Engineering Novel Time Distributions in Gene Regulatory Networks

Seunghee Shelly Jang
Self-Organizing Systems Laboratory, University of Washington

I. Introduction

The 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 differentiation in developing organisms [3]. If we systematically characterize the stochasticity in each context, we get closer to 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 for cancer [4]. 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 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 (e.g. coin-toss), where one of the outcomes is the victory outcome (e.g. a head). When an individual is the first one who lands 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. These individuals also differentiate from their initial state and from the new leader. Thus, it is not far-fetched to imagine that a similar mechanism takes place in developmental differentiation processes.

Let us assume that an individual lands a head for the first time after h tries. Because coin-tosses have probabilistic outcomes, h is a random variable characterized by its probability distribution. Let us assume that the probability distribution of h can be manipulated somehow – with a biased coin, for example – then, the variance of h has an interesting interpretation in the leader election example. If the variance of h is set small, the probability of multiple individuals each ending up with the victory outcome in a short amount of time near the mean 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. We propose to systematically characterize this relationship.

The following proposal is organized as follows. In Section II, we propose a translation of the leader election with a biased coin-toss example into a cellular context, specifically in the development phase. Then we identify the objectives 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. In Section III, the prerequisite backgrounds for the characterization methods, both in theory and experiments are introduced, followed by related works in the field that serve as helpful starting points for the proposed research. In Section IV, some preliminary results, conclusions and recommendations are presented. In Section V, we discuss the plan of work and the tentative schedule for the next year.

II. Approach and Objectives

One can suggest several intracellular analogs to the coin-flip example so long as the mechanism is probabilistic, such as protein dimerization, folding, or saturation. Let us consider the protein saturation example, where there is no protein of interest, X, is present initially and the gene for X gets activated. As the gene begins to express, 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, N, 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 $t = T_c$, n_X reaches N. 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 if we further assume that X saturation determines the state of the cell, then the differentiated distribution of the cell population is affected by the probability distribution of T_c.

There can be a number of ways to manipulate the probability distribution of the T_c in this example. One way is to vary the frequencies of gene activation and deactivation. Another way is to vary 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 (system parameter) of the gene regulation, and feedback or open-loop variations change the qualitative features (structure) of the gene regulation. The relationship between these features of gene regulations and the probability distribution of completion times will help us understand the fundamental design principles employed by nature that uses feedback mechanisms to perform development and differentiation in multicellular organisms. Therefore, we propose the following objectives to characterize this relationship.

– **Synthesize genetic regulatory networks with feedback 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 promoter that is activated by external inputs. For the two feedback mechanisms, a single promoter network that expresses either its own repressor or activator will be synthesized. These synthetic networks will have inducible promoters and a fluorescent protein gene.

The inducible promoters allow us to measure the completion time from the time of induction, and the level of fluorescence emitted by the fluorescent protein is measured to monitor the gene expression activity.

– **Model the three gene networks and analyze the completion time distributions**. Using the Chemical Reaction Network theory, we will propose several models of the gene networks in varying levels of detail. We will 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 structures, and discuss how they make each structure a better or worse suited mechanism for differentiation. Because in addition to structural differences, parametric sensitivity differences will also determine the capacity of each network in differentiation, we propose to investigate the quantitative differences of the networks as well.

– **Iteratively verify predictions made in the models with experiments and modify the models based on the experimental results**. The probability distribution of the completion time in the synthetic 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 level reaches an arbitrary saturation value. On the other hand, flow cytometry reveals the population *distribution* of fluorescence level at each measurement. We will measure the time-series of the distribution and compute the fraction of population that reached the saturation value, which is equivalent to the cumulant distribution distribution of completion time. The experimental results will be used to invalidate and identify the features of candidate models that require modification to attain better fidelity to the system. The modified models, in turn, will be 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 differentiation mechanisms.

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. 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.

