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.

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 was 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 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 (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 to perform development and differentiation in multicellular organisms. Therefore, we propose the following objectives to characterize this 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 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 with 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.

- **Mathematically model the three gene networks**. 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]. 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, biological species tend to be present in much smaller quantities. Thus, biological molecules must be expressed as discrete variables instead of continuous variables. Additionally, the stochasticity of the biochemical interactions become more pronounced in systems with species in small quantities. Therefore, biochemical systems, such as gene regulatory networks, require mathematical descriptions that properly addresses 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).

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

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

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

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 are 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,

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

where X is a 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. To compute the time evolution of cumulants, the extended generator is employed. Let $\psi(X(t))$ be some test function of state $X(t)$, then the expected value of the 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, and λ_j is the reaction rate constant. The cumulant dynamics is then obtained by letting $\psi(X(t)) = G_X(s)$, 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. Biological Background

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 a 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 perform the necessary biological functions [18]. For example, if the processes of transcription and translation were not properly regulated by protein molecules, 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, 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 and these known pairs are used to design and build synthetic gene regulatory networks 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, thus 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, such as 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]. Procrastination 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 coin-flip leader election example - however, the lack of communication between the individuals presents itself in the form of non-differentiated steady-state, i.e. the population distribution is still uniform, and the variance is only observed during the transient process.

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, and DNA polymerase. Through cyclical temperature manipulation of the reaction chamber, the quantity of the desired sequence amplifies exponentially. Then the amplified pieces are digested using restriction enzyme to introduce recognizable sticky (or blunt) ends to each piece. The individual pieces are assembled together using DNA ligases, which are originally derived from DNA repair mechanism. The restriction enzymes were chosen strategically to ensure that when the DNA ligase assembles the pieces together, they are joined in the right order and direction. The product is transformed into a host organism by electroporation before it is ready for assays.

The first usage of green fluorescent protein (GFP) as a marker for gene expression significantly changed the way gene activity is monitored [27]. Modern cellular assay tools such as the flow cytometry, time-lapse fluorescence microscopy, and plate readers are made possible by GFP-fusion proteins. GFP is fused to a protein of interest via DNA recombination and is monitored for fluorescence emission upon excitation. GFP molecules have excitation wavelength of 395nm or 498nm, depending on the organism it was derived from, and emission wavelength of 505nm. In addition to GFP, there are several different types of fluorescence protein that can be used for multicolor live cell imaging to monitor several different gene activities simultaneously.

D. *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 [28]. 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 the Markov chain, they performed Laplace transform to the solution of the CME shown in (2) to obtain the analytical expression [29]. The solution showed that the distribution of the first-passage time approaches limiting behaviors, 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 analyze simulated systems with *varying* transition rates by using randomly generated values and expand their conclusion, this is quite different from state-*dependent* transition rates of feedback mechanisms. It will be interesting to investigate whether a similar conclusion can be drawn from biochemical processes with feedback.

A feedback loop in a gene regulatory network consists of a promoter that are regulated by some TF, which in turn, is expressed by the gene controlled by the promoter. Naturally occurring feedback loops are interesting in themselves, but a synthetic class of hybrid promoters developed to exhibit the programmability of promoters expands the number of possible feedback loops [30]. Hybrid promoters are synthesized by combining multiple operator sites corresponding to different TFs, and the resulting promoters are regulated by more than one type of TF. Additionally, the order in which these operator sites are arranged was shown to affect the expression level of the downstream gene. All the hybrid promoters studied in [30] are inducible by specific inducer chemicals and the concentrations of these inducers are shown to be correlated with the gene expression as well. Therefore, by employing these hybrid promoters, a large variation in assay conditions is available by varying the inducer concentrations and the copy number of genes.

IV. PRELIMINARY RESULTS

A. *Synthetic positive feedback gene network in E. coli*

Design and Construction. Since 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 promoter that is activated by AraC-arabinose complex, and the pLac promoter that is repressed by LacI protein. To create a positive feedback loop, we cloned araC gene downstream of the promoter so that when induced with arabinose, the promoter is activated (Figure 1 (a)). The cloning procedure was conducted using restriction enzymes and DNA ligase. Additionally, we used a strain of *E. coli* that constitutively expresses lacI, to keep the promoter tightly regulated when uninduced by IPTG. Currently, there are several variations of the positive feedback gene network (Figure 1 (b)). Each network has the promoter A12 or D61, and placed in a plasmid with the origin of replication pMB1, pSC101, or pSB3K3[1].

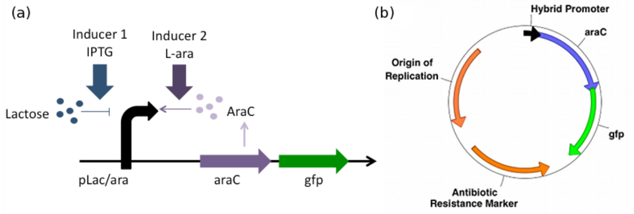

Fig. 1: A diagram of a synthetic positive feedback gene regulatory network. (a) The pLac/ara hybrid promoter is induced by IPTG and arabinose. IPTG inhibits the LacI repression of the promoter, whereas arabinose forms a complex with AraC protein and activates the promoter. (b) The network has a hybrid promoter, controlling the expression of araC and gfp downstream, an origin of replication and an antibiotic resistance marker.

