7-9-2020

An Investigation Of Gene Regulatory Network State Space Variability

Sara Faye Liesman
Illinois State University, Sflwood@hotmail.com

AN INVESTIGATION OF GENE REGULATORY NETWORK STATE SPACE VARIABILITY

SARA FAYE LIESMAN

52 Pages

Genes are segments of DNA that provide a blueprint for cells and organisms to effectively control processes and regulations within individuals. There have been many attempts to quantify these processes, as a greater understanding of how genes operate could have large impacts on both personalized and precision medicine. Gene interactions are of particular interest, however, current biological methods can not easily reveal the details of these interactions. Therefore, we infer networks of interactions from gene expression data which we call a gene regulatory network, or GRN. Due to the robust behavior of genes and the inherent variability within interactions, models incorporating stochasticity are more realistic than those that are only deterministic. These methods are designed to bypass the need for large amounts of data and extensive knowledge about a network. In this work, we extend previous work investigating additional ways to incorporate stochasticity into gene regulatory networks. First, we use a transition function and investigate its inherent variation, then we use a statistical distribution for activating and degrading the states of genes, and finally, we use a new method incorporating spectral density to incorporate stochasticity within a GRN.

KEYWORDS: beta distribution, cell to cell variation, gene regulatory networks, Spectral density, state transition, stochasticity

AN INVESTIGATION OF GENE REGULATORY NETWORK STATE SPACE VARIABILITY

SARA FAYE LIESMAN

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mathematics

ILLINOIS STATE UNIVERSITY

2020

AN INVESTIGATION OF GENE REGULATORY NETWORK STATE SPACE VARIABILITY

SARA FAYE LIESMAN

COMMITTEE MEMBERS:

Olcay Akman, Chair

Pei Geng

Tom Hammond

Daniel Hrozencik

ACKNOWLEDGMENTS

I would like to thank my advisors, Dr. Olcay Akman and Dr. Daniel Hrozencik, for their continuous support and encouragement during the research and writing experience. Their expert advice and support aided my growth as a student throughout this thesis work. Thank you for your patience and all of the opportunities you have given me through the years. In addition to my advisors, I would like to thank the other members of my committee, Dr. Tom Hammond and Dr. Pei Geng, for their insightful comments and suggestions.

I would also like to thank my parents, partner, sisters, brother, and friends for their continuous love and support throughout my master's career. Finally, I would like to thank my peers in the mathematics and biology departments for their friendship and camaraderie that carried me through the years.

S. F. L.

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION

An understanding of organisms on the molecular level requires the knowledge of genes and their interactions. Genes are segments of DNA that aid the regulation of phenotypes in organisms and ultimately dictate the fitness of individuals. Cells function properly when genes and proteins, both within and around cells, function properly [3].

Gene expression is regulated through interaction networks and a series of positive and negative feedback loops [3, 17, 20]. Similar to interaction networks on the macro-level, where organisms interact with each other in order to create an ecosystem, interaction networks also exist on the microscopic level [3, 17, 20]. On the microscopic level, cells, proteins, and molecules interact with one another.

Transcription factors dictate how and when each gene influences cellular activity [22]. Genes help create messenger RNAs (mRNAs) which then work with ribosomes to synthesize proteins. These proteins could then become transcription factors again and assist in the regulation of a new gene [20, 22]. This chain reaction forms a network of gene, protein, and regulator reactions that we call a gene regulatory network. However, current technology and experimental methods do not exist to directly reveal the intricacies of these networks. Therefore, the gene regulatory networks that we know today are networks that have been inferred from gene expression data, often denoted by GRN [8]. Gene expression data provides information about mRNAs, but not necessarily about binding information. Thus, GRNs are used in order to infer the interactions between genes, proteins, and other regulators [8].

An example of a regulatory network can be seen in *Figure* 1a. In this figure we see the p53-Mdm2 network described in [1, 6, 15, 28, 39]. The p53 protein is a tumor suppressing protein that is activated by damage to DNA. Here the term "p53 protein" includes nucleic p53, cytoplasmic p53, and the gene p53. In this network, damage to DNA represses nucleic Mdm2. The repression of Mdm2 allows for p53 proteins to increase. The p53 proteins then help to repair DNA damage, but also decrease nucleic Mdm2 and

increase cytoplasmic Mdm2. Cyctoplasmic Mdm2 causes an increase in nucleic Mdm2, thus there is competition, so to speak, between the p53 protein and cytoplasmic Mdm2 for the increase and decrease of nucleic Mdm2. Once DNA damage has been repaired, nucleic Mdm2 will increase and p53 proteins will decrease.

Although these interactions are dictated by a network, there is still some stochasticity involved between interactions [3, 4, 20, 27, 35]. This stochasticity may be due to concentration levels, binding abilities, or even distances of molecules within cells. In cells, competition between ribosomes and RNase E binding can lead to some stochasticity [27] RNase E signals the degradation of mRNA after transcription, however, mRNA that binds to a ribosome will undergo translation and lead to protein production [27]. Whether mRNA degrades or leads to the production of proteins is thought to be determined by its proximity to either RNase E or a ribosome [27]. Mathematical models have been created in attempts to capture the details of these processes with varying degrees of success. In general, stochasticity is not captured by all inference methods, but likely plays an important role in the formation of GRNs since a large portion of cellular signals are from noise [3].

In order for a cell to properly function, it depends on the coordination of thousands of proteins in different variations interacting at the correct time, place, and in the correct quantity [3, 20, 33]. In order to orchestrate these interactions, regulatory systems exist to help determine when mRNA is produced, how long mRNA should last, how much protein from mRNA should be created, how proteins are arranged and modified, and when they are degraded [3, 20, 33]. However, large amounts of data are needed for gene network inference [3], and limitations in experimental techniques create noisy data sets, so only small interactions have been extracted [32]. Moreover, human interests have introduced a bias into which networks are studied since there is a motivation to study networks related to human diseases [32]. These biological limitations have led to the need for mathematical models to help estimate biologically reasonable parameters such as inputs, time delays, and

A gene regulatory network example of the Mdm2-p53 complex

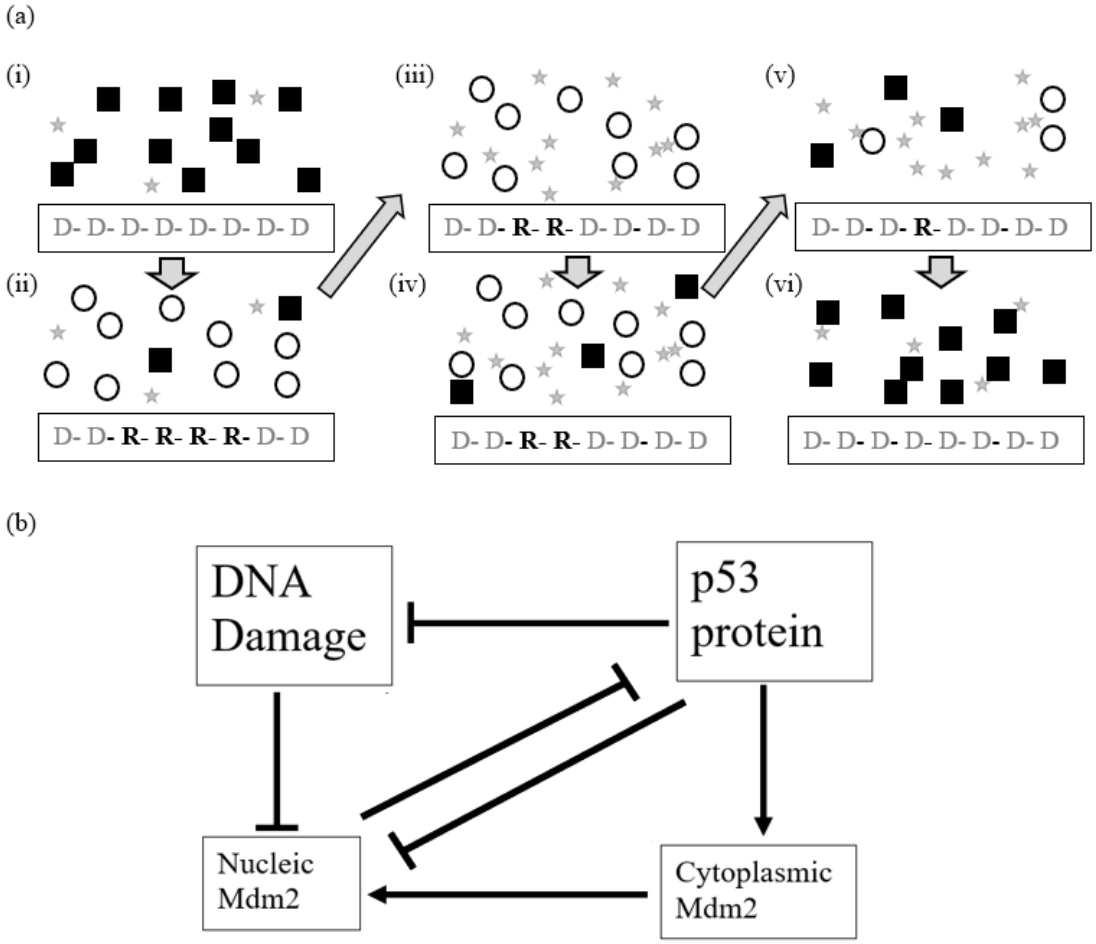


Figure 1: **(a)** A system process with DNA (D), DNA damage (R), Nucleic Mdm2 (squares), Cytoplasmic Mdm2 (stars) and p53 including cytoplasmic, nucleic, and gene p53 (circles). (i) The system is shown with no DNA damage. (ii) The system is shown with DNA damage, a decrease in nucleic Mdm2, and an increase in p53. (iii) p53 has caused an increase in cytoplasmic Mdm2, a decrease in DNA damage, and a decrease in nucleic Mdm2. (iv) Cytoplasmic Mdm2 causes an increase in nucleic Mdm2, while p53 causes a decrease in nucleic Mdm2. (v) A decrease in DNA damage leads to a decrease in p53. (vi) Cytoplasmic Mdm2 decreases and nucleic Mdm2 increases. **(b)** A wiring diagram for (a) recreated from [28]. An arrow from object A to object B ($A \rightarrow B$) indicates that A causes an increase in B. A repression line from object A to object B ($A \dashv B$) indicates that A causes a decrease in B. This process was recreated using information from [1, 28].

genes expressed from feedback loops [3, 17, 19, 23, 32, 33].

Networks are a natural way to model biological systems with interactions [14, 23, 32]. A network is a set of objects that are connected through a set of rules. In the past, networks have been used to model a variety of interactions including protein structure networks, protein-protein interactions, transcriptional regulation, metabolism, and neuronal synaptic connections [32]. In gene regulatory networks the objects, or nodes, in a network graph represent a collection of genes and the edges represent the interactions between sets of genes. Models for GRNs can be dynamical or static, discrete or continuous, and deterministic or stochastic [3, 17]. Some models that have been studied are Bayesian networks, rule-based algorithms, ordinary differential equations, and Boolean networks [17, 23]. *Table* 1 describes some of the possible methods used to study GRNs. The methods discussed here are not exhaustive.

Directed graphs are a common way to represent GRNs [14, 17, 23, 32]. A directed graph consists of a collection of vertices and edges $\{V, E\}$ where each edge is defined by the vertices that it connects (v_m, v_n). This set of objects could then be generalized to represent the interactions between different components in a GRN such that each vertex represents a gene and each edge determines a rule whereby the given gene influences the next gene [17]. In *Figure* 1b we see a directed graph for the p53-Mdm2 complex described by [1].

Directed graphs can be simplified for further abstraction of information through the use of Boolean networks. In a directed graph, each of the nodes or vertices of the graph represent a gene within the network. Unlike a directed graph, each of the nodes in a Boolean network represent the states of genes. These states indicate whether or not a gene in a Boolean network is active (on) or non-active (off). Boolean networks are one of the simplest dynamical models and express variables as either on (1) or off (0) [17, 19]. They were first introduced by Stuart Kauffman in the 1960's [3, 17, 19]. In a network with two variables there would be four possible states and thus four vertices in the directed graph (i.e. 00, 01, 10, or 11). An edge from 01 to 11 would indicate that when component one is

off and component two is on, at the next time step, both components will be on. The benefit of using a Boolean network is that it is able to capture qualitative information about networks without needing detailed information about the parameters that are necessary for detailed quantitative results. Boolean networks are simple enough to be used for large networks, but may be too simple to capture important qualities in a network [17]. Like directed graphs, Boolean networks can be deterministic.

A natural extension of Boolean networks is a generalized logical network which allows for asynchronous state changes and utilizes thresholds on concentrations to determine when a state will change [17]. For example, a system with a variable X_i, concentration of molecules x and a threshold at m could have the following rules:

$$\begin{cases} X_i = 0 & \text{if } x < m \\ X_i = 1 & \text{if } x \geq m \end{cases}$$

Each edge could have its own unique threshold. Also, generalized logical networks can be extended to incorporate stochastic time delays, but still largely remain deterministic [17]. One benefit of generalized logical networks is that they are relatively robust, however, it is difficult to find appropriate scoring functions for the possible networks developed with generalized logical networks and thus inference is difficult [24].

Differential equations have also been used to model GRNs [3, 17]. There are a large number of models that have been developed using differential equations that will not be covered here exhaustively. For example, there are both linear and nonlinear differential equations which represent gene products, mRNAs, and proteins with dynamic and continuous variables [3, 35]. A network with N genes will have $2N$ equations, representing the transcription and translation for each of the N genes [10], and thus many parameters are needed in order to create a model. Often multiple experiments are needed in order to measure gene levels and transcription rates can only be measured as a proportion relative to other rates of transcription [35]. In general, differential equations often require advanced

numerical methods which can be difficult to implement due to missing numerical parameters in models [3].

However, deterministic systems fail to incorporate the stochasticity present in GRNs. In cells where network noise leads to a differentiation of cellular phenotype, incorporating noise and variance in a model will help improve the model outcomes [17, 27].

Bayesian networks incorporate stochasticity. A Bayesian network is created with a directed acyclic graph and a series of conditional probabilities along the edges [17]. Since Bayesian networks utilize statistical principles, they can incorporate noise into their networks in a way that other deterministic models fail to do so [35]. Bayesian networks are able to handle missing data and incorporate prior biological knowledge into the system [35]. However, these Bayesian networks have not been known to arrive at globally optimum solutions and many require large amounts of data even for small networks [3, 17].

Stochastic master equations also incorporate stochasticity and provide more information than deterministic models and Bayesian networks [17]. However, these networks are even more difficult to solve than the rate equations involved in differential equations. They require detailed knowledge about the reactions and their probability density functions, and are computationally expensive [17]. On the other hand, in short time periods, the master equation can provide realistic results and predictions for gene regulation [17].

