MATHEMATICAL BIOLOGY

Analytical distributions for stochastic gene expression

Milton Nogueira

Correspondence:
m.nogueira.da.silva.junior@umail.leidenuniv.nl, Master student, Leiden University,
Full list of author information is available at the end of the article

Abstract

In this essay, we will critically explore how stochastic gene expression has been used to understand gene regulation. More specifically, we will concisely give the description of an analytical framework with which one can provide an expression for the steady-state distribution of the number of proteins in a cell population. More importantly, we will be mostly concerned with the question to what extent this approach sheds light on gene regulation mechanisms.

Introduction

Throughout this essay, our analysis is performed in the context wherein the central dogma, or rather, the central hypothesis of molecular biology is the fundamental principle underlying the flow of sequence information among information-carrying biopolymers. More specifically, if one defines gene as a DNA protein-coding region then the mental mechanism for gene regulation comprises a finite set of regulatory proteins, i.e. the transcription factors, which bind specific sites of DNA in the surroundings of a gene of interest, giving rise to the concept of operator[1]. These transcription factors can either repress[2] the activity of the respective RNA polymerase (RNAP) or facilitate[3] its binding to a fixed DNA sequence, defined as the promoter, which, in this later case, will initiate the process of transcription of DNA into a mRNA.

Mainly driven by diffusion[4], this mRNA will be transported to the cytoplasm wherein it will be bound by a ribosome that will translate it into an amino acid sequence (polypeptide), which will fold into a three-dimensional structure, characterized by a proper function, which, in turn, is defined as a protein.

In this mental mechanism[5], the promoter can be thought as being in one of the states: active or inactive. A striking feature of gene expression, as evidenced

[1] Region of DNA to which the regulatory proteins are bound.

[2] In this case, we identify the proteins involved in the repression of RNA polymerase as the repressor.

[3] Here, another concept emerges: proteins involved in the facilitation of RNAP are thought to be the activator.

[4] Not necessarily true for prokaryotes, seeing that there is no membrane-bound nucleus so DNA is already floating loosely in the cytoplasm.

[5] It might be misleading to use mental mechanism in this context if we rely upon several papers in which one can find irrefutable evidences supporting the falsifiable status of the central dogma, but as the author of this essay is not able to argue to what extent the central dogma is "true" and if the question is relevant in some "complex organism", he chooses to assign the hypothetical status to it.

by several experiments reported in the literature, is that the number of proteins produced by a gene of interest varies in the cell population and over time within a single cell. What is the cause of this variability? To explain single-cell variability over time, we could limit ourselves to a hypothetical scenario, utterly unlikely to happen taking into account sound empirical evidences in the literature, in which a mRNA degrades at the same time as the translated proteins. Assuming that another translation occurs before cell division, so that one has the same cell, and that the process transcription-translation happens instantaneously with respect to the mRNA-protein life time in this hypothetical scenario then it is very unlikely that the same amount of proteins will be produced. If this was the case then we would have no variability in the suggested mRNA-protein life time. However, under randomness of biochemical reactions, it cannot be the case that mRNA molecules are bound by the same amount of ribosomes at each mRNA-protein life time. So far, this suffices as an explanation for single-cell variability in gene expression over time.

On the other hand, to argue cell-to-cell variability, we can draw on the reasoning provided in [1] and limit ourselves to a less likely scenario wherein a cell underwent division and produced two identical daughter cells carrying the same amount of copies of transcription factors with respect to a gene of interest. To promote translation, these transcription factors must perform a random walk (uncorrelated) toward target DNA-sites within each daughter cell causing these cells to have variable phenotypes as regards gene expression.

Hence, this source of noise generated by this cascade of reactions from translation to transcription is said to constitute the intrinsic noise. Moreover, RNA polymerase is a enzyme, a protein, a gene product so it also carries noise which is said to be extrinsic[6]. In this regard, the total noise has two components: intrinsic and extrinsic.

In this essay, we describe an analytical framework of stochastic gene expression in which only intrinsic fluctuations are incorporated. This incorporation is to some extent self-evident taking into account that the description of the central dogma of molecular biology establishes an explicit connection between rates (biochemical parameters) at which the gene-product (mRNA or protein) changes, and noise by means of random timing in the binding and unbinding of the regulatory molecules[7]. On the other hand, the inclusion of extrinsic noise is not straightforward seeing that this is implicit in the description of the central dogma when, for instance, we think about the degree of DNA supercoiling, chromatin organization and epigenetic activity by means of methylation. Yet, extrinsic fluctuations have an important effect on gene regulation[8] process so knowledge of the mechanisms underlying the latter processes would presumably shed light on extrinsic noise and contribute to a better understanding of gene regulation.

[6] Not only the respective RNAP carries extrinsic noise but all the cellular components which interact with the stochastic system comprising the regulation of a gene of interest, and that are not directly involved into the transcription and translation thereof.

