AP3162: Gene-regulatory circuits: Stochastic dynamics

Hyun Youk
Delft University of Technology
(Dated: February 13, 2020)

In this lecture, we will discuss how to model stochastic dynamics of standard gene-regulatory circuits.

I. WHY IS GENE EXPRESSION STOCHASTIC?

In the previous lecture, we discussed equations that describe basic gene-regulatory schemes. The equations that we discussed in lecture 2 were deterministic equations - They were ordinary differential equations whose solutions are deterministic. This means that if we know the concentration of the relevant molecules (mRNA or proteins) in a cell at a particular time point, then the equation tells us what the concentrations of those molecules were in the past and what they will be in the future. Information about the present state is all that one would need to exactly specify the past and the future of the cell. This "clockwork" or "mechanical" view of gene expression and cellular processes in general (such as motion of bacteria), is in fact, not entirely accurate. As we will see below, cellular processes such as gene expression in a cell are stochastic. The typical reason that is often given in the literature for this is that there are very few copies of a molecule involved in typical cellular processes, such as the number of mRNA or protein produced from a gene inside a cell. As an example, a cell may contain between one to a hundred copies of an mRNA from a given gene. Dividing 1 to 100 mRNA molecules by the volume of a cell, say V, gives us: 1/V,2/V,?,100/V. We see that the concentration is not continuous, but rather is discretized in steps of 1/V. Ordinary differential equations from the previous lecture fail to capture these discrete changes in concentration (as they also failed to capture stochastic cell division and death - see lecture note 1). But intuition tells us that if 1/V is "small enough", then we can "smooth out" the discreteness. As we will see, the "small enough" quantitatively means that 1/V is much smaller than the average concentration. In other words, when there are large number of copies of a molecule (mRNA or protein from a gene) inside a cell, we can use the deterministic equations that we derived in the previous lecture. This intuitively makes sense because if we have 1000,000 molecules, then losing or gaining 1 or 2 molecules doesn't make much difference.

But as we mentioned in lecture 1, there is a deeper reason for why gene expression is stochastic. If we watch a single cell over time, we cannot predict exactly when it will transcribe a mRNA from a given gene and when it will translate a protein from each mRNA. This is because we lack the necessary information such as exactly when each RNA polymerase will bind to the DNA, the location and movement of each copy of mRNA from the gene of interest, the location and movement of each ribosome, and many others. The key here is that the amount of information that we would need to construct the movement of all the key molecules inside the cell by using Newton's laws of motion and any other laws of physics and chemistry would be huge. Thus, even if we could see the location of every ribosome and RNA polymerase inside a cell, we would need to know the location of all other molecules inside the cell that collide with them in order to do better than describing their motion as diffusion (the later being a probabilistic description). For the same reason that we can, at best, only give a stochastic description of cell division and death (see lecture note 1), we can only give, at best, a stochastic description of gene expression - the number of copies of mRNA and protein from a given gene over time is a stochastic quantity. As we will see, the stochastic description approaches a deterministic description as the number of copies of the molecule approaches a sufficiently "large" number (to be quantified below).

Below, we will start by building two standard approaches for modelling **stochastic gene expression**. We will then describe the stochastic dynamics of the standard forms of gene circuits - constitutive circuit, positive feedback, negative feedback. We will end by applying stochastic gene expression to bacterial competence.

II. MASTER EQUATION

In lecture note 1, we derived the **Master equation** by using stochastic birth and death of cells. But if we replace "cells" by "molecules", "birth of a cell" by a "creation of a molecule in a cell", and "death of a cell" by "degradation of a molecule in a cell", then we obtain the same Master equation to describe the number of molecules inside a cell. Specifically, let n be the number of copies of a molecule (e.g., mRNA or protein from gene X). Then the probability $P_n(t)$ that there are n copies of the molecule in a cell at time t is given by the following Master equation

$$\frac{dP_n}{dt} = f_{n-1}P_{n-1} + g_{n+1}P_{n+1} - g_nP_n - f_nP_n \tag{1}$$

where $f_n dt$ is the probability that a cell with n copies of the molecule will create another copy of the molecule in time interval $(t,\ t + dt)$, and $g_n dt$ is the probability that a cell with n copies of the molecule will degrade a copy of the molecule in time interval $(t,\ t + dt)$. We can pictorially represent the stochastic creation and degradation of the molecules as a Markov chain (Fig. 1).

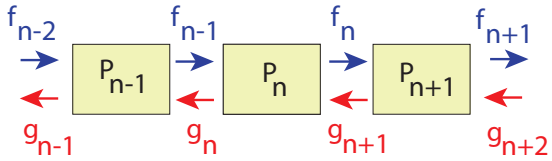

Figure 1. A **Markov chain**: Visualisation of stochastic creation and degradation of molecules in a cell. n is the number of copies of a molecule (e.g., mRNA, protein) inside the cell. $f_n dt$ to be the probability that a cell with n copies of the molecule creates another copy of the molecule within an infinitesimal time interval dt. $g_n dt$ is the probability that a cell with n copies of the molecule degrades one copy of the molecule within an infinitesimal time interval dt.

