Single Molecule Transcription. [1]

Pankaj Mehta

March 28, 2019

In this lecture we discuss some simple aspects of gene expression. We discuss how mRNA production is quite slow and that proteins are produced in bursts. We show how these two elements give a simple prediction for steady-state gene expression profiles.

[1] Please see corresponding slides for list of experimental papers. They are really worth looking at and reading. The derivation here (especially of the Gamma distribution) can be found in this Friedman et al PRL 2006. The derivation of the geometric distribution and how it can be used to simulate Master equations is found in Mehta et al Physical Biology 2008.

Some basic numbers and experimental results

If we are to understand what is going on, let us get a sense of numbers. As shown in the slides, the typical prokaryotic cell has mRNA lifetimes τ_m of order 3-10 minutes. Proteins are usually not actively degraded and instead are "diluted" through cell division. In prokaryotes, cell doubling times are from $20 - 60$ minutes. mRNA is also usually produced in the a bursty fashion. The experimental data shown in Figure 1 shows that transcription of mRNA also proceeds in an intermittent fashion, with periods of transcription (6 min) followed by long pauses.

We can write down a simple differential equation that describes this process (ignoring fluctuations)

$$\begin{aligned} \frac{dm}{dt} &= \alpha_m - \tau_m^{-1} m \\ \frac{dp}{dt} &= \alpha_p m - \tau_p^{-1} p \end{aligned} \tag{1}$$

At steady-state, we have that $\bar{m} = \alpha_m \tau_m$ and $\bar{p} = \alpha_p \tau_p \bar{m} = \alpha_m \tau_p \alpha_p \tau_m$. It will be helpful for the future to define two dimensionless parameters

$$\begin{aligned} \bar{b} &= \alpha_p \tau_m \\ a = &= \alpha_m \tau_p. \end{aligned} \tag{2}$$

It is clear that the $\bar{b}$ is the mean-number of proteins produced from a single mRNA molecule. Experiments show that $\bar{b} > 1$ so that proteins are produced in "bursts". For this reason, we will often call this quantity the mean-burst size. Experiments have also directly measured the distribution of burst sizes (see Figure 2) and found that it is well approximated by a geometric distribution (the discrete version of the exponential distribution). The other quantity a measures the mean-number of bursts (mRNA transcription events) per cell cycle.

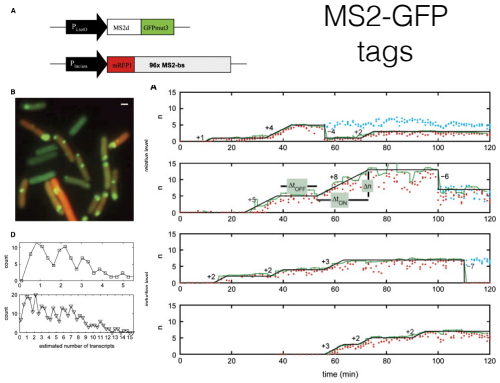
The distribution of protein abundances in a cell is also well described by a gamma distribution (see Figure 3).

$$p(x) = \frac{x^{a-1} e^{-x/\bar{b}}}{\Gamma(a) \bar{b}^a}, \tag{3}$$

where in many cases we can interpret the two parameters of the Gamma distribution exactly as above. Let us now try to derive these results in a simple way.

Figure 1: Data measuring transcription in *E. coli* from Golding et al Cell 2005

Visualizing mRNA molecules



Get kinetics- but have to modify genes

Golding et al Cell. 2005 Dec 16;123(6):1025-36.

.

Geometric Distribution of burst sizes

In this section, we give a simple derivation of why burst sizes are distributed in a geometric manner. We start by considering a single mRNA molecule that can be translated at a rate α_p and can be destroyed a a rate τ_m^{-1}. We know most of the time that nothing happens. In fact, using the exact same intuition as the Gillespie algorithm we can even calculate waiting time distribution between events. However, for calculating the number of proteins produced from each mRNA molecule this distribution plays no role. Instead, we have to ask about the probability that when an event happens, it is the production of protein rather than mRNA degradation. We can write this relative probability as

$$q = \frac{\alpha_p}{\alpha_p + \tau_m^{-1}} = \frac{\bar{b}}{\bar{b}+1}, \tag{4}$$

just like in the second step of the Gillespie algorithm.