[7] This random timing is mostly accounted for by the low copies of the molecules involved.

[8] Definitely in eukaryotes due to their complex cellular system.

However, this is a difficult task and these are subjects of intense research. Why is it difficult? To argue this, we turn ourselves toward the domain of life (Archaea, Bacteria, and Eukarya) with respect to cellular complexity. In fact, between eukaryotes and prokaryotes, there are lots of structural differences ranging from size to complex organization. For example, eukaryotic cells have multiple chromosomes in a compact form in the nucleus while prokaryotic cells have one circular chromosome in the cytoplasm with the absence of a nucleus. Hence, in the later, there is no need of mRNA-Brownian motion for translation to occur. With this simple example, augmented by the intrinsic complexity of the eukaryotic cellular machinery, we can definitely claim that the source of extrinsic noise in eukaryotic cells is inherently much bigger than in prokaryotic cells. Therefore, having described the respective conceptual framework, we will be very specific about the domain of our analysis and the conclusion that we can draw from it as regards the role of stochasticity in gene regulation.

This essay is organized as follows. Firstly, we briefly summarize the technique developed in [2], with which one gets single molecule dynamics in living cells with a subsequent description of the results. Secondly, we concisely go through the rationale of Shahrezaei-Swain‘s analytical framework and we refer to its insights and drawbacks concerning gene regulation augmented by a critical discussion of the simulations. Lastly, we conclude this essay by touching upon some questions risen during the course of this document and by providing a concise discussion hereof.

Probing gene expression in live cells

In 2006, Xiaoliang Sunney Xie and collaborators published two papers, one in Science [2] and one in Nature [3], in which they reported two powerful techniques with which one can actually get single molecule dynamics in living systems. In fact, their techniques allowed them to observe the production of single protein molecules one at a time [9], which, in turn, enabled them to give a quantitative description of gene expression in the cellular environment. In the following subsection, we will be giving a concise explanation of one of these techniques with much more emphasis on what can be concluded from the collected data concerning gene regulation within the respective domain.

E.coli strain $SX4$: the chimeric gene tsr-Venus

In [2], they describe a technique based on the detection of a fluorescent fusion protein (Venus)[10]. In fact, the authors constructed a E.coli strain $SX4$ in which they incorporated a chimeric gene, tsr-Venus, replacing the lacZ-gene, which, in fact, means that the production of tsr-Venus proteins will be observed whenever the regulatory network of lacZ-gene expresses it.

But, why did they choose for this particular fluorescent protein? Because Venus can be slowed down and has a fast maturation time[11]. Next, one has that most

[9] Regarding the domain, their observation was performed in live E.coli cells (Bacteria), i.e. prokaryotic cells.

[10] YFP: yellow fluorescent protein.

[11] YFP has a fast maturation time which is the time interval that it takes for the assembly process to occur: folding, attachment to the inner cellular membrane

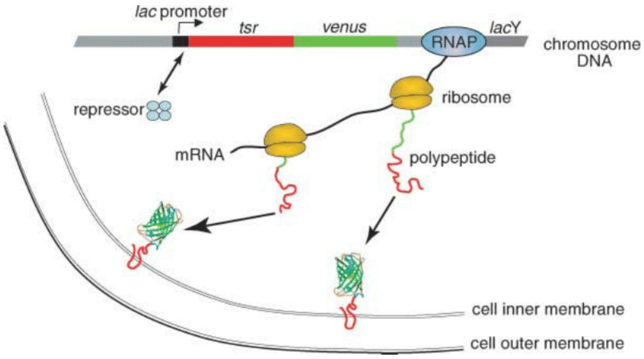

Figure 1 Cartoon showed in [2] regarding their experimental design. Upon unbinding of repressor, RNAP initiates transcription with subsequent binding of several ribossomes to mRNA, which folds into a polypeptide and undergoes localization to the inner cellular membrane and after maturation of fluorophore, it can be easily detected by a fluorescence microscope.

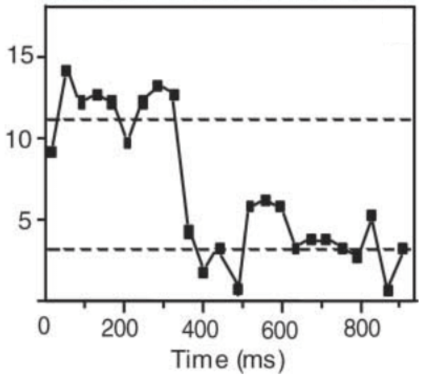

Figure 2 Plot taken from [2] showing the time-trace of photobleaching process of a single tsr-Venus molecule with fluorophore lifetime ≈ 300 ms.