In order to determine whether the models described above are efficient, a comparison between the model results and data is used. Many of the models mentioned create multiple variations and then the task of finding an optimal fit is important for determining which model is appropriate for the GRN [3, 17, 24, 34, 35]. A general comparison between the data obtained for a model and the model itself may be used to provide an estimate about the accuracy of the model [17, 24, 35]. Least square methods and generalized least square methods may also be used to find optimal parameters for GRNs [24, 34, 35]. Scoring functions are also used to optimize models to best fit the data

Table 1: Methods for Gene Regulatory Networks [3, 17, 24]

Type	Qualities	Pros	Cons
Directed Graphs	Static and deterministic	Intuitive	Crude
Boolean Networks	Discrete, dynamic, and deterministic	Simple for large data sets	Crude
Generalized Logical Networks	Discrete, dynamic, and deterministic	Accommodates asynchronous state changes and time delays	Inference is difficult
Differential Equations	Continuous, dynamic, and deterministic	Flexible	Computationally expensive
Bayesian Networks	Static, stochastic, and discrete or continuous	Stochastic	Need large amounts of data
Stochastic Master Equation	Discrete, dynamic, and stochastic	Realistic	Difficult to use

[3, 24]. In order to use these methods large amounts of data are required. In fact, Boolean networks, which are one of the simplest models, require data twice as large as the number of genes in the network [3]. Differential equations can require as many experiments as there are genes in a network [35]. This likely limits the use and comparison of models to small networks. In yeast (*Saccharomyces cerevisiae*) 2355 genes have been identified contributing to regulatory networks [22], this could imply that upwards of 2355 experiments are needed in order to identify the structure and rates of these networks. In addition, transition rates are not easily obtained through experimentation because estimations often rely on linear changes which would imply that there is no natural capacity for transition rates [35]. Combined with the fact that data size is often limited by the cost of experiments [3], it is clear that there is a need to develop more accurate GRNs that consider the effects of

variation and noise.

Akman et al. attempt to address this concern in 2018 by investigating the variation between state transitions in GRNs and noise within a network [2]. The authors use beta distributions to estimate the propensity that state transitions occur in a GRN under ideal conditions. Through their investigation, Akman et al. find that state transitions with high variation lead to a network with high amounts of variation [2]. This is consistent with the literature, whichfinds that short-term fluctuations in protein production can have larger impacts on gene expression [27]. These impacts are likely due to the cascading behavior of GRNs as well as different feedback loops within cells [4, 27]. The authors also use the assumption that propensities among state transitions are not constant and demonstrate how statistical distributions can be used to provide important insight about GRNs. However, the process described in [2] also requires large amounts of data and becomes less stable for large networks. They also assume that an ideal network will have minimum variance among state transition propensities which may be ignoring some key ideas such as:

- Data obtained from a single cell can have large amounts of noise and variation [3]
- Noise from gene expression data can be as large as 30% [3]
- Different feedback loops can amplify or dampen variation in networks [4]
- Genetically identical cells may show cell-cell differences by more than 10% [4]

Despite some shortcomings, Akman et al. have begun to pave the way for introducing more variability into models that will likely require stochastic properties in the near future. In this work we attempt to extend the authors' work and address concerns about noise by investigating additional ways to incorporate variation and noise into future models.

CHAPTER II: MODEL FORMATION AND RESULTS

In this study, we look at intrinsic noise in GRNs and discuss ways to incorporate this noise into future models. During gene and protein interactions transcription factors and chemical signals come into contact with one another and bind in order to cause a reaction such as mRNA synthesis or protein assembly. Traditionally, these reactions and collisions of chemicals in a system were considered deterministic, but due to quantum indeterminacy and lack of mechanical isolation it is now argued that the processes in these systems are more likely stochastic in nature [12]. Here, we assume that we have a well-mixed system with a limited number of molecules per population of molecule so that our systems will incorporate discreteness and stochasticity. In the past, this reaction rate was determined by a set of ordinary differential equations, but when a system involves discreteness and stochasticity, these equations are no longer appropriate [12].

Since our system is both discrete and stochastic, it is most natural to represent our system as a Boolean network with added stochasticity; this representation is similar to the ones described in [2, 28]. In *Figure* 2 we see an example of a Boolean network. Here, each node of the directed graph represents a state of the system and each edge represents the propensity for changing to a state. Recall, that although this graph looks like the wiring diagram, or directed graph, that we saw in *Figure* 1b, the vertices of this graph are the states of genes. Therefore, when we have a two gene system, each node will have two entries. The first entry will represent the state of the first gene, and the second will represent the state of the second gene. These states are either on (1) or off (0). Biologically, genes may be operating at different rates and have more than two states, but mathematically it is possible to abstract information by only considering genes as active or non-active.

Let $G_1, G_2, \cdots, G_n$ represent n genes in a regulatory network. Let $x_i(t)$ be the state of G_i at time t where $i = 1, 2, \cdots, n$ and $t \in [0, T]$. The possible states of our system are 0 or 1, denoted $X_i = 0, 1$ for $i = 1, 2, \cdots, n$. If G_1 is on at time q, $0 \leq q \leq T$, then $x_1(q) = 1$.

If G_1 is off at time q, $0 \leq q \leq T$, then $x_1(q) = 0$. In this study we focus on a predetermined two gene system. Therefore, the state space of this network S is the Cartesian product of each gene's state space, $S = X_1 \times X_2 \times \ldots \times X_n$.

Table 2: State Update Propensities

Update	x_1	x_2
Activation	$p_1^{\uparrow}$	$p_2^{\uparrow}$
Degradation	$p_1^{\downarrow}$	$p_2^{\downarrow}$
Failure to Activate	$1 - p_1^{\uparrow}$	$1 - p_2^{\uparrow}$
Failure to Degrade	$1 - p_1^{\downarrow}$	$1 - p_2^{\downarrow}$

Each edge of the directed graph represents a probability for transition to a different state. These propensities are determined by an update function involving activation and degradation propensities, $p_i^{\uparrow} \in [0, 1]$ and , $p_i^{\downarrow} \in [0, 1]$ respectively. Failure to activate and failure to degrade are represented by $1 - p_i^{\uparrow}$ and $1 - p_i^{\downarrow}$ respectively. Note that $1 - p_i^{\uparrow} \in [0, 1]$ and $1 - p_i^{\downarrow} \in [0, 1]$. These values can be seen in *Table* 2. Although genes are interacting with each other in a network, the propensities for activation and deactivation of genes are each independent. This occurs because genes activate and deactivate independently of one another. Thus, the probability of transitioning from one state to another can be represented as the product of two propensities. Given the independence between gene state propensities, we obtain the value of each edge propensity by finding the product of the appropriate gene propensities for activation, degradation, failure to activate, and failure to degrade. For example:

$$Pr(00 \to 01) = (1 - p_1^{\uparrow}) \times p_2^{\uparrow}, \;\; Pr(00 \to 10) = p_1^{\uparrow} \times (1 - p_2^{\uparrow}),$$

$$Pr(00 \to 11) = p_1^{\uparrow} \times p_2^{\uparrow}, \;\; Pr(00 \to 00) = (1 - p_1^{\uparrow}) \times (1 - p_2^{\uparrow}),$$

$$Pr(11 \to 10) = (1 - p_1^{\downarrow}) \times p_2^{\downarrow}, \;\; Pr(11 \to 01) = p_1^{\downarrow} \times (1 - p_2^{\downarrow}),$$

$$Pr(11 \to 00) = p_1^{\downarrow} \times p_2^{\downarrow}, \;\; Pr(11 \to 11) = (1 - p_1^{\downarrow}) \times (1 - p_2^{\downarrow}),$$

$$Pr(01 \to 00) = (1 - p_1^{\uparrow}) \times p_2^{\downarrow}, \;\; Pr(01 \to 11) = p_1^{\uparrow} \times (1 - p_2^{\downarrow}),$$

$$Pr(01 \to 10) = p_1^{\uparrow} \times p_2^{\downarrow}, \;\; Pr(01 \to 01) = (1 - p_1^{\uparrow}) \times (1 - p_2^{\downarrow}),$$

$$Pr(10 \to 11) = (1 - p_1^{\downarrow}) \times p_2^{\uparrow}, \;\; Pr(10 \to 00) = p_1^{\downarrow} \times (1 - p_2^{\uparrow}),$$

$$Pr(10 \to 01) = p_1^{\downarrow} \times p_2^{\uparrow}, \;\; Pr(10 \to 10) = (1 - p_1^{\downarrow}) \times (1 - p_2^{\uparrow}),$$

These transitions can also be represented in a transition matrix. A general transition matrix is shown in *Table* 3. A more concrete example of a Boolean network with activation and degradation propensities is shown in *Figure* 3

Table 3: Transition Matrix of State Update Propensities

Input/ Output	00	01	10	11
00	$(1 - p_1^{\uparrow})(1 - p_2^{\uparrow})$	$(1 - p_1^{\uparrow})(p_2^{\uparrow})$	$(p_1^{\uparrow})(1 - p_2^{\uparrow})$	$(p_1^{\uparrow})(p_2^{\uparrow})$
01	$(1 - p_1^{\uparrow})(p_2^{\downarrow})$	$(1 - p_1^{\uparrow})(1 - p_2^{\downarrow})$	$(p_1^{\uparrow})(p_2^{\downarrow})$	$(p_1^{\uparrow})(1 - p_2^{\downarrow})$
10	$(p_1^{\downarrow})(1 - p_2^{\uparrow})$	$(p_1^{\downarrow})(p_2^{\uparrow})$	$(1 - p_1^{\downarrow})(1 - p_2^{\uparrow})$	$(1 - p_1^{\downarrow})(p_2^{\uparrow})$
11	$(p_1^{\downarrow})(p_2^{\downarrow})$	$(p_1^{\downarrow})(1 - p_2^{\downarrow})$	$(1 - p_1^{\downarrow})(p_2^{\downarrow})$	$(1 - p_1^{\downarrow})(1 - p_2^{\downarrow})$

In order to determine when a gene will transition, a transition function is used.

$$f(t) = a_0 e^{-kt}, t \geq 0 \tag{II.1}$$

This equation represents the concentration of molecules that are needed in order to trigger a change in the states of genes. In tandem with equation II.1, we also use a threshold m which occurs at time $t = \tau$. This threshold indicates at what concentration we would expect to see a change. For example, if gene 2 is on $x_2 = 1$, then as soon as the concentration of molecules activating that gene decreases below m we would expect the

An example of a Boolean network represented as a directed graph

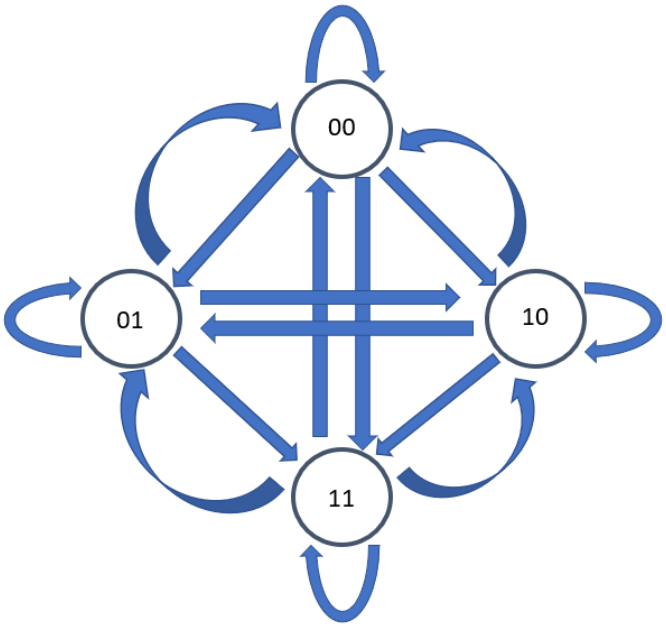
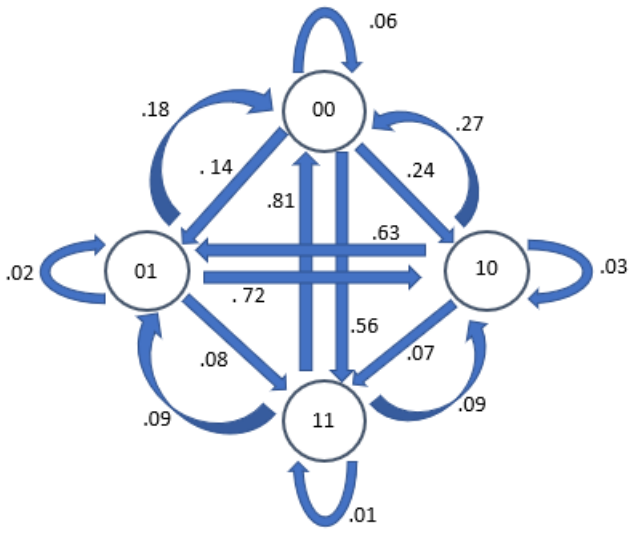


Figure 2: An example of a Boolean network represented as a directed graph

An example of a Boolean network with probabilities and a transition matrix



Four state spaces and two genes in each state space.
0: Represents a gene turned off
1: Represents a gene turned on

	$Gene_1$	$Gene_2$
Activation	.8	.7
Degradation	.9	.9

Transition Matrix

	00	01	10	11
00	.06	.14	.24	.56
01	.18	.02	.72	.08
10	.27	.63	.03	.07
11	.81	.09	.09	.01

Figure 3: An example of a Boolean network with probabilities and a transition matrix

gene to turn off $x_2 = 0$. Likewise, if gene 2 is off $x_2 = 0$, then as soon as the concentration of molecules repressing that gene decrease below m we would expect the gene to turn on $x_2 = 1$. In other words, given a gene in the off state,

$$x_i = \begin{cases} 0 & \text{if } t < \tau \\ 1 & \text{if } t \geq \tau \end{cases}$$

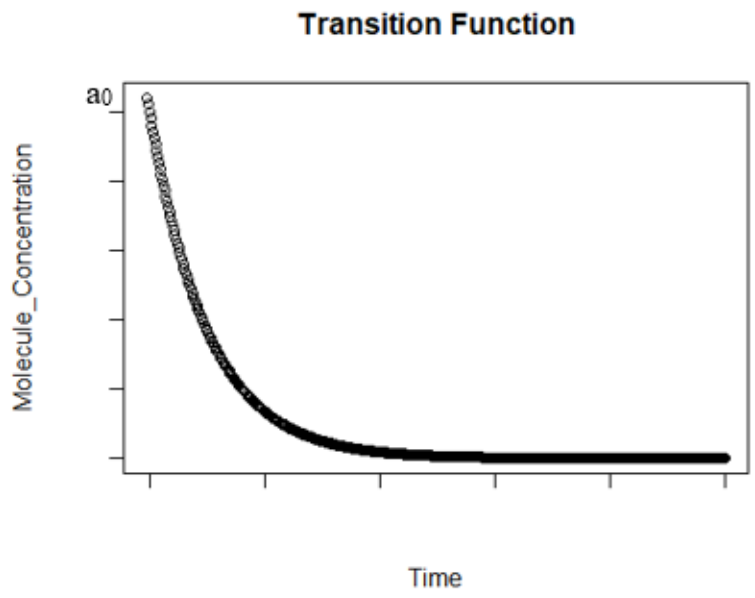
Or given a gene in the on state,

$$x_i = \begin{cases} 1 & \text{if } t < \tau \\ 0 & \text{if } t \geq \tau \end{cases}$$

Here, we use an exponential decay function which implies that transitions occur after a decrease in the proportion of molecules. However, biologically it is possible for transitions to occur after an increase in the proportion of molecules. Although we only use a decaying function for the transition function, we believe that any process which involves an increase in molecules could also be modeled with this exponential decay by changing the way you number the y-axis such that the $\lim_{t\to\infty} f(t) = \kappa$, where κ is the natural carrying capacity of appropriate molecules in a network's transition. In general, we would expect genes that fall below the threshold m to change states, but this is not always the case due to the stochastic nature of these processes. Equation II.1 and the transition matrix together create a system that incorporates both a time delay and stochasticity.

