Chapter Three

Stochastic Gene Expression and Cellular Variability

3.1 BIOLOGICAL CONTEXT AND ONGOING CHALLENGES

A human body includes many trillions of mammalian cells. Humans, like other multicellular organisms, are not comprised of the same cell repeated, a trillion times over (in fact, each of us harbors approximately as many microbial cells as our human cells). Instead, individual cells differentiate into distinct phenotypic states. These states are then heritable from one cell generation to the next, transforming the nature of the organism during development. Lineages of phenotypically convergent cells organize spatiotemporally to give rise to complex structures, functions, and behaviors. Do similar principles apply to bacteria? For example, a single bacteria grown in a 50ml flask filled with rich media can, in the course of a day, give rise to a population of billions. A world in a bottle. Yet, are all of these cells the same? Instead, could many of them have qualitatively different phenotypes than others? The answer is that cells can have markedly different phenotypes even if they are all recent descendants of a common ancestral cell, and that such phenotypic differences are possible even in genetically identical cells.

The previous chapter already described a core concept underlying such emergent differences: bistability. In the genetic toggle switch engineered by Gardner and colleagues, bistability arose due to the nonlinear feedback between two genes and their corresponding transcription factors. Such models suggest a mechanism for incipient forms of differentiation. That is, a cell with many type A repressors will repress expression of type B repressors, thereby remaining in the 'A' state. Similarly, a cell with many type B repressors will repress expression of type A repressors, thereby remaining in the 'B' state. If similar processes subdivide further, it is possible to envision a process of multistability and, perhaps, differentiation in which an eye cell or a wing cell on a fly have a single origin, yet vastly different fates. It is a long way from engineering a synthetic toggle switch to understanding multi-cellular differentiation, but it is a path worth walking on, at least for a few steps. In doing so, the continuous model of bistability leaves a major question unresolved. Cells have a particular history, so given a colony (or lineage) arising from a single cell, in a particular state, shouldn't all the daughter cells remain in precisely the same state? Continuous models of bistability suggest that differences must be set exogeneously. But single-cell observations of GFP reporter levels in *E. coli* with synthetically engineered toggle switches suggest otherwise (Gardner et al. 2000a).

In that particular system, the activity of repressors is controlled by small molecules, i.e., chemical "inducers". These inducers change the activity of transcription factors enabling them to repress, or not, their targets. By varying the levels of IPTG – one widely used inducer – cells switched gene expression from one qualitatively different state to another. Yet, at any given level of IPTG, there was a subpopulation of cells in either state. That is bistability seemed inherent to (nearly) every level of inducer. To borrow the generic language of the prior paragraph, these observations raise a generic question: what made some cells more 'A'-like and others, more 'B'-like? It could be that the cellular population consisted of distinct genotypes which underlies fixed levels of heterogeneity. Yet, it could also be that the very nature of transcription and translation – the process by which information is converted from DNA to RNA into proteins – underlies these cellular-scale differences.

This is the perspective we take here. That cells can have different structures, functions, and behaviors, even if they have identical genotypes. The differences may be amplified or reinforced by nonlinear feedback in the system. Yet the differences can emerge due to noise inside the system, irrespective of whether exogenous factors – and there are many – influence the particular milieu in which a cell lives. This is called 'intrinsic' noise (i.e., arising from stochastic gene expression) to distinguish it from 'extrinsic' noise (i.e., variations arising due to variations in the external environment) (Elowitz et al. 2002a; Süel et al. 2007). To understand the nature of stochasticity in gene expression requires a different kind of model than the continuous models described in the previous chapter. We must turn to models that account for the *discrete* nature of molecules inside cells. That is, the process of transcription yields 1 new mRNA molecule in a cell. Similarly, the process of translations yields 1 new protein molecule in a cell. Such changes also modify the concentration of molecular abundances. A concentration is just the number of molecules divided by a volume. Yet, dividing discrete changes by a continuous volume to yield a concentration does not absolve us of a need to recognize and develop appropriate models of stochastic gene expression. A translation event does not mean that the cell now has 1.23429 more proteins. It has one more. And, if the protein were to degrade, the cell would have one less. Building models that can account for such discrete changes is the focus of this chapter.

But before we do so, it is important to reinforce that such models are necessary to understand pressing biological challenges. To what extent will a population of bacteria be susceptible to antibiotics if the founding cell of a lineage or population is susceptible? What fraction of cells will enter dormancy given stressful conditions, e.g., starvation or exposure to toxins? To what extent will a virus entering a cell initiate a process to kill the cell releasing new progeny or, instead, to integrate its genome with the cell, thereby continuing to divide along with the cell itself? What determines which, if any, of one cell amongst billions or trillions will transform into a cancerous cell, with profound consequences for its local environment and, perhaps, the multicellular host. These questions are but a few of many ongoing challenges in modern biological sciences. The answers

to each may involve theories and principles that cut across scales, including ecological and evolutionary dynamics. Yet as is apparent, modeling stochasticity is a key part to each.

Figure 3.1 provides two examples of intrinsic noise in microbial systems. The first is from the synthetic toggle switch system in which the level of GFP is measured, cell by cell, as a function of increasing levels of IPTG which should induce the system - in theory - to switch from low levels of expression to high levels of expression. Yet, in reality, some of the cells remain in the low state even as some switch to the high state. This finding indicates the potential for intrinsic differences in gene expression to shape the phenotypic state of individuals cells, particularly at intermediate values of an external driver. Even in the 'low' and 'high' states, there is still cell-to-cell variability in GFP. The extent to which the total noise in a system can be compartamentalized into intrinsic and extrinsic sources has been studied in depth (Elowitz et al. 2002a). Figure 3.1 shows one example in which two different (and distinguishable) flourescent proteins were inserted into engineered *E. coli* strains, albeit under the regulatory control of the identical promoter. In theory, if there is not intrinsic differences between cells, then the levels of expression of both proteins should be identical. However, in the absence of high levels of inducer, cells exhibit differences in expression. Only when IPTG levels are increased do the differences between cells diminish - because each cell is induced to its maximum expression.

This chapter will introduce a framework for modeling stochastic gene expression. This framework will provide a baseline of what to expect in systems with 'simple' gene regulation, albeit including the intrinsic variation driven by the discrete nature of molecules. This framework can also be used to probe the limits of our mechanistic assumptions. In doing so, we will build towards trying to compare the predictions of simple models of stochastic gene expression with *in vitro* data in which it is possible to measure molecules one by one.

After establishing key formalisms for describing variation in gene expression, the second part of the chapter will focus in on the very assumptions underlying such models. Measuring gene expression at the level of molecules affords the chance to ask new questions of our quantitative models. In the mid-2000s, Ido Golding, Ted Cox and colleagues developed a new approach to measuring the expression of mRNA, one by one(Golding and Cox 2004; Golding et al. 2005). As we will see, doing so provides a new vista into the fundamental mechanisms of gene expression which we have thus far considered as a memoryless or Markov process. Instead, by carefully examining the dynamics of expression, it will be possible to address whether the expression of mRNA consistent with a model in which transcripts are continuously produced (albeit stochastically) and degraded? Or, instead, must we appeal to a more nuanced model, in which the state of a promoter alternatives between 'On' and 'Off' states such that transcripts appear in bursts? As we will see, the value of quantitative models is often enhanced when they fail. By failing, they provide a baseline for recognizing when measurements point us towards new biological insights rather than just confirming what we thought we knew in advance.