To experimentally measure the probability P_n for each n, we can try two methods. Suppose that there are 1000 genetically identical cells. In one method, we would use a camera that is attached to a microscope to take a snapshot of each of the 1000 cells at the same time, and then count the number of copies of the mRNA in each cell that is transcribed from a gene of interest.We can then plot the histogram of the number of copies of the mRNA per cell. This histogram estimates the $P_n(t)$ for every n at a given time t. This estimate would approach a "true" underlying probability distribution $P_n(t)$ as the number of cells increases from 1000 to infinity. But we do not need to know the exact probability distribution - a simple histogram would still tell us the amount of cell-to-cell variability in the number of copies of the mRNA. A second method would involve focusing on a single cell with a microscope and then recording the number of copies of the mRNA and/or the protein in that cell by following it over time. This would result in a time-lapse movie in which we record the birth and death of every mRNA and/or protein from a single gene. We can then count the copy number of mRNA and/or protein from a gene in that cell at each timeframe of the movie. After filming the cell for a sufficiently long enough time (e.g., until after it divided many times), we can pool together the measurements from all the timeframes to make one histogram of the copy number of mRNA and/or protein. Such a histogram would not tell us $P_n(t)$ but rather the time-independent, **steady-state probability distribution** P_n. In fact, the steady-state probability distribution is typically more important than knowing the $P_n(t)$ at every time t. To see this, note that the cells were already growing (and living) before we switch on the microscope. This means that the cells likely already reached a steady-state distribution before you even started the experiment. An exception occurs for genes whose expression does not begin until we tell them to - this occurs, for example, for inducible genes for which we can add an inducer (i.e., chemical that activates gene expression) to the growth medium of the cells right after switching on the microscope.

We next turn to analyzing the Master equation (Eq. 1) for the standard gene regulations - (1) constitutive, (2) auto-regulatory positive feedback, and (3) negative feedback. Throughout, we will let n be the number of copies of a protein. For a constitutive gene expression, we have $f_n = k$. For an auto-regulatory positive feedback, we can let $f_n = Vn^h/(K + n^h)$ where h is the Hill coefficient. For an auto-regulatory negative feedback, we can let $f_n = V/(K + n^h)$. In all three cases, we will consider a simple degradation scheme, in which a molecule degrades independently of the state of the other molecules and with a constant probability of degradation per unit time (i.e., $g_n = \gamma n$, where γ is a positive constant).

A. Constitutive gene expression - Master equation

We are interested in two quantities: (1) the average copy number $< n >$ of the mRNA or protein in a cell, and (2) the steady-state distribution P_n (i.e., what $P_n(t)$ becomes after a long time). We will now compute both for the case of a constitutive gene expression.

1. Average number of copies of a molecule in a cell

This calculation proceeds the same way as in computing the average number of cells with the Master equation in lecture note 1. By the definition of averages, we have

$$\frac{d < n >}{dt} = k(< n > +1) + \gamma\left[\sum_{n=0}^{\infty} n(n+1)P_{n+1}(t)\right] - k\sum_{n=0}^{\infty} nP_n(t) - \gamma\sum_{n=0}^{\infty} n^2 P_n(t) \tag{2}$$

We can simplify Eq. 2 by using a mathematical trick: $n = (n+1) - 1$ and $n = (n-1) + 1$. We can also simply Eq. 2 by using the fact that $P_{-1}(t) = 0$ and that $\sum P_n(t) = 1$. Then Eq. 2 becomes

$$\frac{d < n >}{dt} = k - \gamma < n > \tag{3}$$

Note that Eq. 3 has the same form as the equation for the constitutive gene expression scheme that we derived in the previous lecture (Eq. 9 in lecture note 2). The only difference is in the meaning. Eq. 9 from the previous lecture was for the protein concentration "p", which we inherently assumed to be deterministic. In other words, we assumed that every cell with the same protein concentration now will have the same concentration at every future instance. On the other hand, Eq. 3 describes the average concentration $< n >$ of a population of cells rather than the concentration in any single cell. In this probabilistic view, two cells with the exactly the same copy number n of a mRNA can have vastly different copy number of mRNA in the future. But the average copy number $< n >$ of the mRNA from the same gene would deterministically change over time (Eq. 3). In particular, we have

$$< n(t) >= \frac{k}{\gamma}(1 + n_0 e^{-\gamma t}) \tag{4}$$

where n_0 is a constant that depends on the initial condition (e.g., n_0 = -1 if $< n(t=0) >= 0$).

