

6.1. Why are cellular processes stochastic?: The “easy” answer

In the previous lecture, we discussed equations that describe basic genetic circuits. All those equations were **deterministic equations**. They were ordinary differential equations whose solutions are deterministic. This means that if you know the concentration of relevant molecules (mRNA or proteins) at a particular time point, then the equation tells you what the concentrations of those molecules were in the past and what they will be in the future. Nothing is left undetermined. 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 molecules involved in typical cellular processes, such as the number of mRNA or protein produced in a cell. As an example, a cell may contain between 1 to 100 mRNA molecules transcribed from a particular gene. Dividing 1 to 100 mRNA molecules by the volume of a cell, say V , gives you: $1/V, 2/V, \dots, 100/V$. 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. But intuition tells you 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 \bar{n}/V , where \bar{n} is the average number of molecules. In other words, when there are large numbers of molecules on average inside the cell, we can use the differential equations that we derived in the previous lecture. This makes an intuitive sense because if you have 1000,000 molecules, then losing or gaining 1 or 2 molecules doesn’t make much difference.

6.2. Why are cellular processes stochastic?: A “deeper” answer

Although a few copies of molecules are involved in a cellular process, it doesn’t logically follow that the process should be non-deterministic (i.e., stochastic). For instance, if we have five billiard balls elastically bouncing off each other and off the sides of a pool table, their motion may seem completely random over time. This collection of balls may even be a chaotic dynamical system. Nonetheless, the laws are completely deterministic. Our inability to calculate the precise trajectory is due to our limitations in math or technical inability to precisely measure the exact position and velocity of each ball (this precision matters for dynamics occurring over a long time scale). But at the end of the day, the Newton’s deterministic laws of motion, completely dictates how each ball moves in the future and how it moved in the past. This is not a stochastic system despite consisting of only five balls.

So just saying that a cellular process like gene expression involves a few copies of mRNA or proteins inside a cell is insufficient to explain why cellular processes are stochastic. The deep and subtle reason, ironically, is that *many molecules* are actually involved in

cellular processes. These molecules include water molecules that bombard the few mRNA molecules inside a cell. Bombardment by these molecules cause each mRNA molecule taking a random walk (collectively, a diffusion) inside a cell. This is a stochastic process because we do not know the position and velocity of every water molecule inside the cell. Moreover, we do not know the position and velocity of every other types of molecules inside the cell that are diffusing around, bombarding with the few mRNA and proteins that are produced from our genetic circuit of interest, and thus causing the mRNA and proteins to take random walks. In short, the randomness is apparent – it is only due to our ignorance of the state of all the particles inside and outside the cells. That is why we cannot describe exactly where each transcription factor will be inside a cell. Thus we cannot describe exactly when each transcription factor or ribosome or RNA polymerase will bind at a particular location on a DNA or a RNA. If we view molecules diffusing inside a cell, simply from a classical physics perspective, then the randomness is really due to our ignorance of what every single molecule is doing. This is precisely what happens in systems of interest to statistical physics, such as a box of ideal gas particles.

6.3. Mathematical descriptions of stochastic gene expression

Now that we know why cellular processes are stochastic, we now develop two approaches for describing stochastic processes: (1) Masters equation approach and (2) Fokker-Planck approach. We will develop and apply both approaches in the context of gene expression. But you can apply them to other cellular processes, for example a random walk of a cell.

6.3 – A. Masters equation approach

The **masters equation** describes the full probabilistic nature of production and degradation of mRNA and proteins. Remember that we do not know when an RNA polymerase that is diffusing inside a cell will bind to a particular promoter on DNA and when a ribosome that is diffusing inside a cell will bind to a particular mRNA molecule. Thus at best we can describe the *probability* that any one of the ribosomes will bind to an mRNA and the probability that any one of the RNA polymerases will bind to our promoter of interest. Let $P_n(t)$ be the probability of a cell having n copies of a certain molecule inside a cell at time t . To be concrete, let's say that the molecule is an mRNA from a particularly gene. But for everything that follows below, the molecule doesn't have to be mRNA. To experimentally measure this probability, we could try two things. Suppose we have 1000 genetically identical cells. If we take a microscope camera and take a snapshot of the 1000 cells at a given time, count the number of mRNA of a given gene in each cell, then we can plot the histogram of the number of mRNA molecules per cell. This histogram estimates the $P_n(t)$ for every n (this estimate becomes exact as the number of cells that we measure approaches infinity – which is