Then the probability of producing exactly b bursts is just given by

$$P_{\text{burst}}(b) = q^b(1-q) = \frac{1}{1+\bar{b}}\left(\frac{\bar{b}}{\bar{b}+1}\right)^b \tag{5}$$

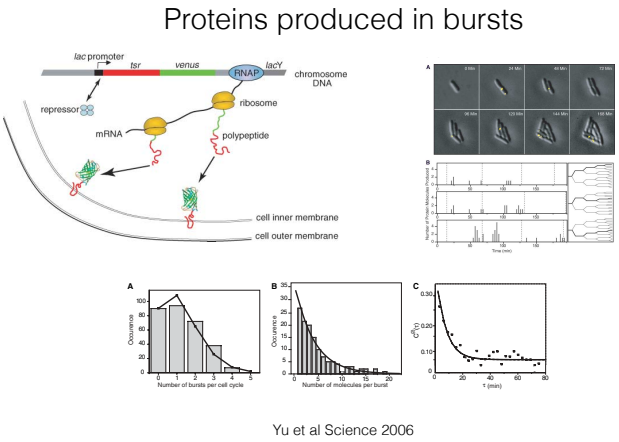


Figure 2: Data measuring bursty protein synthesis. The bottom row shows that the number of bursts (mRNA transcription events) per cell cycle follows a Poisson distribution while the number of proteins per burst follows an exponential or geometric distribution.

This is the famous Geometric distribution. The geometric distribution is just the discrete analogue of the exponential distribution. To see this notice that we can write

$$P_{\text{burst}}(b) = \frac{e^{-\ln(1+1/\bar{b})b}}{1+\bar{b}} \approx \frac{e^{-b/\bar{b}}}{\bar{b}}, \tag{6}$$

where in the second line we have done a Taylor expansion in $1/\bar{b}$ which is valid when $\bar{b} \gg 1$.

Deriving the Gamma Distribution

We can use this to derive the Gamma distribution for protein abundances. To do so, we will write a protein-only master equation. We will call the number of proteins x.

$$\frac{dp(x,t)}{dt} = \alpha_m \int db P_{\text{burst}}(b) p(x-b,t) - \alpha_m p(x,t) + \tau_p (x+1) p(x,t) - \tau_p x p(x,t). \tag{7}$$

The first term on the right is the probability that you have $x-b$ proteins and produce a burst of size b (integrated/summed over all bursts sizes b), the second term is the probability that you have x proteins and produce a burst (of any size), the third term is the probability you have $x+1$ proteins and degrade a protein, and the final term is the probability you have x proteins and degrade a protein. When $x \gg 1$, we can approximate this Master Equation by a

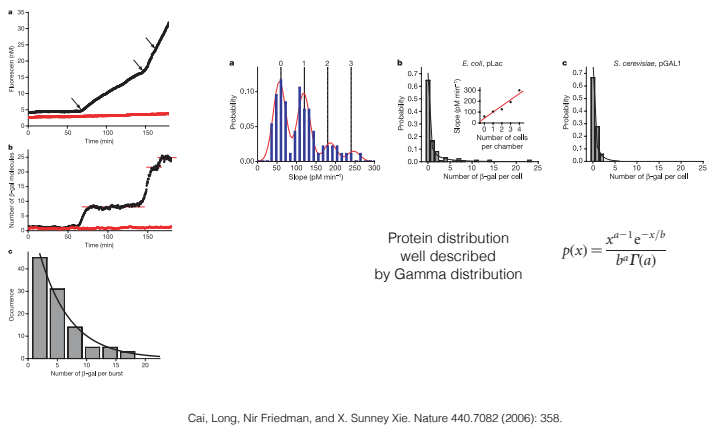


Figure 3: The protein distribution in single cells follows that Gamma distribution and was measured using a very clever experiment by Cai et al. See also this crazy paper from the same group Taniguchi et al Science 2010 where they do this for almost every type of mRNA molecule in *E .coli*.

Fokker-Planck equation of the form

$$\frac{dp(x,t)}{dt} = \alpha_m \int db P_{\text{burst}}(b) p(x-b,t) - \alpha_m p(x,t) + \partial_x \left(\tau_p^{-1} x p(x,t) \right) \tag{8}$$

At steady-state, this just becomes

$$a \int db [P_{\text{burst}}(b) - \delta(x)] p(x-b) = -\partial_x \left(x p(x,t) \right), \tag{9}$$

with $a = \alpha_m \tau_p$ the mean number of bursts per cell cycle. Notice the left hand side is just a convolution of two distributions, an exponential distribution and the distribution we want to solve for. A straight forward calculation shows that

$$\hat{P}_{\text{burst}}(s) = \frac{1/\bar{b}}{s + 1/\bar{b}} \tag{10}$$

Thus, we can take Laplace transform of both sides to get

$$-\frac{as}{s+1/\bar{b}} \hat{p}(s) = s \partial_s \hat{p}(s). \tag{11}$$

This yields

$$\hat{p}(s) = \frac{1}{(s+1/\bar{b})^a}. \tag{12}$$

This is just the Laplace transform of the Gamma distribution.