Murrugarra et al. (2012) and Akman et al. (2018) both utilize a model similar to the one described here. In [28] the edge propensities are kept constant and they work under the assumption that even if a reaction is supposed to occur, there is no guarantee that a transition will take place or even that the correct transition will take place [28]. Akman et al. argued that the propensities discussed in [28] are not likely to remain constant and applied a beta distribution to obtain propensities for state transitions along the edges of a network [2]. These models generally focused on obtaining the propensities of

(a)



(b)

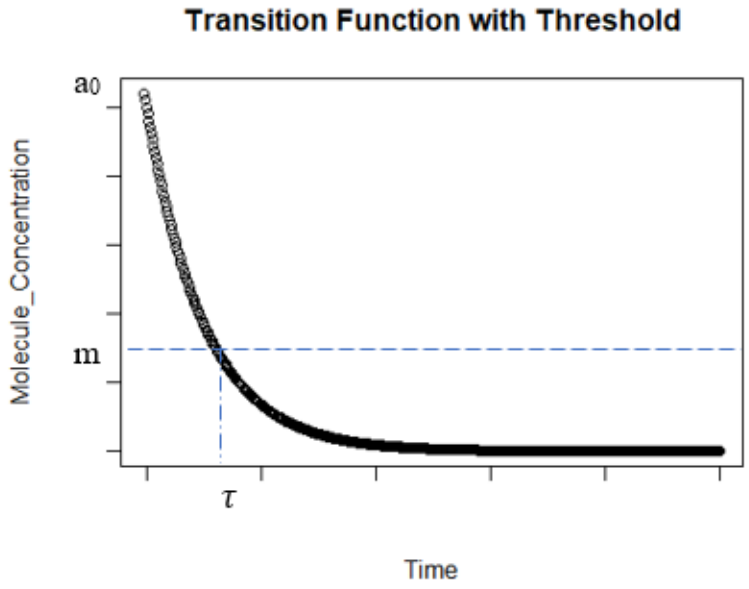


Figure 4: **(a)** The transition function is shown. The plot represents the number of molecules present in system at a time t. **(b)** The transition function is shown with a threshold at m. Once the concentration of molecules dips below the threshold m we expect the gene to transition from one state to another at time τ.

state transitions and overall network variance, but paid little attention to capturing variability in individual gene to gene interactions. Here, we use the same model, but focus on gene to gene variability and propensities. First, we look at the transition function and examine how rates of decay and thresholds impact the variation among gene transitions. Second, we revisit the beta distribution described in Akman et al, but instead of finding transition propensities using the statistical distribution, we use the beta distribution to determine the activation and degradation propensities. Third, we explore the use of spectral density to evaluate variation and propensities in a GRN. Finally, we compare these results and discuss areas that need improvement and future directions we hope to explore.

II.1 Method 1: Transition Function

Previous studies using the transition function and a Boolean network have focused on adding stochasticity into propensities and time delays in between interactions [2, 28]. Variation in Akman et al. [2] was found based on the beta distribution of each transition variability. Here, we investigate the variation inherent to the transition function itself. Recall that the transition function represents the proportion of molecules needed in order to transition between the on and off state of a gene. Once the number of molecules reaches a threshold m at time $t = \tau$, i.e. $X(\tau) = m$, a state to state transition occurs. Arbitrarily, when $X(t) \geq m$ the gene is considered on and when $X(t) < m$ the gene is considered off. This change in state could easily be reversed without loss of generality.

In the transition function, *Equation* II.1, we let $a_0 = 1$ so that the initial concentration of molecules, $f(0)$, can be considered as 100% and the threshold m can represent a percentage of molecules needed in order to activate or degrade a gene. Since the threshold m is determined by the strength of chemical bonds between the molecules and their binding sites, as well as the locations and number of binding sites, and therefore does not vary significantly from cell to cell [4].

In this study, we allow the rate of decay within the concentration of molecules to

vary so that we can measure the effects on variance in a state to state transition. This variance is biologically relevant because some stochasticity in GRNs is likely a direct result of competition between ribosomes and RNase. Ribosomes are responsible for translating mRNA into protein synthesis. RNase E signals the degradation of mRNA after transcription and before translation, however, mRNA that binds to a ribosome will undergo translation and lead to protein production [27]. Whether mRNA degrades or leads to the production of protein is thought to be determined by its proximity to either RNase E or a ribosome [27].

To model the behavior of the transition funciton we use the programming language R. Let $k \sim U(0.1, 10)$, $m = 0.4$, $a_0 = 1$, and $\tau = \frac{-1}{k} * log(\frac{m}{a_0})$. Using this information we find when the transition function will cross the threshold for a given value of k. Then, we replicate this process 1000 times to examine how changes in k affect τ.

After examining the behavior of the transition function, we also investigate how stochasticity plays a role in state to state transitions. In a deterministic model, once the concentration of molecules drops below the threshold m we would expect there to be a transition between states. However, in a stochastic model we only expect a proportion of these changes to happen each time. In this study we assume that only 75% of the genes that were supposed to change states actually have changed states. The choice of 75% was predetermined and compared to other propensities without any significant changes to results. We then record every time 5% of the genes pass the threshold. This process was also simulated in R. Out of the 1000 replicates, each transition function that passed the threshold at a given time τ had a 75% probability of either changing states or remaining the same. This process was repeated for $m = \{0.1, 0.4, 0.9\}$. For each time 5% of the genes pass the threshold, we record how much variation is present in the interval and plot this variation over time.

II.2 Results 1: Transition Function

In *Figure* 5 we illustrate the behavior of the transition function, $X(t) = x_0 e^{-kt}$. The shape of the transition function and the time that the curve passes the threshold m are dependent on the value of the decay rate k. Higher values of k correlate with faster rates of decay and a shorter time τ that is needed for the curve to pass the threshold. Lower values of k correlate with slower rates of decay and a longer time period τ to cross the threshold. Concentrations of molecules that dip below the threshold are expected to change states.

In *Figure* 6a the relationship between the rate of decay k and the time τ that it takes to cross the threshold m is shown. A system that involves molecules with high decay rates, such as $k = 8$, will have faster response times, $\tau \leq 0.2$. In comparison, systems with molecules that take a longer time to decay $k = 1$ will take longer to change from state to state in a system, $\tau > 0.2$. In reality, the rate at which proteins can be produced is limited by the speed of transcription and translation as well as the capacity for ribosomes to assemble amino acids [4].

Although the threshold for state changes to occur does not vary significantly [4], we show how the threshold, m, can affect the time τ to change states in *Figure* 6b. Here, we see that when m is approximately 90% of the molecules present $\tau_{.9}$, will be much smaller than $\tau_{.1}$ when m is approximately 10%. Therefore, when m is high, the rate of decay k will have a smaller impact on the network than when m is low.

Overall, *Figures* 5 and 6 illustrate that GRNs with the ability to change the number of molecules quickly will have faster response times to any environmental changes. In addition, GRNs that only require small changes in the number of molecules in order to trigger a state change will also have faster response times. Fast response times in a system lead to lower fluctuations and thus lower cell-cell variability [4]. These figures show what we expect to occur in a GRN without stochastic behavior. Since it is well known that regulatory networks have stochastic behavior [2, 4, 12, 20, 28], finding ways to incorporate stochasticity and variation within mathematical models may prove beneficial such that

they may be better equipped for capturing variability in the future.

In *Figures* 7, 8, and 9, two sets of points are shown in each graph. In these figures we see the proportion of genes that have changed state by the indicated time step τ on the x-axis. The genes change states at different time steps because k varies uniformly from 0.1 to 10. The light gray set of points represent the deterministic outcome of a gene transition. The color coded sections of each of these figures represents 5% intervals of genes. The majority of the genes transition into a new state within the first quarter of total time steps. In *Figure* 7 it takes roughly 20 time steps for all of the genes to transition. This is considerably more time steps compared to the roughly 7 time steps in *Figure* 8 and roughly 1 time step in *Figure* 9. In *Figure* 10 we see again that when $m = 0.1$ it takes much longer for all of the genes to transition and also that there are more genes with higher variance compared to genes with a threshold of $m = 0.9$.

The variation of each of these 5% intervals is plotted in *Figure* 11. In *Figure* 11a the threshold used is $m = 0.1$ and in *Figure* 11b the threshold used is $m = 0.9$. This figure shows that the longer it takes for a gene to react to its surroundings, the more variation will exist during gene transitions.

II.3 Method 2: Beta Distributed Propensities

The second method we applied in this work was the implementation of the beta distribution for the activation and degradation propensities. The beta distribution was utilized by Akman et al. (2018) to find the transition propensities between the states of genes. Here, we use the beta distribution to determine the activation and degradation propensities for each gene. Therefore, we will have a transition matrix similar to *Table* 3 where,

$$p_i^{\uparrow} \sim Beta(\alpha_{iA}, \beta_{iA}) \text{ and } p_i^{\downarrow} \sim Beta(\alpha_{iD}, \beta_{iD}).$$

Where $i = 1, 2, \cdots, n$ in general, and $i = 1, 2$ for this work with only two genes.

The use of the beta distribution is appropriate for a variety of reasons. First, the

beta distribution has a co-domain $\in [0, 1]$. Second, the beta distribution is commonly used when the set of random variables are probabilities, thus the extension to propensities is not unreasonable. Last, the beta distribution has been used by Wright to model other biological phenomena such as gene frequencies in population dynamics in 1937 [40]. The beta distribution uses two parameters, a shape (α) and a rate parameter (β), but can also be understood intuitively as a number of successes α and failures β. Thus, in a series of N trials where we would expect α successes, we would also expect $N - \alpha = \beta$ failures.

Our calculations for the transition matrix make use of some special functions. One of these functions is the gamma function defined by

$$\Gamma(z) = \int_0^\infty x^{(z-1)} e^{-x} dx.$$

When z is an integer n, the gamma function is also equivalent to

$$\Gamma(n) = (n-1)!.$$

Another function we use is the beta function defined by

$$\mathcal{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt.$$

The beta function is also equivalent to

$$\mathcal{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

One other function we use is the Gauss hypergeometric function denoted

$$HypG_{2F1}(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k x^k}{(c)_k k!}.$$

The transition matrix in *Table* 3 then has the following properties:

1. This matrix is stochastic and thus a transition matrix.

2. $(1-p_i^{\uparrow})$ and $(1-p_i^{\downarrow})$ each have a beta distribution.

3. The probability density function of the product of two beta distributions with shape parameters (a_1, b_1) and (a_2, b_2) for genes 1 and 2 respectively is

$$\Big(\frac{x^{a_2}\Gamma\left[b_1\right]\Gamma\left[a_1-a_2\right]\mathrm{HypG}_{2F1}\left[1-b_2, 1-a_1-b_1+a_2, 1-a_1+a_2, x\right]}{\Gamma\left[a1+b1-c\right]}+$$

$$\frac{x^{a_1}\Gamma\left[b_2\right]\Gamma\left[-a_1+a_2\right]\mathrm{HypG}_{2F1}\left[1-b_1, 1+a_1-b_2-a_2, 1+a_1-a_2, x\right]}{\Gamma\left[-a1+b_2+a_2\right]}\Big)*$$

$$\frac{1}{\left(x\mathcal{B}\left[a_1, b_1\right]\mathcal{B}\left[a_2, b_2\right]\right)}.$$

4. The expected value of the matrix is the expected value of each entry.

5. The expected value of each entry is the product of the expected values of each beta distributed activation or deactivation propensity, denoted E_i for G_i.

6. The variation of each transition between the states of two beta distributions with shape parameters (a_1, b_1) and (a_2, b_2) for genes 1 and 2 respectively is

$$\frac{{E_1}^2\left(1+a_1\right){E_2}^2\left(1+a_2\right)}{\left(E_1+a_1\right)\left(E_2+a_2\right)}-\left(E_1E_2\right)^2.$$

Justification for (1):

A stochastic matrix is an $n \times n$ matrix such that each entry is less than or equal to 1 and each row sum is equal to 1 [7]. In the Transition Matrix, $0 \leq p_i^{\uparrow} \leq 1$ and $0 \leq p_i^{\downarrow} \leq 1$. Therefore, $0 \leq (1-p_i^{\uparrow}) \leq 1$, and $0 \leq (1-p_i^{\downarrow}) \leq 1$. Let x be the product of any of these terms. Then $0 \leq x \leq 1$. Furthermore,

- $(1-p_1^{\uparrow})(1-p_2^{\uparrow})+(1-p_1^{\uparrow})p_2^{\uparrow}+p_1^{\uparrow}(1-p_2^{\uparrow})+p_1^{\uparrow}p_2^{\uparrow}=1$
- $(1-p_1^{\uparrow})p_2^{\downarrow}+(1-p_1^{\uparrow})(1-p_2^{\downarrow})+p_1^{\uparrow}p_2^{\downarrow}+p_1^{\uparrow}(1-p_2^{\downarrow})=1$

- $p_1^{\downarrow}(1-p_2^{\uparrow})+p_1^{\uparrow}p_2^{\downarrow}+(1-p_1^{\downarrow})(1-p_2^{\uparrow})+(1-p_1^{\downarrow})p_2^{\uparrow}=1$
- $p_1^{\downarrow}p_2^{\downarrow}+p_1^{\downarrow}(1-p_2^{\downarrow})+(1-p_1^{\downarrow})p_2^{\downarrow}+(1-p_1^{\downarrow})(1-p_2^{\downarrow})=1$

Justification for (2):

Let $A\sim Beta\left(a_A,b_A\right)$ and $Z=1-A$. The probability density function of A is

$$f_A\left(A\right)=\frac{\left(A\right)^{a_A-1}*\left(1-A\right)^{b_A-1}}{\mathcal{B}\left[a_A,b_A\right]}.$$

The transformation of $Z=1-A$ creates the new pdf

$$f_Z\left(Z\right)=\frac{\left(1-Z\right)^{a_A-1}*\left(1-\left(1-Z\right)\right)^{b_A-1}}{\mathcal{B}\left[a_A,b_A\right]}.$$

Which simplifies to

$$f_Z\left(Z\right)=\frac{\left(1-Z\right)^{a_A-1}*\left(Z\right)^{b_A-1}}{\mathcal{B}\left[a_A,b_A\right]}.$$

This implies that Z has a beta distribution with shape parameters b_A and a_A.

$$Z\sim Beta\left(b_A,a_A\right)$$

Therefore the distribution of $Z=1-A$, where A is a beta distribution, is also a beta distribution.