experimentally impossible). Another method is, instead of taking a snapshot in time, we can just focus on a single cell and make a time-lapse movie of that cell. We will then count the number of mRNA molecule at each time frame in that cell. Then after a sufficiently long enough time, we can plot the histogram of mRNA molecules from a given gene. If the genetic circuit that produces this mRNA is **ergodic**, then the histograms obtained by both methods would be equal. In most cases, we assume that indeed both histograms would be equal and that typical genetic circuits are thus ergodic. Proving that a circuit is so in a mathematically rigorous way is difficult and whether or not a circuit is indeed ergodic would depend on the time scale that you're interested in. If you wait long enough, mutations would occur in the genetic circuit, and thus the circuit would morph into a different circuit. Evidently, the circuit isn't ergodic over this long time scale. But for the purpose of most cell biological studies, we are not interested in how a circuit behaves in evolutionary time scales. Thus we assume ergodicity. Then consider the following "flow chart" of events:

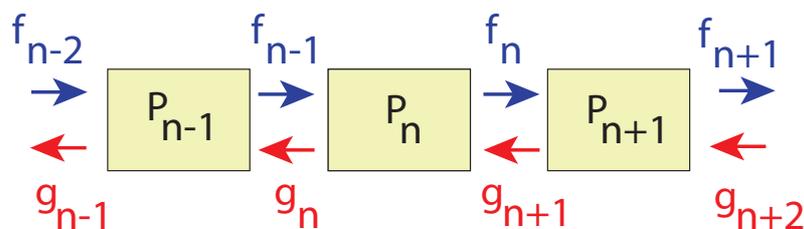


Fig. 6.1. Flow chart of probabilistic creation and destruction of molecules (mRNA, protein, or other biomolecules) in a cell. $f_i dt$ is the probability of a cell having i molecules creating one molecule within time interval dt . $g_i dt$ is the probability of a cell having i molecules destroying one molecule within time interval dt . P_n is the probability of a cell having n molecules (mRNA, protein, or other biomolecules). This "chain" is called a "**Markov chain**". It is just a fancy term to describe a system with probabilistic transitions between different states (i.e., state "n" = cell having n copies of a molecule), that depends only on the current state of the system (i.e., these transition rates depend only on what the value of n is for the cell now, and not what the value of n was before).

Here, f_n is the probability that a cell with n mRNA molecules makes one more mRNA molecule from the gene in time interval dt . g_n is the probability that a cell with n mRNA molecules degrades one mRNA molecule from the gene in time interval dt . By keeping the time interval to be infinitesimally small dt , we ensure that just one but not more than one is created or degraded in this time. Then accounting for flows into and out of the box representing P_n , we have

$$\frac{dP_n}{dt} = f_{n-1}P_{n-1} + g_{n+1}P_{n+1} - g_nP_n - f_nP_n \quad (6.1)$$

For constitutive gene expression, $f_n = k$ and $g_n = \gamma n$. For a positive autoregulation (i.e., one gene that positively feedbacks on itself), we have $f_n = \frac{vn^a}{K+n^a}$ (a sigmoidal function). The

Master equation (6.1) is for every value of n . There are infinitely many equations (one for each P_n) that are coupled to each other (because P_n depends on P_{n-1} and P_{n+1}). We do not know how to solve Masters equation analytically. We will see below computational methods that can simulate how P_n changes over time. Analytically, we can solve for the steady-state for certain types of gene regulation schemes. One such scheme is the constitutive gene expression. Let's see below how to solve for this scheme.