III. Background and Literature Review

Synthetic gene networks are built from borrowed parts, such as natural promoters and transcription factors, and the precision of synthesis is improving with the advance of biotechnology. Though manipulating genetic materials is not a new technology, synthetic biology is different from the traditional genetic engineering in its intention to engineer novel behaviors, such as oscillation or bistability [5], [6]. One of 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 strengthened 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 [7], linear systems theory the stability, controllability, and observability [8], and probability theory the stochastic behaviors in the mesoscopic level of biological molecules [9]. At the same time, increasing efficiency of cloning techniques [10], decreasing cost of DNA synthesis and sequencing [11], and the advance of experimental equipment 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 remainder of this section is divided into two parts to address the fundamentals of each aspect separately.

A. Mathematical Theory

The Chemical Reaction Network theory provides a standardized foundation from which a mathematical description of chemically interacting species inside a fixed volume can be derived [12]. A CRN consists of chemical species (X_i) that interact according to some reactions (R_j), the stoichiometric coefficients of reactants (u_{ij}) and products (v_{ij}) of the reactions, and the rates of these interactions (λ_j). Using the Law of Mass Action, the dynamically changing concentrations of the chemical species of the CRN 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, in they have biochemical interactions, reactant and product species of these interactions, and numerical values for the rates of the interactions. However, the key difference is that whereas chemical systems tend to have a large quantity of each species modeled with continuous concentration unit, biological species tend to be present in much smaller quantities. Additionally, the stochasticity of the biochemical interactions become more pronounced in systems with species in small quantities changing in discrete amounts. Therefore, biochemical systems, such as gene regulatory networks, require mathematical descriptions that properly address the *discrete* copy number of species and the *stochasticity* of interactions.

Using discrete-state continuous-time Markov processes, the stochastic and discrete nature of gene regulatory networks can be modeled [13]. Let the species of a gene regulatory network be denoted by a vector $\mathbf{S} = [S_1, \cdots, S_n]$, and the number of each species denoted by N_i. Then, each discrete state of the system is denoted by the vector $\mathbf{N} = [N_1, \cdots, N_n]$. Because the stochasticity of gene networks forces the description of the system from a deterministic value to a probability distribution over the states, we denote the probability of the system in state $\mathbf{N}$ at time t by $p(\mathbf{N}, t)$. 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 Chemical Master Equation (CME).

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

where the matrix $\mathbf{Q} = [q_{ij}]$, and q_{ij} is the transition rate from state j to state i [14]. The analytical solution of (1) is

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

The matrix exponential, $e^{\mathbf{Q}t}$, makes the computational cost of

the solution (2) prohibitively expensive. Therefore, instead of solving for the probability distribution dynamics analytically, it can be solved numerically to provide exact realizations of CRN with a fixed set of parameters.

The Stochastic Simulation Algorithm (SSA) numerically simulates individual trajectories of the species of a stochastic CRN [15]. 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. By generating a large number of simulations, the time evolution of a stochastic system can be approximated and the dynamics of each species is obtained. It should be clarified that the probability distribution dealt with in the CME corresponds to the *joint* probability distribution of each specific state, $P(\mathbf{N},t)$, whereas the approximated probability distributions obtained using the SSA are the *marginal* probability distribution of each species, $P(N_1,t)$. Though more straightforward for portrayal of probability distribution dynamics of individual species, the SSA algorithm requires that the initial condition and the rate constants be specified *a priori*. Thus, if an analysis requires a different set of parameters or initial conditions, a whole new set of large number of simulations is required.

Another approach to characterizing the evolution of the probability distribution for stochastic biochemical systems is to compute the cumulant dynamic of each species of the system [16]. The cumulants of a random variable is a set of values that characterizes 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,

$$G_Y(s) \quad = \quad \log\left\langle e^{sY}\right\rangle, \tag{3}$$

where Y is a random variable and $\langle\cdot\rangle$ denotes the expected value. The nth order cumulant of Y 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. To compute the time evolution of cumulants, the extended generator is used. Let $\psi(Y(t))$ be some test function of state $Y(t)$, then the expected value of the test function evolves according to the following equation.

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

where the $Y^j(t)$ is the state after the reaction $\mathbb{R}_j : Y(t) \mapsto Y^j(t)$ has occurred, and λ_j is the reaction rate constant. The cumulant dynamics is then obtained by letting $\psi(Y(t)) = G_Y(s)$, and solving the resulting set of ordinary differential equations. In some cases, depending on the reaction order of a CRN, the cumulants of order i depends on the cumulants of order $i+1$, requiring calculation of infinitely many orders of cumulants. In order to obtain a closed-form solution, the cumulants can be truncated or approximated using various methods [17].

