1. Introduction

The critical role of stochasticity in biology has been studied in many contexts - such as creating variations required to survive in hostile environments [**1**], regulating circadian clocks [**2**], and probabilistic differentiation in developing cells [**3**]. If we systematically characterize the stochasticity in each context, we gain the ability to control these biological functions. The ability to control the development of multicellular systems, for one, holds fantastic futures like replacement organs grown *in vitro* or smarter drugs that only target tumors. With such potentials, then, which feature of stochasticity in development should we focus on?

In the development phase of multicellular organisms, an isogenic group of cells differentiates into multiple groups of heterogeneous cells with different epigenetics. This behavior can be likened to a group of people performing a leader election. To fairly elect a leader, the group can repeatedly and separately perform a task that has probabilistic outcomes, where one of the outcomes is the victory outcome. For example, a coin-toss with a head. After some number of tries, if an individual is the first one who ends up with a head among the group, he or she becomes a leader - effectively differentiating him or her from the group. The leader can then send signals to the rest of the group, telling them to stop flipping coins and become followers - again differentiating them from their undecided state, as well as from the new leader. Thus, it is not farfetched to imagine that a similar mechanism takes place inside a developing organism.

Let us assume that an individual ends up with a head for the first time after h tries. Because coin-tosses have probabilistic outcomes, h is also a random variable. And as such, h is characterized by its probability distribution. Let us assume that this probability distribution can be manipulated somehow - by biasing the coin, for example - then, the variance of h has an interesting interpretation in the leader election example. If the variance of h was set small, the probability of multiple undecided individuals each ending up with the victory outcome in a short amount of time is large - in other words, the victory outcomes are closely synchronized, and the group may end up with multiple leaders. However, if the variance is set large, the victory outcomes are asynchronous and the group is less likely to have multiple leaders. Therefore, the probability distribution of h affects the population distribution of differentiated states of leaders and followers.

In the following sections, we discuss the translation of the leader election example into a cellular context, specifically in the development phase, and propose a possible biological equivalent to the biased coin. Then we discuss the ways to characterize the probability distribution of h (or some equivalent random variable in the proper context), both with theoretical analysis by posing the scenario in mathematical language, and with experiments by synthesizing the biological biased-coin equivalent. This discussion of characterization methods is followed by required backgrounds in both theory and experiments, as well as related works in the field that serves as helpful starting points for the proposed research. Some preliminary results are discussed in the last section, along with recommendations and proposed future works.

Key words: Gene regulatory networks, synthetic biology, completion time, probability distribution, development

2. Approach and Objectives

As long as the proposed mechanism is probabilistic, one can suggest several intracellular environment analogs to the coin-flip example, such as protein dimerization, folding, or saturation. Let us consider the protein saturation example, where at $t = 0$, the gene coding for the protein of interest, X, gets activated and there is no X present. The gene begins to express X and the count number of X (N_X) increases as long as the gene remains activated and the rate of X degradation is less than the rate of X synthesis. If the gene is deactivated before N_X reaches the saturation value, then N_X begins to decrease until the gene is activated again. This process of gene activation and deactivation occurs repeatedly until at some time $t = T_c$, the count number of X reaches the saturation value. This time T_c is analogous to h in the coin-flip example, and we call T_c the *completion time* of the protein synthesis process. And as mentioned earlier, the distribution of the differentiated states of the cell population with the gene X is affected by the probability distribution of T_c.

There can be a number of ways to manipulate the probability distribution of the completion time in this example. One way is varying the frequencies of gene activation and deactivation. Another way is varying the mechanism that activates the gene - an open-loop activation from external inputs or a feedback activation/deactivation by X. Frequency variations change the quantitative features of the gene regulation, and feedback or open-loop variations change the qualitative features of the gene regulation. The relationship between the quantitative and qualitative features of gene regulation and the probability distribution of the completion time will help us understand the fundamental design principles employed by nature to perform development and differentiation in multicellular organisms. Therefore, we propose the following objectives to guide the investigation of such relationship.