Assays. Three different types of assay were used to characterize the A12 promoter variant of the positive feedback gene networks 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 in changing concentrations of two inducers.
- *Methods*. We chose a single colony from the agar plate of transformed cells and inoculated LB media. 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 0uM, 10uM, 50uM, 100uM, 500uM, or 1mM IPTG. The plate reader was set to measure the optical density and the fluorescence level of each well every 20

---

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

minutes over 24 hours. The plate was kept in 37C and shaken for 10 minutes before each measurement.

- *Results and Discussion*. The A12 variant of positive feedback gene network had minimal change in the steady-state fluorescence value with varying concentration of arabinose, but significant change with varying concentrations of IPTG. This confirmed that the hybrid promoter response was consistent with the prediction.

– **flow cytometry**

- *Objective*. To measure the population distribution of 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 to inoculate LB cultures with varying concentrations of IPTG and arabinose. 12 different concentrations were tested, and each culture tube contained 0%, 0.1% 1% or 10% arabinose and 0uM, 10uM or 100uM IPTG. In 15 minute intervals, for 4.5 hours, 10uL of culture from each tube was diluted 1:15 in 96 well plates with PBS in each well. 50000 events (cells) from each well were screened using a Accuri C6 flow cytometer. Additional measurements were made at 5 hrs, 6 hrs, 19 hrs and 20.5 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 were observed and again at 135 min. 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. xxxuL 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 an image processing macro tracked each cell, adjusted the focus, and took an image of the cell, over 12 hours.
- *Results and Discussion*. The cells that began to emit high level fluorescence were shown to grow larger than its peers with low level fluorescence. Eventually these cells died, supporting the theory that over-expression of araC is toxic.

B. 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.

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

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

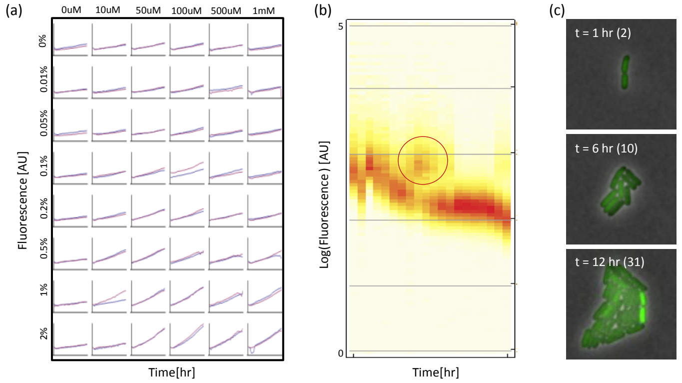

Fig. 2: Preliminary results from three assays - (a) Plate spectrophotometry, (b) flow cytometry, and (c) time-lapse microscopy. (a) Normalized fluorescence over 24 hour in 48 different conditions of IPTG and arabinose are shown. Each subplot has fluorescence level range (y-axis) from 5500 to 16500 [au], 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 distributions 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).

where G_u is an X gene not bound with the transcription factor X, G_b is an X gene bound with X, and X is the transcription factor. The rate constants of reactions are; rate of transcription factor binding (ka), unbinding-to-binding ratio (α), ratio of unbound gene expression to unregulated expression (β_1), ratio of bound gene expression to unregulated expression (β_2), and transcription factor degradation/dilution rate (dx). We simulated 1000 SSA realizations of the CRN to visualize the dynamic of X and approximate the probability distribution of the completion time in this specific condition (Figure 3). We denote the number of X with n_X and set the completion of protein saturation to be when n_X reaches its half steady-state value (N), because it was observed that after a single trajectory with n_X is sufficiently higher than N, the probability of the n_X dropping below N is small. In fact, we have not observed such an event in the simulations. We can be convinced with some confidence that the cell is committed to its fate, after n_X number exceeds N. For future work, we propose to model the system as a Markov process and rigorously determine the probability of the system returning to its initial state after reaching the state of $n_X \geq N$.

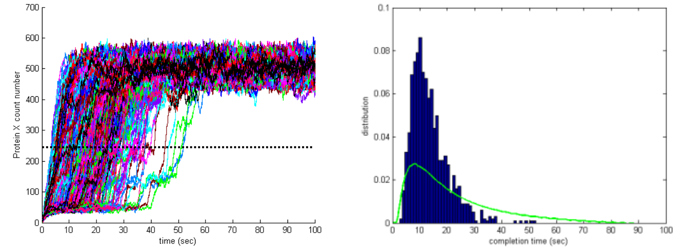

Fig. 3: (a) Thousand trajectories of the CRN obtained from the Stochastic Simulation Algorithm. (b) the approximated probability distribution of completion time (completion is when $n_X > N$, the horizontal black line in (a)). The parametric values used are $[ka, \alpha, \beta_1, \beta_2, ex, dx]$ = $[10^{-3}\log(2)$, 0.1, 10, 1, $10\log(2)$, $\log(2)]$ and the initial condition is $[G_{unboud}, G_{bound}, X]$ = [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. However, since there are reactions of order higher than 1 (i.e. more than one reactant in a given reaction), each ith order cumulant is a function of $i+1$th order cumulants, requiring infinitely many orders of cumulants. Therefore, to obtain a closed-form solution the cumulants with order 3 and higher were truncated. 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

