BE 150: Design Principles of Genetic Circuits

Justin Bois

Caltech

Spring, 2018

Caltech

Donna and Benjamin M.
Rosen Bioengineering Center

This document was prepared at Caltech with financial support from the Donna and Benjamin M. Rosen Bioengineering Center.

11 Bursty gene expression

Last time, we laid out two objectives. First, we aimed to identify the sources of noise and how to characterize them. We summarized statistics about the probability distribution of copy numbers by considering the mean and coefficient of variation. The mean ends up following the same deterministic dynamics we have been considering thus far in the course. The coefficient of variation is a measure of noise, or relative departure from the mean behavior. Now, we will look at the dynamics of the entire probability distribution of copy number, $P(n,t)$.

11.1 Master equations

We will use **master equations** to describe the dynamics of $P(n,t)$. Generally, a master equation is a loss-gain equation for probabilities of states governed by a Markov process.[4] In our case, the "state" is the set of copy numbers of molecular species of interest. The values of n are discrete, so we have a separate differential equation for each $P(n,t)$. Specifically,

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \sum_{n'} \left[W(n \mid n')P(n',t) - W(n' \mid n)P(n,t)\right]. \tag{11.1}$$

Here, $W(n \mid n')$ is the transition probability per unit time of going from n' to n.

The master equation makes sense by inspection and appears simple. The nuance lies in the definition of the transition rates, $W(n \mid n')$. There is also the computational difficulty that n can be very large. In general, solving the master equation is difficult and is usually intractable analytically. Therefore, we *sample* out of the distribution. That is, we can draw many samples of values of n at time points t that are distributed according to the probability distribution that solves the master equation. We can then plot histograms to get an approximate plot of the probability distribution. We can also use the samples to compute moments, giving us estimates of the mean and variance. Generating the samples from the *differential* master equation is done using a **Gillespie algorithm**, also known as a **stochastic simulation algorithm**, or SSA. We will learn how to do these calculations in the next lecture.

11.2 The Master equation for unregulated gene expression

As we have seen in class, unregulated gene expression is described by the macroscale equation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \beta - \gamma n. \tag{11.2}$$

To translate this into stochastic dynamics, we write the production rate as

$$\text{production rate} = \beta\, \delta_{n',n-1}, \tag{11.3}$$

where we have used the Kronicker delta,

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{11.4}$$

[4] A good reference for studying master equations is *Stochastic Processes in Physics and Chemistry* by N. G. van Kampen.

This says that if our current copy number is $n-1$, the probability that it moves to n in unit time is β. Similarly, the decay rate is

$$\text{decay rate} = \gamma\,(n+1)\,\delta_{n',n+1}. \tag{11.5}$$

This says that if we have $n+1$ molecules, the probability that the copy number moves to n in unit time is $\gamma\,(n+1)$. Thus, we have

$$W(n \mid n') = \beta\,\delta_{n',n-1} + \gamma\,(n+1)\,\delta_{n',n+1}. \tag{11.6}$$

We can then write our master equation as

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \beta P(n-1,t) + \gamma\,(n+1)P(n+1,t) - \beta P(n,t) - \gamma\, nP(n,t), \tag{11.7}$$

where we define $P(n<0,t)=0$.

This is a large set of ODEs, and as mentioned in the previous section, solving for $P(n,t)$ is nontrivial and is usually done by Gillespie algorithm. Instead, let's look for a steady state solution of this system of ODEs; i.e., let's find $P(n)$ that satisfies

$$\frac{\mathrm{d}P}{\mathrm{d}t} = 0. \tag{11.8}$$

Even solving for this is difficult. We will instead try a guess-and-check method. We will guess that the steady state distribution is Poisson. We make this guess because it matches an intuitive "story" about the mRNA production. The story of the Poisson distribution is this:

> Rare events occur with a rate λ per unit time. There is no "memory" of previous events; i.e., that rate is independent of time. The probability that k events occur in unit time is Poisson distributed.