Constitutive gene expression: Changes in the average number of mRNA molecules over time

Before solving for the steady-state, let's first ask how the average number $\langle n \rangle$ changes over time. By definition of averages, we have

$$\begin{aligned} \frac{d\langle n \rangle}{dt} &= \frac{d}{dt} \sum_{n=0}^{\infty} n P_n \\ &= \sum_{n=0}^{\infty} n \frac{dP_n}{dt} \end{aligned} \quad (6.2)$$

Substituting the master equation (6.1) into the $\frac{dP_n}{dt}$ in equation (6.2) and then doing some algebra, we obtain

$$\frac{d\langle n \rangle}{dt} = k\{\langle n \rangle + 1\} + \gamma \left\{ \sum_{n=0}^{\infty} n(n+1)P_{n+1} \right\} - k \sum_{n=0}^{\infty} n P_n - \gamma \sum_{n=0}^{\infty} n^2 P_n \quad (6.3)$$

We can simplify above equation further by using a mathematical trick: $n = (n+1)-1$. We also use the fact that $P_{-1} = 0$ and that $\sum P_n = 1$. Then above equation simplifies to

$$\frac{d\langle n \rangle}{dt} = k - \gamma \langle n \rangle \quad (6.4)$$

Note that this equation has the same form as the equation for the constitutive gene expression scheme that we derived in the last lecture. The only difference is in the meaning. In the previous lecture, the equation was for the concentration "x", which we inherently assumed to be deterministic. In other words, we assumed that every cell with the same concentration at a given time will keep on having the same concentration at every instance in the future. On the other hand, equation (6.4) is for the average concentration $\langle n \rangle$. In this probabilistic view, two cells with the exactly the same number n of mRNA can evolve differently in the future. The two cells can have very different values of n at future instants of time. But the "average" number $\langle n \rangle$ of mRNA changes over time deterministically.

Constitutive gene expression: Steady-state solution of the Masters equation

In steady-state, we have $\frac{dP_n}{dt} = 0$ for all n . Thus equation (6.1) becomes

$$0 = f_{n-1}P_{n-1} + g_{n+1}P_{n+1} - g_nP_n - f_nP_n \quad (6.5)$$

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

$$0 = kP_{n-1} + \gamma(n+1)P_{n+1} - \gamma nP_n - kP_n \quad (6.6)$$

Rearranging the terms in the above equation, we obtain

$$-\gamma(n+1)P_{n+1} + kP_n = -\gamma nP_n + kP_{n-1} \quad (6.7)$$

Above equation is true for all values of n ($n \geq 0$). Note that above equation creates a “chain” of equalities, by starting from $n=0$ and moving up to different values of n . At the lowest end of this chain (i.e., at $n=0$) above equation and the fact that $P_{-1}=0$ yields

$$-\gamma P_1 + kP_0 = 0 \quad (6.8)$$

Thus the chain of equalities (6.7) tells us that

$$-\gamma nP_n + kP_{n-1} = 0 \quad (6.9)$$

for all values of n . Now with equations (6.8) and (6.9), we can solve for P_n . First note that equation (6.8) gives us

$$P_1 = \frac{k}{\gamma} P_0 \quad (6.10)$$

Then recursively applying equation (6.9) from n down to $n=0$, we get

$$P_n = \left(\frac{k}{\gamma}\right)^n \frac{P_0}{n!} \quad (6.11)$$

which is equivalent to equation (6.10) when $n=0$. Once we solve for P_0 , we know what P_n is for all values of n . To get P_0 , we use the fact that all probabilities summed together gives us 1. Thus

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

$$\begin{aligned}
 &= P_0 \sum_{n=0}^{\infty} \left(\frac{k}{\gamma}\right)^n \frac{1}{n!} \\
 &= P_0 e^{k/\gamma} \qquad (6.12)
 \end{aligned}$$