Justification of (3):

The pdf of $f\left(x,y\right)$ when $X=A*B$ and $Y=B$ where both A and B are beta distributions. $A\sim Beta\left(a_A,b_A\right)$ and $B\sim Beta\left(a_B,b_B\right)$ is shown below and obtained though a transformation of variables.

$$f\left(x,y\right)=\frac{\left(\frac{x}{y}\right)^{a_A-1}*\left(1-\frac{x}{y}\right)^{b_A-1}*\left(y\right)^{a_B-1}*\left(1-y\right)^{b_B-1}}{\mathcal{B}\left[a_A,b_A\right]*\mathcal{B}\left[a_B,b_B\right]*y}$$

for $0 < x \leq y$ and $0 < y < 1$.

$$\int_x^1 f(x,y)dy =$$

$$\Big(\frac{x^{a_2}\Gamma\left[b_1\right]\Gamma\left[a_1-a_2\right]\mathrm{HypG}_{2F1}\left[1-b_2,1-a_1-b_1+a_2,1-a_1+a_2,x\right]}{\Gamma\left[a1+b1-c\right]}+$$

$$\frac{x^{a_1}\Gamma\left[b_2\right]\Gamma\left[-a_1+a_2\right]\mathrm{HypG}_{2F1}\left[1-b_1,1+a_1-b_2-a_2,1+a_1-a_2,x\right]}{\Gamma\left[-a1+b_2+a_2\right]}\Big)*$$

$$\frac{1}{\left(x\mathcal{B}\left[a_1,b_1\right]\mathcal{B}\left[a_2,b_2\right]\right)}$$

Justification of (4):

Expected value is a positive linear operator. Results follow.

Justification of (5):

The expected value of X is dependent on the shape parameters of $A \sim Beta\left(a_A, b_A\right)$ and $B \sim Beta\left(a_B, b_B\right)$.

$$\int_0^1 \int_0^y x * f\left(x,y\right) dx\ dy = \frac{a_A\Gamma\left[1+a_B\right]\Gamma\left[a_B+b_B\right]}{\left(a_A+b_A\right)\Gamma\left[a_B\right]\Gamma\left[1+a_B+b_B\right]}$$

The expected value of X is also the same as the product of expected values of A and B. This follows naturally because A and B are independent, but this property can also be motivated by the relationship between the shape parameters.

Recall that the expected value of A is $\frac{a_A}{a_A+b_A}$ and therefore, based on a proportional relationship between a and b , the expected value E_A or E_B creates the relationship: $b_A = \frac{1-E_A}{E_A}a_A$ and similiarly $b_B = \frac{1-E_B}{E_B}a_B$

Substituting this expression into the expected value of X for b_A or b_B leads to the expression:

$$\frac{a_A\Gamma\left[1+a_B\right]\Gamma\left[a_B+\frac{1-E_B}{E_B}a_B\right]}{\left(a_A+\frac{1-E_A}{E_A}a_A\right)\Gamma\left[a_B\right]\Gamma\left[1+a_B+\frac{1-E_B}{E_B}a_B\right]}.$$

This expression can be reduced to:

$$\rightarrow \quad \frac{a_A a_B \Gamma\left[a_B\right] \Gamma\left[a_B + \frac{1-E_B}{E_B} a_B\right]}{a_A \left(1 + \frac{1-E_A}{E_A}\right) \Gamma\left[a_B\right] \left(a_B + \frac{1-E_B}{E_B} a_B\right) \Gamma\left[a_B + \frac{1-E_B}{E_B} a_B\right]}$$

$$\rightarrow \quad \frac{a_A a_B \Gamma\left[a_B\right] \Gamma\left[a_B + \frac{1-E_B}{E_B} a_B\right]}{a_A \left(1 + \frac{1-E_A}{E_A}\right) \Gamma\left[a_B\right] a_B \left(1 + \frac{1-E_B}{E_B}\right) \Gamma\left[a_B + \frac{1-E_B}{E_B} a_B\right]}$$

$$\rightarrow \quad \frac{1}{\left(1 + \frac{1-E_A}{E_A}\right) \left(1 + \frac{1-E_B}{E_B}\right)} \quad \rightarrow \quad \frac{E_A E_B}{1}$$

$$\rightarrow \quad E_A E_B.$$

Justification for (6):

The variance of X is also dependent on the shape parameters of $A \sim Beta\left(a_A, b_A\right)$ and $B \sim Beta\left(a_B, b_B\right)$

$$Var\left(X\right) = \frac{\Gamma\left[2 + a_A\right] \Gamma\left[b_B\right] \Gamma\left[b_A\right] \Gamma\left[2 + a_B\right]}{\beta\left[a_A, b_A\right] \beta\left[a_B, b_B\right] \Gamma\left[2 + a_A + b_A\right] \Gamma\left[2 + a_B + b_B\right]} - E\left(X\right)^2$$

As before, we substitute the expressions $b_A = \frac{1-E_A}{E_A} a_A$ and $b_B = \frac{1-E_B}{E_B} a_B$ into the variance of X and we obtain the expression:

$$\frac{\Gamma\left[2 + a_A\right] \Gamma\left[\frac{1-E_B}{E_B} a_B\right] \Gamma\left[\frac{1-E_A}{E_A} a_A\right] \Gamma\left[2 + a_B\right]}{\beta\left[a_A, \frac{1-E_A}{E_A} a_A\right] \beta\left[a_B, \frac{1-E_B}{E_B} a_B\right] \Gamma\left[2 + a_A + \frac{1-E_A}{E_A} a_A\right] \Gamma\left[2 + a_B + \frac{1-E_B}{E_B} a_B\right]} - \left(E_A E_B\right)^2$$

$$\rightarrow \frac{\left(1 + a_A\right)\left(1 + a_B\right)}{\left(1 + a_A + \frac{1-E_A}{E_A} a_A\right) \left(1 + \frac{1-E_A}{E_A}\right) \left(1 + \frac{1-E_B}{E_B}\right) \left(1 + a_B + \frac{1-E_B}{E_B} a_B\right)} - \left(E_A E_B\right)^2$$

$$\rightarrow \frac{{E_A}^2 \left(1 + a_A\right) {E_B}^2 \left(1 + a_B\right)}{\left(E_A + a_A\right)\left(E_B + a_B\right)} - \left(E_A E_B\right)^2.$$

II.4 Results 2: Beta Distributed Propensities

II.4.1 **Effects on Variation** In the following sections we discuss how different parameters and values effect the variation of a state transition. First we see how changing shape parameters of the beta distribution but keeping the expected value constant effect variation. Then we hold the shape parameters constant and see how changes in the expected value effect variation of transitions.

Holding Expected Value Constant

We explore how the variation of transitions is affected by different shape parameters when the expected value of an entry is held constant. In *Figure* 12 we see the relationship between shape parameters and variation when the entry X has 4 different expected values with corresponding gene transition probabilities (a), (b), (c), and (d). These figures reveal that high variation occurs when the shape parameter of one gene is high and the other is low. High variation also appears to be correlated with higher expected values (12c - 12d). Since the shape of beta distributions is dependent on the shape parameters, there is no set distribution for each of the graphs being represented. The variation for each of these figures indicates that variation is highest when the shape parameter of gene 1 α_1 is high and the shape parameter of gene 2 α_2 is low. A closer look reveals that when the expected value of a transition is low for both genes, *Figure* 12a, the maximum variance is less than 0.1. However, when the expected value for both genes is high, *Figure* 12d, the maximum variance is greater than 10.

Holding Shape Parameters Constant

Here we look at how the variation is affected by changes is expected value, but when one of the shape parameters is held constant. Here, the relationship between α, β, and the expected value allow us to look at how changes in expected value affect the variation in a transition between gene states. In *Figure* 13 the x-axis is the expected value of gene 1 and

the y-axis is the expected value of gene 2. We see that in general, when one gene has an expected value of approximately 0.5 and the other has an expected value close to 1 the variance of the transition will be high.

II.4.2 **Variance of the GRN** Given independence between each of the gene transitions in the network, the variance of the entire network will have a covariance of zero and the network's variance will equal the variance of the product of the two beta distributions. The variance of the entire network is shown in *Tables* 4.

Table 4: Variation of the entire GRN

00	01	10	11
$\mathrm{V}((1-p_1^{\uparrow})(1-p_2^{\uparrow}))$	0	0	0
0	$\mathrm{V}((1-p_1^{\uparrow})(1-p_2^{\downarrow}))$	0	0
0	0	$\mathrm{V}((1-p_1^{\downarrow})(1-p_2^{\uparrow}))$	0
0	0	0	$\mathrm{V}((1-p_1^{\downarrow})(1-p_2^{\downarrow}))$

Recall that the variation of each edge of two beta distributions with shape parameters (a_1, b_1) and (a_2, b_2) for genes 1 and 2 respectively is

$$\frac{{E_1}^2\,(1+a_1)\,{E_2}^2\,(1+a_2)}{(E_1+a_1)\,(E_2+a_2)} - (E_1E_2)^2\,.$$

II.5 Method 3: Spectral Density

The third and final method which we used to investigate ways to incorporate variance into gene state transitions involved using spectral density. The general idea behind spectral density is to take a finite set of static data and estimate how the total power is distributed across frequencies. Spectral analysis in particular has been used in a variety of fields such as psychology for heart rates [18], medical fields for fetal heart rates [38], geology for geological formation [5], and bioinformatics for gene prediction [25].

Let $y(t)$ for $t = 0, 1, 2, 3, \cdots, T$ be a discrete-time data sequence from time $t = 0$ to

$t = T$, and assume that $y(t)$ has finite energy such that $\sum_{t=-\infty}^{\infty} y(t) < \infty$. Let $r(q)$ be the autocovariance sequence obtained from the lag q between sampling data such that $r(q) = E[y(t) * y(t-q)]$. Lag is the amount of time between measuring signal output. Then, the power spectral density (PSD) of $r(q)$ is

$$\phi(\omega) = \int_{-\infty}^{\infty} r(q) e^{-i\omega q} dq.$$

Where $\phi(\omega)$ is the PSD and Fourier transform of $r(q)$, ω is the frequency of the signal, and $i = \sqrt{-1}$. Then, $\phi(\omega)$ is the power at different bands of frequency [36]. Therefore, the integral of $\phi(\omega)$ across all possible frequencies is the total power of the signal. Furthermore, since $r(0) = E(y(t)^2)$, and if we assume that $E(y(t)) = 0$, the integral of $\phi(\omega)$ across all frequencies is also the variance [36].

Although we do not have data for the transitions of gene states, we use the transition function to demonstrate how this process would work for variance estimation. In this case, the transition function II.1 is equal to $r(t)$ so that the Fourier transform of $r(t)$ will result in the PSD, and the integral of the PSD will be the variance of the system.

Biologically it is possible for the decay rate to vary [41]. If k varies, it should have probability density function, $g(k)$. This alters the transition function described in [28] and [12] such that the rate of transcription is $a_*(x) = kg(k)x$. Thus, $f_*(t) = a_0 e^{-k_* t}$, where $k_* = kg(k)$. Here, we allow the decay rate k to vary according to a specific statistical distribution.

In order to use a distribution for k, it is necessary to determine the variation of k. However, the decay rates of interaction molecules, like mRNA, are not well studied [41]. In this section, we explore how the variation of interactions is affected by different distributions of k.

To do this we will generate different values of k from different distributions and multiply them by their probability density function, $g(k)$, in order to obtain k_* and $f_*(t)$.

The first moment of $f_*(t)$ is given by the Fourier Transform which is also the power spectral density in this case. The units of frequency that we use are cycles per sampling period. Then, we will calculate the variation of these interactions by integrating the power spectral density function.

We will include the entire time interval for these interactions, but we will only consider frequencies ranging from 0 to 100 periods per time interval. This interval has been chosen because it is unlikely that we will see frequencies greater than 100. For example, $cos(x)$ has two frequencies because it has nonzero energy at two different amplitudes. Based on our function, it seems more likely that we only have one frequency per time interval. Either way, the interval $[0, 100]$ should include our frequency.

II.5.1 **Uniform Distribution** The uniform distribution describes scenarios where every value of k is equally likely to be chosen. Since k is a rate of decay, we let $k \in [0, 1]$. Then, the expected value of $g(k)$ is $E(g(k)) = 1$ and the power spectral density function is

$$\phi_*(\omega) = \frac{a_0}{i\omega + k_*}$$

Let $a_0 = 1$ and find $\displaystyle\int_0^{100} \frac{1}{i\omega + k_*} d\omega$ to find the total variance across frequencies, $Var(\omega)$

$$Var(\omega) = -i \log\left(\frac{100i}{k_*} + 1\right)$$

Using polar coordinates and Euler's Theorem, we can rewrite $Var(\omega)$ as the following

$$Var(\omega) = -i \log\left(\sqrt{\left(\frac{100}{k_*}\right)^2 + 1}\right) - \arctan\left(\frac{100}{k_*}\right).$$

At 100 random values of k_* we obtain variations shown in *Figures* 17 and 18.

II.5.2 **Beta Distributions** We then repeat this process for the beta distribution with different shape parameters. The probability density function of the beta distribution is

$$g(k) = \frac{k^{\alpha-1}e^{-k\beta}\beta^{\alpha}}{\Gamma(\alpha)}.$$

Here, we generate three different beta distributions with shape parameters $(\alpha, \beta) = [(4,1),(2,2),(3,5)]$. Notice that $g(k)$ is not a function of t, so all power spectral density functions will be in the form of

$$\phi_*(\omega) = \frac{a_0}{i\omega + k_*}.$$

Then, variation of each beta distribution, when $\omega \in (0, 100)$, will be of the form

$$-i \log\left(\frac{100i + k_*}{k_*}\right).$$

which can be rewritten as

$$Var(\omega) = -i \log\left(\sqrt{\left(\frac{100}{k_*}\right)^2 + 1}\right) - \arctan\left(\frac{100}{k_*}\right).$$

using a process similar to the one described for the uniform distribution.