Our event here is the production of a gene product. We need these events to happen before decay. The probability mass function of the Poisson distribution is

$$P(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}. \tag{11.9}$$

Let's check if this works. We plug this expression for $P(n)$ into the right hand side of the master equation and see if we can get the expression to be equal to zero.

$$\beta\,\frac{\lambda^{n-1}}{(n-1)!}\,\mathrm{e}^{-\lambda} + \gamma\,(n+1)\,\frac{\lambda^{n+1}}{(n+1)!}\,\mathrm{e}^{-\lambda} - \beta\,\frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda} - \gamma\, n\,\frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda} = 0. \tag{11.10}$$

Division of both sides of the equation by $\lambda^{n-1}\mathrm{e}^{-\lambda}/(n-1)!$ gives

$$\beta + \frac{\gamma}{n}\,\lambda^2 - \beta\,\frac{\lambda}{n} - \gamma\,\lambda = (\beta - \gamma\,\lambda) - \frac{\lambda}{n}(\beta - \gamma\,\lambda) = 0 \tag{11.11}$$

We see that we can get this expression to hold for all n provided $\lambda = \beta/\gamma$. So, the copy number if Poisson distributed with mean β/γ.

It is useful also to consider the moments of the Poisson distribution and compute useful summary statistics of the distribution.

$$\langle n\rangle = \frac{\beta}{\gamma}, \tag{11.12}$$

$$\langle n^2\rangle - \langle n\rangle^2 = \sigma^2 = \frac{\beta}{\gamma}, \tag{11.13}$$

$$\text{noise} = \frac{\sigma}{\langle n\rangle} = \eta = \sqrt{\frac{\gamma}{\beta}}, \tag{11.14}$$

$$\text{Fano factor} = F = \frac{\sigma^2}{\langle n\rangle} = 1. \tag{11.15}$$

11.3 Dynamics of the moments

From the master equation, we can derive an ODE describing the dynamics of the mean copy number of time, $\langle n(t)\rangle$. To do this, we multiply both sides of the master equations by n and then sum over n.

$$\sum_{n=0}^{\infty} n\left[\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \beta P(n-1,t) + \gamma(n+1)P(n+1,t) - \beta P(n,t) - \gamma nP(n,t)\right]$$

$$= \frac{d\langle n\rangle}{\mathrm{d}t} = \beta\sum_{n=0}^{\infty} nP(n-1,t) + \gamma\sum_{n=0}^{\infty} n(n+1)P(n+1,t) - \beta\langle n\rangle - \gamma\langle n^2\rangle, \tag{11.16}$$

where we have used the facts that

$$\langle n\rangle = \sum_{n=0}^{\infty} nP(n,t), \tag{11.17}$$

$$\langle n^2\rangle = \sum_{n=0}^{\infty} n^2P(n,t). \tag{11.18}$$

We have two sums left to evaluate.

$$\sum_{n=0}^{\infty} nP(n-1,t) = \sum_{n=0}^{\infty}(n+1)P(n,t) = \langle n\rangle + 1, \tag{11.19}$$

and

$$\sum_{n=0}^{\infty} n(n+1)P(n+1,t) = \sum_{n=0}^{\infty} n(n-1)P(n,t) = \langle n^2\rangle - \langle n\rangle. \tag{11.20}$$

Thus, we have

$$\begin{aligned}\frac{d\langle n\rangle}{\mathrm{d}t} &= \beta(\langle n\rangle + 1) + \gamma(\langle n^2\rangle - \langle n\rangle) - \beta\langle n\rangle - \gamma\langle n^2\rangle \\ &= \beta - \gamma\langle n\rangle,\end{aligned} \tag{11.21}$$

which is precisely the macroscale ODE we are used to. It is now clear that it describes the mean of the full probability distribution.

11.4 Experimental Fano factors are greater than one

A nice feature of considering the whole probability distribution is that we can predict a value for the noise and Fano factor. Importantly, the Fano factor is invariant to parameters; it is always unity. So, if we see Fano factors of unity in experiment, we know our model for gene transcription is at least plausible.

Gene expression in individual cells typically results in a Fano factor greater than one! There are many ways this could come about, such as explicitly considering the multiple steps involved in making a protein, or by having a switchable promoter, as we will explore in the homework. Today, we will focus on a key experimental results: **gene expression occurs in bursts.**

11.5 Observations of bursty gene expression

In a beautiful set of experiments, Cai, Friedman, and Xie (*Nature*, **440**, 358–362, 2006) discovered that gene expression is bursty. (We will come to understand what we mean by "bursty" means in a moment.)