2. Steady-state solution of the Master equation

In steady-state, we have dP_n/dt=0 for all values of n. Thus the Master equation (Eq. 1) becomes

$$0 = f_{n-1}P_{n-1} + g_{n+1}P_{n+1} - g_n P_n - f_n P_n \tag{5}$$

In particular, for the constitutive gene expression scheme, we have

$$0 = kP_{n-1} + \gamma(n+1)P_{n+1} - \gamma n P_n - kP_n \tag{6}$$

Rearranging the terms in Eq. 6, we obtain

$$-\gamma(n+1)P_{n+1} + kP_n = -\gamma n P_n + kP_{n-1} \tag{7}$$

Above equation is true for all values of n ($n \geq 0$). Eq. 7 generates a chain of equalities, starting from $n = 0$ and moving up to larger values of n. At the lowest end of this chain (i.e., at $n = 0$), Eq. 7 and the fact that P_{-1}=0 yields

$$-\gamma P_1 + kP_0 = 0 \tag{8}$$

Thus, the chain of equalities (Eq. 7) tells us that

$$-\gamma n P_n + kP_{n-1} = 0 \tag{9}$$

for all values of n. Now, with Eqs. 8 and 9, we can solve for P_n. To see this, we first note that Eq. 8 gives us

$$P_1 = \frac{k}{\gamma} P_0 \tag{10}$$

Then, recursively applying Eq. 9 from $n > 0$ down to $n = 0$, we obtain

$$P_n = \left(\frac{k}{\gamma}\right)^n \frac{P_0}{n!} \tag{11}$$

which reproduces Eq. 10 when $n = 0$. Eq. 11 tells us that if we know what P_0 is, then we know what P_n is for any value of n. To determine P_0, we use the fact that summing all the probabilities yields 1. We thus have

$$1 = \sum_{n=0}^{\infty} P_n \tag{12a}$$

$$= P_0 \sum_{n=0}^{\infty} \left(k/\gamma\right)^n \frac{1}{n!} \tag{12b}$$

$$= P_0 e^{k/\gamma} \tag{12c}$$

In Eq. 12c, we used the fact that the summation in Eq. 12b is, by definition, the Taylor expansion of an exponential. From Eq. 12c, we have $1 = P_0 e^{k/\gamma}$, or in other words,

$$P_0 = e^{-k/\gamma} \tag{13}$$

Note that P_0 is a constant (i.e., does not depend on time) because we have computed steady-state probabilities. Plugging Eq. 13 into Eq. 11, we obtain

$$P_n = \left(\frac{k}{\gamma}\right)^n \frac{e^{-k/\gamma}}{n!} \tag{14}$$

This is the steady state distribution of n for a constitutive gene expression (Fig. 2). It is a ubiquitous form of probability distribution called the **Poisson distribution**. The Poisson distribution has the following general form

$$P(n) = (< n >)^n \frac{e^{-<n>}}{n!} \tag{15}$$

An important property of the Poisson distribution (Eq. 15) is that a random variable n follows the Poisson distribution, then its mean $< n >$ and variance $< (n - < n >)^2 >$ are the same. From the steady-state distribution of n (Eq. 14), we see that the mean copy number of the molecule, say the mRNA from a gene, is k/γ. Thus, we have

$$< n >=< (n - < n >)^2 >= \frac{k}{\gamma} \tag{16}$$

Note that Eq. 16 is precisely the steady-state concentration that we would obtain from the deterministic equation for $< n >$ (Eq. 4). Note that the steady-state concentration $< n >$ in Eq. 4 corresponds to the value that $< n >$ approaches after a long time (i.e., $t \to \infty$). Furthermore, the **fractional error**, which is the standard deviation in n (i.e., $\sqrt{< (n - < n >)^2 >}$) divided by the average copy number $< n >$ is

$$\frac{\sqrt{< (n - < n >)^2 >}}{< n >} = \frac{1}{\sqrt{< n >}} \tag{17}$$

Thus, as the copy number of the mRNA (or protein) in a cell increases, the fractional error decreases. The fractional error (Eq. 17) quantifies how important the stochastic dynamics is for the gene of interest. If the cell has a low number of copies of the mRNA (or protein), then the stochastic effect is non-negligible because the average deviation that a cell has from the population-level average is large. This is why one often says that stochastic gene expression is due to a low copy number of the mRNA and/or the protein. But as we discussed at the beginning of this lecture, the deeper answer is that we never have all the necessary information to exactly predict the cell's gene-expression dynamics at every point in time. Although we derived Eq. 17 only fora constitutive gene expression, we would obtain similar outcomes for other types of gene-regulatory schemes.