In the last step, we used the fact that the summation is, by definition, the Taylor expansion of the exponential function. From equation (6.12), we have

$$P_0 = e^{-k/\gamma} \qquad (6.13)$$

Note that it's a constant because the probabilities are in steady-state. Plugging equation (6.13) into equation (6.11), we get

$$P_n = \left(\frac{k}{\gamma}\right)^n \frac{e^{-k/\gamma}}{n!} \qquad (6.14)$$

This is the steady state distribution of n for a constitutive gene expression. It is an ubiquitous form of probability distribution called the **Poisson distribution**. Poisson distribution has a general form

$$P(n) = x^n \frac{e^{-x}}{n!} \qquad (\text{Poisson distribution: 6.15})$$

If a random variable n follows the Poisson distribution (equation 6.15), then its mean $\langle n \rangle$ and its variance $\langle (n - \langle n \rangle)^2 \rangle$ are the same. Namely, they are both equal to x . Plotting the steady state distribution of n (equation 6.14), it follows that the average number of mRNA and the variance are

$$\langle n \rangle = \langle (n - \langle n \rangle)^2 \rangle = \frac{k}{\gamma} \qquad (6.16)$$

Note that this is exactly the average concentration that we expect from the deterministic equation and equation (6.4). Note that the **fractional error**, namely the standard deviation in the number of mRNA (i.e., $\sqrt{\langle (n - \langle n \rangle)^2 \rangle}$) divided by the average number of mRNA in a cell (i.e., $\langle n \rangle$) is

$$\frac{\sqrt{\langle (n - \langle n \rangle)^2 \rangle}}{\langle n \rangle} = \frac{1}{\sqrt{\langle n \rangle}} \qquad (6.17)$$

Thus as the number of mRNA in a cell increases, the fractional error decreases. Equation (6.17) quantifies how important stochastic effects are in a cell. If the cell has a low number of

a given molecule (e.g., mRNA, protein, metabolites), then the stochastic effect is non-negligible because the average deviation that a cell has from the population-level average behavior is large. This is the reason for the “easy” answer given in section 6.1. for why gene expression is stochastic. Although we derived the equation (6.17) only for constitutive gene expression, we would obtain similar outcomes for other types of genetic circuits.

