.0859	0.0213	0.7852	3.0054	0.1541	0.5904
red	0.0863	0.0205	0.7897	3.0054	0.1538	0.5859
cyan	0.0852	0.0221	0.7772	3.0054	0.1572	0.6085
magenta	0.0873	0.0200	0.5555	1.4885	0.1626	0.6594

3.4. Model selection and comparison

We fit our the mathematical models to experimental IMT data from multiple cell lines (AT1 and MCF) and experimental conditions using MATLAB's MLE (maximum likelihood estimate) routine. To determine which of the three models is best, we use the Akaike information criterion for finite sample sizes ($AICc$), which measures the relative quality of statistical models for a given set of data [26]. The $AICc$ value of a model is given by the formula $AICc = 2k - 2ln(L) + 2k(k+1)/(n-k-1)$, where k is the number of estimated parameters of the model, and L is the optimized value of the likelihood function. The preferred model is the one with the minimum $AICc$ value.

The figures below (Figures 8-12) show the best fits of the model two-stage, three-stage and EMG models to the data sets. Tables 4 and 5 give the best-fit parameters for the two- and three-stage models. The final column of Table 5 indicates if the best fit parameter set corresponds to a two-stage model. That is, for some data

sets the best fit model consists of a two-part stochastic process with a deterministic lag (flag=1). Tables 6-10 provide the $AICc$, ML (maximum likelihood), and $AICp$ values for each model and data set; $AICp$ is given by the quantity $e^{(AIC_m - AIC_i)/2}$, where $AICc_i$ is the $AICc$ value of the ith model and AIC_m is the minimum $AICc$ value. $AICp$ can be interpreted as the relative probability of model i [26]. We note that, with the exception of the CHX data, in each case the two-stage model performs considerably better than both the more complex three-stage model or EMG model. Because CHX inhibits protein synthesis across the board, the inability of the model to describe this data set may indicate that CHX substantially alters checkpoint function. That is, under these conditions, the cell cycle checkpoints no longer behave as bistable switches.

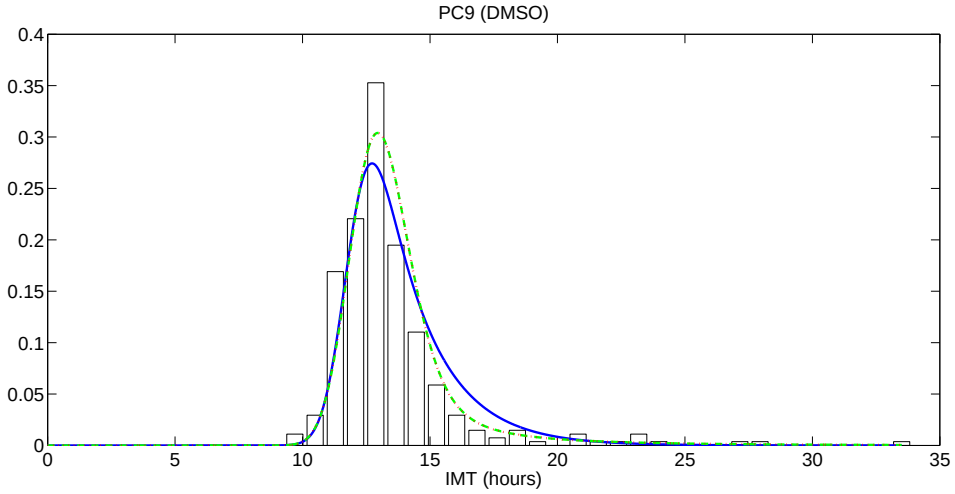


Figure 8: MCF cells (DMSO). Blue: EMG model, Green: Two-stage model, Red: Three-stage model