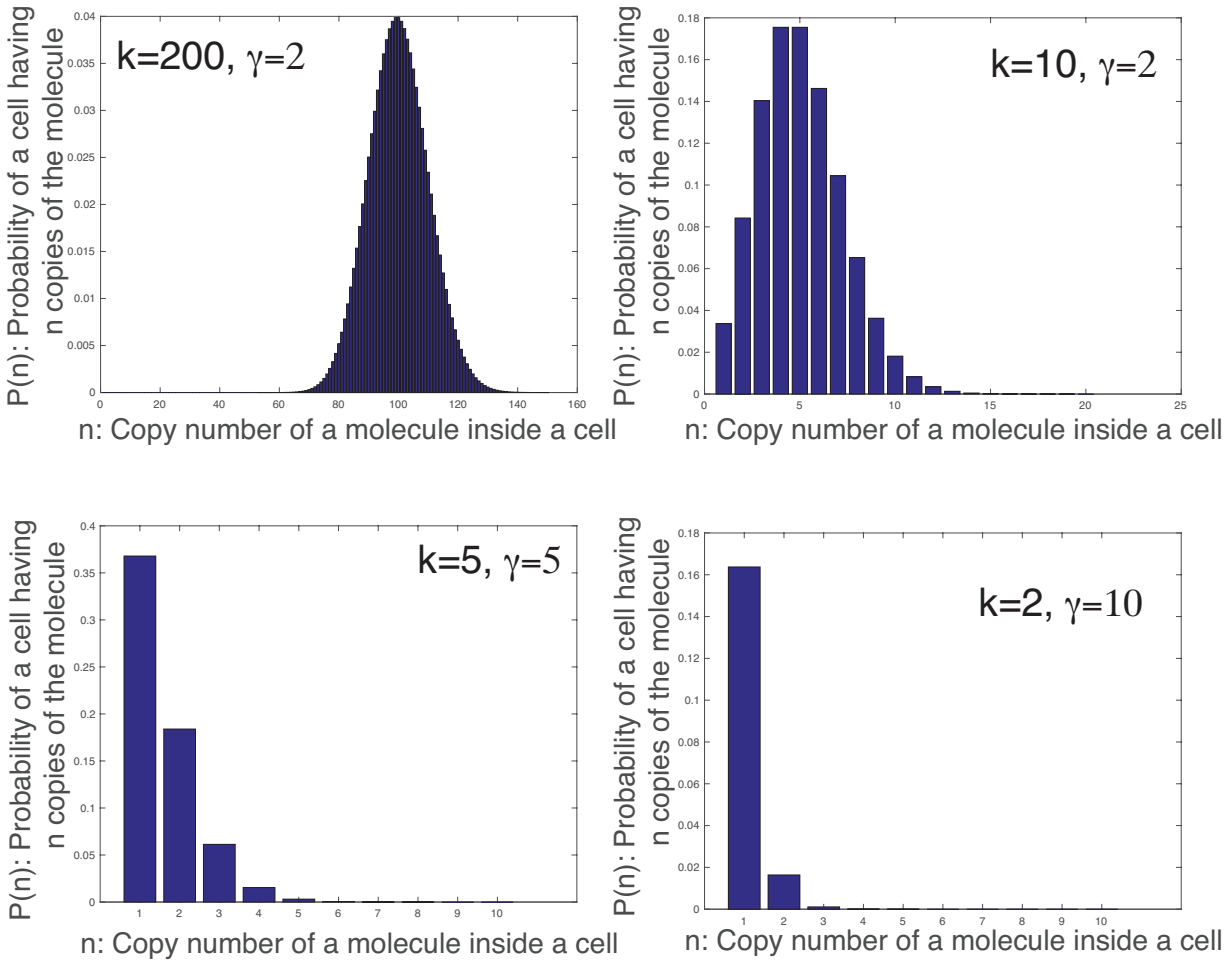

Figure 2. **Steady-state probability distributions for a constitutive gene expression.** Histograms of $P(n)$ given by Eq. 14 for different values of the production rate k and degradation constant γ.

3. Simulating the Master equation

We can simulate the stochastic gene-expression dynamics by using a type of Monte Carlo-simulation algorithm called the **Gillespie algorithm** (Fig. 3). Problem set 1 derives the Gillespie algorithm. The basic idea behind the Gillespie algorithm is that we can simulate the transitions between the different "boxes" in the Markov chain (Fig. 1) by using the f_n to calculate the time that we need to wait for the creation of another copy of the mRNA (or protein) and using the g_n to calculate the time that we need to wait for the degradation of one copy of the mRNA or the protein. The Gillespie algorithm works for almost any form of gene-regulatory circuits.

III. FOKKER-PLANCK EQUATION

Our second approach for modeling stochastic gene expression uses the **Fokker-Planck equation**. The main idea behind the Fokker-Planck equation is that we solve the Master equation by turning its discrete values of n into continuous values of n. To see how this modification can help us, note that the Master equation is difficult to solve, even for just the set of steady-state probabilities $\{P_n\}$, because there are infinitely many P_n's, with each P_n depending on the other P_n's. As we will see, the Fokker-Planck equation results from turning the infinite number of equations for the infinite set $\{P_n\}$ into a single equation - an equation for $P(n)$, where n is now a continuous variable.