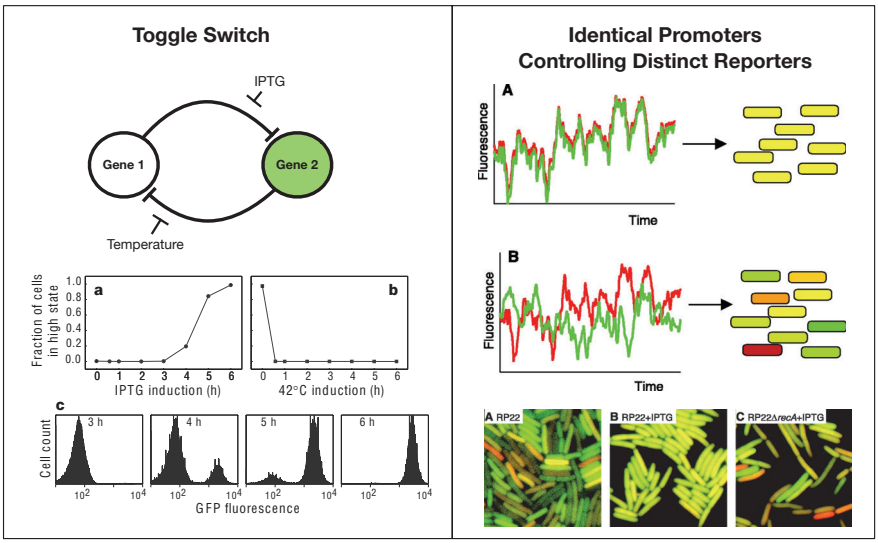

Figure 3.1: Stochasticity leads to variation in cellular level outcomes, whether in bistable switches (left) or in identically controlled reporters (right). (Left) a toggle switch is a bistable feedback loop where two genes mutually inhibit each other, and the activity of the two genes are controlled by external factors - IPTG and temperature. When IPTG is high, the system turns 'On' (as evidenced by the expression of GFP (green flourescent protein), and when temperatures are increased then the sytem turns 'Off'. However, as shown in panel (c), individual cells do not move uniformly between states, instead there are marked differences in cellular state associated with the bistability of the circuity. (Right) In the event that the same promoter is used to control distinct reporters that express GFP and RFP, it is possible to evaluate the extent to which the expressoin of a promoter is identical (as in the top image) or variable (as in the bottom image). The correlation between identical promoters is one way to measure intrinsic noise. Images reproduced from (Gardner et al. 2000a; Elowitz et al. 2002a)

3.2 STOCHASTICITY IN GENE REGULATION

3.2.1 Continuous and discrete paths

Rather than beginning with a model of stochastic gene regulation, it is worthwhile to revisit deterministic models that pose a useful starting point and, eventually, a relevant contrast. Consider a simple gene regulatory system in which gene X positively regulates gene Y, in other words:

$$X \to Y. \tag{3.1}$$

Assume that the inducer is present so that the transcription factor X is active. A mathematical model of the dynamics can be written in terms of continuous changes in the concentration of Y:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \beta(x) - \alpha y \tag{3.2}$$

where $\beta(x)$ is the concentration-dependent production rate of y proteins. If x is present at or near saturation levels such that $\beta(x) = \beta$, then the solution is known (as was discussed in the previous chapter):

$$y(t) = \frac{\beta}{\alpha}\left(1 - e^{-\alpha t}\right) \tag{3.3}$$

In this model, every cell will eventually converge to the same limiting concentration, β/α. Moreover, small deviations from this equilibrium concentration will relax back exponentially to equilibrium with a rate constant of α. In essence, this model predicts that a target gene under control of a single transcription factor will behave identically, irrespective of initial conditions.

Yet, this model assumes that the concentrations of y change continuously. Instead, to whatever extent such a model remains valid, it is important to keep in mind that it is an approximation to a process in which the number of y molecules inside each cell changes in a sequence, e.g.,

$$\text{Cell 1: } 0 \to 1 \to 2 \to 1 \to 2 \to 3 \ldots \tag{3.4}$$

This particular sequence includes four production events and one degradation event. But this is not the only path that could yield a cell with 3 molecules after 5 events, e.g.,

$$\text{Cell 2: } 0 \to 1 \to 0 \to 1 \to 2 \to 3 \ldots \tag{3.5}$$

$$\text{Cell 3: } 0 \to 1 \to 2 \to 3 \to 2 \to 3 \ldots \tag{3.6}$$

In the end, each outcome is the same, but the processes are different. In addition, whatsoever the underlying state, this is not the only final outcome possible with 5 events, e.g.,:

$$\text{Cell 3: } 0 \to 1 \to 2 \to 1 \to 0 \to 1 \ldots \tag{3.7}$$

These sequences suggest a route forward: develop models that describe paths in a "molecular state space". That is, instead of quantifying cells in terms of their continuous concentration of molecules, consider quantifying cells in terms of n, the discrete number of molecules. If we further restrict ourselves to individual synthesis and degradation events, then the dynamics of cells could be represented by the following kinds of reactions:

$$\boxed{n-1} \underset{\text{degradation}}{\overset{\text{synthesis}}{\longleftrightarrow}} \boxed{n} \underset{\text{degradation}}{\overset{\text{synthesis}}{\longleftrightarrow}} \boxed{n+1} \tag{3.8}$$

This state space spans all the non-negative numbers, $n = 0, 1, 2, \ldots$. With this in mind, the next question becomes, are sequences enough? For example, consider two cells that do not have Y proteins. Then, an inducer is added to the system, such that the transcription factor X initiates gene expression in both. Even if the experimental conditions are tightly controlled such that both cells are exposed to exactly the same amount of inducer, the two cells need not produce the new Y protein at the same time. These differences suggest that paths differ in both the *discrete number* of molecules as well as the *timing* at which events take place. Timing of individual events constitutes the second component of stochasticity we must account for in understanding gene expression.

3.2.2 Timing between individual events

Consider an event, like transcription, translation, or degradation, that takes place at a rate r per unit time. For example, if an event takes place at $r = 3$ hrs^{-1}, then there should be approximately 3 events in every hour, or one event every 20 minutes. In a large period of time T, there should be rT events on average. But, what about a small period of time? For example, how many events should take place in a minute if $r = 3$ hrs^{-1}. The value of rT in this case is $3 \cdot \frac{1}{60} = \frac{1}{20}$ or 0.05. It would seem intuitive to think about this value of 0.05 as the average number of events. But in any given cell, the event either takes place or does not. If this is true for each cell, then perhaps the value of 0.05 should be interpreted as a probability of an event taking place. This interpretation becomes more evident as we consider ever smaller intervals of time, ΔT. Formally, rates of biological processes should be thought of as *probabilities per unit time*. This definition may seem counter-intuitive and suggests the need for additional explanations.

Probability rates, i.e., the probability per unit time, can exceed 1, just as $r = 3$ hrs^{-1} does in this example. These rates must be combined with time to yield a probability. And, of course, probabilities of any particular event must lie between 0 and 1, inclusive. Let's make this connection now. Denote the probability that the event takes place in some very small time interval $\mathrm{d}t$ as $P_{\text{event}} = r\mathrm{d}t$ and the probability that the event does not take place in some very small time interval $\mathrm{d}t$ as $P_{\text{noevent}} = 1 - r\mathrm{d}t$. Here $\mathrm{d}t \ll 1/r$, so it would be appropriate to consider time intervals much much less than 1/3 of an hour given $r = 3$ hrs^{-1}. In any small increment, the chances of an event taking place will be small, and as $\mathrm{d}t \to 0$, the chances will become vanishingly small. Given these definitions, what is the probability that it will take a finite duration T before

an event takes place? Formally, what we mean to ask is: what is the probability that the event takes place in a very small time interval between T and $T + \mathrm{d}t$ and not in any small time interval before then?

To answer this question, consider dividing up the timeline between 0 and T into very small increments, each of duration $\mathrm{d}t$. There are $N = T/\mathrm{d}t$ such increments. What is the probability that the event did not place in the first increment? The answer is simply: $1 - r\mathrm{d}t$. If no events to take place in the interval $(0, T)$, then the same logic should apply for all of the N increments. This is equivalent to the probability of tossing a biased coin, in which it turns up "tails" (no event) N times, each with probability $1-r\mathrm{d}t$. The probability of such a sequence of non-events is $(1 - r\mathrm{d}t)^N$. Now, the probability that the event takes place between $(T, T+\mathrm{d}t)$ is $r\mathrm{d}t$. The product of these two probabilities represents the probability that one must wait a time interval of T before observing an event which takes place at a rate r:

$$P_{\text{event}}(T, T + \mathrm{d}t) = (1 - r\mathrm{d}t)^{\frac{T}{\mathrm{d}t}} r\mathrm{d}t \tag{3.9}$$

$$= \left(1 - \frac{rT}{T/\mathrm{d}t}\right)^{\frac{T}{\mathrm{d}t}} r\mathrm{d}t \tag{3.10}$$

$$= e^{-rT} r\mathrm{d}t \tag{3.11}$$

where the last step utilizes the definition of an exponential

$$e^a \equiv \lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n \tag{3.12}$$

and noting that in this example $a = rT$ and $n = T/\mathrm{d}t$. The key result is that an event which can occur, or not, at any moment in time at a fixed rate will be distributed exponentially in time. That is, events are more likely to take earlier than later, indeed, exponentially more likely! This run counters to intuition that events should be spaced somewhere near the average interval. For example, if events occur at a rate of $r = 3$ hrs^{-1} then the average interval should be 1/3 hrs, or 20 minutes. Yet, if were to make many such samples of a process and then "bin" the measurements into minute intervals, we would find a surprising result: the most likely waiting period was less than 1 minute! The technical appendices show that the expected waiting time is $\langle T \rangle = 1/r$, even if the peak of the exponential distribution is near 0 (see Technical Appendices for details).