of the fluorescent proteins, including Venus, diffuse fast in the cytoplasm what can have a negative effect on the measurements when using a fluorescence microscope. To argue this, we take into account the time interval during which they collect data, i.e the time interval of a measurement, which is equal to $\Delta t_{data} = 100$ ms, and the diffusion distance of a single YFP in the cytoplasm during this time interval, which is $\approx 1\mu m$. Hence, the diffusion distance of a single YFP coincides with the diameter of a E.coli, which, in turn, means that during data collection, a single YFP could be anywhere in the cytoplasm. However, Venus (YFP) can be slowed down by attaching to it a membrane localization sequence, tsr, which means that after folding, Venus will go toward the inner cellular membrane and will attach to it what facilitates detection by a fluorescence microscope as we see in figure 1.

and Venus fluorophore maturation. In fact, we will see later in this essay that this is on the time scale of minutes what is compatible with reported time scales for translation and transcription in E.coli cells, showing its suitability to probe the dynamics of gene expression and regulation in the respective domain.

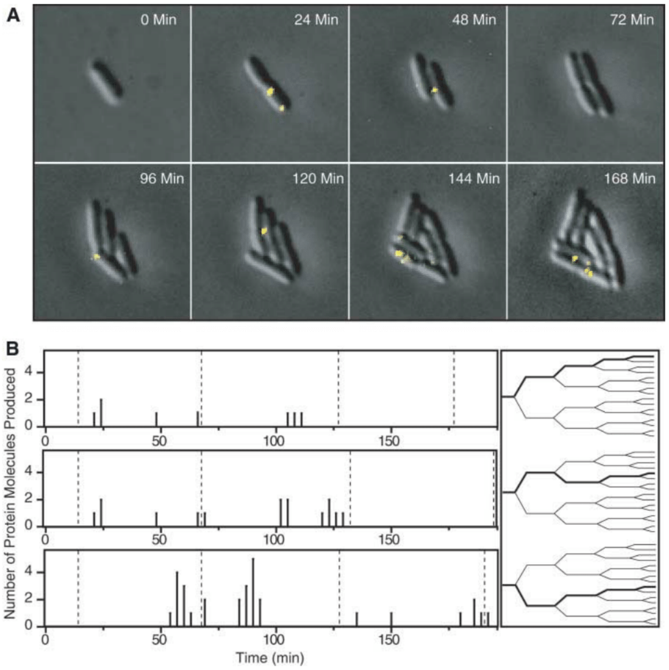

Figure 3 Data collected from [2]. Here, the dashed lines represent cell division phenomenon along three cell lineages. As we see in the data, the number of bursts varies per cell cycle and the number of proteins varies per burst as well. Moreover, one can also see that there is a slightly increase of the burst frequency per cell cycle.

What can we extract from the design strategy of the experiment tell us as regards its own "efficiency"? As we seen in figure 2, in which they plot the intensity of the fluorescent signal of a single protein tsr-Venus molecule along time, it takes around 300 ms for the fluorescent signal to disappear. Every three minutes [12] they used the fluorescence microscope on the cells during a time interval of 1200 ms, which means that data was collected during $\Delta t_{data} = 100$ ms and after that they kept shining light on the cell during 1100 ms so as to kill the fluorescent signal. In fact, they actually calculated the fraction of proteins that could survive this photobleaching process what is $\approx 1\%$. Therefore, their technique guarantees that they were actually detecting the expression of new proteins at each time.

What did they conclude qualitatively and quantitatively from the experiments? Throughout this argument, we refer to and rely on the data shown in figure 3. In fact, they firstly observed that proteins are produced in *bursts* randomly in time, and each burst event arises, in average, from a single tsr-Venus mRNA molecule. Secondly, the number of bursts also varies per cell cycle. However, What is a *burst*? A burst is a event through which a limited amount of proteins is produced. By applying statistical methods to the collected data, they arrived at $a = 1.2 \pm 0.3$ for the average number of bursts per cell cycle[13] with $\tau_{cell} = 55 \pm 10$ min being

[12] This dwell time of 3 min was chosen to avoid damaging the cells.

[13] If we take into account that $\tau_{cell} = 55 \pm 10$ then $a = 1.2 \pm 0.3$ shows compatibility with the in vitro observed time interval, in E.coli cells, for lacZ repressor dissociation from lacZ operator $[20 \quad min, 50 \quad min]$.

the average cell division time. How did they conclude that a single burst event arises, in average, from a single transcription event, or rather, a single tsr-Venus mRNA molecule? In fact, relying on a dimensional-based argument, one arrives at $\frac{n_{mRNA}}{a\frac{1}{d_0}}\tau_{cell} = 1.4 \pm 0.42$ for the average number of tsr-Venus mRNA molecules per burst, where $n_{mRNA} = 0.037 \pm 0.013$ is the average number of tsr-Venus mRNA molecules at steady state[14], and $\frac{1}{d_0} = 1.5 \pm 0.2$ min is a convenient representation for the cellular lifetime of tsr-Venus mRNA molecule.