To turn the discrete values of n into continuous values, we will use the following technique. Let Δn be some "typical" variation in the copy number n. For concreteness, we can let Δn be the standard deviation in n. Now suppose that

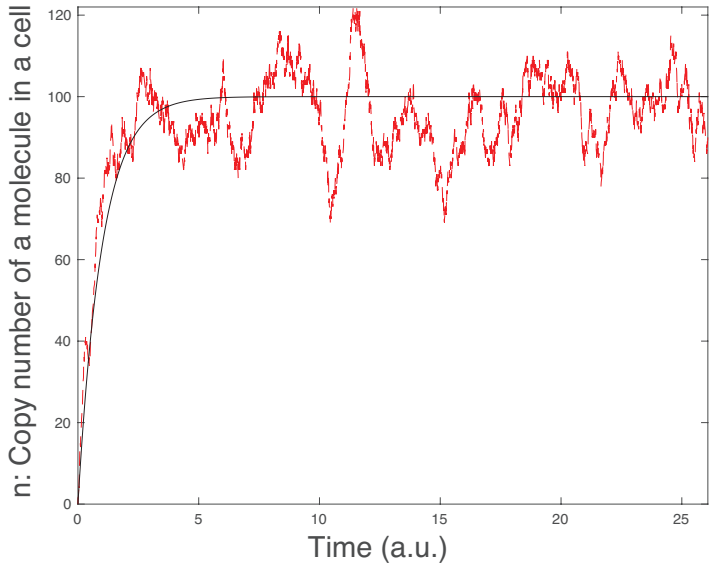

Figure 3. **Using the Gillespie algorithm to simulate the stochastic, constitutive gene expression that is dictated by the Master equation (Eq. 1).** $k = 100$ and $\gamma = 1$, where $f_n = k$ and $g_n = \gamma n$ in Eq. 1. Red curve shows the stochastically changing copy number n of the mRNA in a cell as a function of time (computed by the Gillespie algorithm - see problem set 1). Black curve is the deterministically changing value of n (computed using Eq. 3).

$\Delta n/ < n >$ is sufficiently small. For a constitutive gene expression, by taking Δn to be the standard deviation in n, $\Delta n/ < n >$ is sufficiently small means that the fractional error, $1/\sqrt{< n >}$, is sufficiently small according to Eq. 17. Consider some arbitrary, smooth function $L(n)$. Then the Taylor expansion of $L(n)$ around any value n_0 is

$$L(n_0 + \Delta n) = L(n_0) + \frac{\partial L}{\partial n}\Delta n + \frac{1}{2}\frac{\partial^2 L}{\partial n^2}(\Delta n)^2 + ... \tag{18}$$

where the derivatives are evaluated at n_0 and Δn is a typical deviation in n. Note that here we are assuming that n is a continuous variable. If Δn is sufficiently small, we can ignore all the terms in Eq. 18 that are beyond the second order term. Now, the main idea that we use to derive the Fokker-Plank equation is that if $\Delta n/ < n >$ is small, then we can just boldly Taylor expand all the functions of n in the Master equation (Eq. 1) up to and including second order in n as done in Eq. 18. Specifically, using the Taylor expansion, we obtain

$$f(n-1)P(n-1) \approx f(n)P(n) - \frac{\partial}{\partial n}(f(n)P(n)) + \frac{1}{2}\frac{\partial^2}{\partial n^2}(f(n)P(n)) \tag{19a}$$

$$g(n+1)P(n+1) \approx g(n)P(n) + \frac{\partial}{\partial n}(g(n)P(n)) + \frac{1}{2}\frac{\partial^2}{\partial n^2}(g(n)P(n)) \tag{19b}$$

where we have set Δn=1 (i.e., a typical variation in n is by one molecule). Substituting Eqs. 19a and 19b into the Master equation (Eq. 1) and then ignoring all terms that involve third and higher-order derivatives, we obtain

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\left\{(f(n) - g(n))P(n) - \frac{1}{2}\frac{\partial}{\partial n}(f(n) + g(n))P(n)\right\} \tag{20}$$

This is the **Fokker-Planck equation**. It is a continuous version of the Master equation in which we treat the n as a continuous variable. The term inside the curly brackets in Eq. 20 is called the **net probability flux**, which we denote by $J(n,t)$. We can thus rewrite Eq. 20 more compactly as

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial J(n,t)}{\partial n} \tag{21}$$