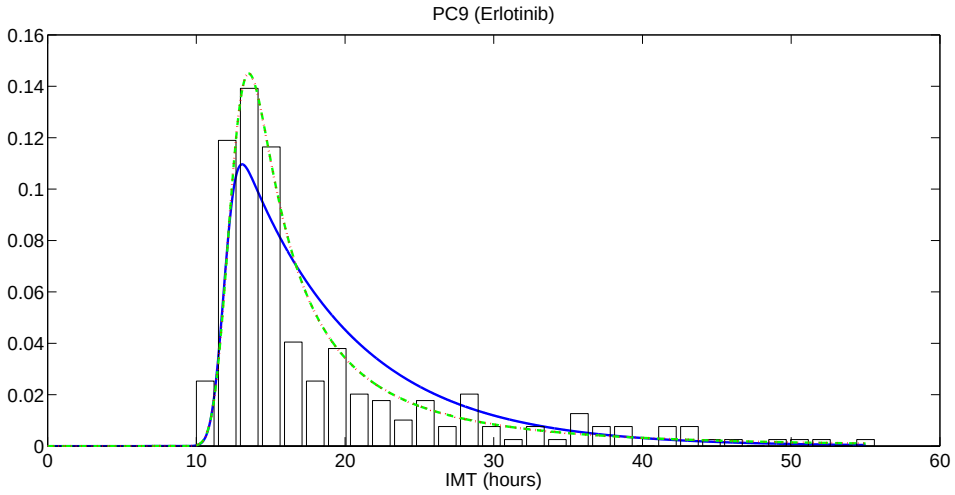


Figure 9: MCF cells (Erlotinib). Blue: EMG model, Green: Two-stage model, Red: Three-stage model

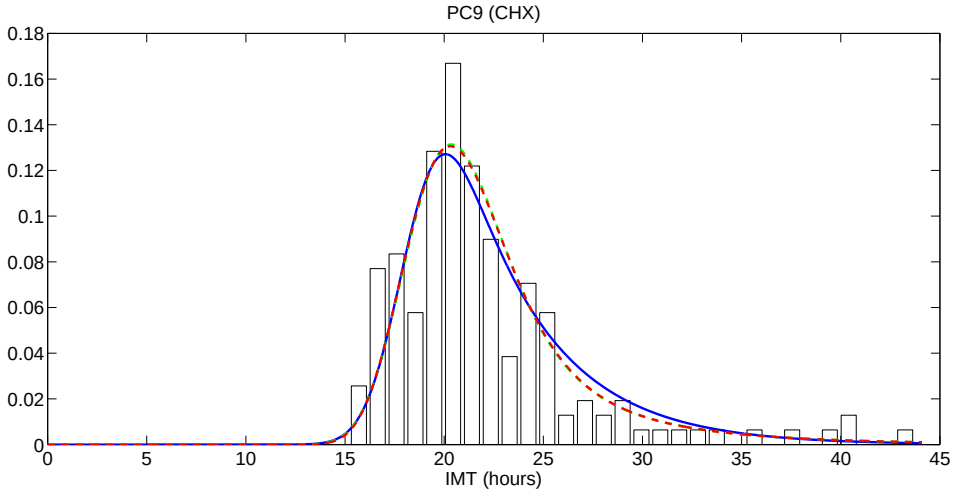


Figure 10: MCF cells (CHX). Blue: EMG model, Green: Two-stage model, Red: Three-stage model

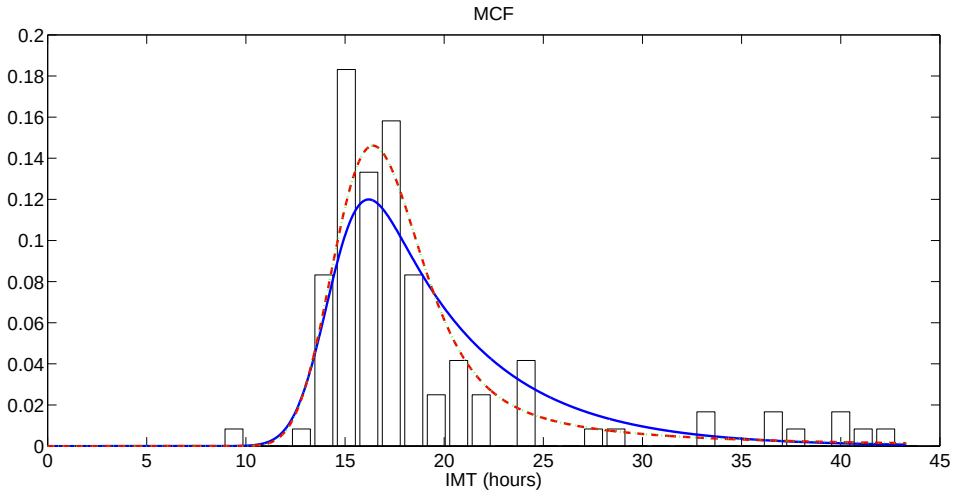


Figure 11: MCF cells. Blue: EMG model, Green: Two-stage model, Red: Three-stage model

