Stockholms universitet

Statistical modelling and inference of single-cell gene expression profile

Lai Mei Yip Lundström

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2022:4
Matematisk statistik
Juni 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2022:4**
http://www.math.su.se

Statistical modelling and inference of single-cell gene expression profile

Lai Mei Yip Lundström*

June 2022

Abstract

Single-cell ribonucleic acids (scRNA) sequencing technologies have made it possible to measure genetic information at the cellular level, thereby facilitating the characterisation of a cell by its gene expression levels. This thesis sets out to model the messenger RNA (mRNA) transcriptional and degeneration process of a given gene in a cell by means of a bivariate Markov Chain as well as to derive its stationary distribution. The steady-state stationary distribution of the number of mRNA molecules that are synthesized by a given gene in a cell is approximated using perturbation techniques and the parameters are inferred using maximum likelihood. The stationary distributions of the different genes with the inferred parameters then form the gene expression profile of the cell. The result is that the negative binomial distribution is shown to be the exact solution to the simpler problem in the perturbative solution. Furthermore, it is shown that biologically relevant quantities such as a gene's mRNA transcriptional frequency and transcriptional size are related to the parameters of the negative binomial distribution. In addition, a comprehensive study of the probability currents in the Markov Chain has also found them to be closely connected to the mean of the distribution. The model is then applied to scRNA sequencing data and the results are presented and discussed.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: laimei.yip@gmail.com. Supervisor: Chun-Biu Li.

Acknowledgement

I would like to express my sincerest gratitude to my supervisor Chun-Biu Li and co-supervisor Tobias Wängberg for their guidance in making this thesis possible.

Contents

1 Introduction

The human species is one of the most complex living organisms that have evolved in the evolutionary process at large. The relentless quest to understand and decipher the human body is taken to new heights with rapid advancements in single-cell sequencing technologies [28], which are used to obtain single-cell genomic information. In particular, single-cell ribonucleic acids (scRNA) sequencing technologies have enabled the microscopic measurement of the number of messenger RNA (mRNA) molecules that are synthesised by a given gene in a cell. This process of synthesising mRNA molecules is called transcription and the number of such molecules detected in the cell is used to quantify the gene's expression level. Hence, the availability of scRNA sequencing data has enabled the study of the transcriptome of cells at the individual level, which in turn enables the construction of a cell profile based on the expression levels of the genes.

This thesis sets out to model the mRNA transcriptional and degeneration process of a given gene in a cell by means of a bivariate Markov Chain [1] as well as to derive its stationary distribution. This process is believed to be a series of random biochemical reactions [3], which often satisfy the Markov property. The steady-state stationary distribution of the Markov process is approximated using perturbation techniques [8] and the parameters are inferred using maximum likelihood. The essense of perturbation theory is to break a complex problem down into a simpler problem whose exact solution can be derived and other parts that are dependent on a small, real-valued parameter and dimensionless ϵ such that the behaviour of the complex problem can be observed as ϵ goes to zero. Intuitively, if ϵ is small enough, then the solution to the simpler problem serves as a good approximation to the complex problem. The reason for using perturbation techniques to approximate the stationary distribution is precisely because its exact distribution is a Poisson-beta mixture distribution, which cannot be solved analytically. Moreover, the ϵ parameter is easily identified after taking into account the physical properties of mRNA transcription such that it is possible to cast the problem into a perturbation framework . The result is that the negative binomial distribution is shown to be the exact solution to the simpler problem in the perturbative solution, which lends support to the common use of this distribution in biology literature [26] [27] and computational tools that model RNA sequencing data such as `DESeq` [10], `edgeR` [9] and `NBPSeq` [11]. Furthermore, it is shown that biologically relevant quantities such as a gene's mRNA transcriptional frequency and transcriptional size are related to the parameters of the negative binomial distribution. In addition, a comprehensive study of the probability currents in the Markov Chain has also found them to be closely connected to the mean of the distribution.