Note that Eq. 21 is just the **transport equation** in one dimension. To understand the meaning of the net probability flux $J(n,t)$, consider the Markov chain picture (Fig. 1) in which we consider a chain of boxes, each representing the copy number of the mRNA or protein that a cell can have. We said earlier that we can view P_n as the fraction of genetically identical cells that have n copies of the mRNA or the protein. To visualize this pictorially, we can envision "putting" cells into a box labeled n if those cells have n copies of the molecule at a given time t. Then at a later time $t+dt$, we will have to transfer some cells to the box labeled $n-1$ (to the left of the box labeled n in Fig. 1) if those cells have degraded a molecule during time interval $(t, t+dt)$ and transfer some cells to the box labeled $n+1$ (to the right of the box labeled n in Fig. 1) if those cells have created a molecule during time interval $(t, t+dt)$. The transfer to the box labeled $n-1$ represents a probability flux to the left (let's denote this as J_-) while the transfer to the box labeled $n+1$ represents a probability flux to the right (let's denote this as J_+). Then the net probability flux $J(n,t)$ is $J_+ - J_-$. According to the Fokker-Planck equation, n is now continuous. This means that we have a continuous line of n-values, at that there is a net current $J(n,t)$ at each point on this line (each point representing a particular value of n). In light of this interpretation, the Fokker-Planck equation (Eq. 21) is just a statement that if we sum up the probabilities $P(n)$ over all values of n, then we should always get one.

A. Steady-state solution of the Fokker-Planck equation

Solving the Fokker-Planck equation (Eq. 21) by hand is difficult. But as was the case with the Master equation, finding the time-independent, steady-state probability $P(n)$ of the Fokker-Planck equation is easier and one that we can do by hand. The major advantage of the Fokker-Planck equation is that we can, in fact, find the steady-state probability $P(n)$ for *any* gene regulatory circuit where as we can find the steady-state solutions of the Master equation only for simple enough gene circuits by hand. This is one of the main advantages of the Fokker-Planck equation over the exact, Master equation. In steady state, the Fokker-Planck equation (Eq. 21) becomes

$$\frac{\partial P(n,t)}{\partial t} = 0 \tag{22}$$

This means that the net probability current $J(n,t)$ is a constant. In fact, it must be zero at all times and for all values of n. The reason is that when $n=0$, we must have $J(0,t)=0$ since no probability current can flow past $n=0$ either from $n>0$ into $n<0$ or from $n<0$ to $n>0$. There are no cells with negative copy numbers of the mRNA or protein. Since $J(n,t)$ is a constant (the same value at all n) at steady state, we must indeed have $J(n,t)=0$ in steady-state. Thus, from the definition of $J(n,t)$ (the curly bracket in Eq. 20), we have

$$(f(n)-g(n))P(n) = \frac{1}{2}\frac{\partial}{\partial n}(f(n)+g(n))P(n) \tag{23}$$

To simplify above equation, we define $H(n) = (f(n)+g(n))P(n)$. Then Eq. 23 becomes

$$\frac{f(n)-g(n)}{f(n)+g(n)}H(n) = \frac{1}{2}\frac{dH}{dn} \tag{24}$$

We can solve for $H(n)$ in Eq. 24 by separating variables as follows:

$$2\frac{f(n)-g(n)}{f(n)+g(n)} = \frac{1}{H(n)}\frac{dH}{dn} \tag{25}$$

Integrating Eq. 25 over n and realizing that the right side of Eq. 25 is just $dlog(H)/dn$, we obtain

$$H(n) = Aexp\left(\int_0^n 2\frac{f(x)-g(x)}{f(x)+g(x)}dx\right) \tag{26}$$

where A is a constant to be determined. From the definition of $H(n)$, we can now solve for $P(n)$:

$$P(n) = \frac{A}{f(n)+g(n)}exp\left(\int_0^n 2\frac{f(x)-g(x)}{f(x)+g(x)}dx\right) \tag{27}$$

Eq. 27 is the **steady-state solution of the Fokker-Planck equation**. It holds for any arbitrary gene circuits since we did not assume any functional forms for f and g. From Eq. 27, we see that A isa normalization factor that ensures that summing the probabilities yields one (i.e., $\int P(n)dn = 1$). The integral inside the exponential looks like a **potential energy function**. Motivated by this, let us define

$$U(n) = -\int_0^n 2\frac{f(x) - g(x)}{f(x) + g(x)}dx \tag{28}$$

Then we can rewrite the steady-state solution of the Fokker-Planck equation as

$$P(n) = \frac{A}{f(n) + g(n)}exp(-U(n)) \tag{29}$$

The main idea here is that $exp(-U(n))$ is like the Boltzmann factor from statistical mechanics. $U(n)$ is like the energy measured in units of kT, where k is the Boltzmann constant and T is the temperature. In this case, there is no real "temperature" or the Boltzmann factor. So this is just an analogy. Nonetheless, this analogy gives us an intuition for what $P(n)$ should look like for certain genetic circuits. For instance, in problem set 1, you are asked to look at a bistable, auto-regulatory positive feedback circuit. For this circuit, you will find that $U(n)$ represents a double-well potential, and that $P(n)$ would be bimodal. This is what we would expect for a double-well potential in physics as well.