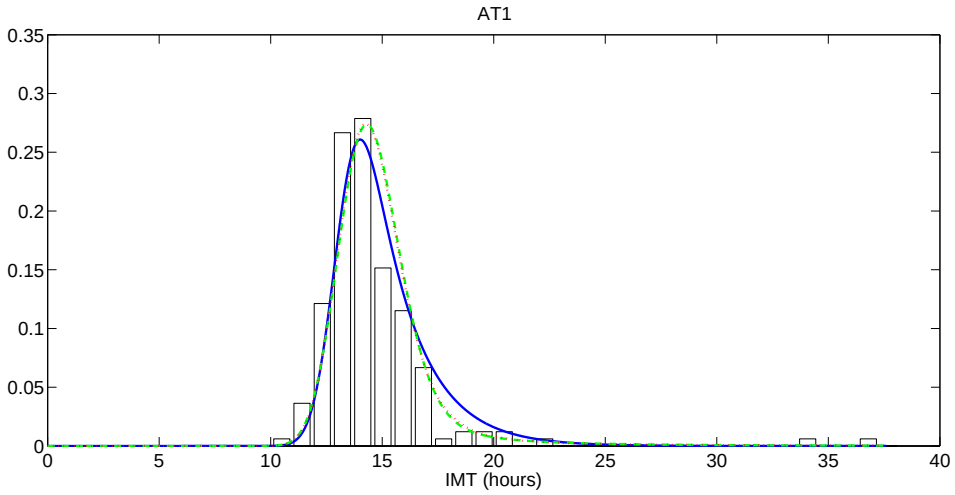


Figure 12: AT1 cells. Blue: EMG model, Green: Two-stage model, Red: Three-stage model

Table 4: **Two-stage model parameters**

data set	μ_1	σ_1	μ_2	σ_2
MCF (DMSO)	0.9087	2.3140	0.0788	0.0239
MCF (Erlotinib)	0.1292	0.4860	0.0861	0.0207
MCF (CHX)	0.2629	0.6254	0.0535	0.0256
AT1	1.2412	3.6619	0.0703	0.0243
MCF	0.2533	1.0166	0.0644	0.0310

Table 5: **Three-stage model parameters (MCF cell line)**

data set	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3	flag
DMSO (MCF)	0.9087	2.3142	0.1088	0.0329	0.2860	0.0866	0
Erlotinib (MCF)	0.1615	0.6072	0.6473	2.4344	0.0861	0.0207	0
CHX (MCF)	0.0668	0.0363	0.2671	0.6388	0.2654	0	1
AT1	1.8001	1.8022	0.0726	0.0236	1.4672	4.7161	0
MCF	0.7489	3.0054	0.3828	1.5362	0.0645	0.0310	0

Table 6: **MCF (DMSO) AIC Analysis (n = 343)**

	MLE	AICc	AICp
EMG	-683.0432	1372.2	0
One-Stage	-746.8349	1497.7	0
Two-Stage	-665.8195	1339.8	1
Three-Stage	-665.8258	1343.9	0.1259

Table 7: **MCF (Erlotinib) AIC Analysis (n = 267)**

	MLE	AICc	AICp
EMG	-825.3567	1656.8	0.0001
One-Stage	-881.9015	1767.8	0
Two-Stage	-815.4143	1639.0	1
Three-Stage	-815.4160	1643.2	0.1241

Table 8: **MCF (CHX) AIC Analysis (n = 164)**

	MLE	AICc	AICp
EMG	-453.6005	913.3510	0.5510
One-Stage	-468.8970	941.8685	0
Two-Stage	-451.9537	912.1590	1
Three-Stage	-451.9010	914.1817	0.3647