B. Biology Background

Within a single cell resides a genome, a chain of DNA molecules, containing all the genetic information the cell needs to harvest energy, reproduce and survive. Though mighty in its information content, the genome alone cannot make a living organism. It requires molecular machinery that actualizes the information in a useful form. Thus, DNA is transcribed into RNA, RNA is translated into protein, and proteins perform the necessary biological functions [18]. The gene regulatory function of proteins are crucial, such that without the proper regulation of transcription and translation, the entire genome would be uniformly transcribed and translated all the time and mean a disaster for the cell. The intricately connected networks of gene regulation exists to ensure that each cell is viable and functional.

The two major components of gene regulatory mechanisms are promoters and transcription factors (TF). TFs are protein complexes that binds to the promoter of a gene to regulate the expression. Promoters are sequences of DNA that are found at the 5'- end of genes and serve as the recognition sites for RNA polymerase to initiate RNA synthesis. The promoter sequences contain operators that serve as binding sites for specific TFs. A large number of TF and promoter pairs have been identified, and synthetic gene networks are designed and built by arranging them in specific configurations [19], [20]. For example, the critical structure of stress response in *B. subtilis* were identified by synthesizing a gene network with identical promoters and TFs, but with one of the two feedback loops (coupled positive and negative feedbacks) removed [21]. This synthetic version, 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.

There are other examples where feedback mechanisms are observed. A class of gene networks that give rise to stochastic state switching (e.g. cancer and developmental differentiation) has been consistently shown to contain positive feedback loops [22], [23], [24], [25]. Another type of behavior that arises from gene networks with positive feedback loops is procrastinating differentiation [26]. It refers to the phenomenon observed in isogenic cells, that when triggered for specific response (e.g. sporulation, apoptosis), the response times of the cells widely vary within the microcolony. This phenomenon is closely related to the leader election example – however, the lack of communication between the individuals presents itself in the form of unimodal steady-state after some transient multimodal distribution.

C. Biological Experiments

Gene network synthesis procedure can be broken into two major steps. The first step is acquiring the desired DNA sequences (e.g. promoters, transcription factor genes), and the second step is joining these pieces together in the right order. Natural promoters and genes are obtained from the host organism's genome through Polymerase Chain Reaction (PCR). In this process, the desired sequence is isolated and amplified by using two short pieces of single-stranded DNA that are complementary to the 3'- and 5'- ends of the desired sequences. Through cyclical temperature manipulation of the reaction chamber, the copy number of the desired sequence amplifies exponentially. Then the amplified pieces are digested using restriction enzyme to introduce recognizable ends to each piece, and assembled

together using DNA ligase. The restriction enzymes were chosen strategically to ensure that when the DNA ligase assembles the pieces together, the right order and direction are maintained. The product is transformed into a host organism by electroporation before it is ready for assays.

Modern cellular assay tools such as the flow cytometry, time-lapse fluorescence microscopy, and plate readers are made possible by the Green Fluorescent Protein (GFP) [27]. GFP can be fused to or co-expressed with a protein of interest, and by measuring the level of fluorescence emitted when excited at the appropriate wavelength the gene activity is monitored. For example, in a flow cytometer, cells suspended in liquid culture are passed through a cylindrical passage where they are subject to excitation. The intensity and amount of scattered light is used to compute the relative size, internal complexity and fluorescence intensity of cells.

D. Related Work

Using the Laplace transform, the CME of 1-dimensional complex biochemical processes can be solved [28]. In this work, a kinetic proofreading (KPR) process was modeled by a Markov chain with an absorbing state that corresponds to the completion of the proofreading process. The time derivative of the absorbing state probability is equal to the probability distribution of completion time [29], and the exact solution can be computed by using the Laplace transform. The solution showed that the distribution approaches to Γ- or exponential distribution depending on the direction of the bias imposed by the transition rates. However, the solution and the conclusion are limited to an open-loop system where the transition rates are independent of the states. The Laplace transform approach discussed in this work, after significant modification, may prove to be helpful in analyzing systems with feedback – those with state-dependent transition rates.