Case I: $\alpha = 4$**,** $\beta = 1$

The first case for beta distributions that we investigate is when the shape and rate parameters are $\alpha = 4$ and $\beta = 1$. This distribution can be seen in *Figure* 14
Then, we obtain the probability density function:

$$g(k) = \frac{k^3 e^{-4}}{6}.$$

From here, we find the expected value of k, k_*, and the PSD, $\phi_*(\omega)$:

$$k_* = \frac{k^4 e^{-4}}{6}$$

$$\phi_*(\omega) = \frac{6}{6i\omega + k^4 e^{-4}}.$$

The resulting variation can be seen in *Figure* 19a.

Case II: $\alpha = 2$**,** $\beta = 2$

The second case for beta distributions that we investigate is when the shape parameters are $\alpha = 2$ and $\beta = 2$. *Figure* 15 shows the behavior of this distribution. The probability density function under these parameters is

$$g(k) = 4ke^{-2k}.$$

Then, we can find the expected value of k, $E(k) = k_*$, and the PSD, $\phi_*(\omega)$:

$$k_* = 4k^2 e^{-2k}$$

$$\phi_*(\omega) = \frac{1}{i\omega + 4k^2 e^{-2k}}$$

The resulting variation can be seen in *Figure* 19b.

$\alpha = 3$**,** $\beta = 5$

The third and final case that we investigated for the beta distribution was when the shape and rate parameters were $\alpha = 3$ and $\beta = 5$. From here, we obtain the probability density function:

$$g(k) = 62.5k^2 e^{-5k}$$

Then, the expected value of k, $E(k) = k_*$, and the PSD, $\phi_*(\omega)$ are:

$$k_* = 62.5k^3e^{-5k}$$

and

$$\phi_*(\omega) = \frac{1}{i\omega + 62.5k^3e^{-5k}}.$$

The resulting variation can be seen in *Figure* 19c.

II.6 Results 3: Spectral Density

Figures 17 and 18 show the behavior of variation for different values of k_*. The real variation is shown in *Figure* 17. This variation is linear and negative. The magnitude of variation is greater for lower values of k_*. In *Figure* 18 variation takes on a logarithmic shape and is also negative. The magnitude of variation decreases as k_* increases.

Figure 19 shows the variation from three different beta distributions, $\alpha = 4$ and $\beta = 1$, $\alpha = 2$ and $\beta = 2$, and $\alpha = 3$ and $\beta = 5$. Similar to the uniform distribution, there is a real and complex component to each of these distributions. The real component for each of these distributions is also linear, and the complex component is logarithmic. All variation shown is negative. All variation decreases as the magnitude of k_* decreases.

The transition function illustrated with different rates of decay

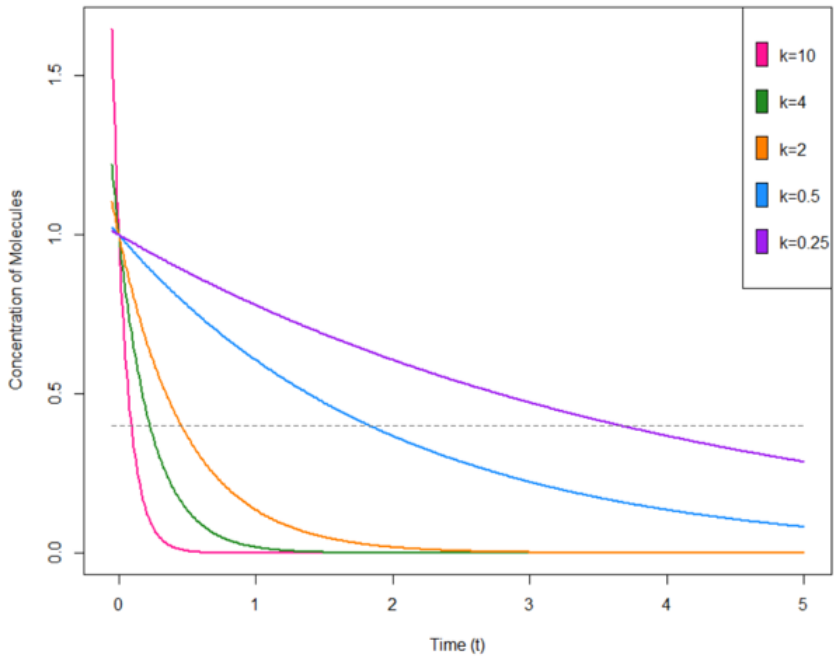


Figure 5: The transition function, $X(t) = x_0 e^{-kt}$, illustrated with different rates of decay, k, and a threshold, $m = 0.4$

An interaction between decay rates and the time it takes to cross the threshold

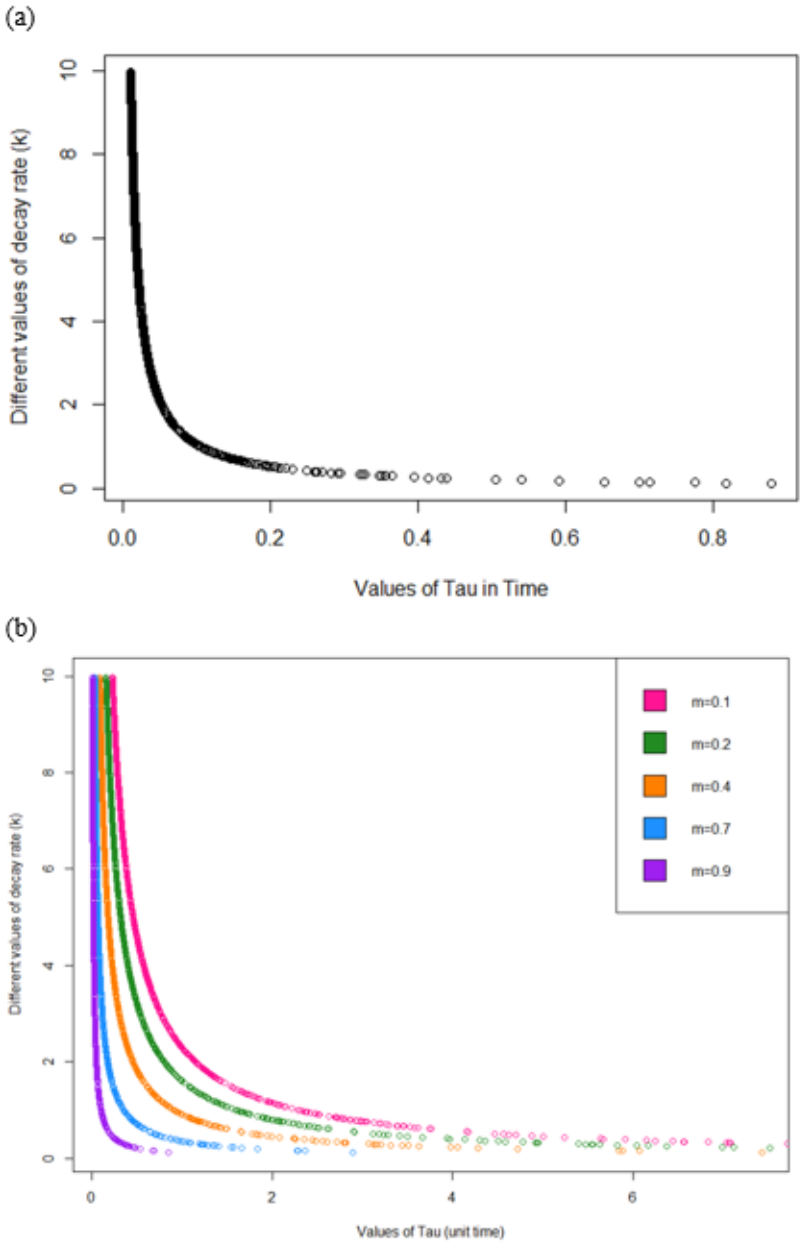


Figure 6: **(a)** How different decay rates, k, and time, τ, to cross the threshold, $m = 0.4$, interact. Tau is the time it takes for the proportion of molecules to cross a threshold that triggers a state change in the network. **(b)** How different decay rates, k, and time, τ, to cross a threshold, m, interact. This behavior is shown for different values of m to illustrate how m may impact each network.

The proportion of genes that have changed states at time τ when $m = 0.1$

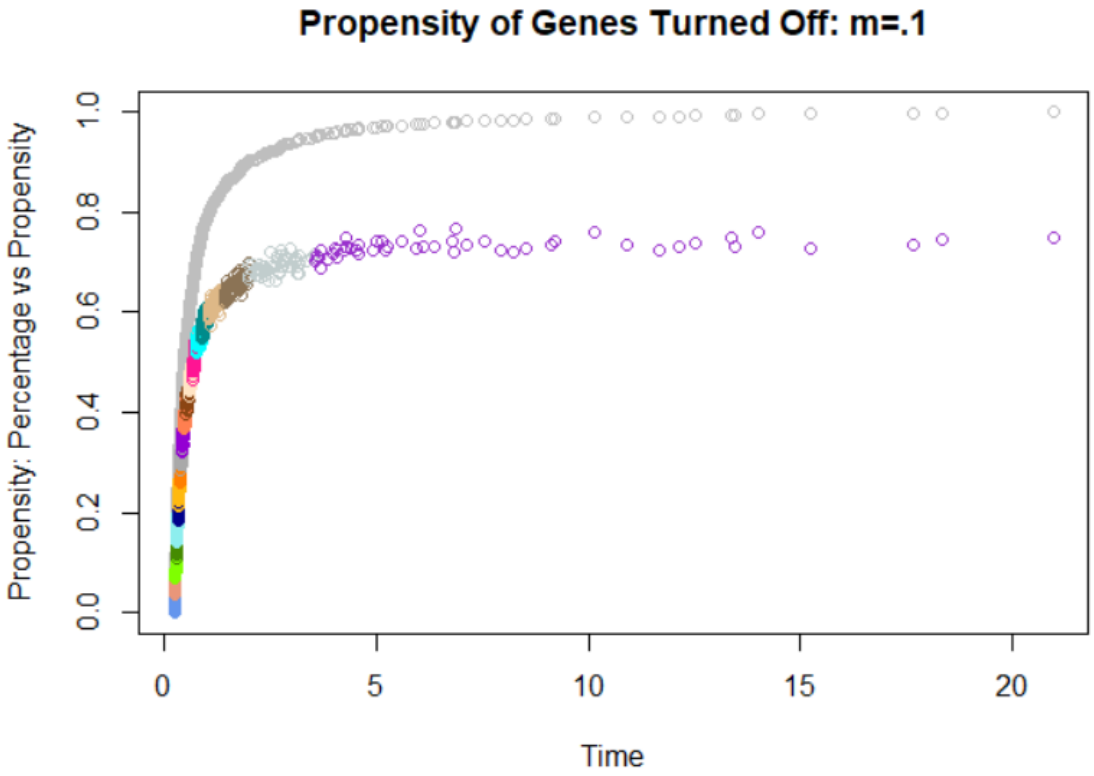


Figure 7: The proportion of genes that have changed states at time τ when $m = 0.1$

The proportion of genes that have changed states at time τ when $m = 0.4$

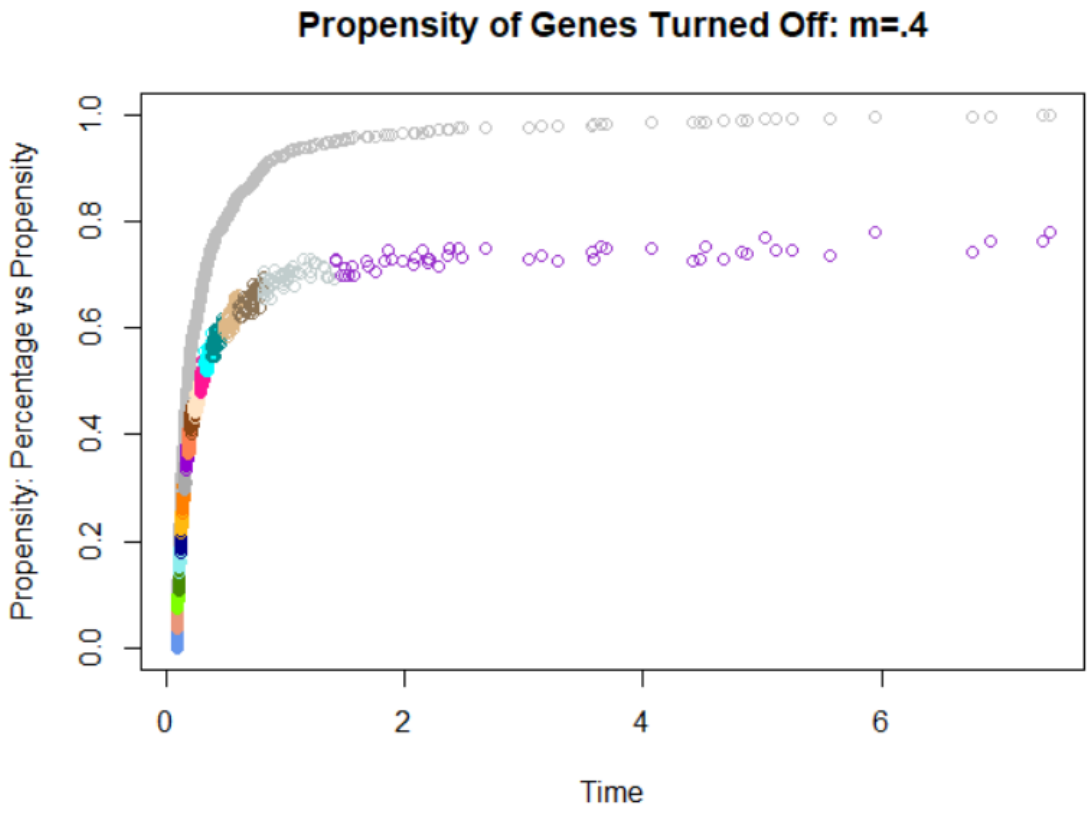


Figure 8: The proportion of genes that have changed states at time τ when $m = 0.4$