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

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), the probability distribution of completion time is approximated as follows.

$$\begin{aligned} f(N,t) &= \frac{\partial}{\partial t}F(N,t) \\ &= \frac{1}{2}\left(\frac{-\sqrt{\kappa_{n_X n_X}(t)}\kappa'_{n_X}(t)-\kappa'_{n_X n_X}(t)(N-\kappa_{n_X}(t))}{\kappa_{n_X}(t)}\right)\times \\ &\quad \exp\left(-\frac{(N-\kappa_{n_X}(t))^2}{\kappa_{n_X n_X}(t)}\right) \end{aligned} \tag{7}$$

Using this cumulant truncation method, completion time probability distributions of two different sets of parameters and initial conditions were approximated (Figure 3(b)). The normalized version of the curve shows a good approximation of the median, however the distribution in the most populated points are under-estimated, and the right-tail end are over-estimated.

V. PLAN OF WORK

A. Experiments

We will repeat the concentration variation assays discussed in the preliminary result section for the D61 hybrid promoter. And the hypothesis of AraC/GFP toxicity will be investigated further with the use of araC-knockout strain of *E. coli*. Because D61 promoter has lower expression level compared to A12 when fully induced, the negative effect is expected to be subdued. Also, to confirm whether the cells are affected by araC toxicity, we will transform the gene network into strain of *E. coli* that has its araC gene from its genome removed [32]. Then similar assays will be conducted to find the optimal concentrations that will delay the accumulation of AraC. Additionally, a different hybrid promoter with a lower level of expression will be switched in to test whether more viable cells can be obtained.

We plan on conducting a series of single-cell tracking time-lapse fluorescence microscopy experiments of the gene networks in varied concentrations of the inducers to obtain time series data of fluorescence level that resemble Figure 3 (a), and subsequently an approximated completion time probability distribution that resembles Figure 3 (b). The mean and the variance of this distribution are expected to vary with changing conditions of the inducers, which is closely related to varying the parameters ka and α in the model (IV-B). Additionally, we will use other hybrid promoters to study the effect of β_1 and β_2 variations in the completion time probability distribution.

In addition to the positive feedback gene networks, we will design and analyze negative feedback gene networks and open-loop gene networks. The focus of the design will be to keep the details, such as the inducers, the plasmids, and the *E. coli* strains, as equivalent as possible, and vary only the structure of the feedback. This is so that there are no external bias affecting the assay results. The analysis processes will be consistent with those of the positive feedback gene networks (i.e. plate spectrophotometry, flow cytometry, and time-lapse microscopy).

B. Theory

The approximation method introduced in the Preliminary Results section will be investigated further. Though a fair approximation, this approach assumes that n_X is a continuous random variable. Therefore, an approximation with a probability mass function of a discrete random variable equivalent to the Gaussian function will be explored to obtain a similar solution. An alternative way of approaching the problem might lie in identifying the upper and lower bounds of the mean and the variance of the completion time, instead of identifying the exact solution. Using a variety of tools (matrix norms, asymptotics, etc,) we will obtain the analytical solutions for the limits and their parametric sensitivity. The purpose of this analysis is to identify the parameters that would result in the largest observable difference in the experiments with the synthetic gene regulatory system. Additionally, we will study what possible type of completion time probability distributions can be engineered using these gene regulatory networks (e.g. Gaussian, uniform, etc.).

Parallel with the experimental objectives, we will analyze negative feedback and open-loop systems. We expect to observe qualitative differences in these probability distributions, and will discuss the reasons for these differences. Furthermore, we will discuss the connection between the frequency of observation of each type of feedback in development/differentiation processes and the qualitative features of the probability distributions.

VI. SCHEDULE AND REQUIRED RESOURCES

The schedule of work is shown in Figure 4. 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.

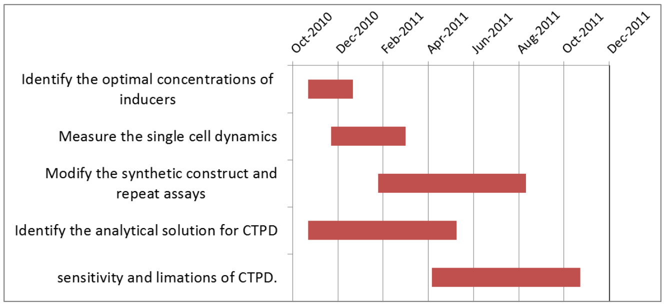

Fig. 4: Schedule of work

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.