This result makes it evident that a model of stochastic gene expression should account for both the state of cells and the potential differences in timing for events that could change such states. Such a model would seem to be possible from an algorithmic perspective, that is, we might be able to simulate an individual cell amongst many. For example, here is "pseudo-code" for a numerical simulation that one could plausible envision writing that also matches the intuition of many biologists for how cells work:

```
Simulating the stochastic gene expression of a single cell
```

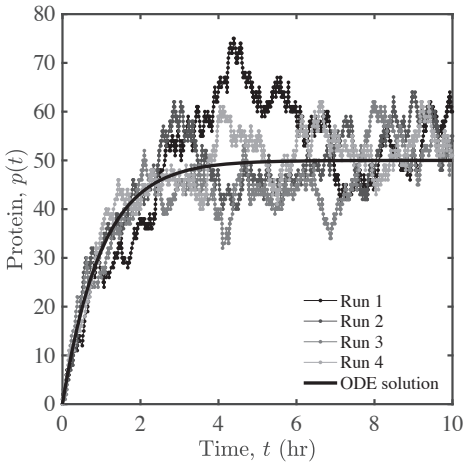

Figure 3.2: Realized stochastic gene expression dynamics compared to the ODE solutions for a fixed expression level $\beta = 50$ nM/hr and degradation rate $\alpha = 1$/hr. The ODE dynamics are that of Eq. (3.2). The stochastic dynamics utilize a Gillespie algorithm to simulate the processes of production and degradation (for more details see the computational laboratory associated this this chapter).

```
Specify initial state of cell, e.g., number of proteins
While measurements are taking place, do the following:
   Calculate event rates based on the current protein count
   Randomly select the waiting time until the next event
   Identify which event type took place at the specified time
   Update the state of the cell and time based on the event
        production: add one protein
        degradation: subtract one protein
Continue
```

This pseudo-code provides the starting point for the workhorse of many stochastic gene expression frameworks. In formal terms it is termed the Gillespie algorithm and is described in detail in the computational lab guide accompanying this chapter (Gillespie 1977). The Gillespie algorithm is a powerful tool and paradigm. Yet, for the example of $X \to Y$, a quantitative bioscientist should want to know what to expect from the simulation even before beginning. As we will find, the stochastic simulations hew close to, but not exactly on, the expectations from the ODE dynamics (see Figure 3.2).

It is evident that individual cells may have different states. Hence, it is worthwhile to think not about predicting, in absolute terms, the state of a cell, but instead predicting, in relative terms, the probability that a cell is in one particular state out of many. To do so, we will consider the concept of an *ensemble*, that is many systems prepared identically, but whose fates may differ

precisely because of the nature of the stochastic process. Here, these systems will be cells, and the processes will be the production and degradation of proteins. The stochastic simulation framework generalizes to other systems, e.g., that of individuals and the processes of birth and death (see next Chapter). Irrespective of the application, consider the following definition: $P_n(t)$: probability that there are exactly n molecules in the system at time t. In the case of stochastic gene expression of a protein Y, $P_{30}(10) = 0.01$ means there is a 1% probability of finding exactly 30 Y proteins in a randomly chosen cell at time 10 hrs. In practice, if the ensemble included 100,000 cells, then one would expect to find 1,000 cells with 30 Y proteins at time 10 hrs. Now, what happens if one of the proteins in one of these cells degraded at some point between 10 and 10.01 hrs – and no other events took place? Then, $P_{30}(10.01)$ would decrease. This type of *observational* approach belies the fact that with so many events taking place, it might be possible, in advance, to predict how such probabilities should change. That is indeed our goal: to characterize the dynamics of $P_n(t)$ given knowledge of mechanisms taking place that shape the transitions between different states. As it turns out, this goal is non-trivial, but even if we can't solve entire goal, we can make progress on quantifying other measurable features of such dynamics, including the mean and variance. The means and variance of gene expression states will become rather useful in comparing expectations to observations.

3.3 CHARACTERIZING DYNAMICS OF INDIVIDUAL CELLS, GIVEN STOCHASTIC GENE EXPRESSION

3.3.1 Getting to a full model of stochastic gene expression

This section outlines how to construct a principled mathematical model of stochastic gene expression. In doing so, the dynamics will modify $P_n(t)$, the probability that a cell has n proteins at time t. The two processes - protein production and degradation - change these probabilities in different ways. First, a cell in the ensemble with $m < n$ proteins may change to one with $n = m + 1$ proteins given production. Second a cell in the ensemble with $m > n$ proteins may change to one with $n = m - 1$ proteins given degradation. Hence production and degradation provide distinct routes to *increase* $P_n(t)$. But the same processes also provide distinct routes to *decrease* $P_m(t)$. First, a cell in the ensemble with n proteins may change to one with $m = n + 1$ proteins given production. Second a cell in the ensemble with n proteins may change to one with $m = n - 1$ proteins given degradation. In general, we can write these changes in probability as

$$\begin{aligned} P_n(t + \Delta t) \quad &= P_n(t) + \Sigma_{m \neq n} P_m(t) W_{mn} \\ &\quad - \Sigma_{m \neq n} P_n(t) W_{nm} \end{aligned} \tag{3.13}$$

where Σ_{mn} (Σ_{nm}) denotes the rate per unit time of transitioning from a state m to n (n to m). In essence one can think of this as the sum of transition probabilities conditioned upon the current state of the ensemble.

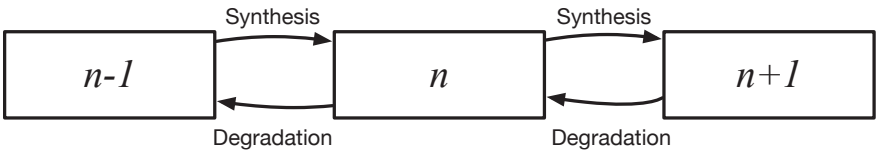

Figure 3.3: Transitions between states with n molecules depending on synthesis and degradation. The synthesis of molecules increases the value of n and the degradation of molecules decreases the value of n. The Master Equation considers the change in probability of any particular system state n as a function of current probabilities and the rates of synthesis and degradation.