Through a clever experimental setup (I encourage you to read the paper), Cai, Friedman, and Xie were able to get accurate counts of the number of β-galactosidase molecules in individual cells over time. They found that the number of molecules was constant over long stretches of time, and then the number suddenly increased (Fig. 17). This shows that expression of the β-*gal* gene happens in bursts. In each burst, many molecules are made, and then there is a period between bursts where no molecules are made.

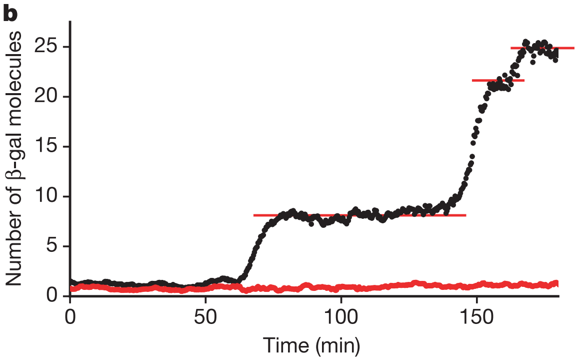

Figure 17: Number of β-galactosidase molecules present in a single cell over time (black) and a blank background (red). Take from Cai, Friedman, and Xie *Nature*, **440**, 358–362, 2006.

Cai and coworkers also found that the number of molecules produced per burst was geometrically distributed, as shown in Fig. 18. Recall, the "story" behind the Geometric distribution.

> We perform a series of Bernoulli trials with success probability p_0 until we get a success. We have k failures before the success. The probability distribution for k is Geometric.

The probability mass function for the Geometric distribution is

$$P(k; p_0) = (1 - p_0)^k p. \tag{11.22}$$

Considering this story, this implies that a burst is turned on, and then it turns off by a random process after some time.

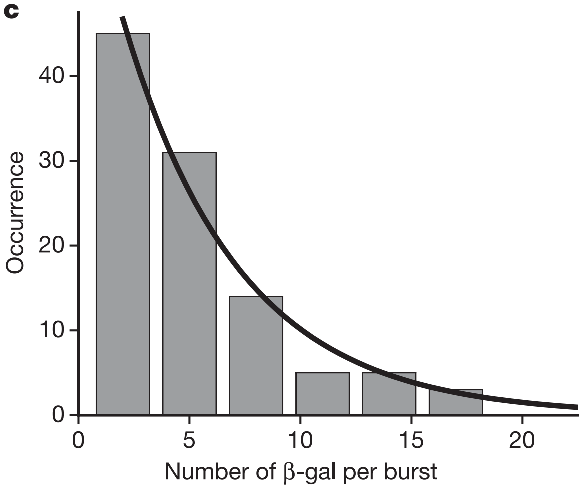

Figure 18: Histogram of burst sizes. The histogram is well-described by the geometric distribution. Take from Cai, Friedman, and Xie *Nature*, **440**, 358–362, 2006.

11.6 Master equation for bursty gene expression

We also see from Fig. 17 that the time scale of the bursty gene expression is much shorter than the typical decay time. So, per unit time, we can have many gene products made, not just a single gene product as we have considered thus far. So, we can re-write our transition probabilities per unit time as

$$W(n \mid n') = \beta' \, \xi_{n-n'} - \gamma\,(n+1)\,\delta_{n+1,n'}, \tag{11.23}$$

where ξ_j is the probability of making j molecules in a burst and β' is the probability per unit time of initiating production of molecules. If $\xi_j = 0$ for all $j \neq 1$, then $\beta' = \beta$, and we have the same equations as before. So, our model is only slightly changed; we just allow for more transitions in and out of state n.

We just have to specify ξ_j. We know that the number of molecules produced per burst is geometrically distributed, so

$$\xi_j = p_0(1-p_0)^j. \tag{11.24}$$

Now, our master equation is

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \beta' \sum_{n'=0}^{n-1} p_0(1-p_0)^{n-n'} P(n',t) + \gamma\,(n+1)P(n+1,t)$$

$$- \beta' \sum_{n'=n+1}^{\infty} p_0(1-p_0)^{n'-n} P(n,t) - \gamma\, n P(n,t). \tag{11.25}$$