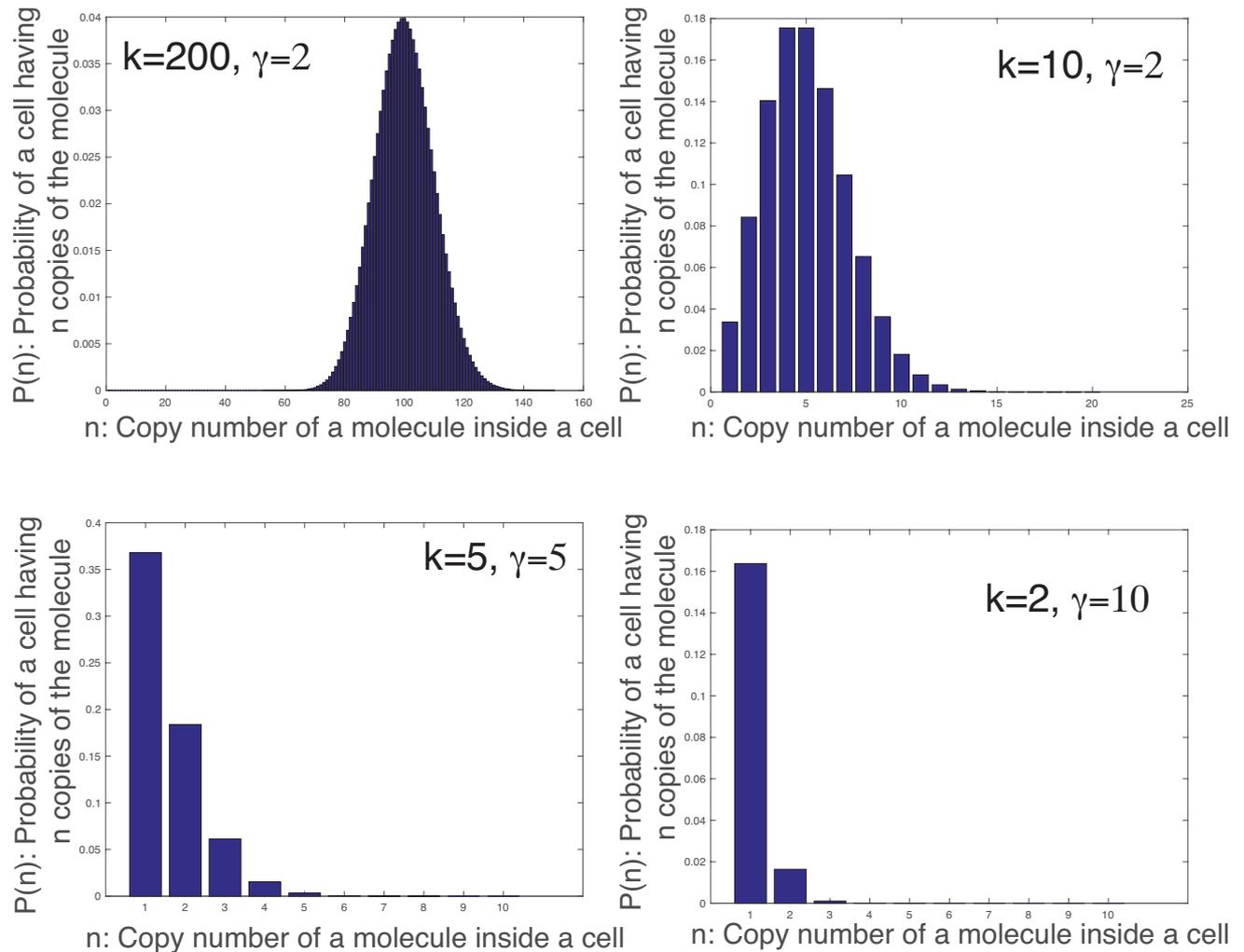


Fig. 6.2. Steady-state distribution of mRNA level is the Poisson distribution (equation 6.14). See the MATLAB code “PoissonMastersEquation.m” (on course website).

We can simulate this by using a type of Monte Carlo simulation called the Gillespie algorithm. Problem set #1 describes how you can simulate the Gillespie algorithm. The basic idea behind the Gillespie algorithm is we can simulate the transitions between the different “boxes” in the Markov chain (Figure 6.1) by using the f_n to calculate the time you need to wait for a molecule to be created and using the g_n to calculate the time you need to wait for a molecule to be degraded in a cell. The Gillespie algorithm works for complex gene regulation schemes so we do not need to restrict ourselves to only constitutive gene expression.