Eq. 3.14 is known as "The Master Equation" (for an extended treatment see (van Kampen 2001)). The Master Equation fits on one line, but describes an infinite number of equations, one for each value of $n = 0, 1, 2 \ldots$. To make the application self-evident, consider the case of simple gene regulation such that $X \rightarrow Y$. In this case, Eq. 3.14 can be written in terms of the 4 ways that probability "flows" in and out of $P_n(t)$ (see Figure 3.3):

$$\begin{aligned} P_n(t+\Delta t) &= P_n(t) + \beta P_{n-1}(t)\Delta t \qquad (3.14) \\ &\quad +(n+1)\alpha P_{n+1}(t)\Delta t \\ &\quad -\beta P_n(t)\Delta t - \alpha n P_n(t)\Delta t \qquad (3.15) \end{aligned}$$

Recall that synthesis occurs at a rate of β per unit time, irrespective of the current protein abundances. In addition, degradation occurs at a rate of α per unit time per protein. Hence, to get the total transition rates per unit time, α must be multiplied by the relevant number of proteins - that is $(n+1)$ for degradation events that increase $P_n(t+\Delta t)$ and n for degradation events that decrease $P_n(t+\Delta t)$. Note that $\mathrm{d}P_n(t)/\mathrm{d}t = \frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}$ in the limit that $\Delta t \rightarrow 0$. Hence, the dynamics for this system is:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = \beta P_{n-1} + (n+1)\alpha P_{n+1}(t) - \beta P_n(t) - \alpha n P_n(t) \qquad (3.16)$$

If only we knew how to solve this last equation... In fact, there are ways, but those lay outside the scope of this current set of chapter notes. Instead, let's imagine that we did, in which case it would be possible to calculate – at any moment in time – the expected mean and variance of protein abundance within cells. The average number of proteins is defined as follows:

$$\langle n(t) \rangle = \sum_{n=0}^{\infty} n P_n(t) \qquad (3.17)$$

The variance in the number of proteins is defined as follows:

$$\begin{aligned}\text{Var}[n(t)] &= \left\langle [n(t) - \langle n(t)\rangle]^2 \right\rangle \\ &= \sum_{n=0}^{\infty} \left[n - \sum_{m=0}^{\infty} mP_m(t)\right]^2 P_n(t) \\ &= \sum_{n=0}^{\infty} n^2 P_n(t) - \left[\sum_{n=0}^{\infty} nP_n(t)\right]^2 \end{aligned} \tag{3.18}$$

This last equation is the standard formulation of the variance written out explicitly, i.e., $\text{Var}[n(t)] = \langle n^2\rangle - \langle n\rangle^2$. Note the crucial point, we began this chapter discussing how to describe stochastic gene in gene expression. Here, we are a step closer to a partial victory: a description of the expected state of cells as well as variance between them. If we knew $P_n(t)$, then calculating such dynamics would be trivial. But, as shown next it's possible to figure out these dynamics even without knowing $P_n(t)$.

3.3.2 Deriving the mean and variance of stochastic cellular dynamics

It seems difficult to know what to do when faced with a challenging question. Questions are often posed like puzzles in which answers seem apparent, only after you know the answer. Yet, how does one figure out the answer in the first place? That is what makes research different than puzzle. We may be interested in randomness in gene expression, but hopefully the process of understanding should not be imbued with a similar kind of randomness. To the extent that many possible mathematical approaches can help you logically reason through a problem, I advocate for doing the simplest thing. Write down an expression that clearly recapitulates your question, however simplified it may seem. There may be many things to know about the ensemble of cells, but first and foremost: what is the average number of proteins found in a randomly chosen cell, $\langle n(t)\rangle$? The mean is a dynamic variable and, in principle, should be governed by a predictable set of rules. Hence, rewrite the equation for the mean and consider its derivative in time (removing the parenthetical notation of time for P_n):

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\langle n(t)\rangle = \sum_{m=0}^{\infty} mP_m(t)\right) \tag{3.19}$$

which becomes

$$\frac{\mathrm{d}\langle n\rangle}{\mathrm{d}t} = \sum_{m=0}^{\infty} m\frac{\mathrm{d}P_m}{\mathrm{d}t} \tag{3.20}$$

because the time derivative of $mP_m(t)$ is $m\mathrm{d}P_m/\mathrm{d}t$. The change in probabilities is given by the master equation, yielding:

$$\begin{aligned}\frac{\mathrm{d}\langle n\rangle}{\mathrm{d}t} &= \sum_{m=0}^{\infty} \beta m P_{m-1} - \sum_{m=0}^{\infty} \beta m P_m \\ &+ \sum_{m=0}^{\infty} \alpha m(m+1)P_{m+1} - \sum_{m=0}^{\infty} \alpha m^2 P_m \end{aligned} \tag{3.21}$$

The Technical Appendices explain how to reduce these sums, term by term. They are worth reviewing. It makes the answer more fun, because as it turns out the dynamics of the mean number of molecules is nothing other than:

$$\frac{\mathrm{d}\langle n\rangle}{\mathrm{d}t} = \beta - \alpha\langle n\rangle. \tag{3.22}$$

This equation seems straightforward. Indeed, we've seen it before. This equation is analogous to the equation for maximal production of a target gene at a rate β with per-protein degradation at a rate α, as presented in the prior chapter on *deterministic* regulation of proteins. Yet now, instead of assuming that all cells are identical, this equation describes the dynamics of the *average* number of proteins in an ensemble of cells. The solution to this equation is:

$$\langle n(t)\rangle = \frac{\beta}{\alpha}\left(1 - e^{-\alpha t}\right) \tag{3.23}$$

In this case, the average number of proteins increases from 0, at the start of addition of a signal, towards the carrying capacity of β/α. Hence, despite the stochasticity inherent in the dynamics, we have not lost all sense of predictability. Rather, the mean is predicted to increase towards the same carrying capacity as predicted in a deterministic model. But, unlike the deterministic model, not all cells are predicted to have the same protein abundances. Variation will remain due to the stochastic nature of gene expression. Rather than repeating the derivation for the variance, let's examine the variance of protein abundances at equilibrium.

The master equation for the gene regulatory network, $X \rightarrow Y$ operating at saturation is:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = \beta P_{n-1} + (n+1)\alpha P_{n+1}(t) - \beta P_n(t) - \alpha n P_n(t) \tag{3.24}$$

If the ensemble of cells is at steady state, then there should be no change in $P_n(t)$, that is, $\frac{\mathrm{d}P_n(T)}{\mathrm{d}t} = 0$ for all values of n. Denote P_n^* as the equilibrium distribution of probabilities such that this steady state condition is satisfied. As described in the technical appendices setting this condition yields a recursive relationship such that

$$P_n^* = \frac{\lambda^n e^{-\lambda}}{n!} \tag{3.25}$$

where $\lambda = \beta/\alpha$. Examples of the results of stochastic simulations of simple gene expression are shown in Figure 3.4, in which it is evident that the resulting distributions of proteins are in fact Poisson distributed. This finding is a powerful result and worth the time to examine the derivation. In other words, even in an environment with the same production rate, the same degradation rate, individual cells would nonetheless differ in their expression levels. The distribution describes probabilities that cells differ in their *discrete* number of proteins. It also has the property that the variance is equal to the mean. This then completes our description of the state of the system, at least at steady state. Note

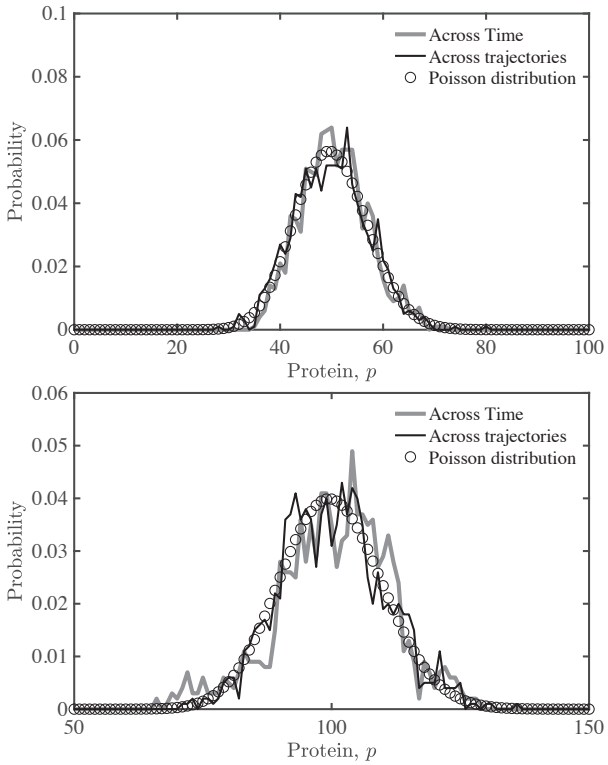

Figure 3.4: Emergence of Poisson distributions via a stochastic gene expression process, given β = 50 nM/hr (top) and β = 100 nM/hr (bottom), both have $\alpha = 1/\text{hr}$.

that convergence to the steady state is a more complicated dynamic, nonetheless it reinforces the central message: that stochastic dynamics elucidate features of the continuous dynamics we have already described as well as shed light on new features that are not present in the continuous model.