- **Synthesize single-gene networks in *E. coli***. Three different mechanisms of gene regulation will be studied in this research - open-loop, positive feedback and negative feedback. The synthetic gene network corresponding to an open-loop mechanism will have a single gene that is activated by some external inputs. For the two feedback mechanisms, a single gene network that expresses either its own repressor or activator will be synthesized, to correspond to a negative or positive feedback mechanism, respectively. All of the gene networks will be synthesized with inducible promoters and fluorescence protein gene. The inducible promoters allow us to measure the completion time by setting the initial time to when the promoter is induced, and the level of fluorescence emitted by the fluorescence protein is used to monitor the gene activity.
- **Mathematically model the three gene networks**. Using the Chemical Reaction Network theory, we will model the interactions among the gene network species. We apply a variety of stochastic analysis tools to the models in order to characterize the completion time, its probability distribution, and sensitivity to parameter variations and structural variations. Such analysis tools include the Chemical Master Equation (CME), the Stochastic Simulation Algorithm (SSA), and cumulant and moment dynamics. We will identify the qualitative differences of the gene networks arising from the difference in structure, and discuss how they can make each structure a better or worse suited mechanism used in development processes. In addition to the

qualitative features, quantitative features regarding the change in parameters will be investigated. The limitations on the probability distribution of completion time placed by physically feasible parameter values may also explain why a certain structure is more frequently observed in development than others.

– **Iteratively verify predictions made in the models with experiments and modify the models based on the experimental results**. The gene activity is monitored by measuring the level of fluorescence emitted by the synthesized fluorescence protein. The probability distributions of the completion time in these synthetic gene networks will be approximated using cellular assays, such as time-lapse microscopy or flow cytometry. Time-lapse microscopy allows us to monitor the individual trajectory of fluorescence level in a *single cell* and the time at which the fluorescence level reaches some saturation value. On the other hand, flow cytometry reveals the *distribution* of fluorescence level at each measurement. Therefore, by measuring the fluorescence distributions at multiple times, we can study the distribution dynamics of the fluorescence level. And from the dynamics, we will approximate the fraction of the population that has reached the saturation value at each measurement time. The experimental results will be used to invalidate some of the candidate models and point out the features that require modification to attain better fidelity to the actual systems The modified models, in turn, are used to design experiments that will better highlight the key features of the systems. The mathematical model predictions obtained from this iterative process will identify the salient features of development process and allow us to synthesize gene networks with the complexity comparable to the naturally occurring examples.

The following section will provide a broad overview of the fundamentals in both theory and experiments to accomplish our objectives. Two specific related works are discussed afterwards, each with a focus on theory and experiments respectively. These works were chosen based on their close proximity to the objectives of the proposed research, and served as a foundation for obtaining the preliminary results that are discussed in a later section.

3. Background and literature review

3.1. Overview. As the biotechnology steadily advances, researchers are able to synthesize gene regulatory networks with increasing precision and success. These synthetic gene networks are built from borrowed biological components of natural genetic regulatory parts, such as promoters and transcription factors. Though manipulating genetic materials is not a new technology, synthetic biology is different from traditional genetic engineering in its intention to engineer novel behaviors, such as oscillation or bistability [**4, 5**]. The underlying objectives of these synthesis-based approach to biology is to identify and isolate the salient features of complex gene networks and discover the nature's design principles. And synthetic biology is strenghthened by two complementary approaches of mathematical theory and biological experiments. A well-established study of differential equations is used to analyze the dynamics of the systems [**6**], linear systems theory the stability and controllability [**7**], and probability theory the stochastic behaviors in the mesoscopic level of biological molecules [**8**], to name a few. At the same time, increasing efficiency of cloning techniques [**9**], decreasing cost of DNA synthesis and sequencing [**10**], and the advance of experimental equipments all contribute to engineering biological test beds for verifying hypotheses obtained from mathematical theories. As the objectives of the proposed research spans both theory and experiments, the rest of the background section is divided into two sections to address the fundamentals of each aspect separately.

3.2. Theory. The theory of Chemical Reaction Network was originally developed to provide a standardized foundation from which a mathematical description of chemically interacting species inside a fixed volume can be derived [**11**]. The CRN of a given system contains chemical species (X_i) that interact with respect to some reaction (R_j), the stoichiometric coefficients of reactants (u_{ij}) and products (v_{ij}) of the chemical reactions, and the rates of these interactions (λ_j). From this description, using the Law of Mass Action, the dynamically changing concentrations of the chemical species are modeled by a set of ordinary differential equations. This method translates smoothly into the context of biological interactions inside a cell. Cellular environments are no different from the environments inside a chemical processing plant, such that they have biochemically interactions, reactant and product species of these interactions, and numerical values that describe the rates of the interactions. However, the key difference is that whereas chemical systems tend to have a large quantity of each species, the quantity of biological molecules tend to be present in much smaller quantities. Thus, chemical species can be expressed as continuous variables, whereas biological molecules must be expressed as discrete variables. Additionally, the stochasticity of the biochemical interactions become more pronounced in a system with species in small quantities. Therefore, biochemical systems, such as gene regulatory networks require mathematical description that properly addresses the discrete copy number of species and the stochasticity of interactions.