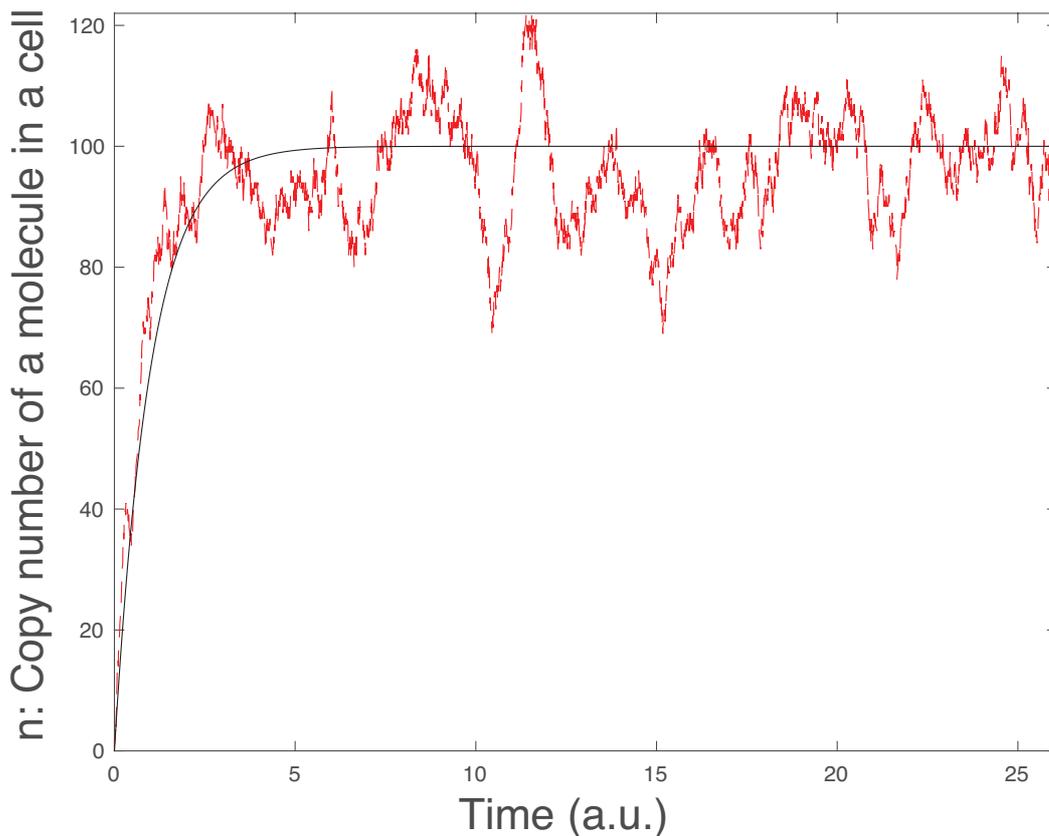


Fig. 6.3. Gillespie simulation for a constitutive gene expression. $k=100$ and $\gamma=1$. Red curve shows the stochastically changing copy number n of the molecule in a cell as a function of time (computed using the Gillespie algorithm – see Problem set #1). Black curve is the deterministically changing value of n over time (computed from the deterministic equation for the constitutive gene expression).

6.3 – B. Fokker-Planck equation

A second approach that is often used for modelling stochastic gene expression and cellular processes is the **Fokker-Planck approach**. The main idea of the Fokker-Planck approach is to facilitate the solving of the Master equation by turning the discrete n 's into continuous n 's. The idea here is that the Master equation is difficult to solve, even for steady states, for many genetic circuits because there are infinitely many P_n 's, with each function P_n dependent on the other P_n 's. Fokker-Planck approach turns these infinitely many P_n 's into a single function, which we call $P(n)$, by smoothing out the discrete values of n . To do this, suppose that $\frac{\Delta n}{\langle n \rangle}$ is typically small, where Δn is a “typical” deviation (not necessarily the standard deviation) in the number of molecules and $\langle n \rangle$ is the average number of molecules in a cell. For constitutive gene expression, if we take Δn to be the standard deviation, then the previous statement is equivalent to saying that $\frac{1}{\sqrt{\langle n \rangle}}$ is small, according to equation (6.17). Consider some arbitrary function $L(n)$. Then the Taylor expansion of $L(n)$ around some value of n_0 is

$$L(n_0 + \Delta n) = L(n_0) + \frac{\partial L}{\partial n} \Delta n + \frac{1}{2} \frac{\partial^2 L}{\partial n^2} (\Delta n)^2 + \dots \quad (6.18)$$

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 very small, we can ignore all the terms in the Taylor series that are higher than the 2nd order term. In some classes, you probably have kept only up to and including the 1st order term. Up to which order terms you keep is somewhat arbitrary. In this case, for our purpose, keeping all terms up to and including the 2nd order term will serve us well. Now, the main idea that we use to derive the Fokker-Plank equation is that if $\frac{\Delta n}{\langle n \rangle}$ is typically small, then we can just boldly Taylor expand all the functions of n in the Master equation (equation 6.1) using Taylor series shown in equation (6.18). Using the Taylor expansion, we have

$$f(n-1)P(n-1) = f(n)P(n) - \frac{\partial}{\partial n} (f(n)P(n)) + \frac{1}{2} \frac{\partial^2}{\partial n^2} (f(n)P(n)) + \dots$$

$$g(n+1)P(n+1) = g(n)P(n) + \frac{\partial}{\partial n} (g(n)P(n)) + \frac{1}{2} \frac{\partial^2}{\partial n^2} (g(n)P(n)) + \dots$$

where $\Delta n=1$. Substituting above expansions into the Master equation (6.1) and ignoring the terms that involve derivatives of order 3 or more, we obtain