What is their argument for this one-to-one relation: a single tsr-Venus mRNA molecule per burst? They argue that the tsr-Venus repressor rapidly rebinds the exposed tsr-Venus operator[15], what shows compatibility with a likely scenario. In fact, under the assumption that an endogenous fluorescent fusion protein might not change the function of the native protein, if we take into account that a unicellular organism (E.coli) enters into different phases of the cell life cycle for which some proteins are more needed than others, then slow transitions of tsr-Venus promoter between active and inactive states could also have been observed in their data. Thirdly, they also observed that the number of tsr-Venus proteins varies per burst. An argument for that, relies on the variability of the amount of ribosomes bound to tsr-Venus mRNA molecule.

Next, their data also shows a temporal spread for individual bursts what they ascribed to the maturation time. In fact, they did estimate an average spread of 7.0 ± 2.5 min what can be seen as another justification for their choice of the fixed dwell time. Bearing in mind the lifetime of tsr-Venus mRNA molecule (1.5 ± 0.2 min), If we consider a scenario in which tsr-Venus mRNA molecule is being sequentially bound by ribosomes then this dwell time (3 min) suits to consecutively detect many tsr-Venus proteins produced by a single burst, what is compatible with the reported duration range of each burst ($3 - 15$ min).

Now, what is the distribution of the number of bursts per cell cycle then? To answer this question, they plotted the histogram seeing in figure 4(A), which shows that the data fits a *Poisson distribution*, which, in turn, suggests that a burst event, or rather, a tsr-Venus mRNA translation occurs independent of each other with a constant rate[16] $a = 1.2 \pm 0.3$.

What about the distribution of the number of proteins per burst? As shown in the histogram plotted in figure 4(B), they fitted an exponential distribution for the number of tsr-Venus proteins per burst with length scale $b = 4.2 \pm 0.5$ being the average number of proteins per burst. From now on, we refer to a and b as the *burst*

In fact, as they argue, for the lower bound dissociation time, we can think of a likely scenario wherein, due to small copy numbers of transcription factors, there can either be no RNAP to bind the respective exposed operator or there might occur some failure in the translation process itself.

[14] This number suggests then that, at steady state, the true-distribution of tsr-Venus mRNA molecules over a population of E.coli SX4 cells is presumably peaked around 0.

[15] Here, we bear in mind that there is no tsr-Venus operator, but lacZ operator instead. This is just a convenient semantic assignment for the author of this essay not to get confused when analysing the respective literature.

[16] It has been introduced before as the average number of bursts per cell cycle.

frequency and *burst size* respectively. Knowing that the geometric distribution is the discrete analogue of the exponential distribution reveals an intrinsic characteristic of the translation process, i.e the competition between ribonuclease and ribosome for tsr-Venus mRNA binding.

In fact, the data in figure 4(B) fits a geometric distribution with ribosome binding probability[17] of $\rho = 0.81 \pm 0.05$. Now, estimating the average number of proteins per cell in the cell population boils down to the calculation of the product $ab \approx 5.0 \pm 0.8$, what is close to the experimentally measured number of 4.1 ± 1.8 tsr-Venus protein molecules per cell in a population comprised of approximately 300 cells.

However, this intriguing number of 4.1 ± 1.8 tsr-Venus protein molecules per cell raises the question of whether they plotted the histogram of the number of tsr-Venus protein molecules per cell in the cell population so as to fit the data to some well-known distribution? In [2], they did not report any attempt to find the steady state distribution of the number of tsr-Venus proteins per cell[18].

Nonetheless, in [3], as shown in figure 5, when applying another technique to living E.coli cells, known as *microfluidic based assay* [19], under other conditions[20], they plotted the histogram of the number of β- *galactocidase* proteins per cell and they fitted a gamma distribution to the data with parameters $a = 0.16$ and $b = 7.8$, which are close to the estimated values $a = 0.11 \pm 0.03$ bursts per cell cycle and $b = 5 \pm 2$ β-*galactosidase* proteins per burst.

Returning to the former publication [2], is it possible that we could have guessed a gamma probability mass function as the steady state distribution for the number of tsr-Venus proteins per cell? In fact, as we have argued above, fitting a Poisson distribution for the number of bursts per cell cycle suggests that the bursts occur randomly in time and are independent of each other, or rather, transcription occurs randomly and independently.

Moreover, each burst event is geometric distributed with ribosome probability binding given by $\rho = \frac{b}{1+b}$. Therefore, if X_i and Y denote random variables representing the number of proteins per burst and the number of proteins per cell, with X_i independently and identically distributed, then performing an convolution-based reasoning yields

$$\left(\bigwedge_{i=0}^{a} X_i \sim Geom(\rho) \right) \wedge \left(Y \sim \sum_{i=0}^{a} X_i \right) \Rightarrow Y \sim NB(a, b), \tag{1}$$

what means that the number of tsr-Venus protein molecules per cell in the cell population is supposed to be negative binomial distributed with parameters a and b.