The discrete values of biochemical molecules inside gene regulatory networks allow us to model the systems as discrete-state continuous-time Markov processes [**12**]. Let the species of an arbitrary gene regulatory networks be denoted by a vector $\mathbf{S} = [S_1, \cdots, S_n]$, and the number of each species are denoted by X_i. Each discrete state of the system is then denoted by the vector $\mathbf{X} = [X_1, \cdots, X_n]$. And because the stochasticity of gene networks forces the description of the system from a deterministic value to a probability distribution over the states, we define the probability of the system in state $\mathbf{X}$ at time t to be $p(\mathbf{X}, t)$. The vector of the probabilities of all the states is $\mathbf{p}(t)$ and the probability vector, given some initial distribution $\mathbf{p}_0$, evolves according to the following master equation.

$$\dot{\mathbf{p}}(t) = \mathbf{Q}\mathbf{p}(t). \tag{1}$$

The above equation is the Chemical Master Equation, and the matrix $\mathbf{Q} = [q_{ij}]$ contains the rates of system transitions from state j to state i [**13**]. The analytical solution of (1) is

$$\mathbf{p}(t) = e^{\mathbf{Q}t}\mathbf{p}_0. \tag{2}$$

Do not be mislead by the elegantly simple form of the solution, as the matrix exponential, $e^{\mathbf{Q}t}$, requires an infinite sum of high computational cost. Instead of solving for the probability distribution dynamics analytically, a numerical alternative exists, where the exact realizations of the corresponding CRN can be obtained.

The Stochastic Simulation Algorithm (SSA) was developed to numerically simulate individual trajectories of the species of an arbitrary stochastic chemical reaction network [**14**]. The method employs the fact that 1) each rate of a chemical reaction is the

inverse of the mean waiting time for the reaction, and 2) the probability of a reaction is equal to the ratio of the reaction propensity to the sum of all reaction propensities. Then, the time evolution of a stochastic system can be approximated by generating a large number of simulations of the system and studying the dynamics of each species. It should be clarified that the probability distribution dealt with in the CME corresponds to the *joint* probability distribution of each specific state, $P([X_1 = x_1, \cdots, X_n = x_n], t)$, whereas the approximated probability distributions obtained via SSA are the *marginal* probability distributions of all species, $P(X_1, t)$. Thus, it is more straightforward with the SSA to portray the time-evolution of each species probability distribution. However, this numerical algorithm requires that the initial condition and the rate constants be specified *a priori*, which means that if an analysis requires a different set of rates or initial conditions, a whole new set of large number of simulations is required to study the specific condition.

An alternative approach of characterizing the evolution of the probability distribution for stochastic biochemical systems is to compute the cumulant dynamic of each species of the system [**15**]. The cumulants of a random variable are set of values that characterizes the shape of the corresponding probability distribution. For example, the second order cumulant of a random variable is its variance and is representative of the width of the probability distribution. The cumulants are computed using the cumulant generator function,

$$G_X(s) = \log\left\langle e^{sX}\right\rangle, \tag{3}$$

where X is the random variable and $\langle\cdot\rangle$ denotes the expected value. The nth order cumulant of X is computed by taking the nth derivative of (3) with respect to s and setting $s = 0$. Usually, no more than the first four cumulants are computed for a given species, because cumulants of order five or higher have no straightforward interpretation related to the probability distribution characteristics. The time evolution of these cumulants requires an additional function called the extended generator. Let $\psi(X(t))$ be some test function of state $X(t)$, then the expected value of this test function evolves according to the following equation.

$$\begin{aligned} \frac{d\langle\psi(X(t))\rangle}{dt} &= \langle L\psi(X(t))\rangle \\ &= \sum_{j=1}^{m} \lambda_j \left(\psi(X^j(t)) - \psi(X(t))\right), \end{aligned} \tag{4}$$

where the $X^j(t)$ is the state after the reaction $\mathbb{R}_j : X(t) \mapsto X^j(t)$ has occurred, λ_j is the reaction rate constant, and L is the extended generator. The cumulant dynamics is then obtained by letting $\psi(X(t)) = G_X(s)$, and solving the resulting set of ordinary differential equations. An interesting connection exists between the cumulant dynamics and the differential equation obtained by using the Law of Mass Action, such that the first order cumulant dynamics is equal to the deterministic dynamics predicted by Mass Action kinetics. In fact, this is not surprising because the first order cumulant its the mean of the population and Law of Mass Action predicts the average behavior of the population.