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n} \left\{ (f(n) - g(n))P(n) - \frac{1}{2} \frac{\partial}{\partial n} (f(n) + g(n))P(n) \right\} \quad (6.19)$$

This is the **Fokker-Planck equation**. It is a continuous version of the Master equation (6.1) in which we treat n to be a continuous variable. The term inside the curly brackets in equation (6.19) is called the **probability flux**. It describes the flow of states (the “current” flows between the boxes in Fig. 6.1). Labeling the whole thing inside the curly bracket as $J(n,t)$, we can rewrite equation (6.19) as

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n} J(n,t) \quad (6.20)$$

This equation is identical to the **transport equation** that you may have seen in physics.

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

Solving equation (6.20) by hand (i.e., finding $P(n,t)$) is difficult. But as we did with the Master equation, we can solve for the steady-state. Importantly, we can solve equation (6.20) for the

steady-state of *any* genetic circuit whereas in the case of the Master equation (6.1), we could only do so for the constitutive gene expression. In steady state, equation (6.20) becomes

$$\frac{\partial P(n, t)}{\partial t} = 0 \quad (6.21)$$

This means that $J(n, t)$ is a constant. In fact, this constant must be zero. The reason is that at $n=0$, we must have $J(0, t)=0$ (at all times). This is because no current can flow past $n=0$ towards the negative values of n (n must be positive). Moreover, no current can flow past $n=0$ from the negative side of n towards the positive side of n (since there are no states with negative values of n). Since $J(n, t)$ is a constant (same value at all n) at steady state, we must have $J(n, t)=0$ in steady-state. Thus from the definition if $J(n, t)$, we have

$$(f(n) - g(n))P(n) = \frac{1}{2} \frac{\partial}{\partial n} (f(n) + g(n))P(n) \quad (6.22)$$

By defining

$$H(n) = (f(n) + g(n))P(n) \quad (6.23)$$

Equation (6.22) becomes

$$\frac{(f(n) - g(n))H(n)}{f(n) + g(n)} = \frac{1}{2} \frac{dH(n)}{dn} \quad (6.24)$$

We can solve for $H(n)$ by separating variables,

$$\frac{2(f(n) - g(n))}{f(n) + g(n)} = \frac{1}{H(n)} \frac{dH(n)}{dn} \quad (6.25)$$

and then integrating to get

$$H(n) = A \cdot \exp\left(\int_0^n 2 \frac{f - g}{f + g} d\tilde{n}\right) \quad (6.26)$$

where A is a constant to be determined. Then from the definition of $H(n)$, we have

$$P(n) = \frac{A}{f + g} \cdot \exp\left(\int_0^n 2 \frac{f - g}{f + g} d\tilde{n}\right) \quad (6.27)$$

This is the **steady state solution of the Fokker-Planck equation**, for an *arbitrary gene expression scheme*. Here, A is the normalization factor (it should be a constant so that $\int_0^\infty P(n)dn = 1$). Thus indeed, the Fokker-Planck equation is easier to solve for the steady-state than the discrete Master equation (6.1). The integral inside the exponential behaves like a **potential energy** function $U(n)$. Specifically, thinking of $U(n) = -\int_0^n 2\frac{f-g}{f+g} d\tilde{n}$, we have

$$P(n) = \frac{A}{f+g} \cdot \exp(-U(n)) \quad (6.28)$$

The idea is that the $\exp(-U(n))$ is like the Boltzmann factor in statistical mechanics with the $U(n)$ being 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 useful 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. You will find that $U(n)$ represents a double-well potential, and that $P(n)$ would be bimodal. This is what we would expect too for a double well potential in physics.