In essence, the equilibrium is characterized by noise, such that the variance is equal to that of the mean number of proteins. This also means that the relative error decreases with more expression, and explains how intrinsic noise appears to diminish in strongly regulated systems. The standard error is equal to the standard deviation divided by the mean. For a Poisson distribution, the variance scales like the mean, $\bar{y}$, such that the standard deviation scales like $\bar{y}^{1/2}$, and the standard error scales like $\bar{y}^{-1/2}$. A system with 400 molecules will have 5% standard error compared to a system with 20 molecules which will have 22% standard error. This finding also implies that the basic mechanisms of transcription and translation implies that there should be intrinsic noise in gene regulatory circuits (see Figure 3.1) and baseline expectations should be that

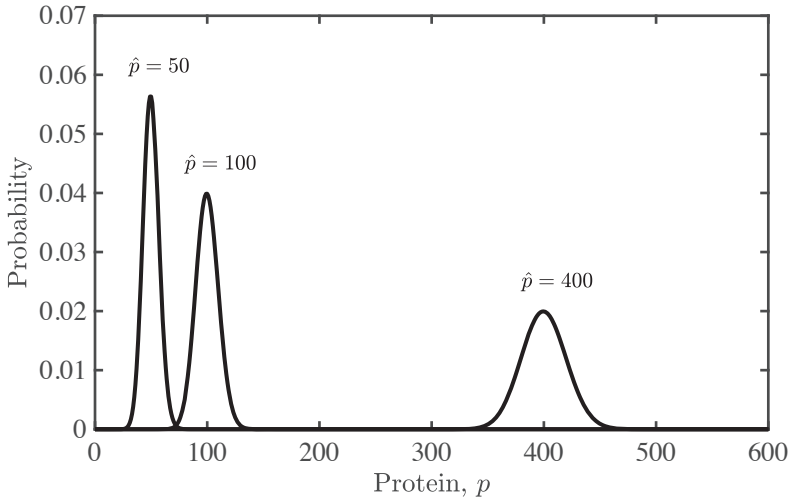

Figure 3.5: Expected Poisson distributions have variances that scale with mean, here for $\hat{p}$ = 50, 100 and 400. Hence, although the overall magnitude of variability increases, the relative variation actually decreases with increasing mean. For example, given $\hat{p}$ = 100 there should be a standard deviation of 10 and therefore 10% noise (10/100) is typical. In contrast, given $\hat{p}$ = 400, there should be a standard deviation of 20, and therefore 5% noise (20/400) is typical. Hence, intrinsic noise should decrease with the increasing strength of a promoter.

such noise scales with the expression level such that the variance is proportional to the mean. The scaling of the mean and variance is yet another hallmark of intrinsic noise; and will prove critical in the next section as we explore whether gene expression can be described as a Poisson process, i.e., memory-less.

3.4 IS GENE EXPRESSION "BURSTY"

Thus far, this chapter has reviewed the processes by which the process of simple gene expression of a target gene Y controlled by a transcription factor X can lead to variation in expression including variation in steady state values. Implicitly, such a a model also suggests that the timing between expression events (e.g., mRNA transcription or protein translation) should be exponentially distributed, which is a characteristic of a Poisson process (not to be confused ith a Poisson distribution). A cell undergoing a Poisson process of gene expression should produce molecules exponentially distributed in time, with a rate set by the production – whether of mRNA or proteins. Such a model can potentially both the variability and repeatiblity of gene expression in genotypically identical cells. For example, cellular gene expression is expected to be variation, such that the variance in molecules is equal to the mean. Yet, perhaps this assumption does

not fully describe cellular dynamics. If not, how would we know? The advantage and utility of quantitative models of living systems fully comes to fruition when we ask models to make predictions that, when tested, show signs of failure. The failure of an otherwise consistent quantitative model may point to new principles or mechanisms that govern the underlying dynamics.

In the mid-2000s, Golding, Cox and colleagues developed a single molecule reporter system designed to measure the expression and abundance of mRNA molecules one at a time (Golding and Cox 2004; Golding et al. 2005). Because mRNA can't be "seen" per se, the key insight was to develop a system to enhance the signal associated with a single molecule. As explained in detail later in this chapter, a reported system was engineered such that each mRNA molecule had an appended sequence that was not translated, but instead, served as an attachment site for another protein complex that included GFP. Hence, when mRNA molecules were transcribed, a part of the mRNA could become rapidly bound with GFP proteins, leading to sharp green foci in the cell that could be measured directly with microscopy. This gene reporter system then alllowed the team to track the expression of mRNA between cells and over time. Figure 3.6 shows a series of trajectories of mRNA measured one by one in a synthetic reporter system. The shape of these trajectories suggest something unusual, insofar as there appear to be periods in which there is significant activity (the periods of steady accumulation) and other periods of relative stasis. This difference does not seem to be consistent with a memory-less (i.e., Markovian) system of gene expression which has been the assumption used throughout this chapter.

In looking at the trajectories of Figure 3.6, it is important to keep in mind the context in the mid-2000s and try to see them with fresh eyes. The study of stochastic gene expression advanced rapidly in the late 1990s as new methods to see and count variation at the scale of individual cells afforded new opportunities to quantify and characterize the quantitative rules of cellular life. Part of the focus of these studies were to identify features of cellular gene regulatory networks so as to identify "motifs", i.e., patterns of regulation that seemed to occur more frequently than expected by chance (Milo et al. 2002). In doing so, perhaps those motifs would also shed light on functional features of gene regulatory networks. This was a productive path, e.g., see the work on the feed-forward loop and principles of gene expression networks more generally (Alon 2007). Other research focused on engineered systems, like the toggle switch introduced in the prior chapter (Gardner et al. 2000a). Yet a different approach altogether was to shed light on the basic, stochastic nature of the component processes, including transcription, translation, and binding (Elowitz et al. 2002b).

One might think that such basic biology is best revealed by observation. Perhaps. But observation coupled to quantitative models provides an even more powerful lens. Let us return to the basic model of stochastic gene expression and consider a single cell with no proteins. In that case, production should occur at a rate β, implying that the time before the first production event should be exponentially distributed with a mean time of $1/\beta$. Recall how Luria and Delbrück also used information from those experiments in which no resistant

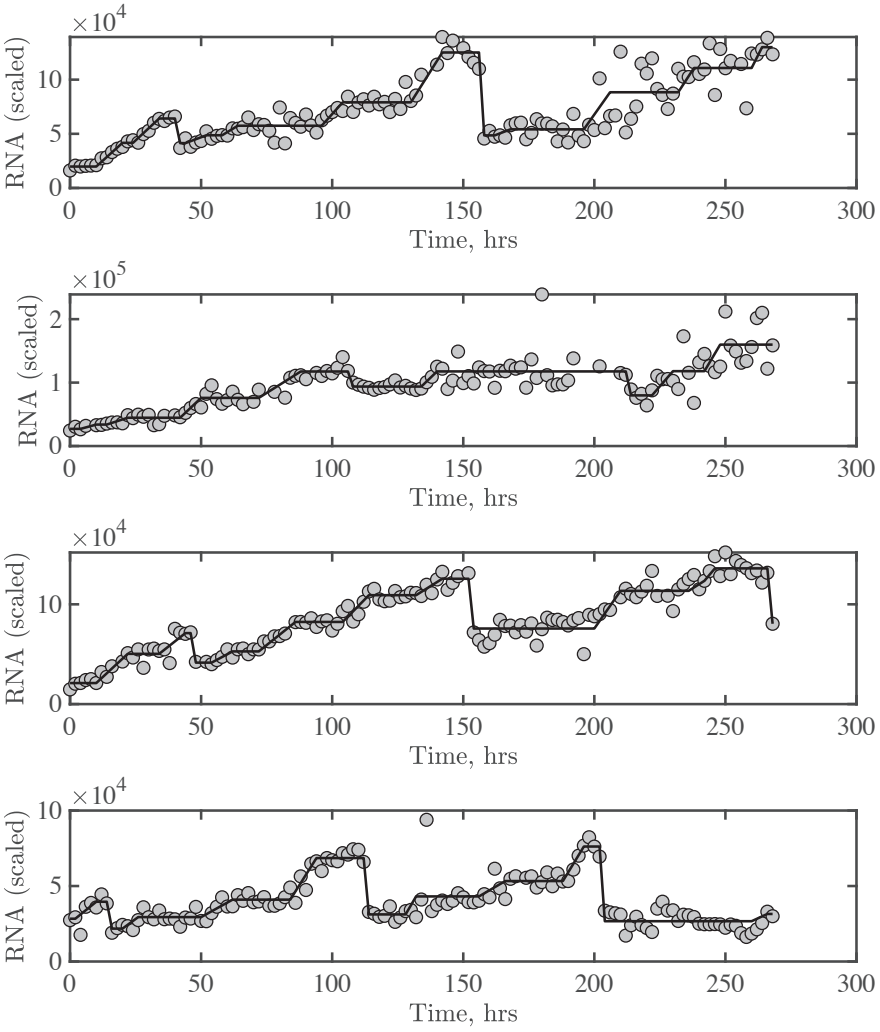

Figure 3.6: Individual trajectories of mRNA in a synthetic reporter system, exhibiting periods of transcriptional bursts and inactivity. This data provides the basis for comparing a Poisson model of gene activity with a two-state model (i.e., with both On and Off states). Data courtesy of Ido Golding, original analysis in (Golding et al. 2005).