A type of feedback loops in gene regulatory networks involves a promoter that are regulated by the protein that acts as the regulator for the promoter (*transcriptional regulation*). Naturally occurring feedback loops are interesting in themselves, but a synthetic class of hybrid promoters developed to exhibit the programmability of promoters expands the possible design space [30]. Hybrid promoters are synthesized by combining multiple operator sites corresponding to different TFs, and often have tighter regulation of uninduced leaky expression and larger range of expression. The range of expression is controlled by varying the inducer concentrations and owing to this feature, a fine-tune control over synthetic gene networks is possible.

IV. Preliminary Results

A. Approximation of the probability distribution of completion time

Using the basic understanding of the gene expression mechanism, we expressed the feedback gene regulatory network with the following Chemical Reaction Network.

$$G_u + X \underset{\alpha\ ka}{\overset{ka}{\rightleftharpoons}} G_b \qquad G_u \xrightarrow{\beta_2 ex} G_u + X$$

$$X \xrightarrow{dx} \phi \qquad G_b \xrightarrow{\beta_1 ex} G_b + X$$

where X is the transcription factor, G_u is an X gene not bound with X, and G_b is an X gene bound with X. The rate constants of reactions are: rate of transcription factor binding (ka), unbinding-to-binding ratio (α), ratio of unbound gene expression to basal expression (β_1), ratio of bound gene expression to basal expression (β_2), and degradation/dilution rate of X (dx). We simulated 1000 SSA realizations of the CRN to visualize the dynamics of X and approximate the probability distribution of the completion (Figure 1). We denote the number of X with n_X, and set the completion of protein saturation to be when n_X reaches the half steady-state value, N, because it was observed that after n_X is sufficiently higher than N, the probability of the n_X dropping below N is small. It can be said with some confidence that the cell is committed to its fate after n_X exceeds N.

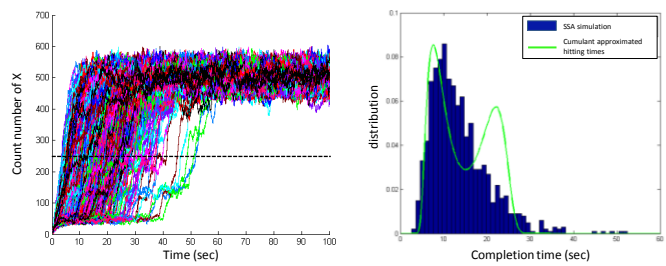

Fig. 1: (a) Thousand trajectories of the CRN obtained from the Stochastic Simulation Algorithm. (b) the approximated probability distribution of completion time (completion is when $n_X > N$, the horizontal black line in (a)). The parametric values used are $[ka, \alpha, \beta_1, \beta_2, ex, dx]$ = [$10^{-3}\log(2)$, 0.1, 10, 1, $10\log(2)$, $\log(2)$] and the initial condition is $[G_u, G_b, X]$ = [5, 0, 0]. The green curve is the approximated completion time distribution using the cumulant truncation method.

We applied the extended generator method to compute the cumulant dynamics of the system. Because the reaction order of the CRN is 2 (because of the bimolecular reaction), each ith order cumulant is a function of $i+1$th order cumulants, requiring infinitely many orders of cumulants. Therefore, the cumulants with order 3 and higher were truncated to obtain a closed-form solution. This is equivalent to assuming that the population of each species has a Gaussian distribution at all times. And since a Gaussian random variable is distributed by,

$$g(\hat{n}_X) = \frac{1}{\sqrt{2\pi\kappa_{n_X n_X}}} \exp\left(-\frac{(n_X - \kappa_{n_X})^2}{\sqrt{2\kappa_{n_X n_X}}}\right), \tag{5}$$

the fraction of X above N as a function of time is given by

$$F(N,t) = \frac{1}{2} \pm \frac{1}{2}\mathrm{erf}\left(\frac{N - \kappa_{n_X}(t)}{\sqrt{2\kappa_{n_X n_X}(t)}}\right) \tag{6}$$

where κ_{n_X} and $\kappa_{n_X n_X}$ are the first and the second order cumulants of n_X. This function is an approximation of the cumulative distribution of the completion time. By taking the derivative of (6) with respect to time, the probability distribution of completion time is approximated as follows.