[17] This is the probability of a tsr-Venus mRNA molecule producing at least one tsr-Venus protein, which is calculated by using the expression $\rho = \frac{b}{1+b}$.

[18] To justify this, as the author of this essay has a very limited background in the decision-making process in the respective field then he can only speculate about the suitability of the data seeing that getting insight into the probability distribution of the number of tsr-Venus proteins per cell in the cell population was a natural step in their rationale.

[19] For this technique, one has β- *galactocidase* as a reporter protein.

[20] The burst frequency a and the burst size b depend on the conditions.

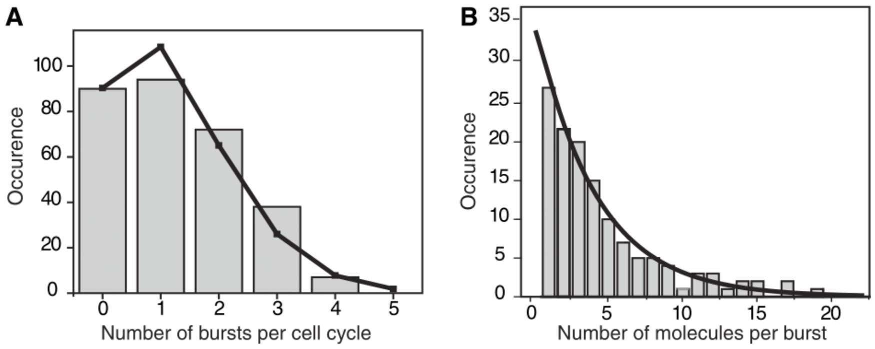

Figure 4 Histograms taken from [2]. A) Here we see that the histogram of the number of bursts per cell cycle fits a Poisson distribution with average $a \approx 1.2$. B)Here we see that the histogram of the number of proteins per burst event fits a exponential distribution with average $b \approx 4.2$.

But, why is it intuitive? Because a NB-distribution models an stochastic event in which one is interested in knowing the probability of a particular number of Bernoulli trials to have a fixed number of successes. Indeed, what is the probability of having a certain number of tsr-Venus proteins in the cell given that a and b are the estimated burst frequency and burst size. Moreover, as the gamma distribution is the continuous analogue of the NB-distribution then a continuous version of (1) reads

$$\left(\bigwedge_{i=0}^{a} X_i \sim Exp(b)\right) \wedge \left(Y \sim \sum_{i=0}^{a} X_i\right) \Rightarrow Y \sim \Gamma(a,b), \tag{2}$$

with a clearer intuitive explanation.

In fact, as the lifetime of a protein in E.coli cells, here conveniently denoted by $1/d_1$, is on the time scale of a cell cycle, i.e. $1/d_1 \approx 30$ min, then we add a bursts per life cycle with size b, which gives ab for the total number of proteins during the course of a cell division. However, in the cell population, we must add a-exponentially distributed bursts with length scale b, which amounts to a gamma distribution, as seen in figure 6, with mean ab and variance ab^2.

Moreover, for completeness, we can give an argument for the probability distribution of the number of tsr-Venus mRNA molecules per cell in the cell population. In fact, it should be Poisson distributed as we want to know how many translations have occurred during the trs-Venus mRNA lifetime. In this scenario, if we take into account that there is degradation then a Poisson distribution is definitely a strong candidate. Now, we ask ourselves if there is an analytical framework in which one can derive the expression of the probability distribution for the number of tsr-Venus protein molecules per cell in the cell population? What about the probability distribution for the number of tsr-Venus mRNA molecules per cell in the cell population? From now to the end of this essay, we will be entirely concerned with these questions.

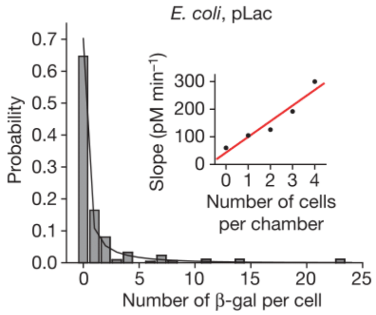

Figure 5 Histogram taken from [3] showing that the number of β-*galactosidase* per cell in the cell population fits a gamma distribution with $a = 0.16$ bursts per cell cycle and $b = 7.8$ proteins per burst.