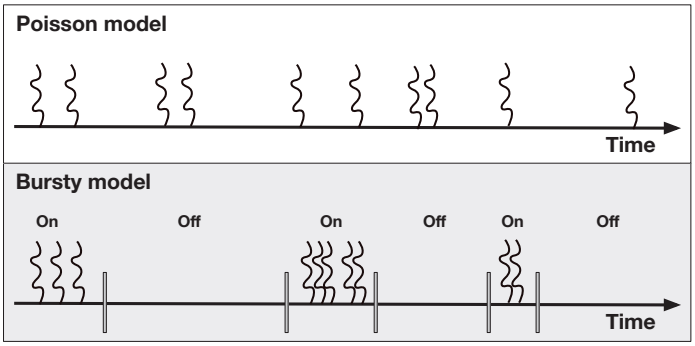

Figure 3.7: The Poisson model and 'bursty' model of transcription. (Top) In the Poisson model, transcripts are generated at a constant rate with an exponential distribution between events. (Bottom) In the bursty model, no transcripts are generated in Off periods, whereas transcripts are generated in bursts during On periods. Note that these two schematics include the same number of total transcripts, albeit distributed in starkly different ways.

colonies emerged as a means to test the dependent and independent hypotheses for mutation. If experimentalists had methods to measure individual transcriptional events as they took place then those time distributions could reveal something about production rates, β. Moreover, observing differences in transcripts between cells could also provide information on the ratio of β/α. Comparing the variance to that predicted under the 'Poisson' hypothesis is analogous to the efforts of Luria and Delbrück to utilize the magnitude of variation to characterize the validity of hypotheses. Yet, the evidence the Golding and colleagues observed suggested that transcription was not strictly Poisson. Instead, the cells seemed to switch between two states, what they term Off and On. In Off states, transcription would not occur. Instead, in On states transcription would occur, and the simplest model of transcription assumes that such events are independent of one another, i.e., are well described by the kind of model we just analyzed. Figure 3.7 provides a conceptual schematic underlying the two alternative hypotheses. The rest of this chapter will explain how individual molecules were measured and then break down different levels of evidence to try and address the two hypotheses: is mRNA transcription bursty or memory-less?

3.4.1 Reporter system to measure mRNA

The Golding paper introduced a reporter system to measure mRNA molecules, one by one as a means to evaluate the burstiness of transcription. This major advance was developed over time, and is described in an earlier paper by the same team (Golding and Cox 2004). The key methodological insight is to combine two promoters into a cassette that is then integrated into *E. coli* cells. A

schematic of the system and expected output is shown in Figure 3.8. One set of genes is controlled by the TetO promoter. When active, this produces a fusion protein comprised of MS2d and a GFP. Hence, a cell will exhibit strong, diffuse green fluorescence after induction. However, this not the only gene in the reporter system. The other component is controlled by the Lac/Ara promoter. When active, then transcription will yield a mRNA including both the mRNA for mCherry (red) and a long mRNA that is comprised of 96 repeats of the same sequence: a MS2 binding site. Hence, given induction via IPTG, then a long mRNA molecule will be transcribed. If MS2d is present, it will rapidly bind to these mRNA molecules. The other part of the fusion protein - GFP - will be localized revealing 'foci', ideally one per mRNA. Later the first part of the mRNA signal will be translated and there will also be a red background signal, corresponding to the production of mRFP1 proteins. Altogether, this system provides a means to potentially count mRNA (via the accumulation of green foci) and the relationship between mRNA and proteins (by contrasting the number of green foci with the strength of the protein signal (measured in the red channel).

3.4.2 Evidence of individual-level mRNA molecule detection

Before evaluating evidence for burstiness, it is essential to ask: does the detection system measure individual mRNA molecules? Qualitative lines of evidence are documented in Golding et al. (Golding et al. 2005) and one of the most critical components is worth reviewing. First, given the detection system, the integrated GFP within segmented foci should provide a rough approximation of the number of mRNA. For example, if g is the GFP intensity per foci and G_{tot} is the total intensity measured in foci per cell, then presumably $n = G_{tot}/g$. Such intensities discount the background GFP which is not bound to a mRNA molecule. When measured across cells, this intensity provides an estimate of the unknown number of transcripts. Figure 3.8 (top) shows the result of estimating transcripts per cell from approximately 84 different cells from a given experiment. Notably, the estimates peak at discrete values. This suggests that the intensity is both sufficiently uniform and resolvable at *individual* mRNA molecules, and that the system can be used to discriminate a difference of a single mRNA molecule between cells. That is, frankly, quite remarkable. The technical feat in this Cell paper was described the year before (Golding and Cox 2004). But beyond being a technical feat, the impact of this technology was made self-evident when the ability to measure RNA dynamics in living *E. coli* cells was put into service of answering a scientific question. And, to answer this question requires that we combine a quantitative model of gene expression with this new type of measurement.

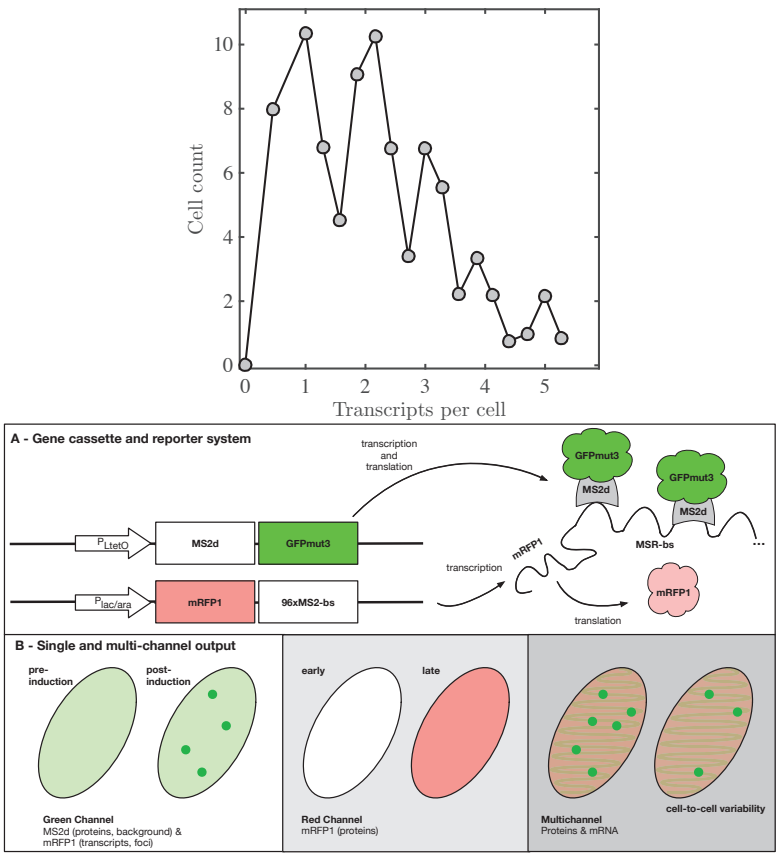

Figure 3.8: Gene reporter system and output. (Top) Integer-valued peaks for measurements of transcripts per cell in a mRNA detection system in *E. coli*. (Bottom-A) The two promoter reporter system, including a binding site for a fusion protein on RNA. (Bottom-B) Expected output via the gene channel, red channel, and multi-channel, including cell to cell variation. For more details, see the description in the text and refer to (Golding and Cox 2004; Golding et al. 2005).