$$\begin{aligned} f(N,t) &= \frac{\partial}{\partial t} F(N,t) \\ &= \frac{1}{2}\left(\frac{-\sqrt{\kappa_{n_X n_X}(t)}\,\dot{\kappa}_{n_X}(t) - \dot{\kappa}_{n_X n_X}(t)(N - \kappa_{n_X}(t))}{\kappa_{n_X}(t)}\right) \times \\ &\quad \exp\left(-\frac{(N - \kappa_{n_X}(t))^2}{\kappa_{n_X n_X}(t)}\right) \end{aligned} \tag{7}$$

where $\dot{\kappa}$ denotes the time derivative. The method requires numerical solution to the nonlinear ODE of cumulants and numerical differentiation of the cumulative distribution of completion time. Also, depending on the choice of the threshold, the integral of $f(N,t)$ does not approach 1 as $t \to \infty$. But in such cases, we are able to compute the exact error. Using this cumulant truncation method, completion time probability distributions of two different sets of parameters and initial conditions were approximated (Figure 2). For the first set, the curve (cumulant truncation) agrees well with the histogram (SSA), but in the second set, the curve shows a bimodal distribution where the first mode is a fair approximation of the histogram, whereas the second mode is unseen in the histogram.

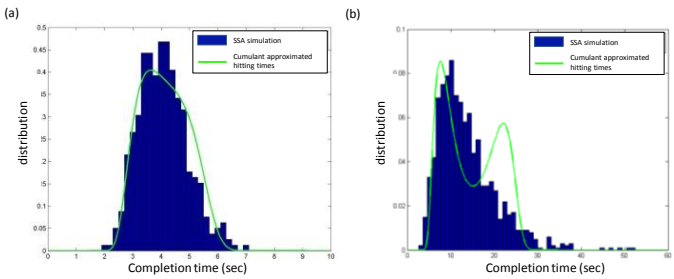

Fig. 2: Completion time probability distribution approximated using the SSA and the cumulant truncation method. (a) $[ka, \alpha, \beta_1, \beta_2, ex, dx]$ = $[10^{-2}\log(2), 0.1, 1, 0.001, 10^{1.5}\log(2), \log(2)]$ and the initial condition is $[G_{unboud}, G_{bound}, X]$ = [9, 1, 1]. (b) Identical to the condition used in Figure 1.

B. Synthetic positive feedback gene network in E. coli

Design and Construction. Because even a small amount of output can trigger the rapid activation of a positive feedback mechanism, we employed promoters that have minimal leaky expression and give maximal control over the range of expression [31]. We selected two hybrid promoters from [30] – A12 and D61 – that have operator sites from the pBAD and pLac promoters. The promoters are activated by AraC-arabinose complex, and repressed by LacI protein. To create a positive feedback loop, we cloned the araC gene downstream of the promoter so that when induced with arabinose, the promoter is activated (Figure 3 (a)). Additionally, we used a strain of *E. coli* that constitutively expresses lacI, to keep the promoter tightly regulated when uninduced by IPTG. The recombinant DNA was obtained using restriction enzymes and DNA ligase. Currently, there are several variations of the positive feedback gene network (Figure 3 (b)). Each network has the promoter A12 or D61, and placed in a plasmid with the origin of replication pMB1, pSC101, or pSB3K3[1].

Assays. Three different types of assay were used to characterize an A12 promoter variant of the positive feedback gene network in varying concentrations of IPTG and arabinose. The following is a brief summary of each assay, including the objectives, methods, results, and conclusion.

– **plate spectrophotometry**

· *Objective*. To confirm the response behavior of the hybrid promoter predicted in the original paper [30] in changing concentrations of two inducers.