The proportion of genes that have changed states at time τ when $m = 0.9$

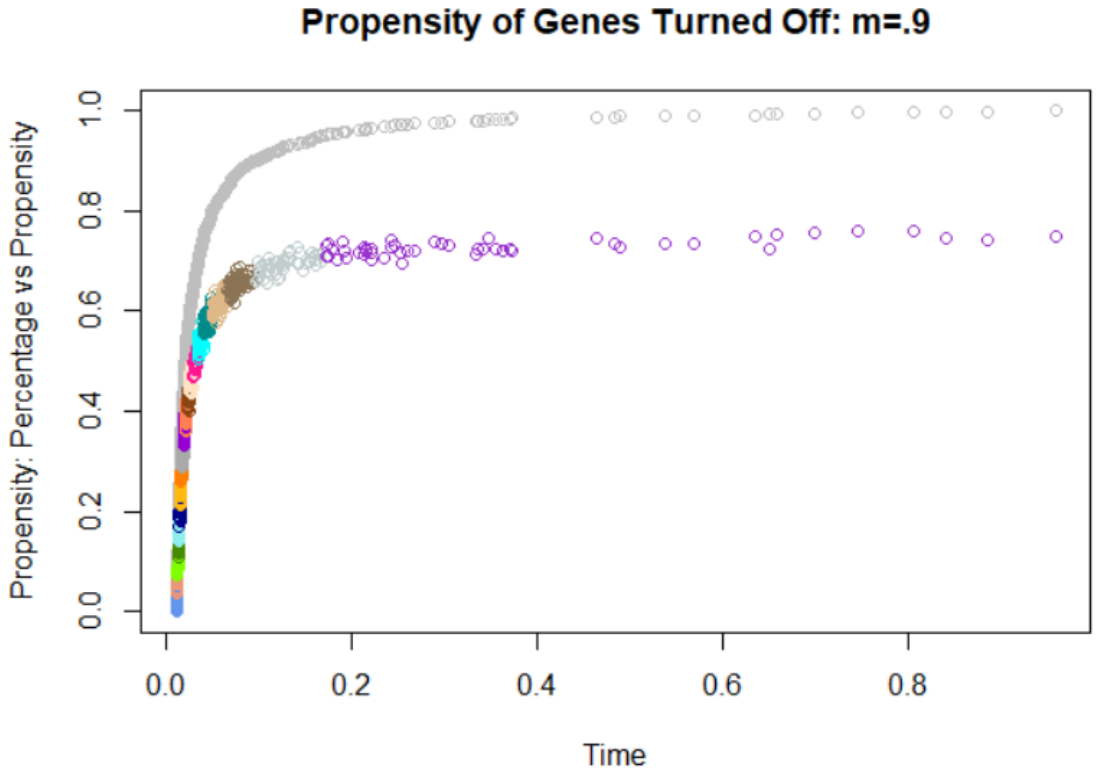


Figure 9: The proportion of genes that have changed states at time τ when $m = 0.9$

Graphs of the variation at every fifth percentile for $m = 0.1$ in black, $m = 0.4$ in red, and $m = 0.9$ in blue

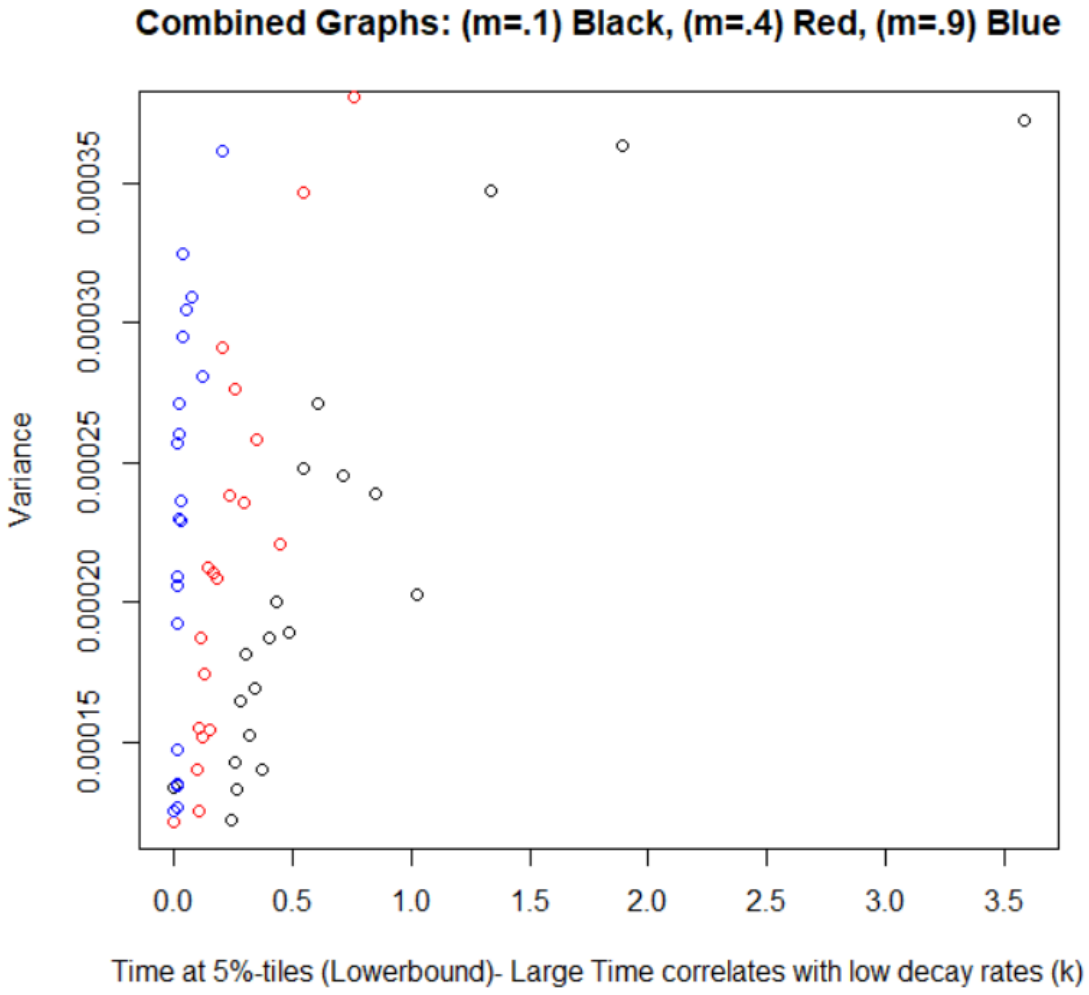


Figure 10: Graphs of the variation at every fifth percentile for $m = 0.1$ in black, $m = 0.4$ in red, and $m = 0.9$ in blue

A plot of how variation is affected by the time it takes for a transition to occur

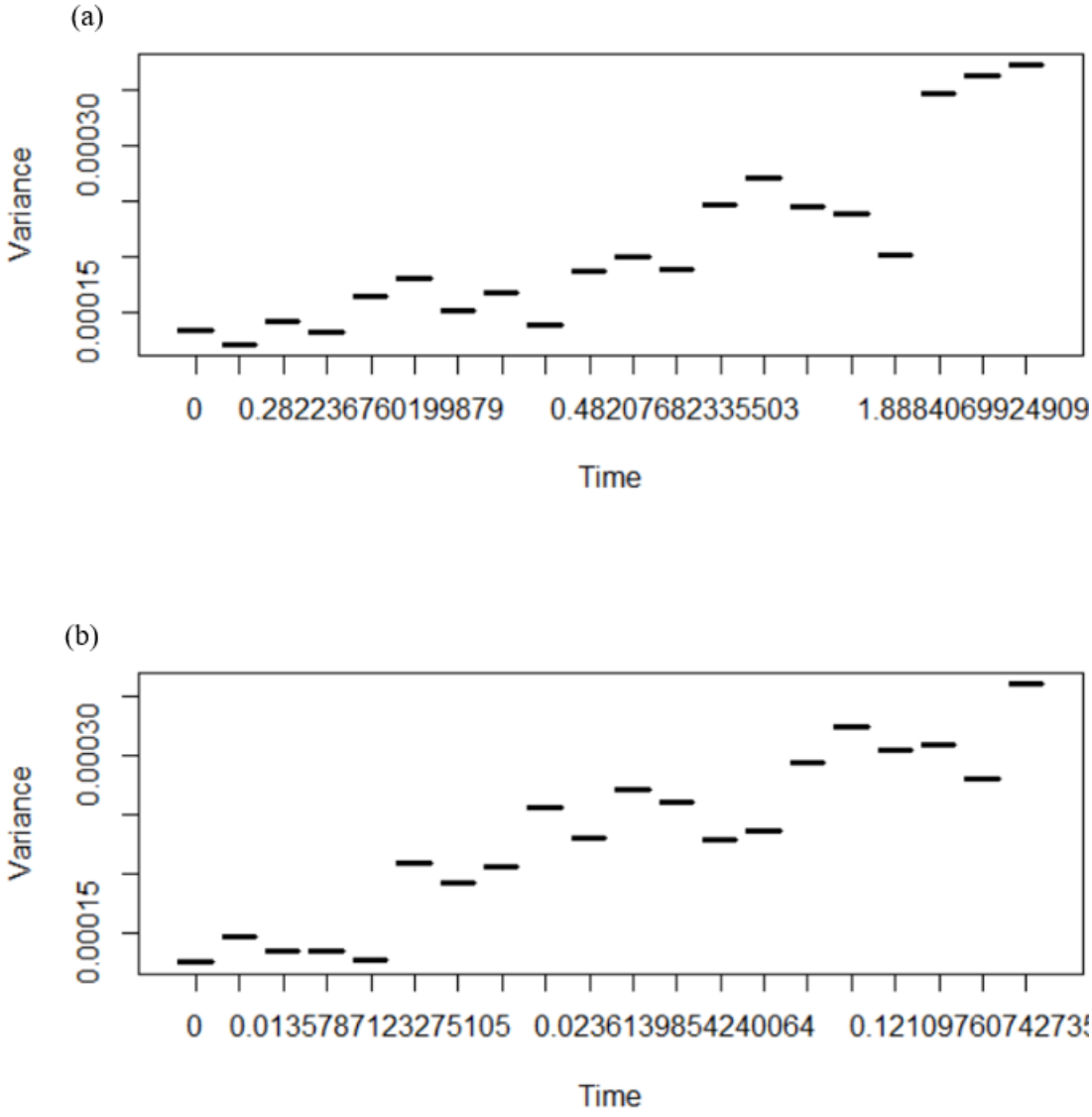


Figure 11: How variation changes with respect to the time it takes for the transition function to cross the threshold. **(a)** Threshold $m = 0.1$ **(b)** Threshold $m = 0.9$

Variation when the expected value of beta distributed propensities is held constant and shape parameters vary

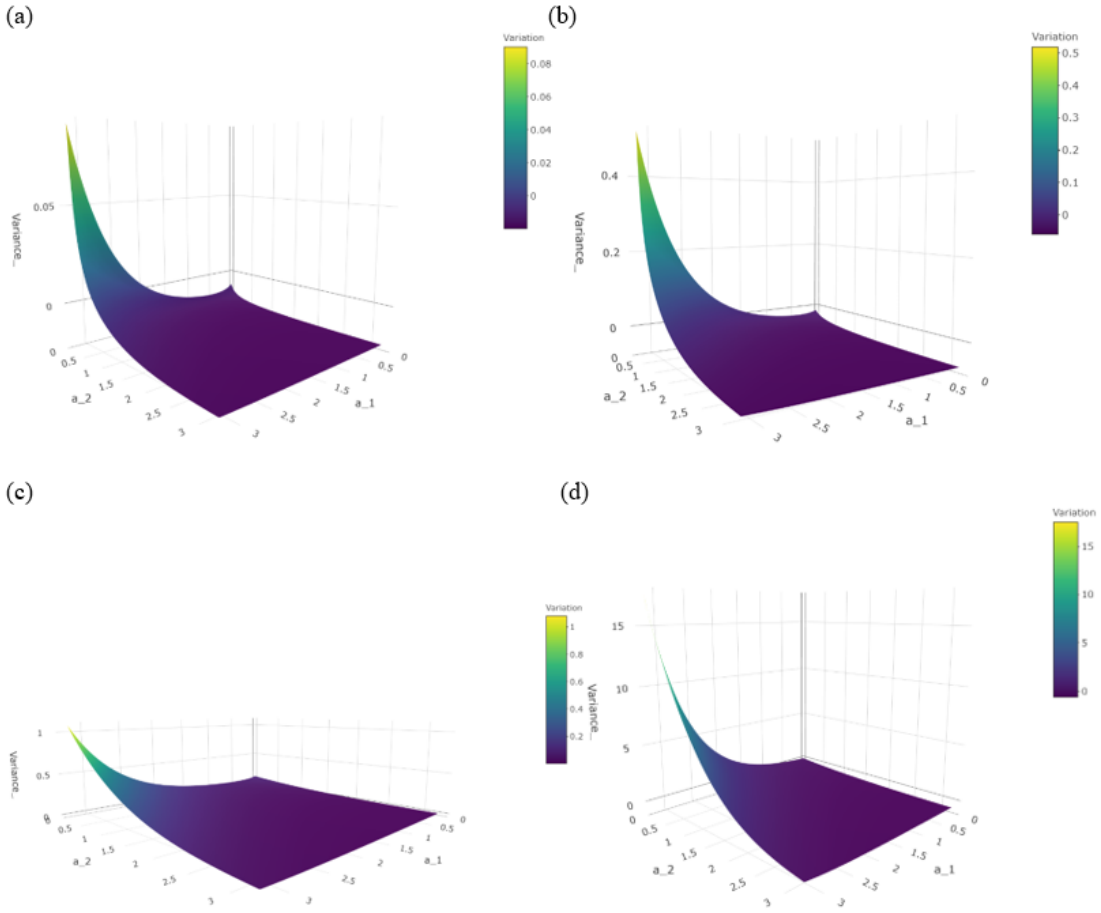


Figure 12: **(a)** $E(X) = 0.02$, the expected value of the transition in gene 1 is 0.2, and the expected value of the transition in gene 2 is 0.1. **(b)** $E(X) = 0.06$, the expected value of the transition in gene 1 is 0.3, and the expected value of the transition in gene 2 is 0.2. **(c)** $E(X) = 0.56$, the expected value of the transition in gene 1 is 0.8, and the expected value of the transition in gene 2 is 0.7. **(d)** $E(X) = 0.63$, the expected value of the transition in gene 1 is 0.9, and the expected value of the transition in gene 2 is 0.7.