3.3. Experiments. Within a single cell resides a genome, a chain of DNA molecules, that contains all the genetic information the cell needs to harvest energy, reproduce and survive. The genome alone, though mighty in its information content, cannot make a living organism. It requires molecular machinery that actualizes this information in useful form, thus is the function of RNA and protein. DNA is transcribed into RNA, and in turn the RNA is translated into protein, and proteins are the true workers of biological functions [**16**]. The role of protein molecules as the regulators of genomic information transfer is the most critical with regards to the viability of an organism. If the processes of transcription and translation were not properly regulated, in other words if the entire genome was uniformly transcribed and translated, it would mean a disaster for the cell. Therefore, there exists intricately connected networks of gene regulation that allows cells to allocate energy, respond to its environment and procreate.

The two major components of gene regulatory mechanisms are promoters and transcription factors (TF). TFs are protein complexes that act either as a repressor or an activator by binding to the promoter of a gene. Promoters are short sequence of DNA that are located at the 5'-end of a gene and are recognized by RNA polymerase to initiate an RNA synthesis. A bacterial promoter has two short 6 basepair long sequences that are conserved in most promoters, called the consensus sequences. The rest of the promoter sequences are composed of operators that serve as binding sites for specific TFs. A large number of TF and promoter pairs have been identified in metabolic pathways, signal transduction pathways, and developmental regulatory pathways. The known pairs of TF and promoter are used to design and build synthetic gene regulatory networks by arranging them in specific configurations [**17, 18**]. For example, the critical structure of stress response in *B. subtilis* were identified by synthesizing the same gene network, but with one of the two feedback loops (coupled positive and negative feedbacks) removed [**19**]. The synthetic network, when transformed inside cells, prohibited the cells from leaving their competence state, showing that the removed feedback is critical to the overall mechanism of *B. subtilis* stress response.

Feedbacks are not observed just in this specific example of transient differentiation. In fact, feedback mechanisms are frequently observed in a number of gene regulatory network classes. A class of gene networks that give rise to stochastic state switching, such as cancer and developmental differentiation, has been consistently shown to contain positive feedback loops [**20, 21, 22, 23**]. Another class of behavior that arises from containing positive feedback loops in the gene regulatory networks is procrastinating differentiation [**24**]. Procrastination refers to the phenomenon observed in isogenic cells, that when triggered for specific response (e.g. sporulation, apoptosis), the response time of each cell widely vary within the microcolony and results in varied states of the population. This phenomenon is an example of the relationship between the probability distribution of completion time and the distribution of differentiated state mentioned in the previous section.

3.4. Related Works. One way of deriving the analytical expression of completion time probability distribution is to solve the CME of the system in Laplace domain [**25**]. In this work, a kinetic proofreading (KPR) process was modeled by a Markov chain with an absorbing state, where the absorbing state corresponded to the completion of the proofreading process that required sequential intermediate steps. In recognition that the completion time is essentially the first-passage time of Markov chain, they performed Laplace transform to the solution of the CME shown in (2) to obtain the analytical expression [**26**]. The solution showed that the distribution of first-passage time approaches a limiting behavior, depending on the direction of the bias imposed

by the transition rates - forward to the sink state, or backward to the initial state. However, the solution and the conclusion is limited to an open-loop system where the transition rates are independent of the states. Though the authors simulate systems with site-*varying* transition rates by using randomly generated values and expand their conclusion, this is quite different from site-*dependent* transition rates of feedback mechanisms. It will be interesting to investigate whether a similar conclusion can be drawn from biochemical processes with feedback.

Another example is the library of hybrid promoters created in combinatoric fashion using operating sites from multiple natural promoters [**27**]. Promoters usually have more than one operating sites that are recognized by transcription factors. The authors isolated these operating sites from natural promoters and constructed hybrid promoters through random DNA assembly. The expression level of these hybrid promoters were characterized in consistent conditions, which allows the construction of reliable synthetic gene networks.