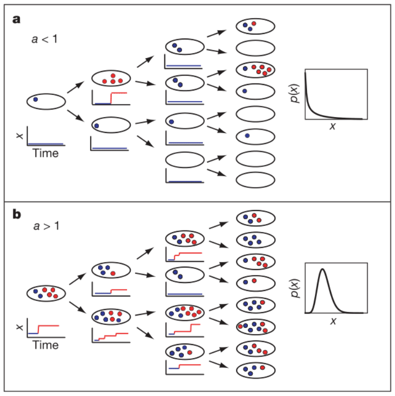

Figure 6 Can the distribution of the number of proteins per cell be peaked around zero? Yes, and here we see that it is determined by how many mRNA molecules are translated during the course of a cell division process (burst frequency a) regardless the burst size b. For $a < 1$, most of the cells will have no protein, or rather, the distribution will be peaked at 0. For $a > 1$, then it is very likely that more cells will contain at least a single copy of a protein so the distribution will be peaked around a number different from zero, which, of course, will depend on the burst size.

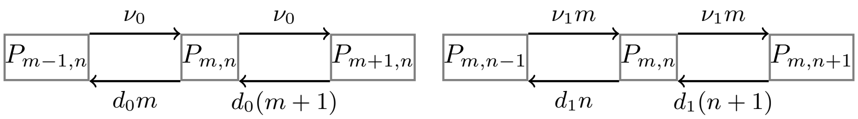

Figure 7 The Markov chain showing the transition probabilities regarding the two stage model.

Shahrezaei-Swain‘s approach: exploiting differences in time scale

In 2008, V. Shahrezaei and P. Swain, grounded in the results of single-cell experiments, including [2] and [3], applied the adiabatic reduction technique[21] to the stochastic counterpart of the birth-death model for gene expression which enabled them to arrive at an expression for the probability distribution of the number of protein molecules per cell in the cell population.

In the two stage model, as shown in figure 8(A), the promoter is assumed to be always in the active state. Drawing upon the inherent Markov process in this model, displayed in figure 7, leads us to the following birth-death master equation

$$\begin{aligned}\frac{\partial P_{m,n}}{\partial t} &= \nu_0(P_{m-1,n} - P_{m,n}) + \nu_1 m(P_{m,n-1} - P_{m,n})\\ &+ d_0[(m+1)P_{m+1,n} - mP_{m,n}]\\ &+ d_1[(n+1)P_{m,n+1} - nP_{m,n}],\end{aligned} \tag{3}$$

[21] The reduction of the dimension of a dynamical system based on differences in time scales.

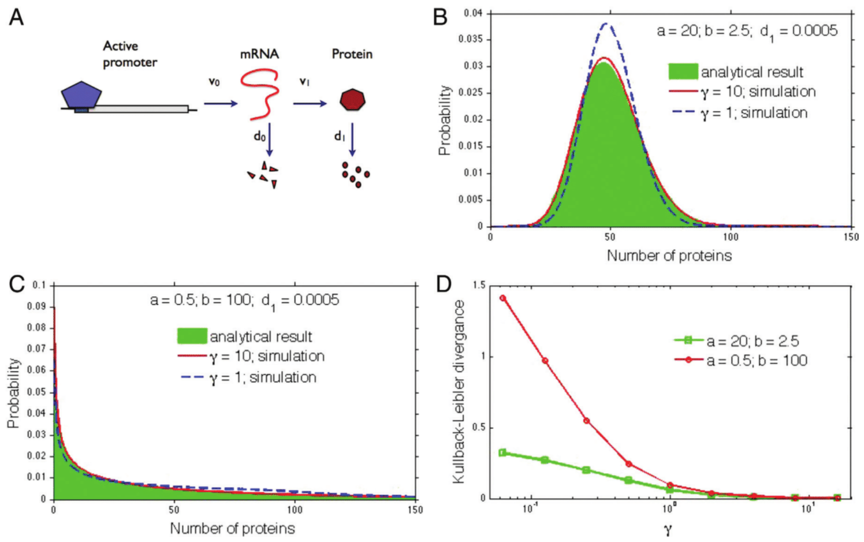

Figure 8 A) Two stage model with the promoter being always active. B) A simulation comparing the analytical solution of the two stage model and the numerical simulation of the master equation by using the Gillespie algorithm. Here, we see that the higher is γ, the better is the fitting. Moreover, consistent with what we discussed earlier in figure 6, for $a > 1$, one has that the distribution is peaked at a positive number. C) Here, for $\gamma = 10$, one sees a "perfect match". Moreover, as $a < 1$, with a high $b = 100$, ones sees that it is peaked at 0 what is consistent with the discussion in figure 6. D) Here, the KL divergence quantifies the effects of small γ. As we see, for γ around 10, one has a perfect match, whereas for $\gamma < 1$, one sees a high divergence. This high divergence is due to the fact that, in this case, proteins are being degraded while being produced during the lifetime of a mRNA, so the probability distribution describing the number of proteins per mRNA cannot be geometrically distributed which, in turn, implies that the probability distribution of the number of proteins in the cell cannot be the negative binomial what is reflected in this high divergence effects for lower γ.