This thesis is organized as follows: Section 2 provides a detailed description of the model, including the derivation of the stationary distribution using generating functions as well as how it is approximated using perturbation techniques. The discussion on transcriptional frequency and size as well as probability currents are also in this section. Section 3 presents the results of maximum likelihood inference on a dataset of mRNA counts sequenced from pyramidal cells taken from the hippocampus CA1 of the mouse cerebral cortex. Section 4 contains a discussion of the results and possible extensions of this thesis.

2 Modelling mRNA transcription and degeneration in a cell

Gene expression is a complex process that involves two main sub-processes. First, genetic information that is contained in the DNA is decoded by an enzyme which then goes on to synthesize mRNA molecules. This process is called transcription. Second, the mRNA molecules are translated into protein molecules. Figure 1 shows a simple schematic picture of the gene expression process.

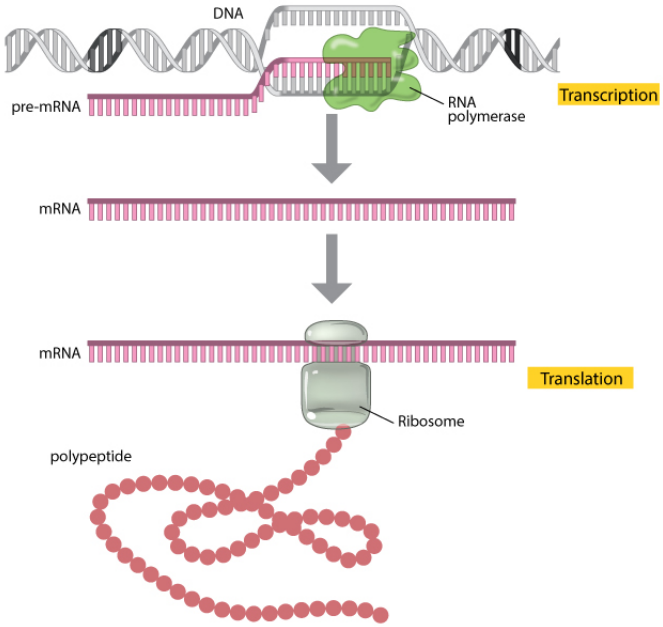

Figure 1: Gene expression: The information in DNA is transferred to a messenger RNA (mRNA) molecule by way of a process called transcription. During transcription, the DNA of a gene serves as a template for complementary base-pairing, and an enzyme called RNA polymerase II catalyzes the formation of a pre-mRNA molecule, which is then processed to form mature mRNA. The resulting mRNA is a single-stranded copy of the gene, which next must be translated into a protein molecule. *Source: Clancy, S. and Brown, W. (2008) Translation: DNA to mRNA to Protein. Nature Education 1(1):101*

Various mathematical models exist to model the gene expression process. A common element between many of these models, however, is the assumption of stochasticity in gene expression, which is also observed experimentally. Specifically, the stochasticity comes from the fact that gene expression is essentially a series of inherently random biochemical reactions whose measurable products are mRNA and protein molecules [4]. And since DNA does not exist in huge abundance in a cell, variability in the number of mRNA and protein molecules synthesised by the same genes can be used to differentiate seemingly similar cells. Furthermore, some researchers model only the mRNA synthesis process [3] [14] [5] while some model both the mRNA and protein synthesis processes [16] [15]. This thesis chooses to use a stochastic model to model the synthesis of mRNA molecules. For the sake of brevity, from now on, mRNA molecules are called mRNAs.

As mentioned in the introduction, scRNA sequencing technologies have enabled the collection of gene expression information at the single cell level, which makes it possible to model stochasticity in gene expression at the cellular level and apply the model that is described in the rest of this section. Briefly and simplified, biological samples are carefully collected and the viable cells are isolated into single cells. Next, these isolated individual cells are lysed to capture the mRNAs. Each molecule is uniquely identified, tagged to the cell of origin and subsequently sequenced. By DNA sequencing, it means to identify exactly the sequences of nucleotides in the mRNA and then match it against sequencing libraries to identify which gene it comes from.