Table 9: **AT1 AIC Analysis (n = 182)**

	MLE	AICc	AICp
EMG	-368.7090	743.5528	0.0003
One-Stage	-402.4623	808.9916	0
Two-Stage	-359.5311	727.2882	1
Three-Stage	-359.4373	731.3546	0.1309

Table 10: **MCF AIC Analysis (n = 106)**

	MLE	AICc	AICp
EMG	-304.8850	616.0053	0.0169
One-Stage	-323.6136	651.3437	0
Two-Stage	-299.7228	607.8416	1
Three-Stage	-299.7235	612.2955	0.1079

Table 11: **Two-stage model part lengths. *Although the parts are numbered in this table, the order in which they occur is not determined.**

data set	mean	variance	mean	variance
	part 1*	part 1	part 2	part 2
MCF (DMSO)	1.1004	2.6714	12.6904	1.0805
MCF (Erlotinib)	7.7399	10.4651	11.6144	0.8193
MCF (CHX)	3.8037	4.6395	18.6916	2.0688
AT1	.8372	2.5867	14.1844	1.2928
MCF	3.9479	7.9744	15.5280	1.8968

Table 12: **Three-stage model part lengths. *Although the parts are numbered in this table, the order in which they occur is not determined.**

data set	mean	variance	mean	variance	mean	variance
	part 1*	part 1	part 2	part 2	part 3	part 3
MCF (DMSO)	1.1005	2.6716	9.1912	0.9168	3.4965	0.5662
MCF (Erlotinib)	6.1920	9.3556	1.5449	4.6745	11.6144	0.8193
MCF (CHX)	14.9701	2.1025	3.7439	4.6276	3.7679	0
AT1	0.5555	0.7462	13.7741	1.2064	0.6816	2.6537
MCF	1.3353	4.6373	2.6123	6.4862	15.5039	1.8924

3.5. Biological interpretation

Our results suggest that the mammalian cell cycle is best conceptualized as a two-part stochastic process, wherein one part is short and highly variable and the other is long and much less variable (Tables 11-12). It may be possible to interpret these "parts" in biological terms. A comparison of the IMT distributions for MCF cells cultured with the growth factor signaling inhibitor Erlotnib to those cultured with DMSO suggests that the short, variable part of the cell cycle may be growth-factor dependent, while the consistently long part of the cell cycle may be refractory to growth-factor signaling. Additional experiments involving the perturbation of growth-factor signaling can help to validate this hypothesis. However by interfacing this simple model of the cell cycle with a more mechanistic model of the biological networks that control the cell cycle we can also link the parts of the IMT distribution to specific biological molecules. Below we motivate and describe a procedure for establishing such an interface.

In progressing through the cell cycle, a cell undergoes discrete and irreversible phenotypic changes. The most apparent of these changes is that of DNA replication and the associated phenotypes, G1 and G2, are assigned to cells with one, respectively two, sets of DNA. Another fundamental phenotypic change that eukaryotic cell undergoes is the activation of the E2F transcription factor [15].

Experimental work has shown that these changes are controlled by complex networks of cell cycle proteins and mathematical modeling has shown that these networks enable cells to undergo irreversible change [15, 18]. More specifically, these networks exhibit multiple steady states (which correspond to distinct phenotypes), and there exists a range of parameter values over which both steady states are stable. This final property implies that critical biological variables must obtain a critical threshold

value in order for the cell to move between phenotypes and that, having undergone a phenotypic change, the cell is resistant to reverting to its previous phenotype. Because our stochastic model assumes that checkpoint passage is governed by a random variable y, it is natural to relate y to the critical biological variables of a mathematical model of the checkpoint. In so doing, we will relate stochastic variations in the value of the biological variables, which may be attributed to fluctuations in protein concentrations [23, 2], to the stochastic process that governs y. In order to make the relationship more concrete, we will illustrate it with a simple model of the R-point from [19].