describing the evolution of the probability distribution of having m mRNA molecules and n protein molecules at time t, with ν_0 and ν_1 being the probability per unit time of transcription and translation respectively, whereas d_0 and d_1 denote the probability per unit time of mRNA and protein degradation. Next, by introducing the generating function

$$F(z^{'}, z) = \sum_{m,n} z^{'m} z^{n} P_{m,n}, \qquad (4)$$

they reduce the infinity system (3) of stochastic dynamical equations to a single dimensionless partial differential equation

$$\frac{\partial F}{\partial \nu} - \gamma \left[b(1+u) - \frac{u}{\nu} \right] \frac{\partial F}{\partial u} + \frac{1}{\nu} \frac{\partial F}{\partial \tau} = a \frac{u}{\nu} F, \qquad (5)$$

with $a = \nu_0/d_1$, $b = \nu_1/d_0$, $\gamma = d_0/d_1$ (parameter carrying difference in time scale), $\tau = d_1 t$ (time in protein life time), $u = z^{'} - 1$ and $\nu = z - 1$ denoting the variable carrying the dynamics of mRNA and protein molecules respectively. Here, one must

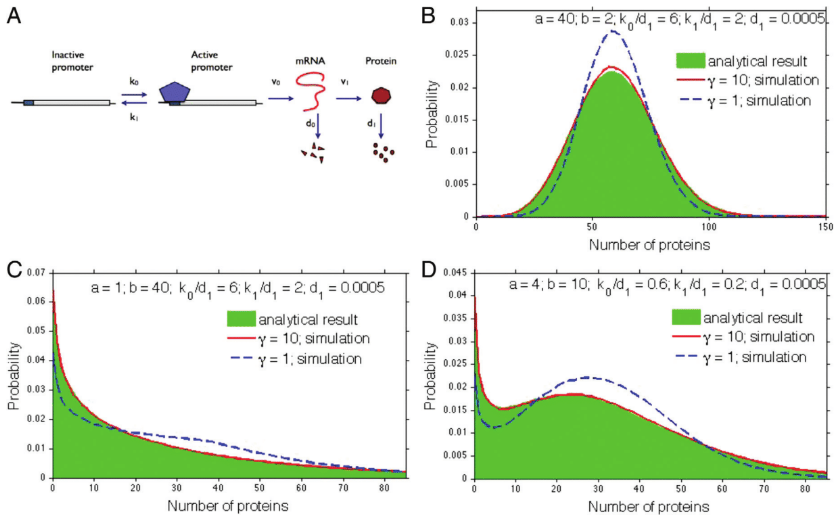

Figure 9 A)The three stage model including the transition of promoter between active and inactive states. B) and C) Similar results (unimodal behavior) for slightly high transition rates. D)Here, for small transition rates, one sees the bimodal behavior.

notice that the parameters of interest naturally emerges from the model with the intended meaning, i.e., a and b, or rather, the *burst frequency* and the *burst size*.

Moreover, under this reduction, $F(u, \nu)$ can be interpreted as the joint probability distribution indeed. Drawing on the method of characteristics, they reduced (5) to a system of ordinary differential equations, and under[22] $\gamma \gg 1$, or rather, that the protein lifetime $1/d_1$ is much longer than the mRNA lifetime $1/d_0$, they arrived at the solutions

$$\begin{aligned} u(\nu) &\backsimeq \frac{b\nu}{1-b\nu}, \\ \nu &= \nu_0 e^{\tau}, \end{aligned} \tag{6}$$

with $\tau > 0$. Hence, under $\gamma \gg 1$, we have a situation amenable to the quasi steady-state approximation.

In fact, for the most part of a protein lifetime, one has that the variable carrying the dynamics of mRNA molecules u is at steady state, or rather, the mRNA probability distribution is at steady state. Next, eliminating the fast variable u from (5) allowed them to arrive at the ODE

$$\frac{dF}{d\nu} \backsimeq \frac{ab}{1-b\nu}F, \tag{7}$$

[22] This assumption is consistent with empirical evidences. In fact, as we mentioned earlier in this essay, tsr-Venus mRNA lifetime was estimated to be ≈ 1.5 min while a protein lifetime in $E.coli$ is ≈ 30 min.

whose solution reads

$$F(z,\tau) = \left[\frac{1-b(z-1)e^{-\tau}}{1+b-bz}\right]^{a}, \tag{8}$$

which, in turn, can be conveniently expanded in z.