4. Preliminary Results

4.1. Construction and assay of positive feedback gene networks. Even regulated gene regulatory networks often exhibit some basal leaky expression without explicit activation. And a system with positive feedback can quickly switch from an *off* state to an *on* state with only a small amount of the output because of the autocatalysis. Therefore, we wanted promoters that had minimal leaky expression and gave maximal control over the range of expression. In [**28**], hybrid promoters made from multiple operating sites of natural promoters were shown to have tighter regulation of leaky transcription than its natural peers. Based on this result, we selected two hybrid promoters from [**27**] - A12 and D61 - based on their response inducers and the AND-gate score. The promoters have operating sites from a pBAD promoter that is activated by AraC-arabinose complex, and a pLac promoter that is repressed by LacI protein. Therefore, by inserting the araC gene downstream of the promoters, we were able to construct a positive feedback loop. In addition, GFP gene was inserted so that the expression level can be monitored using fluorescence imaging techniques. Ideally, the circuit would require both IPTG (relieves LacI suppression) and arabinose (forms the activator AraC-arabinose complex) in order to express the output proteins. Currently, we have four different variations of the positive feedback gene network (shown in Figure 1 (c)) that are sequence verified and transformed into 3 different strains of *E. coli*: DH5α, BL21 LacI^q, and Keio JW0063-1. Each of four gene circuits either has A12 or D61 promoter and pMB1, pSC101, or pSB3K3 origin of replication[1].

A preliminary cellular assay using microscope fluorescence imaging - cells grown on agar with inducer - showed that the cells with the positive feedback network has slower growth rate than the cells with open-loop network (1 (a)). A similar observation was made in a flow cytometer assay - cells were induced at maximal concentration of IPTG and L(+)-arabinose, and the fluorescence distribution of 50,000 cells were measured at 15 minute interval (1 (b)). About two hours after the induction, a separate population of cells with fluorescence level an order of magnitude larger was observed. However, this population disappeared 30 minutes afterwards. We hypothesize that this population corresponds to the cells with activated positive feedback, and that the over-synthesis of protein is affecting the cell viability - in short, AraC/GFP toxicity is suspected.

[1]Each has a copy number of $\sim 10^0, \sim 10^1$ and $\sim 10^2$ respectively

4.2. Approximation of completion time distribution.

$$
\begin{aligned}
G_u + X &\underset{\alpha\ ka}{\overset{ka}{\rightleftharpoons}} G_b \\
G_b &\xrightarrow{\beta_1 ex} G_b + X \\
G_u &\xrightarrow{\beta_2 ex} G_u + X \\
X &\xrightarrow{dx} \phi
\end{aligned}
\tag{5}
$$

The Chemical Reaction Network shown on the right corresponds to the positive feedback genetic network discussed in the previous section. The three chemical species are gene without a transcription factor bound to it (G_u), gene with a transcription factor (G_b), and the transcription factor (X). The parameters are rates of transcription factor binding (ka), unbinding-to-binding ratio (α), ratio of unbound gene expression to a basal expression (β_1), ratio of bound gene expression to a basal expression (β_2), and transcription factor degradation/dilution rate (dx). One thousand SSA realizations of (5) are shown in Figure 2(a). For each trajectory, we locate the time at which the copy number of X reaches half of its steady-state value (N) and set this as the response time of the network to inducers. This response time distribution is shown to have asymmetric shape (Figure 2 (b)).

For sensitivity analysis of response time distribution, analytical solution for (2) is required. However, the solution becomes more unwieldy as the size of the state-space grows. Therefore, to investigate the response time distribution as a function of the system parameters, we devised a crude approximation method using truncated cumulant dynamics. First, we compute the first and second order cumulants of X, $\kappa_X(t)$ and $\kappa_{XX}(t)$ by assuming at all times t, X is distributed normally. Then, using the normal distribution function, we compute the fraction of the ensemble population with X count number less than (N/2) as a function of time. This turns out to be an approximation of the cumulative distribution of the response time. By taking the derivative of this function, we are able to approximate the response time distribution. Though a fair approximation, the function lacks mathematical rigor. For example, the normal distribution assumption applies to continuous variables, and an alternative method of approximating the response time distribution is warranted.