IV. APPLICATION OF NOISY GENE EXPRESSION: EXCITABLE GENE-REGULATORY CIRCUIT CONTROLLING CELL FATES

This section refers to G. Suel et al., *Nature* (2006).

In this section, we consider another way of modelling stochastic gene expression: adding a noise term by hand (without deriving it from anywhere) to a deterministic equation that models the gene circuit. This phenomenological approach has the advantage that the math becomes easier than solving the Master or the Fokker-Planck equation while accounting for noise in gene expression (one typically approximates the noise by is a **white noise term** in this approach). We will consider this approach in the context of **bacterial competence**.

When the soil bacterium, *Bacillus subtilis*, runs out of nutrients, it can enter into a **competent** state in which it uptakes any foreign DNA that it finds outside. The rationale here is that perhaps the foreign DNA contains genes that can help the bacterium survive in the nutrient-limited environment. One reason why this phenomenon is interesting is that only a fraction of the *B. subtilis* cells in a population enter into the competent state, despite all the cells in the population being genetically identical and living in the same nutrient-limited environment. So what is special about this fraction of cells (approximately 3.6% according to Suel et al.)? Suel et al. investigated this phenomenon by labeling the key proteins involved in competence with fluorescent proteins. Using these genetically engineered bacteria, Suel et al. used a fluorescence microscope to make time-lapse movies of individual cells entering or exiting the competent state (Fig. 4A). From these movies and a mathematical model of the gene circuit that controls the cells' entrance into the competent state (Fig. 4B-C), Suel et al. deduced that noise in the expression of two key genes can explain why a small fraction of the genetically identical cells enter into the competent state. In this section, we summarize their mathematical model (for their beautiful experiments, please read G. Suel et al., *Nature* (2006)).

When starved of nutrients, the *B. subtilis* cells decide whether to become competent or form a spore. The spore would "wake up" when there are enough nutrients again outside. Interestingly, only about 3.6% of the vegetative cells (i.e., happily dividing cells) enter into the competent state. The gene-regulatory circuit that controls this decision involves three key genes: *comK*, *comS*, and *comG*. ComK and ComS are transcription factors that together determine the expression level of *comG*. When the expression level of comG is high, then the cell more likely enters into a competent state. Thus by having a gene encodes a fluorescent protein be controlled by the promoter of comG ("P_{comG}" in Fig. 4A-B), and then measuring the fluorescence of individual cells, Suel et al. could determine, in individual cells, how much ComG each cell was producing (i.e., the more fluorescent the cell, the more ComG it was making, and thus more likely to enter a competent state).