How the expected value of beta distributed propensities affects variation of transitions

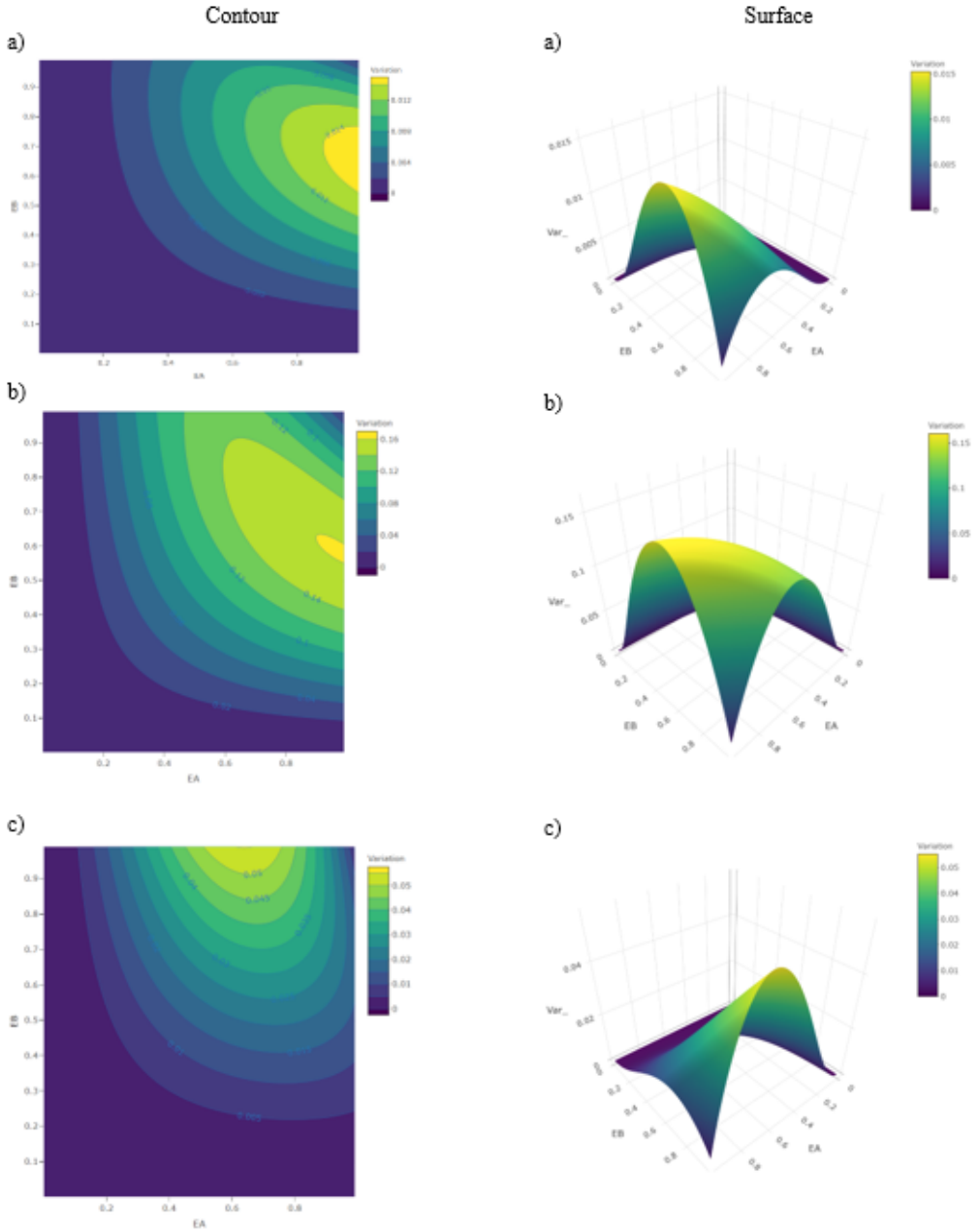


Figure 13: Three surfaces measuring variation when the expected value varies. Different shape parameters are used for each set of figures: **(a)** $a_a = 20$ and $a_b = 9$ **(b)** $a_a = 0.5$ and $a_b = 0.3$ **(c)** $a_a = 2$ and $a_b = 9$.

The beta distribution with shape parameters $\alpha = 4$ and $\beta = 1$

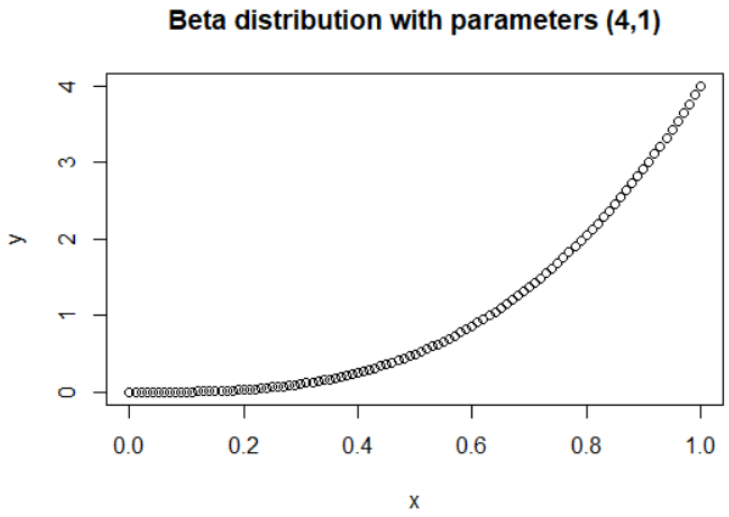


Figure 14: The beta distribution with shape parameters $\alpha = 4$ and $\beta = 1$

The beta distribution with shape parameters $\alpha = 2$ and $\beta = 2$

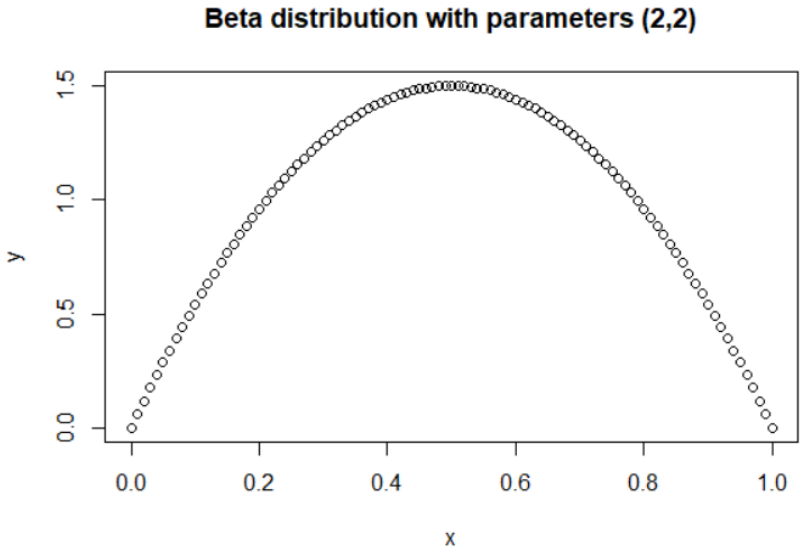


Figure 15: The beta distribution with shape parameters $\alpha = 2$ and $\beta = 2$

The beta distribution with shape parameters $\alpha = 3$ and $\beta = 5$

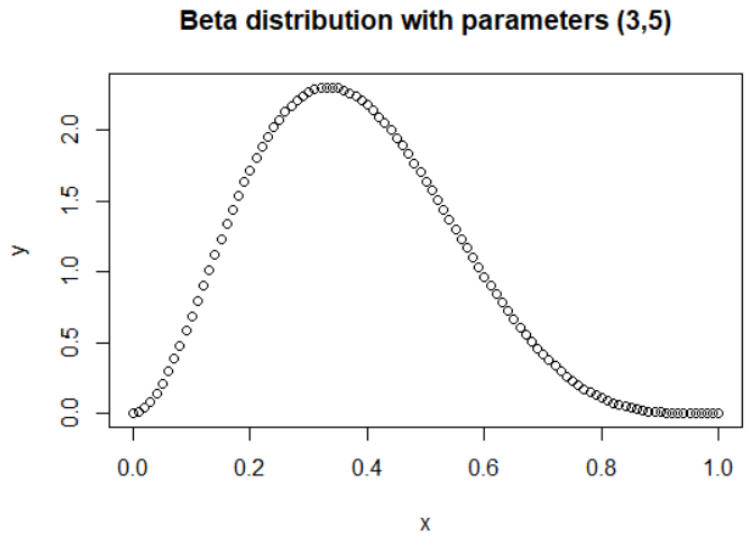


Figure 16: The beta distribution with shape parameters $\alpha = 3$ and $\beta = 5$

Real variation from spectral density is plotted when the transition rate has a uniform distribution

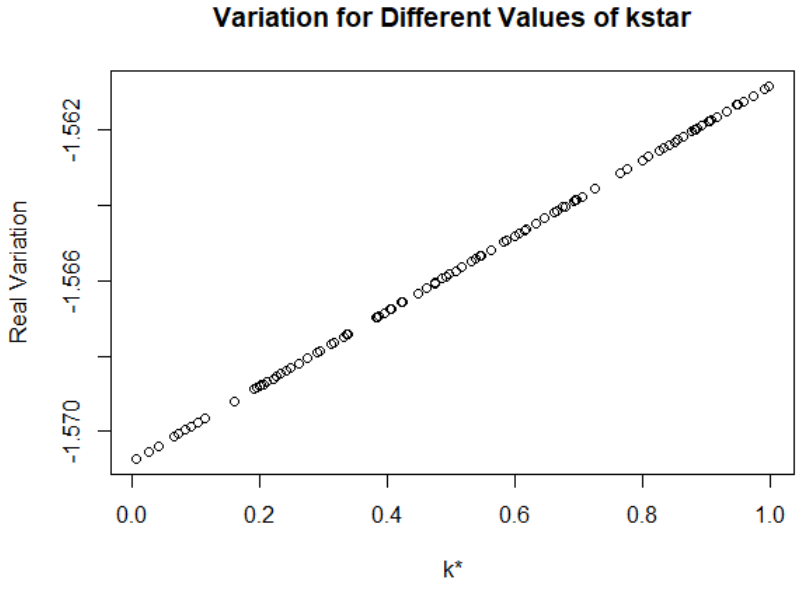


Figure 17: A plot of k_* from a uniform distribution vs. the real variation of the interaction

Complex variation from spectral density is plotted when the transition rate has a uniform distribution

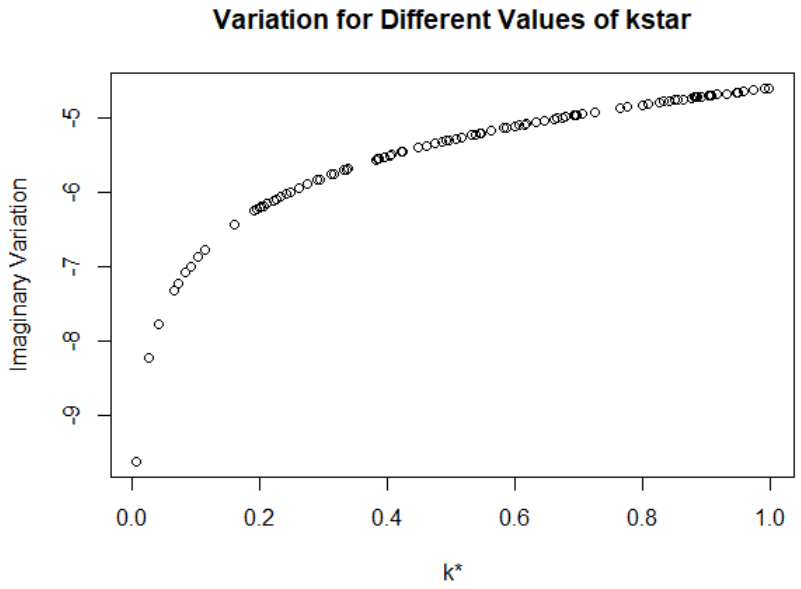


Figure 18: A plot of k_* from a uniform distribution vs. the complex variation of the interaction

Variation from spectral densities of beta distributed transition rates

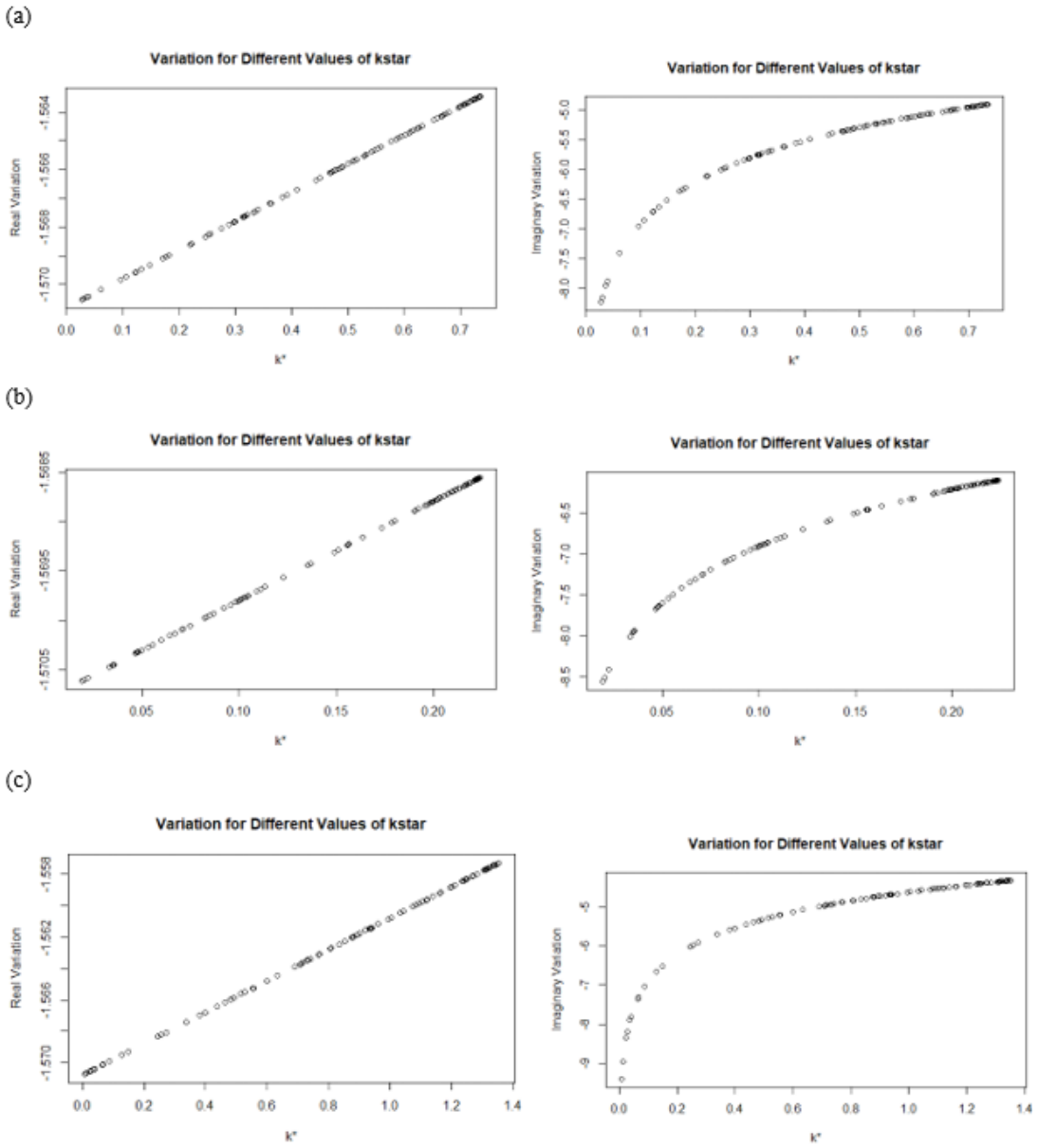


Figure 19: Variation measured from different spectral densities and beta distributions of k_*. All distributions contain real (left) and imaginary (right) components. **(a)** $k_* \sim Beta(\alpha = 2, \beta = 2)$ **(b)** $k_* \sim Beta(\alpha = 4, \beta = 1)$ **(c)** $k_* \sim Beta(\alpha = 3, \beta = 5)$

CHAPTER III: DISCUSSION

III.1 Method 1: Transition Function

In *Figure* 5 we show how the rate of decay effects the time required for the curve to pass the threshold m. We see that larger values of k lead to a shorter time tau for the curve to pass the threshold. On the other hand, smaller values of k create a longer time period to cross the threshold. We also look at the relationship between the rate of decay and the time that it takes to cross the threshold in *Figure* 6. Again, higher rates of decay indicate a faster response time for change in gene states. We also found that higher thresholds are more responsive to change than lower thresholds and also less affected by changes in decay rates.