In fact, being under $\gamma \gg 1$ entails that[23] $P_{m,n} \backsimeq P_{0,n}$, so they capitalized on the definition (4) and compared this with the expansion of (8). Upon doing so, they arrived at the following expression

$$P_n(\tau) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^{n}\left(\frac{1+be^{-\tau}}{1+b}\right)^{a} \times 2^{F_1\left(-n;-a;1-a-n;\frac{1+b}{e^{\tau}+b}\right)}, \tag{9}$$

which describes the temporal evolution of the probability distribution of proteins per cell on the protein time scale, where

$$2^{F_1\left(-n;-a;1-a-n;\frac{1+b}{e^{\tau}+b}\right)} = \sum_{k=0}^{n}(-1)^k(-a)_k\frac{\Gamma(n+1)}{\Gamma(n-k+1)}\frac{(b)_k}{(1-a-n)_k}\frac{\left(\frac{1+b}{e^{\tau}+b}\right)^{k}}{k!}, \tag{10}$$

with $(a)_k = \Gamma(a+k)/\Gamma(a)$ and so forth. For $\tau \gg 1$, they arrived at the expression for the steady state distribution

$$P_n = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{b+1}\right)^{n}\left(1-\frac{b}{1+b}\right)^{a}, \tag{11}$$

which is the negative binomial distribution as expected in (1).

How to get the expression for the probability distribution of the number of mRNA molecules per cell in the cell population? In fact, initially, there is no protein, so one has a master equation describing the probability of having m mRNA-molecules at time t on mRNA time scale. To find the steady state distribution, which is Poisson, we draw on a well-known argument by applying the equilibrium condition to the flux of probabilities described in figure 7.

However, to derive the time dependent Poisson distribution for the number of mRNA in the cell on mRNA lifetime, one needs to work a bit harder on that. However, in [4], the authors provided an expression, which reads

$$P_m(t) = \frac{\lambda(t)^m}{m!}e^{-\lambda(t)}, \tag{12}$$

with

$$\lambda(t) = \lambda_{ss}(1-e^{-d_0 t}), \tag{13}$$

with $\lambda_{ss} = \nu_0/d_0$ being the steady state number of mRNA molecules in the cell. This expression is consistent with an argument provided earlier in this essay.

[23] This is consistent with the estimation of tsr-Venus mRNA molecules at steady state reported in [2], in fact, $n_{mRNA} = 0.037 \pm 0.013$.

Regarding the simulations, as we see in figure 8(B), they numerically implemented the Gillespie algorithm to the equation (3) and compared with the analytical solution in (11). If $\gamma \gg 1$ then the analytical solution accurately predicts the solution of (3). If $\gamma < 1$ then the divergence effect takes over, and the analytical solution provided in (11) is a poor approximation for the solution of (3).

In addition, the authors also considered a more realistic model, the three stage model, figure 9(A), where the promoter can be active and inactive. By applying the same technique, they could derive a expression for the probability distribution of the number of proteins in the cell. More importantly, they showed in the simulations, as seen in the figure 9(D), that a bimodality may emerge which is achieved by slow transitions between active and inactive states of the promoter.

Conclusion

What can this result tell us about gene regulation mechanisms? If we assume that the equation (3) is a reasonable representation for gene expression in prokaryotic organisms then their results states that if $\gamma \gg 1$, i.e., if a protein lifetime is much greater than a mRNA lifetime then the negative binomial distribution is an accurate approximation for the simulated distribution of (3). On the other hand, if $\gamma < 1$ then the analytic and simulated solution diverge significantly from each other with higher divergence effects for $\gamma << 1$. In fact, $\gamma \gg 1$ implies that all the proteins produced by a single mRNA, remain after mRNA degradation, what strongly suggests a geometric burst. On the other hand, if $\gamma < 1$ then some of the protein are degraded while others are being translated so it is not the case that a single mRNA leaves a geometric burst of proteins behind.

Therefore, under the representation assumption, their approach sheds light on the mechanism of gene regulation in prokaryotic organisms what is supported by experimental data as we have seen through the course of this essay. However, with respect to eukaryotic organisms, complexity in the regulation process turns it into a very challenging target. Hence, regarding eukaryotes, it is very unlikely that the gamma distribution gives a good approximation for the steady state distribution. We perhaps could say that, in this context, the major drawback of their approach is that it predicts that independent of the cell life phase, bursts per cell cycle will have the same average, which is unlikely to be true if we take cell growth into account. In fact, as Ji Yu et al [2] reported, we see in figure 3 that burst frequency depends on the cell cycle what they ascribe to DNA replication during cell growth.

References

1. Brian Munsky, Gregor Neuert, Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science* **336**, 183-187, 2012.
2. Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, X. Sunney Xie. Probing gene expression in live cells, one protein at a time. *Science* **311**, 1600-1603, 2006.
3. Long Cai, Nir Friedman, X. Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358-362, 2006.
4. Vahid Shahrezaei and Peter S. Swain. Analytical distribution for stochastic gene expression. *PNAS* **105**, n0.45, 2008.