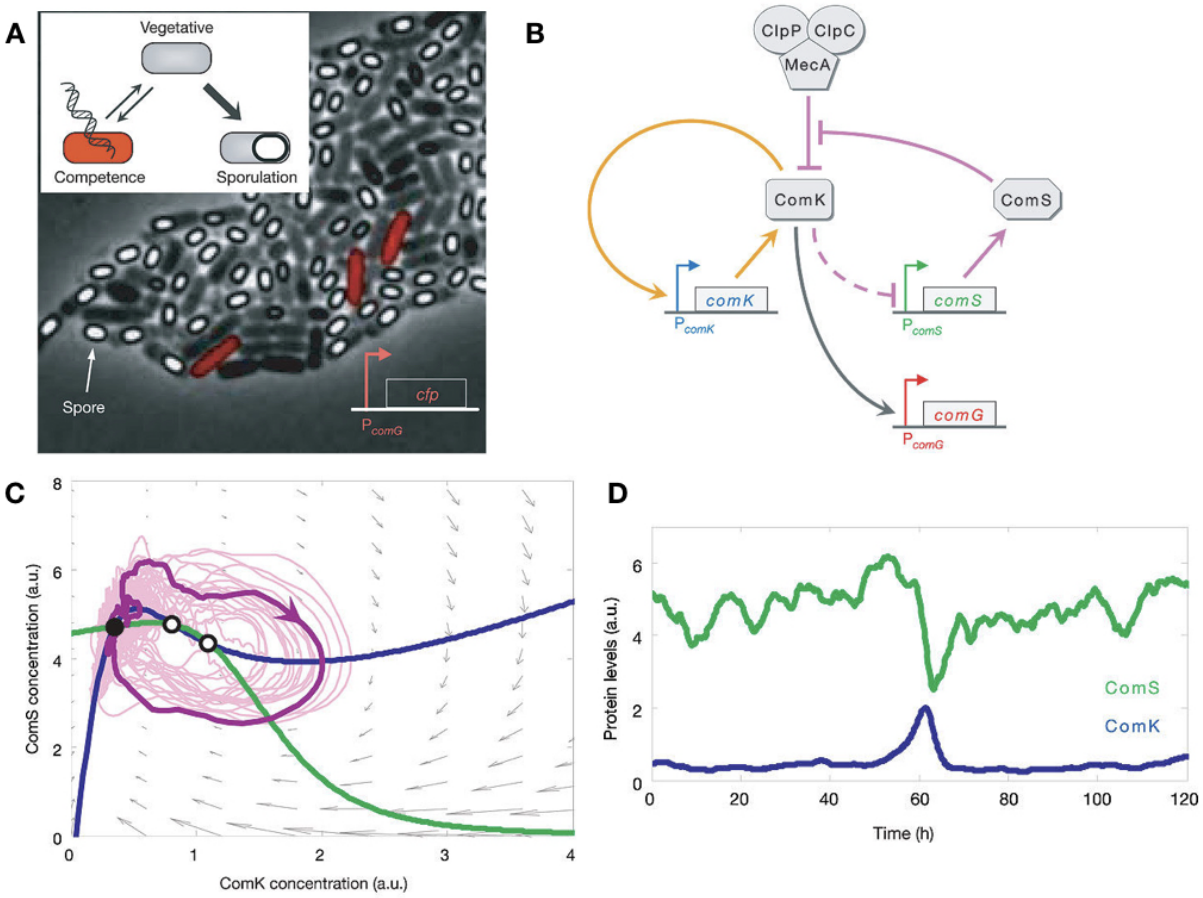

Figure 4. **Noise induces *B. subtilis* to enter and exit the competent state.** Figures taken from G. Suel et al., *Nature* (2006). (**A**) When starved of nutrients, each vegetative *B. subtilis* cell must decide whether to enter the competent state (marked red) or sporulate (white ovals). Only a small fraction (about 3.6 %) of the genetically identical cells enter the competent state. (**B**) The gene-regulatory circuit that governs the entrance to and exit from the competent state. ComK and ComS are the main transcription factors of interest in this study. They regulate the expression of *comG*. When there are many ComG proteins inside a cell, the cell more likely enters into a competent state. (**C**) Analysis of the genetic circuit in (B) using nullclines and vector fields (this plot is called a **phase portrait** in the language of dynamical systems). Blue is the nullcline for dK/dt (K is the concentration of ComK) and green is the nullcline for dS/dt (S is the concentration of ComS). Grey arrows show the vector field, which one obtains from analyzing the nullclines. Black circle is the stable fixed point (represents vegetative growth). The two white circles represent unstable fixed points. The many pink trajectories show "excursion trajectories" of individual cells that were initially at the stable fixed point and were pushed out by noise. The single purple trajectory shows a representative trajectory of a single cell. (**D**) Stochastic simulation of the gene-circuit in (B). Green and blue curves show the amount of ComS and ComK respectively. During competence, the ComK level is high while ComS level is low - the two amounts are negatively correlated.

$$\frac{dK}{dt} = a_k + \frac{b_k K^n}{k_0^n + K^n} - \frac{K}{1+K+S} \tag{30a}$$

$$\frac{dS}{dt} = \frac{b_s}{1+(K/k_1)^p} - \frac{S}{1+K+S} + \xi(t) \tag{30b}$$

where K and S are the concentrations of CoK and ComS protein, respectively. a_k is the basal (i.e., minimal) rate of production of ComK and b_k is the fully activated rate of production of ComK. k_0 is the half-saturation concentration - the concentration of ComK required to produce half the fully activated rate. n and p are Hill coefficients. b_s is the maximal production rate of ComS protein. $\xi(t)$ is a random noise (white noise) term that represents the random fluctuations in the amount of ComS. We can analyze Eqs. 30a and 30b by finding their nullclines (see lecture note 2), as Suel et al. have done (Fig. 4C). By plotting the vector field around the nullclines (as in lecture note 2), we find

that the system has one stable fixed point and two unstable fixed points. If a cell is sitting at the stable fixed point and there is a sufficient amount of noise ξ, then the cell would be pushed out and follow the vector field (purple and pink trajectories in Fig. 4C). Suel et al. found that the vector fields allow the cells to go in an "orbit" around the two unstable fixed points. The orbit starts from the stable fixed point and then come back to the stable fixed point. Crucially, these orbits allow the cells to take "excursions" into regions of the ComS-ComK phase space where the ComK level is high and ComS level is low (pink and purple trajectories that go below the green and blue nullclines in Fig. 4C). Suel et al. found that these regions correspond to the ComG level being high (makes sense from the circuit diagram in Fig. 4B). As mentioned earlier, higher amounts of ComG promote the cells to become competent. The fact that noise drives the cells into the phase space that correspond to being competent explains why only a small fraction of genetically identical cells in the population become competent.