We can simplify the second sum by noting that can be written as a geometric series,

$$\sum_{n'=n+1}^{\infty} p_0(1-p_0)^{n'-n} = p_0 \sum_{j=1}^{\infty}(1-p_0)^j = p_0\left(\frac{1-p_0}{1-(1-p_0)}\right) = 1-p_0. \tag{11.26}$$

Thus, we have,

$$\begin{aligned}\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} &= \beta' \sum_{n'=0}^{n-1} p_0(1-p_0)^{n-n'} P(n',t) + \gamma(n+1)P(n+1,t) \\ &\quad - \beta'(1-p_0)P(n,t) - \gamma n P(n,t). \end{aligned} \tag{11.27}$$

As we might expect, an analytical solution for this master equation is difficult. We are left to simulate it using SSA.

11.7 The steady state distribution for bursty expression is Negative Binomial

We can try the same guess-and-check method as before to get the steady state distribution. We will guess that the distribution is Negative Binomial. We guess this because it has the right story.

> We perform a series of Bernoulli trials with a success rate p_0 until we get r successes. The number of failures, n, before we get r successes is Negative Binomially distributed.

Bursty gene expression can give mRNA count distributions that are Negative Binomially distributed. Here, "success" is that a burst in gene expression stops. So, the parameter p_0 is related to the length of a burst in expression (lower p_0 means a longer burst). The parameter r is related to the frequency of the bursts. If multiple bursts are possible within the lifetime of mRNA, then $r > 1$. Then, the number of "failures" is the number of mRNA transcripts that are made in the characteristic lifetime of mRNA.

The Negative Binomial distribution is equivalent to the sum of r Geometric distributions. So, the number of copies will be given by how many bursts we get before degradation. This suggests that $r = \beta'/\gamma$. We can then write the Negative Binomial probability mass function as

$$P(n; r, p_0) = \frac{(n+r-1)!}{n!(r-1)!} p_0^r (1-p_0)^n, \tag{11.28}$$

where $r = \beta'/\gamma$. You can plug this in to the master equation to verify that this is indeed a steady state.

We can put this in more convenient form. Instead of using p_0 to parametrize the distribution, we can instead define the **burst size** b as the mean of the geometric distribution describing the number transcripts in a burst,

$$b = \frac{1-p_0}{p_0}. \tag{11.29}$$

Recall that r is the typical number of bursts before degradation (or dilution), so r is a **burst frequency**. So, for convenience, we write the steady state distribution as

$$P(n; r, b) = \frac{(n+r-1)!}{n!(r-1)!} \frac{b^n}{(1+b)^{n+r}}. \tag{11.30}$$

Strictly speaking, r can be non-integer, so

$$P(n; r, b) = \frac{\Gamma(n+r)}{n!\,\Gamma(r)!} \frac{b^n}{(1+b)^{n+r}}, \tag{11.31}$$

where $\Gamma(x)$ is the gamma function.

The Negative Binomial is interesting because it can be peaked for $r > 1$, but has a maximum at $n = 0$ for $r < 1$. So, for a low burst frequency, we get several cells with zero copies, but we can also get cells with many. This might have interesting implications on the response of a group of cells to a rise in lactose concentration and also drop in other food sources. If $r < 1$, we then have a simple explanation for the "all-or-none" phenomenon of induction observed 60 years ago in *E. coli* (Novick and Weiner, *PNAS*, 1957). Cells are either fully induced or not at all; with $r < 1$, many cells have no β-galactosidase at all.

The known summary statistics of the negative binomial distribution are also useful.

$$\langle n \rangle = rb \tag{11.32}$$

$$\langle n^2 \rangle = rb(1 + b + rb) \tag{11.33}$$

$$\sigma^2 = rb(1 + b) \tag{11.34}$$

$$\text{Fano factor} = F = 1 + b \tag{11.35}$$

$$\text{noise} = \eta = \sqrt{\frac{1 + b}{rb}}. \tag{11.36}$$

So, we can tune the burst frequency and burst size to control variability. For example, we could keep $\langle n \rangle$ constant by increasing the burst size b while decreasing the burst frequency r (big, intermittent bursts), which would result in increased noise. If, instead, we decreased the burst size while increasing the burst frequency (short, frequent bursts), we get reduced noise. This gives a design principle, **bursty gene expression enables cells to regulate the mean and cell-cell variability of protein levels by controlling burst frequency and burst size.**