[1]Each has an approximate copy number of $10^0, 10^1$, and 10^2, respectively

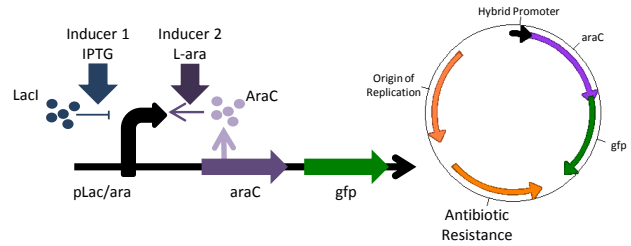

Fig. 3: A diagram of a synthetic positive feedback gene regulatory network. (a) The pLac/ara hybrid promoter is induced by IPTG and arabinose. IPTG inhibits the LacI repression of the promoter, whereas arabinose forms a complex with AraC protein and activates the promoter. (b) The network has a hybrid promoter, controlling the expression of araC and gfp downstream, an origin of replication and an antibiotic resistance marker.

· *Methods*. We inoculated LB media with a single colony from an agar plate of cells transformed with an A12 gene network. The culture was diluted 1:150 in PBS with 48 different concentrations of IPTG and arabinose in a 96 well plate (duplicates were made for each condition). Each well contained 0%, 0.01%, 0.05%, 0.1%, 0.5%, 1% or 2% arabinose and 0μM, 10μM, 50μM, 100μM, 500μM, or 1mM IPTG. The plate reader was set to measure the optical density and the fluorescence level of each well every 20 minutes over 24 hours. The plate was kept in 37C and shaken for 10 minutes before each measurement.

· *Results and Discussion*. The rate of fluorescence level increase was slower and the steady-state expression level was lower for arabinose induction conditions relative to IPTG (Figure 4 (a)). This confirmed that the hybrid promoter response was consistent with the prediction.

– **flow cytometry**

· *Objective*. To measure the distribution of the fluorescence level of cells from the time they are induced until steady-state is reached, and to observe the transient in the mean and the variance of the fluorescence level.

· *Methods*. We used the same colony from the spectrophotometry assay to inoculate LB media with varying concentrations of IPTG and arabinose. 12 different concentrations were tested; each culture tube contained 0%, 0.1% 1% or 10% arabinose and 0μM, 10μM or 100μM IPTG. In 15 minute intervals, for 4.5 hours, 10μL of culture from each tube was diluted 1:15 in 96 well plates with PBS in each well. 25000 events (cells) from each well were screened using an Accuri C6 flow cytometer. Additional measurements were made at 5 hrs, 6 hrs, 20.5 hrs and 21.5 hrs after induction.

· *Results and Discussion*. At 30 minutes after induction, the mean of the fluorescence level distribution was higher compared to the initial mean value at the time of induction. From 45 min to 105 min after induction, the mean fluorescence level decreased. At 120 min, two distinct populations of fluorescence level – a median value at $10^{2.5}$ and another at $10^{2.9}$ – were observed and again at 135 min (Figure 4 (b)). However, the bimodal distribution disappeared abruptly in the next measurement and no more higher fluorescence population was observed. We hypothesize that the cell population with activated positive feedback are suffering from AraC/GFP toxicity.

– **time-lapse microscopy**

- *Objective*. To confirm whether the high fluorescing cells – the cells with activated positive feedback – have different viability compared to the lower fluorescing cells.
- *Methods*. An agar plate with 10% arabinose and 1mM IPTG was prepared. An overnight culture was diluted in the morning and grown for 3 hours to reach log-growth phase. 2uL of cells were transferred onto the agar plate. Using the microscope, 10 sparsely populated areas were selected. Within each area, a single cell was marked for tracking, and every 10 minutes for 12 hours an image processing macro tracked each cell, adjusted the focus, and took an image of the cell.
- *Results and Discussion*. The cells that began to emit high level fluorescence were shown to grow larger than its peers with low level fluorescence (Figure 4 (c)). Eventually these cells died, supporting the theory that over-expression of araC is toxic.