3.5 THE GEOMETRY OF BURSTS

What is a hallmark signal of burstiness? What is a hallmark signal of genes that are not bursty? Like the work of Luria and Delbrück, it is critical to ask quantitative questions of alternative mechanistic models. In this case, if transcription is not bursty, then there should be an equal chance of transcription at any particular moment. This represents the default, 'Poisson' process model of transcription. As was shown earlier in this chapter, Poisson process model exhibit exponentially distributed waiting times between events such that the steady state level of mRNA should have a variance equal to that of the mean. This implies that if one could measure mRNA one at a time, then the time between each transcriptional event should be exponentially distributed with a rate equal to the average transcriptional rate. However, in a "bursty" transcription model, then there are periods with many transcripts produced in a relatively short period of time inter-spaced with periods in which no transcripts are produced. What does such a mechanism say about the timing between burst and non-burst periods and the size of the bursts? Moreover, how big should bursts be, if they exist? Just as in the case of Luria and Delbrück, understanding the predictions of the non-Poisson model take a bit more effort.

3.5.1 Hallmark of burstinees I: Time before mRNA appears

Consider the production of mRNA in an induced cell. The average dynamics given a Poisson model of transcription can be derived via the Master equation given a production rate β and a decay rate α. As shown earlier in this chapter and in the Technical Appendices, the expected dynamics are:

$$\langle n(t) \rangle = n^* \left(1 - e^{-\alpha t}\right) + n_0 e^{-\alpha t} \tag{3.26}$$

such that if there are no mRNA molecules in the cell before induction the dynamics of mRNA should follow:

$$\langle n(t) \rangle = n^* \left(1 - e^{-\alpha t}\right). \tag{3.27}$$

This equation can be directly compared to experiment. Notice that when $t \approx 0$, then $\langle n(t) \rangle \approx \beta t$, such that equilibrium should be $n^* = \beta/\alpha$. This also implies that the deviation of the average expression from the equilibrum, $\delta(t) \equiv n^* - n(t)$, should therefore decay exponentially given a rate, α: $\delta(t) = n^* e^{-\alpha t}$. Hence by measuring the initial production of mRNA and its relaxation to equilibrium it is possible to estimate both the prdocution rate and decay rate. Golding et al. (Golding et al. 2005) estimated these constants as $\hat{\beta} = 8.4$ hr and $\hat{\alpha} = 0.83$/hr. This implies there should be approximately a steady state of 10 mRNA molecules in the system at any given time. Such a Poisson model can in fact fit the data; but just like fitting a Poisson distribution estimate of a selection-dependent mutation rates, these estimates are insufficient to distinguish between the alternative hypotheses. But, they yield a clue.

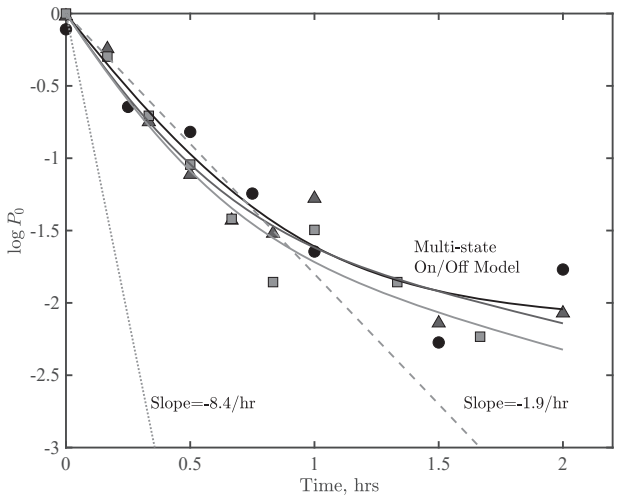

Figure 3.9: Dynamics of the time to first appearance of mRNA are inconsistent with a Poisson production hypothesis. The steeper slope with rate 8.4/hr corresponds to that derived from an estimate of the effective transcription rate. The shallower slope is the estimate reported in (Golding et al. 2005). The solid lines and associated symbols denote the best fit curve assuming a multi-stage process as described in the text. An extended description of this multi-stage model is included in the Technical Appendices.

Given that the production rate is 8.4 molcules/hr, one would anticipate that it should take approximately 0.12 hrs (or 7 minutes) for the first mRNA molecule to appear. Formally, this is equivalent to a waiting time problem, how long should we wait until the system generates a molecule given a single process (transcription) that takes place at a fixed rate? The waiting time for a Poisson process with rate β is simply the exponential distribution, $P_0(t) = \beta e^{-\beta t}$. Instead, Golding et al. found that the observed time of the first mRNA transcipt was exponentially distributed but with the wrong characteristic time! Instead of 7 minutes, the average waiting time was approximately 30 minutes (see Figure 3.9). Yet, there is even more to learn from these curves. Despite the original claims of a slower, exponential decay, the data itself seems to have longer, non-exponential tails. What could give rise to the slower dynamics?

One explanation is that not all the genes in every induced cell was ready to transcribe. In a bursty model of gene trascription, the gene switches between Off and On states, corresponding to where transcription is silenced and when it

is active, respectively. Hence, instead of a stochastic system with one state, we must consider an addition expression state, which we will denote as $B = \{0, 1\}$, where 0 and 1 denote the transctionally inactive and active states, respectively. Hence, the complete model can be described in terms of (B, m). The transitions in this system can be catalogued as:

$$\text{Transcription}: \ (1,m) \overset{\beta}{\rightarrow} (1,m+1) \tag{3.28}$$

$$\text{Degradation}: \ (B,m) \overset{\alpha m}{\rightarrow} (B,m-1) \tag{3.29}$$

$$\text{Gene turn On}: \ (0,m) \overset{k_{on}}{\rightarrow} (1,m) \tag{3.30}$$

$$\text{Gene turn Off}: \ (1,m) \overset{k_{off}}{\rightarrow} (0,m) \tag{3.31}$$

These four events correspond to transcription in the active state, degradation of mRNA (irrespective of the expression state), switching from Off to On, and switching from On to Off. Note that because the rate of degradation is proportional to m, these rules ensure that m remains non-negative. This slight change in specification has profound consequences on expected dynamics.

In order to expand the model, we must also specify the switching rates k_{on} and k_{off} for the transitions from 0 to 1 and from 1 to 0 respectively. As a result, a cell may initially be in the On state. If so, then a mRNA molecule will be transcribed if and only if the mRNA is initiated prior to the switch to the Off state. Otherwise, the system will have to wait until a switch back to the On state before a mRNA can be transcribed. In essence, one can then catalogue the sequence of gene expression states before the first mRNA transcript as follows, where * denotes the first measured transcriptional event:
1*
101*
10101*
...
01*
0101*
010101*
...
The first set of sequences denote a situation where an On gene leads to a transcription event, then one in which the gene goes from On to Off to On before a transcription, etc. The second set of sequences denote a situation an Off gene turns On and then is followed by a transcription event, then one in which the gene goes from Off to On to Off to On and then transcription, etc. It should be apparent that these multiple events would necessarily delay the average time and potentially drive the system away from purely exponential waiting times..