This simple model can be interpreted as describing the interplay between the transcription factor E2F3a, which supports the G1/S transition in proliferating cells, and its regulator, the retinoblastoma protein (Rb), which is detailed in [25, 15], and summarized here: Growth factor signaling leads to Myc-induced transcription of Cyclin D which phosphorylates and deactivates Rb. This results in the activation of E2F3a and the transcription of cyclins (including cyclin E), which bolster pro-mitotic signaling, in part, by helping to phosphorylate Rb. The mathematical model tracks the cellular concentration of Cyclin E and the cellular concentration of phosphorylated Rb. The author's show that the system exhibits bistability, and that the transition between steady-states is controlled by the concentration of a starter kinase, which we consider to be Cyclin D. In particular, Cyclin D levels must reach a threshold value in order for the cell to switch to a promitotic state, so that the concentration of Cyclin D is a bifurcation parameter. Since cellular protein concentrations are approximately log normally distributed [8] this bifurcation parameter is expected to be log normally distributed as well. The random variable y, on the other hand is normally distributed at any point in time. Hence, it is reasonable to identify y with the natural logarithm

of the bifurcation parameter. For convenience, however, we choose to normalize y so that $y = 0$ corresponds to the basal condition of the cell and $y = 1$ corresponds to checkpoint passage. For this reason, we instead define y as follows:

$$y = \frac{\log(b) - \log(b_0)}{\log(b^*) - \log(b_0)},$$

where b_0 denotes the basal value of the bifurcation parameter and b^* denotes the threshold value of the bifurcation parameter. Thus, if we attribute variability in the time it takes a cell to divide to stochastic fluctuations in protein concentrations, we can identify the random variable y with Cyclin D. Moreover, the stochastic process that governs y is consistent with the expectation that protein concentrations be log normally distributed.

This method can be used more generally to relate stochasticity in the time it takes a cell to pass a checkpoint to the concentrations of specific biological molecules, provided a suitable mathematical model.

3.6. Conclusions and future directions

This work investigates the application of multi-part stochastic models to the study of IMT data. We have (i) developed and tested procedures for performing maximum likelihood estimation of model parameters using IMT data and MATLAB mle.m, (ii) explained how these models can be linked to intracellular protein concentrations, and (iii) analyzed IMT variability within our model framework. In regard to the first item, we identified convergence to local maxima, in which one part of the cell cycle corresponds to a Dirac delta function, as an obstacle to fitting convolution models to IMT data. To overcome this difficulty, we proposed a method for approximating very concentrated distributions. The threshold at which the approximation is applied is

linked to the accuracy of the data. On applying this method to simulated data, we found that the parameters of the two-part stochastic model can accurately estimated, while those of the three-part model are more difficult to estimate. In particular, for parameters like those fit to the Erlotinib-treated MCF cells, our results suggest that the parameters of the three-part model are not identifiable. However, our results also suggest that across multiple cell lines, the cell cycle is best conceptualized as a two-part stochastic process.

In the future, it would be interesting to consider data from irradiated cells in order to see if it a third damage induced checkpoint becomes apparent. Our results also suggest that the majority of variability in the time it takes a cell to divide can be attributed to a single stochastic process that, in the cell lines studied here, is (i) short and highly variable in length and (ii) sensitive to signaling through EGFR. These results are consistent with previous work, which attributes IMT variability to variability in the time to complete G_1 pre synthesis [4], the length of which is dependent on c-Myc signaling [9]. Specifically, it was proposed that a loss of c-Myc lengthens $G_1 - ps$ by slowing rates of protein synthesis [9]. This proposal is consistent with our stochastic model, which attributes variability to fluctuations in protein concentrations. In summary, the two-part stochastic model presented here provides a tractable, accurate, and biologically meaningful description of variability in the time it takes a cell to divide that is well suited for the analysis of IMT data. Nonetheless, under some conditions, e.g. the presence of DNA damage, a three-part stochastic model may be preferable.