Later, in *Figures* 7, 8, and 9, we include stochasticity while using the transition function and by using a binomial distribution where there is a 75% probability that genes which are supposed to change states actually do change states. Akman et al. argue that the use of the binomial distribution to determine the outcome of genes likely overestimates the variation of transitions [2], but this distribution was used for the sake of a preliminary investigation and could easily be changed in the future. Overtime, we see that the number of genes that transition has a propensity of 75%, which is the given probability of the binomial distribution used to determine the outcome of each gene.

In Figure 11, we see that as time increases, the variation present in transitions also increases. This is true regardless of the threshold value. This would indicate that cells with genes that can respond quickly to their surroundings are more likely to have lower variability between cells. This is consistent with literature that indicates that fast response times in a system lead to lower fluctuations and thus lower cell-cell variability [4]. It has been shown that frequent transcripts with fewer proteins per transcript lead to networks with less variation, whereas, less frequent transcripts and larger protein yield per transcript results in more noise [27]. Low protein yield per transcript is energy inefficient [27]. High protein yield per transcript leads to more variation but is energy efficient [27].

III.2 Method 2: Beta Distributed Propensities

In *Figure* 12 we find that large differences in shape parameters and high expected values for transitions lead to the most variation in gene state transitions. We also find that a mix of intermediate and high expected values leads to greater variation than low expected values in *Figure* 13. There is currently no known biological explanation for this behavior in the literature. We hypothesize that genes that are acting effectively, and thus have a high expected value or high propensity and are working appropriately, are able to handle more variation without a decrease in fitness compared to genes that are inefficiently working. It may also be unlikely to see a mixture of intermediate and high propensities in a system, since most systems were likely selected to have lower variation through evolution.

One of the benefits to using the beta distributed activation and degradation propensity method is that no simulation of the network is needed. There is also a natural extension to the beta distributed activation and degradation propensity method. When shape parameter $a_B = a_A + b_A$, the distribution is a bivariate beta distribution according to a lemma discussed in Krysicki (1999) [21] and Nadarajah and Kotz (2005) [30]. Krysicki also showed that this property extends to multiple independent beta random variables [21]. "If U_1, U_2, $\ldots U_p$ are independent beta random variables with shape parameters (a_i, b_i), $i = 1, 2, \ldots, p$ and if $a_{i+1} = a_i + b_i$, $i = 1, 2, \ldots, p-1$, then the product $U_1 * U_2 * \ldots U_p$ is also a beta random variable with shape parameters $(a_1,\ b_1 + \ldots + b_p)$" [21].
Bivariate Beta Distribution:

$$f(x,y) = \frac{\Gamma(a+b+c)}{\Gamma(a)\,\Gamma(b)\,\Gamma(c)} x^{a-1} y^{b-1} (1-x-y)^{c-1}$$

Where $a, b, c > 0,\ x \geq 0,\ y \geq 0$ *and* $x + y\ \leq\ 1$.

Nadarajah and Kotz discuss the relationships among two independent beta distributions using this lemma but leave the relationships created from multiple independent beta distributions for future work [30]. Although we limit ourselves to two

genes for the majority of our work, it is more realistic to consider cases with more than 2 genes interacting and so exploring these properties between 3 and more beta distributions may yield interesting results. In fact, the bivariate beta distribution has been used to model proportions of alleles in population dynamics [40]. Therefore, future use of the bivariate beta distribution may lead to promising results and insights.

III.3 Method 3: Spectral Density

In *Figures* 17, 18, and 19 we see the behavior of variation for a variety of distributions. In all cases, the real component of the PSD is linear and the complex component logarithmic. This indicates that the distribution of k does not have a large impact on the variation of transitions. There is some literature that agrees with the results shown here. Stochasticity in gene expression is not strongly dependent on the statistical distribution of transcription initiation [27]. Stochasticity will occur when there are a limited number of promoters for gene regulation in the cell and the time interval for transcription time is longer [27].

All of the variation represented through this method is less than zero. This is likely due to the fact that the Fourier transform involves complex values and that $i^2 = -1$. Interpretation of negative variation is not common, but there is some evidence that the magnitude of variation is more important than the sign [9]. Likewise, complex, or imaginary, variation is also difficult to interpret. The resulting negative and complex variation for all cases is a possible downfall to this modeling process and future investigation is needed to see if this behavior is prevalent across all distributions.

CHAPTER IV: CONCLUSION

Collectively, each of these modeling techniques provides different qualitative information about gene state transitions. Integrating these models and the information that they provide into future models could help to create more accurate models overall. However, these models do not capture the larger picture of interwoven networks that each impact one another. Separating networks mathematically does not provide an accurate view of an organism's complex system of networks each impacting one another. In the future, integration of multiple networks will help to build more realistic models.

REFERENCES

[1] Abou-Jaoudé, W., Ouattara, D. A., & Kaufman, M. (2009). From structure to dynamics: frequency tuning in the p53-Mdm2 network I. Logical approach. *Journal of theoretical biology 258*, 561–577.

[2] Akman, O., Comar, T., Harris, A. L., Hrozencik, D., & Li, Y. (2018). Dynamics of gene regulatory networks with stochastic propensities. *International Journal of Biomathematics*, *11*(3), 1–13. https://doi.org/10.1142/S1793524518500328

[3] Alakwaa, F. M. (2015). Modeling of Gene Regulatory Networks: A Literature Review. *Journal of Computational Systems Biology*, *1*(1), 1–8. https://doi.org/10.15744/2455-7625.1.102

[4] Alon, U. (2006). An Introduction to Systems Biology: Design Principles of Biological Circuits (1st ed.). Boca Raton, FL: Chapman & Hall/CRC.

[5] Asadzadeh, S., & Roberto de Souza Filho, C. (2016). A review on spectral processing methods for geological remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 47*, 69–90.

[6] Ciliberto, A., Novak, B., & Tyson, J. J. (2005). Steady States and Oscillations in the p53/Mdm2 network. *Cell Cycle, 4*(3), e107–e112.

[7] Dobrow, R. P. (2016). Introduction to Stochastic Processes With R (1st ed.). Hoboken, New Jersey: John Wiley and Sons.

[8] Emmert-Streib, F., Dehmer, M., & Haibe-Kains, B. (2014). Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology, 2*, 1–7. https://doi.org/10.3389/fcell.2014.00038

[9] Firth, B. D., & Menezes, R. de. (2004). Quasi-variances.*Biometrika, 91*(1), 65–80.

[10] Gehrmann, E., & Drossel, B. (2010). Boolean versus continuous dynamics on simple two-gene modules. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 82*(4). https://doi.org/10.1103/PhysRevE.82.046120

[11] Ghosh, I. (2019). On the Reliability for Some Bivariate Dependent Beta and Kumaraswamy Distributions: A Brief Survey. *Stochastics and Quality Control*, 1–7. https://doi.org/10.1515/eqc-2018-0029

[12] Gillespie, D. T. (2007). Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, *58*(1), 35–55. https://doi.org/10.1146/annurev.physchem.58.032806.104637

[13] Ginsburg, G. S., & Phillips, K. A. (2018). Precision medicine: From science to value. *Health Affairs, 37*(5), 694–701. https://doi.org/10.1377/hlthaff.2017.1624

[14] Gowda, T., Vrudhula, S., & Kim, S. (2009). Modeling of gene regulatory network dynamics using threshold logic. In G. Stolovitzky, P. Kahlem, & A. Califano (Eds.) *The challenges of systems biology: Community efforts to harness biological complexity*, *1158*. New York: Wiley-Blackwell.

[15] Hu, W., Feng, Z., & Levine, A. J. (2012). The Regulation of Multiple p53 Stress Responses is Mediated through MDM2. *Genes and Cancer, 3*(3–4), 199–208. https://doi.org/10.1177/1947601912454734

[16] Jambunathan, M. V. (1954). Some Properties of Beta and Gamma Distributions. *The Annals of Mathematical Statistics*, *25*(2), 401–405. https://doi.org/10.1214/aoms/1177728800

[17] Jong, H. De. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*, *9*(1), 67–103.

[18] Jorna, P. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, *34*(2–3), 237–257.

[19] Kauffman, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, *244*, 177–178.

[20] Klymkowsky, M., & Cooper, M. (2015). Biofundamentals 2.0.

[21] Krysicki, W. (1999). On some new properties of the beta distribution. *Statistics and Probability Letters*, *42*, 131–137. https://doi.org/10.1515/dema-1999-0318

[22] Lee, W. P., & Tzou, W. S. (2009). Computational methods for discovering gene networks from expression data. *Briefings in Bioinformatics, 10*(4), 408–423. https://doi.org/10.1093/bib/bbp028

[23] Lipshtat, A, Neves,S. R., & Iyengar, R. (2009). Specification of spatial relationships in directed graphs of cell signaling networks. In G. Stolovitzky, P. Kahlem, & A. Califano (Eds.) *The challenges of systems biology: Community efforts to harness biological complexity 1158*. New York: Wiley-Blackwell.

[24] Manioudaki, M., Tzamali, E., Reczko, M., & Poirazi, P. (2009). Bioinformatics for Systems Biology (1st ed.; S. Krawetz, ed.). https://doi.org/10.1017/CBO9781107415324.004

[25] Marhon, S. M., & Kremer, S. C. (2011). Gene prediction based on DNA spectral analysis: A literature review. *Journal of Computational Biology, 18*(4).

[26] Mathur, S., & Sutton, J. (2017). Personalized medicine could transform healthcare (Review). *Biomedical Reports, 7*(1), 3–5. https://doi.org/10.3892/br.2017.922

[27] McAdams, H., & Arkin, A. (1997). Stochastic Mechanisms in Gene Expression. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(3), 814–819.

[28] Murrugarra, D., Veliz-Cuba, A., Aguilar, B., Arat, S., & Laubenbacher, R. (2012). Modeling stochasticity and variability in gene regulatory networks. *Eurasip Journal on Bioinformatics and Systems Biology*, *2012*(1), 5. https://doi.org/10.1186/1687-4153-2012-5

[29] Nadarajah, S. (2005). Reliability for some bivariate beta distributions. *Mathematical Problems in Engineering*, *2005*(1), 101–111. https://doi.org/10.1155/MPE.2005.101

[30] Nadarajah, S., & Kotz, S. (2005). Some bivariate beta distributions. *Statistics: A Journal of Theoretical and Applied Statistics*, *39*(5), 457–466. https://doi.org/10.1080/02331880500286902

[31] Olkin, I., & Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics and Probability Letters*, *96*(January), 54–60. https://doi.org/10.1016/j.spl.2014.09.013

[32] Przulj, N. (2011). "Biological networks uncover evolution, disease, and gene functions," In P. Pevzner & R. Shamir (Eds.) *Bioinformatics for Biologists*, pp. 291–314. Cambridge: Cambridge University Press.

[33] Schwartz, R. (2011). "Regulatory network inference," In P. Pevzner & R. Shamir (Eds.) *Bioinformatics for Biologists*, pp. 315–343. Cambridge: Cambridge University Press.

[34] Song, M., Lewis, C. K., Lance, E. R., Chesler, E. J., Yordanova, R. K., Langston, M. A., Bergeson, S. E. (2009). Reconstructing generalized logical networks of transcriptional regulation in mouse brain from temporal gene expression data. *Eurasip Journal on Bioinformatics and Systems Biology*, *2009*. https://doi.org/10.1155/2009/545176

[35] Stark, J., Brewer, D., Barenco, M., Tomescu, D., Callard, R., & Hubank, M. (2003). Reconstructing gene networks: What are the limits? *Biochemical Society Transactions,31*(6), 1519–1525. https://doi.org/10.1042/bst0311519

[36] Stoica, P., & Moses, R. (2005). Spectral analysis of signals (T. Robbins, ed.). Upper Saddle River, New Jersey: Prentice-Hall.

[37] Thibos, L. N. (2014). Fourier analysis for beginners (6th ed.). Retrieved from http://hdl.handle.net/2022/21365

[38] Van Laar, J., Porath, M., Peters, C., & Oei, S. (2008). Spectral analysis of fetal heart rate variability for fetal surveillance: review of the literature. *Acta Obstetricia et Gynecologica Scandinavica*, *87*(3), 300–306.

[39] Waning, D. L., Lehman, J. A., Batuello, C. N., & Mayo, L. D. (2010). "Controlling the Mdm2-Mdmx-p53 circuit". *Pharmaceuticals*, *3*(5), 1576–1593. https://doi.org/10.3390/ph3051576

[40] Wright, S. (1937). The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the United States of America*, *23*(6), 307–320. https://doi.org/10.1126/science.85.2212.504

[41] Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., & Darnell, J. E. (2003). Decay rates of human mRNAs: Correlation with functional characteristics and sequence attributes. *Genome Research*, *13*(8), 1863–1872. https://doi.org/10.1101/gr.1272403

[42] Yates, C. A., Ford, M. J., & Mort, R. L. (2017). A Multi-stage Representation of Cell Proliferation as a Markov Process. *Bulletin of Mathematical Biology*, *79*(12), 2905–2928. https://doi.org/10.1007/s11538-017-0356-4