Figure 3.9 reveals that a Poisson process model of transcription lead to exponentially distributed waiting times. But, that is not what is observed. The

waiting times seem both longer than anticipated based on the maximum transcriptional rate and also non-exponential. To diagnose this observation note that using a fast switching model delays the waiting time, even if the observed waiting times are exponential. Indeed, in the limit of very rapid switching (here 10 per hr), then the gene state switches back rapidly between On and Off states. In that event the probability of being in an On state can be denoted as p, whose dynamics must satisfy

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -pk_{off} + (1-p)k_{on} \tag{3.32}$$

with an equilibrium probability $p^* = \frac{k_{on}}{k_{on}+k_{off}}$. For example, the equilibrium probablity $p^* = 1/2$ in the event that the transition rates are equal. In the event that these rates are very fast (compared the gene expression), this finding implies that the rate of production should be slowed down by a factor of p^* such that the effective waiting period is longer than that expected in a Poisson model. In essence, rapid switching changes the waiting time but not the shape of the distribution. In contrast, when switching is very slow (i.e., slower than the time to transcribe conditional upon the gene being in the On state), then the distribution changes shape and is characterized by short waiting times if the gene starts in the On state but very long waiting times if the gene starts in the Off state. In that case, the longer tail is a result of waiting for the gene to switch between Off and On states, corresponding to where transcription is silenced and when it is active. The Technical Appendices derives an approximate distribution of such a multi-state model (see Figure 3.9 for the consequences). The original publication of this data noted that the average waiting time before the first appearance of mRNA was delayed relative to expectations. But, with the benefits of hindsight, it is critical to note that the observed data seems to have a longer-than-exponential tail. That longer-than-exponential tail is a hallmark of a multi-state model! But it's not the only hallmark, for more we need to turn to the size of bursts.

3.5.2 Hallmark of burstinees II: Variation in numbers of mRNA produced

In a bursty model of gene trascription, the gene switches between Off and On states, corresponding to where transcription is silenced and when it is active, respectively. We have shown that in the limit of fast switching, the system should exhibit exponentially distributing waiting times, albeit slower than that expected in a Poisson model. And, in the limit of slow swiching, then waiting times should no longer be exponentially distributed, and in fact in very slow switching should be characterized by two rates: the (fast) transcriptional rate and the (slow) switching rate. However, once switched, we expect that many transcription events will happen in close succession. How big are these bursts?

To understand the nature of transcriptional bursts, it is important to revisit the principles underlying deciding which of two random events occurs first. To do so, consider a system in which there are two processes that occur at rates r_1

and r_2. If $r_1 = r_2$, then it seems apparent that both should have an equal chance. But when $r_1 \neq r_2$, then it would seem that the event with the higher rate is more likely to occur first. But how much more likely? Formally, the probability that process 1 occurs first can be broken down in terms of conditional probabilities. That is, it could be that process 1 occurred in a small interval of time $(0, \mathrm{d}t)$ and process 2 did not. Or, it could be that process 1 occurred in a small interval of time $(\mathrm{d}t, 2\mathrm{d}t)$ before process 2 occurred. And so on.

Formally, the probability that event 1 occurs before event 2 can be written as the following integral:

$$P_{event-1} = \int_0^\infty \mathrm{d}t \quad \overbrace{r_1 e^{-r_1 t}}^{\text{(event-1 occurs)}} \quad \overbrace{e^{-r_2 t}}^{\text{(event-2 has not yet occurred)}} \tag{3.33}$$

As shown in the technical appendices, the result of this integral is:

$$P_{event1} = \frac{r_1}{r_1 + r_2}. \tag{3.34}$$

In essence, the probability of an event is equal to its relative contribution to the total rate at which events occur. This equation generalizes, so that $P_{\text{event i}} = \frac{r_i}{\sum_{j=1}^{s} r_j}$ where s denotes the number of process types. With this in mind, how many transcripts should be produced in a single burst? To answer this, assume that production during On states occurs at a rate β and a switch from an On to an Off state occurs at a rate k_{off}. Hence the probability that n transcripts are produce in one burst is

$$p(n) = \overbrace{\left(\frac{\beta}{\beta + k_{off}}\right)^n}^{n \text{ production events}} \times \overbrace{\left(\frac{k_{off}}{\beta + k_{off}}\right)}^{\text{off event}} \tag{3.35}$$

In essence, the two processes are competing. For there to be n transcripts produced in one burst, then the production process must take place before the off process occurs, precisely n times. The total probability is analogous to getting the sequence $\overbrace{1, 1, \ldots 1}^{n \text{ times}}, 0$ given a biased Boolean coin that comes up with a 1 with probability $\beta/(\beta + k_{off})$ and 0 with a probability of $k_{off}/(\beta + k_{off})$. As shown in the Appendices, these burst distributions are geometric distributions. The geometric distribution has the feature that when $\beta > k_{off}$ then the probabilities of having many events in a burst becomes quite flatter and flatter. This is precisely what is seen in Figure 3.6 in which long trajectories of activity alternate with periods of inactivity. These kinds of periods also explain why there is far more variation in the stationary levels of mRNA than would be expected given a Poisson model. If the process of expression includes additional states this variation must be included in understanding the total observed variation in mRNA levels in a cell. Quantifying the extent of that gap is a story for you to explore (in the homework).

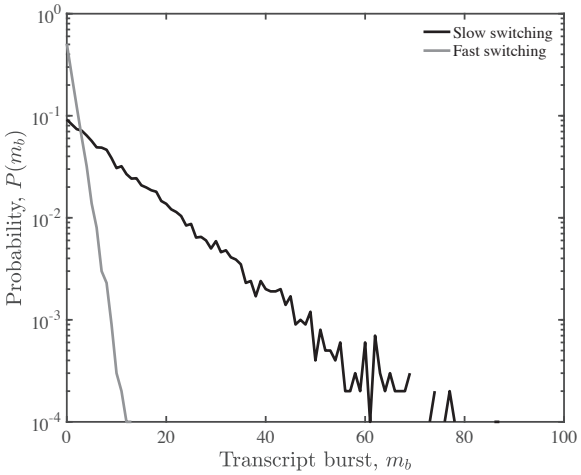

Figure 3.10: Transcript burst distributions given fast and slow switching. In both cases β = 10/hr. In the fast switching model, k_{off} = 10/hr. In the slow switching model, k_{off} = 1/hr.

3.6 TAKE-HOME MESSAGES

- Cellular gene expression is a stochastic process arising, in part, from the fact that individual mRNA are transcribed and then translated into proteins.
- For a Poisson process, the timing between one event and the next should be exponentially distributed.
- Stochastic gene expression dynamics are expected to generate Poisson distributions at steady state, even as the average dynamics is well described by a deterministic solution, in which the average expression converges rapidly to the mean with a time-scale given by the cellular growth rate.
- Poisson distributions have the feature that their variance is equal to their mean, implying that the relative error (i.e., the standard deviation divide by the mean) decreases like the inverse of the square root of the mean. In practice, the relative error goes down (and stochasticity becomes less important) as the mean increases.
- An engineered reported system can be used to measure individual expression of mRNA.
- Measurements of single mRNA dynamics reveals that multiple lines of evidence for bursty transcription that diverge from those expected in a Pois-

son process, including: (i) longer time to the first appearance of mRNA; (ii) increased variance in measurements.

- The bursty dynamics of mRNA lead to longer periods of activity and longer periods of inactivity than would be expected given a Poisson model where all the periods are exponentially distributed.
- A model of bursty transcription implies that expression switches between On and Off states, implying geometric distributions of expression rather than exponential.

3.7 PROBLEM SET

This problem set builds upon the developing toolkit you have accumulated, including computational concepts explored in the computational laboratory guide associated with this chapter. The laboratory guide includes the following major concepts:

- Finding the time to the next random event, given one or more concurrent Poisson processes.
- Combining multiple discrete events together to simulate a stochastic trajectory governed by one or more reactions.
- Modeling stochastic gene expression, including a basal level of production and a fixed decay/degradation.
- Sampling a stochastic trajectory over discretely-spaced time points.
- Assessing the statistics of an ensemble of stochastic trajectories.

With these tools in hand, the following problems are intended to deepen understanding of stochastic gene expression, including the opportunity to analyze data from *in vitro* studies. The overall objective of this problem set is to explore the principles by which cellular individuality emerges, one molecule at a time.

Problem 1. Autoregulation and memory:

Write a program to simulate an autoregulatory, positive feedback loop where X activates itself via Boolean logic. Assume that the protein dilution rate is α, the max production rate is β_+, the basal level is β_-, and the half-saturation concentration is K. Here you will explore the dynamics that unfold when expression is noisy. Throughout, assume that β_- = 20 nM/hr, K = 30nM, and α = 1 /hr. Identify the critical value of β_+ beyond which you expect the long-term dynamics to exhibit bistability? Provide evidence in support of this critical β_+^c.