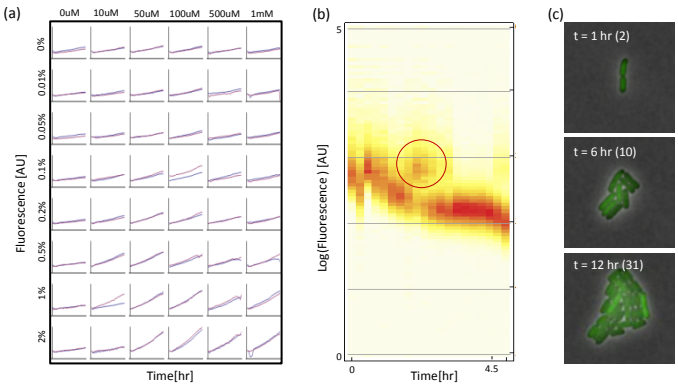

Fig. 4: Preliminary results from three assays - (a) Plate spectrophotometry, (b) flow cytometry, and (c) time-lapse microscopy. (a) Normalized fluorescence over 24 hour in 48 different conditions of IPTG and arabinose are shown. Each subplot has fluorescence level range (y-axis) from 5500 to 16500 [au], and time (x-axis) from 0 to 24 hours. The inducer concentrations of IPTG and arabinose are shown in the top row and the first column. (b) Fluorescence distribution time-series measured over 4.5 hours in 15 minute intervals. The emergence and disappearance of high fluorescence population is indicated with a red circle. (c) Three sample images from a time-lapse microscopy – 1 hr, 6 hr, and 12 hr (cell population shown in parenthesis).

V. Plan of Work

A. Theory

We propose to model the system discussed in the previous section as a Markov process model and rigorously determine the probability of the system returning to its initial state after reaching the state of $n_X \geq N$ to test the validity of assuming the completion as an absorbing state. Also, the approximation method introduced will be investigated further. Though fair, the Gaussian assumption approach requires n_X to be a continuous random variable. Therefore, an approximation with a discrete probability mass function equivalent to the Gaussian function will be explored to obtain a similar solution. Furthermore, the error shown in Figure 2 (b) needs to be corrected, possibly through using different numerical algorithms for solving the cumulant dynamics ODE.

An alternative way of approaching the problem is to identify the upper and lower bounds of the mean and the variance of the completion time. Using a variety of tools (matrix norms, asymptotics, etc,) we will obtain the analytical solutions for the limits and their parametric sensitivity. This analysis will help us identify the parameters that would result in the largest observable difference in the experiments with the synthetic gene network. Additionally, we will study the possibility of engineering the gene networks to obtain arbitrary shape of completion time probability distributions, such as Gaussian or uniform. Furthermore, we will analyze negative feedback and open-loop systems using analogous approach. We will compare and contrast the features of the completion time probability distributions and discuss how each network is suited for different differentiation steps.

B. Experiments

We plan on conducting single-cell tracking time-lapse fluorescence microscopy experiments in varied concentrations of the inducers to obtain time series data of fluorescence level that resemble Figure 1 (a), and approximate the completion time probability distribution that resembles Figure 1 (b). Additionally, we will tune the gene network with respect to the system parameters used in the CRN. For example, we can change the *ex* value – the rate of gene expression – by tuning the ribosome binding site of the RNA or the operator sites of the promoter. The tuning and estimation of parameters will help us systematically reconcile the model and the actual system. As with the theoretical approach, we will synthesize negative feedback and open-loop gene network with focus on keeping the extraneous details – the inducers, plasmids, and *E. coli* strain – as equivalent as possible between the three variations to minimize external bias affecting the analysis.

We will repeat the concentration variation assays for the D61 hybrid promoter and other hybrid promoters with varied expression range. To test the hypothesis of AraC/GFP toxicity we will employ an araC-knockout strain of *E. coli* to eliminate as much background effect as possible and repeat the assays [32]. Because the inducer concentration will affect the accumulation rate of AraC, we will investigate whether an optimal concentration exists in order to keep the cell viable for a reasonable duration.

VI. Schedule and Required Resources

The schedule of work is shown in Figure 5. The biological equipments and computational software required for the research are funded by the Molecular Programming Project, part of the National Science Foundation's Expedition in Computing program.