Future work should focus on parameter identifiability for the three-stage model, and the possibility of improving the method of parameter estimation for both models. In particular, the results with the synthetic data suggest that for both the two- and

three-stage models, the rate of convergence (as a function of the number of data points), is somewhat slow. A possible direction of future research would be to adapt the method to include censored data, i.e. data from cells that do not divide in the course of the experiment. Such a method could improve the reliability of parameter estimates by effectively increasing the size of the data set. A second direction for future research is to determine the distribution of the maximum likelihood estimators of the parameters in order to perform inference. This step is especially important for testing statistical significance in an experimental setting. While this work focuses solely on dividing cells, this modeling framework has the potential to be applied to other cell fate decisions such as differentiation and death. For example, the model could be adjusted to handle two thresholds, one corresponding to division and the other to death, and our framework could provide distributions of decision times for both. In the future we hope that this research work will provide experimentalists with reliable tools for the analysis of IMT data.

4. CONCLUSION

Information flow in mammalian cells is a very complex and variable process. This work presents two complementary approaches to address the temporal variability in cell cycle entry. The first part of this thesis adapts an existing model of the Rb-E2F network to pinpoint potential molecular sources of variability in IMT. Analysis of this simple mathematical model of the Rb-E2F network suggests that regulation of the RP node is the primary source of variability in R time. Because of this, it is reasonable to believe those molecular components embedded into the RP node, specifically the retinoblastoma protein, are key players in determining IMT. While analysis of this simple model provides insight into the molecular components driving variable cell cycle entry, it does not provide a means of analyzing IMT data. The second approach of modeling the cell cycle presented in this thesis complements the first approach in that it provides numerical methods for the analysis of IMT data, under the assumption that IMT variability is the result of molecular stochasticity.

Specifically, in the second part of this thesis we apply multi-part stochastic models to the study of IMT data, develop and test procedures for performing maximum likelihood estimation of model parameters, explain how the random variables associated with the model can be linked to intracellular protein concentrations, and analyze IMT variability within our model framework. To compare the models' ability to fit IMT data, we use model selection theory, specifically the Akaike information criterion, to determine that across all of our cell lines, the cell cycle is best described as a two-part stochastic process. Furthermore, this work suggests that much of the variability in the time it takes a cell to divide can be attributed to a single stochastic, growth factor dependent process, which is short and highly variable in length relative to the other part of the cell cycle. The stochastic model presented in this thesis provides a

tractable, accurate, and biologically meaningful description of variability in the time it takes a cell to divide that is well suited for the analysis of IMT data. Collectively, the approaches presented in this provide a greater understanding of how mammalian cells process information and the noise sources involved in temporal variability in cell cycle entry.

BIBLIOGRAPHY

[1] Golubev A. Exponentially modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. *Journal of Theoretical Biology*, 262:257–266, 2010.

[2] Marusyk A. and Polyak K. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta*, 1805, 2010.

[3] Smith J. A. and Martin L. Do cells cycle? *P. Natl. Acad. Sci. U.S.A.*, 70(4):1263–1267, 1973.

[4] Zetterberg A. and Larsson O. Kinetic analysis of regulatory events in G1 leading to proliferation or quiescence of Swiss 3t3 cells. *Cell Biology*, 82:5365–5369, 1985.

[5] Zilman A., Ganusov V. V., and Perelson A. S. Stochastic models of lymphocyte proliferation and death. *PLoS ONE*, 5, 2010.

[6] Amati B., Alevizopoulos K., and Vlach J. Myc and the cell cycle. *Frontiers in Bioscience*, 3:250–268, 1998.

[7] Pardee A. B. A restriction point for control of normal animal cell proliferation. *PNAS*, 71:1286–1290, 1974.

[8] Furusawa C., Suzuki T., Kashiwagi A., Yomo T., and Kaneko K. Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *Biophysics*, 1:25–31, 2005.

[9] Schorl C. and Sedivy J. M. Loss of prooncogene c-myc function impedes G1 phase progression before and after the restriction point. *Molecular Biology of the Cell*, 14:823–835, 2003.

[10] Shannon C. A mathematical theory of communication. *The Bell System Technical Journal*, 27:623–656, 1948.

[11] OpenStax CNX. Openstax, Biology. `http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.53`. Accessed: 2016-06-15.

[12] Guardavaccaro D. and Pagano M. Stabilizers and destabilizers controlling cell cycle oscillators. *Mol Cell*, 22:1–4, 2006.

[13] Giacinti G. and Giordano A. Rb and cell cycle progression. *Oncogene*, 25:5220–5227, 2006.