4.3. Completion Time Probability Distribution. In this paper, *completion* is defined as the instance where the population of a chemical species reaches some saturation value, $N/2$, and the *completion time* is defined as

$$
t_i = \min_t \{ t \in T : X_i(t) \geq \chi \}. \tag{6}
$$

The stochastic nature of gene regulatory networks dictates that the CTPD is the more appropriate quantity of interest. From the CTPD, we can derive the fraction of the population that is *completed* at any given time.

cumulative distribution of the completion monotonically increasing $F_i : \mathbb{R} \to [0,1]$ right

$$
F_i(\chi) = 1 - P(X_i(t) < \chi) \tag{7}
$$

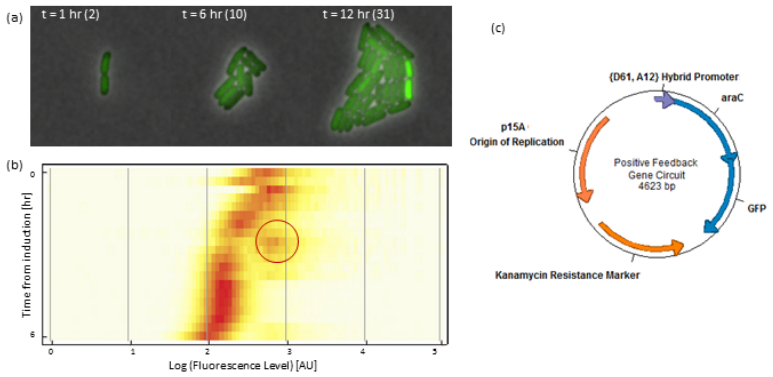

FIGURE 1. Preliminary experimental results and construction of a positive feedback gene network. (a) Microscope imaging data over 12 hour duration. The doubling time ($\sim$ 5 hr) is significantly reduced compared to open-loop system. (b) Flow cytometer data showing the fluorescence distribution over 6 hours. The red circle indicates the separate population with higher fluorescence level, which is observed for less than an hour before disappearing altogether. (c) The construction of positive feedback gene networks. Each circuit consists of a hybrid promoter, araC gene and gfp gene.

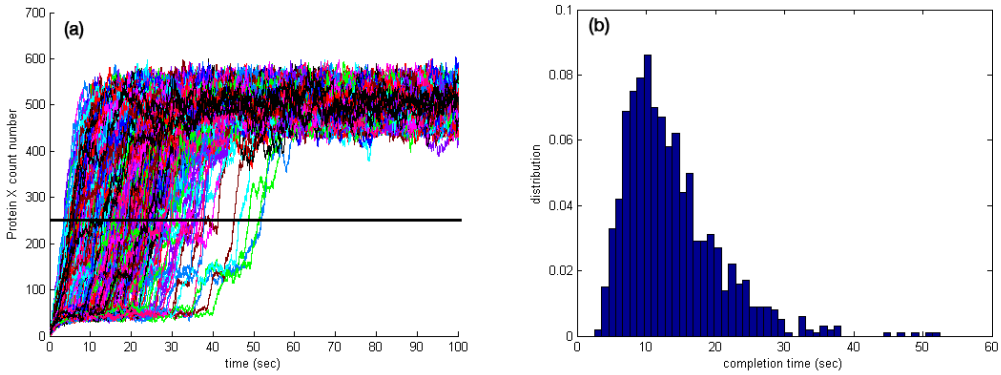

FIGURE 2. Thousand realizations of Stochastic Simulation Algorithm of positive feedback network CRN, and completion time distribution. The parameters used are $[k,\alpha,\beta_1,\beta_2,ex,dx] = [10^{-3}\log(2), 0.1, 1, 10, 10\log(2), \log(2)]$ and the initial condition is $[G_{unboud}, G_{bound}, X]$ = [5, 0, 0]. The system is considered complete,when the protein X reaches one-half of its steady-state value, where the steady-state value is predicted by solving the Mass Action Kinetics ODE model. The completion time distribution (b) corresponds to the time distribution at the dotted line in (a).

probability distribution of the completion $f_i : \mathbb{R} \to [0,1]$

$$f_i(t) \quad = \quad P(X_i(t) = \chi) \tag{8}$$

completion time cumulative distribution

$$P(\chi,t) \quad = \quad \frac{N_{Tot}\text{fraction of population with } X_i(t) \geq \chi}{N_{Tot}} \tag{9}$$

In the previous section, the cumulants of chemical species and the dynamics of the cumulants were obtained directly from an arbitrary stochastic chemical reaction network. Computing the probability distribution of a random variable from the cumulants is an alternative form of a popular problem in mathematics called *the classical moment problem*, where the probability distribution of a random variable is computed from a sequence of moments. The classical moment problem does not have a definitive solution, but various approximative algorithms exist.