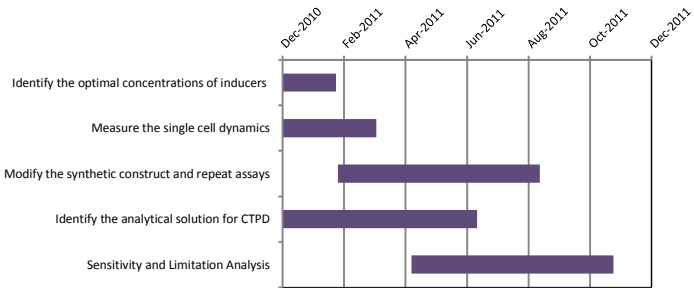

Fig. 5: Schedule of work

References

[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, T. Elston, W. Blake, and J. 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*, vol. 80, p. 6689, 1983.
[4] R. Losick and C. Desplan, "Stochasticity and Cell Fate," *Science*, vol. 320, pp. 65–68, 2008.
[5] M. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, 2000.
[6] T. Gardner, C. Cantor, and J. Collins, "Construction of a genetic toggle swithch in Escherichia coli," *Nature*, vol. 403, no. 20, 339-342 2000.
[7] M. Hirsch, S. Smale, and R. Devaney, *Differential equations, Dynamical systems, and an introduction to chaos*. Elsevier Academic Press, 2004.
[8] N. Nise, *Control Systems Engineering*. Wiley, 2000.
[9] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.
[10] D. Gibson and et. al., "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, 2010.
[11] R. Carlson, "The changing economics of DNA synthesis," *Nature Biotechnology*, vol. 27, 2009.
[12] M. Feinberg, "Lectures on Chemical Reaction Networks," Online - Departments of Chemical Engineering and Mathematics, The Ohion State University, 1979.
[13] D. Stroock, *An Introduction to Markov Processes*. Springer, 2005.
[14] D. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A: Statistical Mechanics and its Applications*, vol. 188, pp. 404–425, 1992.
[15] ——, "Exact Stochastic Simulation of Coupled Chemical Reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, 1977.
[16] J. Hespanha, "Modeling and Analysis of Stochastic Hybrid Systems," *IEEE Proc - Control Theory and Applications, Special Issue on Hybrid Systems*, vol. 153, pp. 520–535, 2007.
[17] ——, "Moment closure for biochemical networks," *Proc of the Third Int. Symp. on Control, Communications and Signal Processing*, 2008.
[18] B. Alberts, *Molecular Biology of the Cell (Chapters 1 - 7)*, 4th ed. New York : Garland Science, 2002.
[19] M. Elowitz, A. Levine, E. Siggia, and P. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–1186, 2002.
[20] T. Knight, "Idempotent Vector Design for Standard Assembly of Biobricks," MIT, Tech. Rep., 2003.
[21] G. Suel, J. Garcia-Ojalvo, L. Liberman, and M. Elowitz, "An excitable gene regulatory circuit induces transient cellular differentiation," *Nature*, vol. 440, pp. 545–550, 2006.
[22] 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.
[23] M. Levine and E. H. Davidson, "Gene regulatory networks for development," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 4936–4942, Apr 2005.
[24] 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*, vol. 295, pp. 1669–1678, Mar 2002.
[25] 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.*, vol. 100, pp. 13 356–13 361, Nov 2003.
[26] A. Eldar and M. Elowitz, "Functional roles for noise in genetic circuit," *Nature*, vol. 467, 2010.
[27] M. Chalfie, Y. Tu, G. Euskirchen, W. Ward, and D. Prasher, "Green Fluorescent Protein as a Markoer for Gene Expression," *Science*, vol. 263, no. 5148, pp. 802–805, February 1994.
[28] G. Bel, B. Munsky, and I. Nemenman, "The simplicity of completion time distributions for common complex biochemical processes," *Physical Biology*, vol. 7, no. 1, 2010, bel, Golan Munsky, Brian Nemenman, Ilya.
[29] S. Redner, *A Guide to First-Passage Processes*. Cambridge University Press, 2001.
[30] R. Cox, M. Surette, and M. Elowitz, "Programming gene expression with combinatorial promoters," *Molecular Systems Biology*, 2007.
[31] 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.
[32] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection," *Mol. Syst. Biol.*, vol. 2, p. 2006.0008, 2006.