The result is a gene expression matrix that contains the amount of genetic material expressed by every gene in individual cells. Figure 2 shows a schematic picture of scRNA sequencing data analysis pipeline.

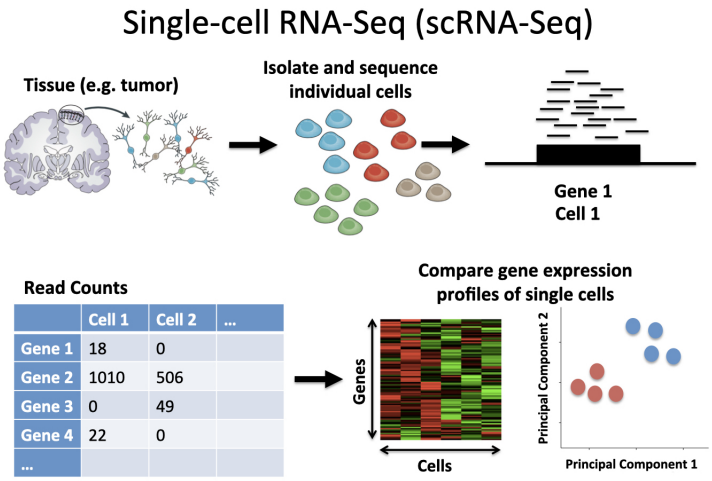

Figure 2: A schematic view of how scRNA sequencing data is collected and analysed. *Source: https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/*

This rest of this section covers details of the stochastic model. Section 2.1 explains the application of a bivariate Markov Chain to model mRNA synthesis and degeneration. Section 2.2 outlines the stationary distribution of the Markov Chain and shows how the stationary distribution is derived using generating functions. Section 2.3 details perturbation techniques being applied to approximate the stationary distribution. Section 2.4 shows the derivation of the probability distributions of the mRNA transcriptional frequency and transcriptional size as well as how their means are related to the parameters of the unperturbed problem in section 2.3. Finally, 2.5 discusses the probability currents of the stationary distribution.

2.1 A bivariate Markov Chain

Similar to the model of gene product synthesis in [1], this thesis models the time evolution of a gene's mRNA synthesis and degeneration as a bivariate continuous-time Markov Chain. In essence, a gene's mRNA synthesis and degeneration process is modelled as a birth and death process that is embedded in a two-state Markovian switching process. For a given gene in a single cell, it is assumed that it switches between two states, namely the *active* state and the *inactive* state. The time it takes to go from being active to inactive is exponentially distributed at a rate of λ. When the gene is in the inactive state, the time it takes to jump to the active state is exponentially distributed at a rate of γ.

When a gene is in the *active* state, it is assumed that it is capable of synthesizing mRNAs and it does so at a constant rate of μ. This means that the inter-arrival time of a mRNA is exponentially distributed with parameter μ. Furthermore, the lifetime of a mRNA is exponentially distributed at a rate of δ before it degenerates to something that is not measurable in the system. The lifetimes and inter-arrival times of all mRNAs are independent and identically distributed. Finally, when a gene is in the *inactive* state, it is assumed that it is not capable of synthesizing mRNAs and only the degeneration of mRNAs can take place.

Figure 3 shows a kinetic scheme that represents the bivariate Markovian Chain of mRNA synthesis and degeneration in cells. Define a bivariate continuous-time stochastic process $\{X(t), t \geq 0\}$. The state space of this process is