The assumption that at any given time the population is distributed normally implies that the cumulants of order three and higher are all equal to zero. Thus the set of ordinary differential equations of the cumulant dynamics contains only the first and second order cumulants of the chemical species, making the task of ODE solving manageable. Additionally, the normal distribution assumption allows us to solve for the probability distribution of the chemical species given the cumulants. For example, let the first and second order cumulants of a chemical species X_i at time τ be $\kappa^{(1)}_{X,i}(\tau)$ and $\kappa^{(2)}_{X,i}(\tau)$. Then the probability distribution of X at the time is

$$P(X_i = \chi, \tau) \quad = \quad \frac{1}{\sqrt{2\pi\kappa^{(2)}_{X,i}(\tau)}}\exp\left(-\frac{(\chi - \kappa^{(1)}_{X,i}(\tau))^2}{2\kappa^{(2)}_{X,i}(\tau)}\right). \tag{10}$$

The normal distribution has a well-characterized description of the population distribution with respect to its mean and standard deviation. This relationship is represented by the erf function such that the fraction within $[\kappa^{(1)} - n\sqrt{\kappa^{(2)}}, \kappa^{(1)} + n\sqrt{\kappa^{(2)}}]$ is equal to erf$\left(\frac{n}{\sqrt{2}}\right)$. Using this relationship, we solve for the distance between the mean (μ) and the saturation level (χ) with

respect to the standard deviation and derive the following equation

$$\begin{aligned} \text{distance between } \kappa^{(1)} \text{ and } \chi &= n \\ &= \frac{\chi - \kappa^{(1)}}{\sqrt{\kappa^{(2)}}} \end{aligned} \tag{11}$$

Using this, we compute the *completed* fraction of the population, X_i above χ, at any given time t. Because the normal distribution is symmetric and the fraction given by the erf function includes the population both above and below the mean, we divide the fraction by two and add to or subtract from $1/2$ in order to avoid incorrect counting. Finally, the cumulative and probability distribution of completion time, $F(\chi,t)$ and $f(\chi,t)$, are given in (12) and (13). $\dot{\kappa}_X^{(1)}(t)$ and $\dot{\kappa}_X^{(2)}(t)$ are the time derivative of the first and second order cumulants of X.

Even in the cases where the analytical solutions of the first and second order cumulants are not available, numerical approximations are easily obtained through brute force Euler integration. In the following section, the algorithm is verified with the SSA and compared with previous algorithms for computing CTPD.

5. Plan of Work, Schedule and Required Resources

5.1. Plan of Work & Schedule. The long-term goal of this project is to study feedback gene regulatory networks and its role in developmental process. We believe that there exists universal features with in the structure and the relative magnitude of the system parameters that give rise to differentiation of isogenic cells into multiple specialized groups. The proposed project has distinctive goals in both theory and experiments. It is important that the two endeavors are complementary to each other. The schedule, therefore, will have individual estimated duration, but will be concurrent. The 12 months starting from December 2010 are divided into four quarters of 3 months each.

5.2. Required Resources. The experimental side of the research requires basic biological laboratory equipments, a flow cytometer, a plate reader, and a microscope. Some examples of additional components required to enhance the cellular assays are software for microscope fluorescence image processing where a single cell can be tracked through a long duration of experiments, or multiple filters for imaging of multiple fluorescent protein tags. External resources, not required in residence, are DNA sequencing and synthesis facility,

$$F(\chi,t) = \frac{1}{2} \pm \frac{1}{2}\mathrm{erf}\left(\frac{\chi - \kappa_X^{(1)}(t)}{\sqrt{2\kappa_X^{(2)}(t)}}\right) \quad (12)$$

$$f(\chi,t) = \frac{d}{dt}F(\chi,t) = \pm\frac{1}{2}\left(\frac{-\sqrt{\kappa_X^{(2)}(t)}\dot{\kappa}_X^{(1)}(t) - \dot{\kappa}_X^{(2)}(t)(\chi - \kappa_x^{(1)}(t))}{\kappa_X(t)}\right)\exp\left(-\frac{(\bar{x} - \kappa_X^{(1)}(t))^2}{\kappa_X^{(2)}(t)}\right) \quad (13)$$