[14] Yao G., Tan C., West M., Nevins J. R., and You L. Origin of bistability underlying the mammalian cell cycle. *Molecular Systems Biology*, 7(485), 2011.

[15] Yao G., Lee T. J., Mori S., Nevins J. R., and You L. A bistable Rb-E2F switch underlies the restriction point. *Nature Cell Biology*, 10:476–482, 2008.

[16] Hartwell L. H. and Weinert T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246:629–634, 1989.

[17] Hogg R. V., McKean J. W., and Craig A. T. Introduction to mathematical statistics. chapter 3, page 172. Pearson, Boston, 2013.

[18] Tyson J.J. and Novak B. Regulation of the eukaryotic cell cycle: Molecular antagonism, hysteresis, and irreversible transitions. *J. Theor. Biol.*, 210:249–263, 2001.

[19] Tyson J.J. and Novak B. Irreversible transitions, bistability and checkpoint controls in the eukaryotic cell cycle: A systems-level understanding. In *Handbook of systems biology*, chapter 14, pages 265–285. Elsevier, San Diego, 2013.

[20] Tyson J.J. and Diekmann O. Sloppy size control of the cell division cycle. *J Theor Biol*, 118(4):405 – 426, 1986.

[21] Mateyak M. K., Obaya A. J., and Sedivy J. M. c-myc regulates cyclin d-cdk4 and -cdk6 activity but affects cell cycle progression at multiple independent points. *Mol Cell Biol*, 19(7):4672–4683, 1999.

[22] Vermeulen K., Van Bockstaele D., and Berneman Z. The cell cycle: a review of regulation, dergeulation and therapeutic targets in cancer. *Cell Proliferation*, 36:131–149, 2003.

[23] Sandip Kara, William T. Baumann, Mark R. Paul, and John J. Tyson. Exploring the roles of noise in the eukaryotic cell cycle. *PNAS*, 106(16):6471?6476, 2009.

[24] Folks J. L. and Chhikara R. S. The inverse gaussian distribution and its statistical application – a review. *J. R. Statist. Soc. B*, 40(3):263–289, 1978.

[25] Humbert P. O., Verona R., Trimarchi J. M., Rogers C., Dandapani S., and Lees J. A. E2f3 is critical for normal cellular proliferation. *Genes and Development*, 14:690–703, 2000.

[26] Burnham K. P. and Anderson D. R. Multimodel inference. *Soiciological Methods and Research*, 33(2):261–304, 2004.

[27] Dougherty R. Claude Shannon. `https://www.nyu.edu/pages/linguistics/courses/v610003/shan.html`. Accessed: 2016-06-15.

[28] Leander R., Allen E. J., Garbett S. P., Tyson D. R., and Quaranta V. Derivation and experimental comparison of cell-division probability densities. *J Theor Biol*, 359(0):129 – 135, 2014.

[29] Tyson D. R., Garbett S. P., Frick P. L., and Quaranta V. Fractional proliferation: A method to deconvolve cell population dynamics from single-cell data. *Nat Meth*, pages 923–928, 2012.

[30] Quaranta V., Tyson D. R., Garbett S. P., Weidow B, Harris M. P., and Georgescu W. Trait variability of cancer cells quantified by high-content automated microscopy of single cells. *Methods Enzymol*, 467:23–57, 2009.

[31] Wong J. V., Dong P., Nevins J. R., Mathey-Prevot B., and You L. Network calisthenics: Control of E2F dynamics in cell cycle entry. *Cell Cycle*, 10:18:3086–3094, 2011.

[32] Harbour J. W. and Dean D. C. The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes Development*, 14:2393–2409, 2000.

[33] Tom Serge Weber, Irene Jaehnert, Christian Schichor, Michal Or-Guil, and Jorge Carneiro. Quantifying the length and variance of the eukaryotic cell cycle phases by a stochastic model and dual nucleoside pulse labeling. *PLOS Computational Biology*, 10:e1003616, 2014.

[34] Leung J. Y., Ehmann G.L., Giangrande P. H., and Nevins J. R. A role for Myc in facilitating transcription activation by E2F1. *Oncogene*, 27:4172–4179, 2008.