$$\Omega = \{(A, m), (I, m); m \in \mathbb{N}\},$$

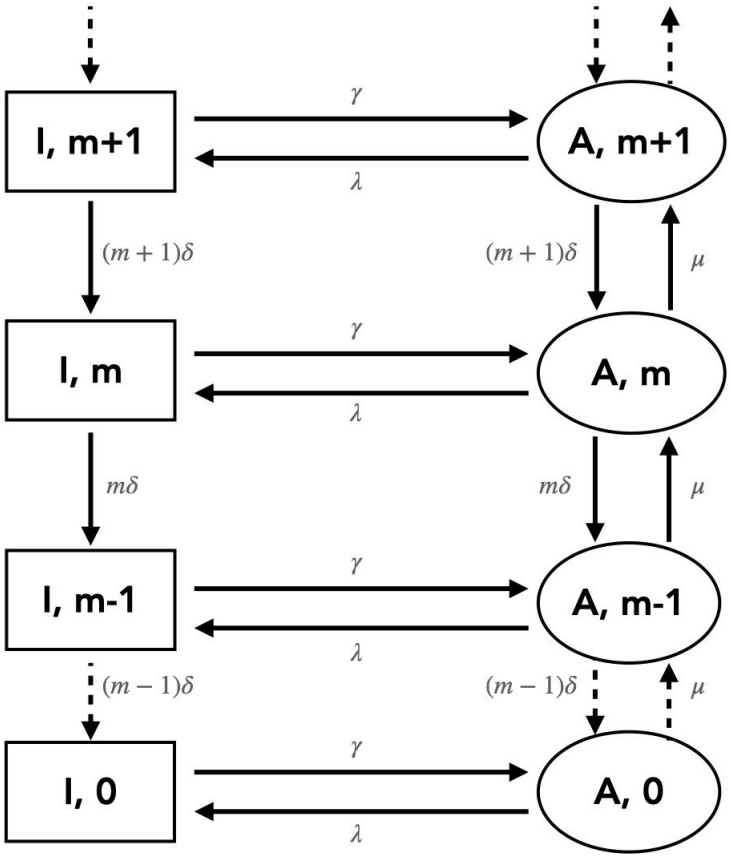

Figure 3: Kinetic scheme of a gene's mRNA synthesis and degeneration in a cell. m denotes the number of mRNAs in the system.

where A denotes active and I denotes inactive. The number of mRNAs in the system is denoted as m. The kinetic scheme shows that if the process is currently in state (A, m), then for all $s \geq 0$, $X(t+s)$ can jump to

- $(A, m+1)$ at a rate of μ;
- $(A, m-1)$ at a rate of $m\delta$, since there are m mRNAs in the system and all of them degenerate at the same rate δ;
- (I, m) at a rate of λ.

On the other hand, if the process is currently in the state (I, m), it can jump to

- $(I, m-1)$ at a rate of $m\delta$;
- (A, m) at a rate of γ.

Furthermore, the transition probabilities in this Markov Chain are time-homogeneous. This means that, for any pair of states i and j,

$$P(X(t+s) = j | X(t) = i),$$

is independent of t. In other words, the probability that the process jumps to state j at time $t+s$ is independent of how much time the system has been in state i.

2.2 Stationary distribution and probabilities

The asymptotic behaviour of this bivariate continuous-time Markov process is of main interest in this thesis, particularly the stationary distribution of the number of mRNAs in the system. Ycart and Peccoud [1] have shown that when the degeneration rate δ is strictly positive, the stationary distribution exists. In this thesis, it is assumed that δ is strictly positive.

By stationarity, it means that the continuous-time Markov Chain has transition probabilities that converge to limiting values that is independent of the initial state. These limiting probabilities, also commonly known as stationary probabilities, satisfy this condition: for a given state, the rate at which the system leaves it must equal to the rate at which the system enters it. There are two types of stationary states in Markov Chain: equilibrium and steady-state. For the equilibrium stationary state to hold, the incoming probability flow must equal the outgoing probability flow between all pairs of states in the Markov Chain, that is, the net probability flow is zero. In the case of steady-state stationarity, non-zero net probability flow can exist between pairs of states. The Markov Chain in this thesis can only have steady-state stationarity as the vertical probability flow on the left side of the kinetic scheme travels in only one direction, that is, for a pair of states (I, m) and $(I, m-1)$, there is only probability flowing from (I, m) into $(I, m-1)$ but not in the opposite direction.