	Theory	***Experiment***
1st quarter	Using a variety of tools (matrix norms, asymptotics, etc), identify analytical solutions for upper and lower bounds of the probability distribution cumulants with some reasonable error.	Identify the optimal concentration of inducers for activating the hybrid promoters. Plate reader assay allows measurements over large number of inducer conditions (up to 96 conditions at a time). Ideally, we would want to identify the effect of metabolic strain caused by positive feedback.
2nd quarter	Study the sensitivity of the bounds with respect to system parameters.	Using the optimal condition, study the single cell dynamics of positive feedback gene networks using fluorescence microscope imaging. The fluorescence time series data will resemble the SSA realizations shown in Preliminary Results section - from which the distribution of the response time is derived.
3rd quarter	Investigate different probability distributions of response time that can be engineered via altering the structure of gene regulatory networks.	Repeat the experiments using variations of the current construct (e.g. different RBS, promoters) and initial conditions (e.g. copy number of plasmids). Using the data, we can invalidate some candidate models and make modifications.
4th quarter	Theorize what different distributions in response time would mean biology, specifically in the developmental context.	

Bibliography

[1] D. Wolf, V. Vazirani, and A. Arkin. Diversity in times of adversity: probabilistic strategies in microbial survival games. *Journal of Theoretical Biology*, 2005.

[2] M. Kaern, TC Elston, WJ Blake, and JJ Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 2005.

[3] T. Suda, J. Suda, and M. Ogawa. Single-cell origin of mouse hemopoietic colonies expressing multiple lineages in variable combinations. *PNAS USA*, 80:6689, 1983.

[4] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403, 2000.

[5] T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle swithch in Escherichia coli. *Nature*, 403(20), 339-342 2000.

[6] M. Hirsch, S. Smale, and R. Devaney. *Differential equations, Dynamical systems, and an introduction to chaos*. Elsevier Academic Press, 2004.

[7] N. Nise. *Control Systems Engineering*. Wiley, 2000.

[8] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[9] D. Gibson and et. al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 2010.

[10] R. Carlson. The changing economics of DNA synthesis. *Nature Biotechnology*, 27, 2009.

[11] M. Feinberg. Lectures on Chemical Reaction Networks. Online - Departments of Chemical Engineering and Mathematics, The Ohion State University, 1979.

[12] D. Stroock. *An Introduction to Markov Processes*. Springer, 2005.

[13] D. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188:404–425, 1992.

[14] D. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 81(25), 1977.

[15] J. Hespanha. Modeling and Analysis of Stochastic Hybrid Systems. *IEEE Proc - Control Theory and Applications, Special Issue on Hybrid Systems*, 153:520–535, 2007.

[16] B. Alberts. *Molecular Biology of the Cell (Chapters 1 - 7)*. New York : Garland Science, 4th edition, 2002.

[17] M. Elowitz, A. Levine, E. Siggia, and P. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[18] T. Knight. Idempotent Vector Design for Standard Assembly of Biobricks. Technical report, MIT, 2003.

[19] G. Suel, J. Garcia-Ojalvo, L. Liberman, and M. Elowitz. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440:545–550, 2006.

[20] A. Kashiwagi, I. Urabe, K. Kaneko, and T. Yomo. Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced Attractor Selection. *PLoS ONE*, 2006.

[21] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U.S.A.*, 102:4936–4942, Apr 2005.

[22] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, M. J. Schilstra, P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, H. Bolouri, and E. H. Davidson. A genomic regulatory network for development. *Science*, 295:1669–1678, Mar 2002.

[23] V. F. Hinman, A. T. Nguyen, R. A. Cameron, E. H. Davidson, and E. H. Davidson. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 100:13356–13361, Nov 2003.

[24] A. Eldar and M. Elowitz. Functional roles for noise in genetic circuit. *Nature*, 467, 2010.

[25] G. Bel, B. Munsky, and I. Nemenman. The simplicity of completion time distributions for common complex biochemical processes. *Physical Biology*, 7(1), 2010. Bel, Golan Munsky, Brian Nemenman, Ilya.

[26] S. Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001.

[27] R. Cox, M. Surette, and M. Elowitz. Programming gene expression with combinatorial promoters. *Molecular Systems Biology*, 2007.

[28] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 1997.