Let M be a non-negative and integer-valued random variable that denotes the number of mRNAs in the cell. Define $P_I(m)$ and $P_A(m)$ as the stationary probabilities that there are m mRNAs in the cell when the gene is inactive, respectively active. This implies that the marginal stationary probability that there are m mRNAs in the cell is $P(m) = P_I(m) + P_A(m)$. If the stationary probabilities exist, then the rate at which the system leaves the state (I, m) (and (A, m)) must equal the rate at which it enters (I, m) (and (A, m)). In essence, they must satisfy the following conditions according to the transitions above:

$$\begin{aligned}(\gamma + m\delta)P_I(m) &= (m+1)\delta P_I(m+1) + \lambda P_A(m);\\ (\lambda + \mu + m\delta)P_A(m) &= (m+1)\delta P_A(m+1) + \mu P_A(m-1) + \gamma P_I(m).\end{aligned} \tag{1}$$

Furthermore, it can be shown that the long-run probabilities that the gene is inactive, respectively active, are $\frac{\lambda}{\lambda+\gamma}$ and $\frac{\gamma}{\lambda+\gamma}$. That is,

$$\begin{aligned}P_I &= \frac{\lambda}{\lambda+\gamma};\\ P_A &= \frac{\gamma}{\lambda+\gamma}.\end{aligned}$$

The proof can be found in the appendix.

The set of equations in (1) shows a linear recursive relationship in both $P_I(m)$ and $P_A(m)$. Generating functions are useful in finding a solution to this kind of general linear recurrences. Hence, define

$$\begin{aligned}g_I(z) &= \sum_{m=0}^{\infty} z^m P_I(m);\\ g_A(z) &= \sum_{m=0}^{\infty} z^m P_A(m)\end{aligned} \tag{2}$$

as the generating functions of $P_I(m)$ and $P_A(m)$, respectively. Then it follows that

$$\begin{aligned}g_I(1) &= \sum_{m=0}^{\infty} P_I(m) = P_I = \frac{\lambda}{\lambda+\gamma};\\ g_A(1) &= \sum_{m=0}^{\infty} P_A(m) = P_A = \frac{\gamma}{\lambda+\gamma}\end{aligned}$$

as stated above. Furthermore, since $P(m) = P_I(m) + P_A(m)$, then

$$g(z) = g_I(z) + g_A(z) = \sum_{m=0}^{\infty} z^m P(m)$$

is the probability generating function of $P(m)$ and

$$g(1) = \sum_{m=0}^{\infty} P(m) = 1.$$

Multiplying both sides of the equations in (1) with z^m and summing over all m, they can be written in terms of the generating functions in (2) as such:

$$\begin{aligned}
\gamma g_I(z) + \delta \sum_{m=0}^{\infty} m z^m P_I(m) &= \delta \sum_{m=0}^{\infty} (m+1) z^m P_I(m+1) + \lambda g_A(z); \\
(\lambda + \mu) g_A(z) + \delta \sum_{m=0}^{\infty} m z^m P_A(m) &= \delta \sum_{m=0}^{\infty} (m+1) z^m P_A(m+1) + \mu \sum_{m=0}^{\infty} z^m P_A(m-1) + \gamma g_I(z).
\end{aligned} \tag{3}$$

The following results are useful to help express (3) in terms of the generating functions:

$$\begin{aligned}
\sum_{m=0}^{\infty} m z^m P_I(m) &= z g_I'(z); \\
\sum_{m=0}^{\infty} (m+1) z^m P_I(m+1) &= g_I'(z); \\
\sum_{m=1}^{\infty} z^m P_A(m-1) &= z g_A(z).
\end{aligned} \tag{4}$$

Substituting (4) into (3) and after some simple algebra results in the following set of differential equations:

$$\begin{aligned}
\delta(z-1) g_I'(z) &= -\gamma g_I(z) + \lambda g_A(z); \\
\delta(z-1) g_A'(z) &= \gamma g_I(z) - \lambda g_A(z) + \mu(z-1) g_A(z).
\end{aligned} \tag{5}$$

Recalling that $g(z) = g_I(z) + g_A(z)$, it is instructive to see that adding the two equations in (5) together yields the following result:

$$\delta(z-1) g'(z) = \mu(z-1) g_A(z) \iff g'(z) = \frac{\mu}{\delta} g_A(z), \quad \text{for } z \neq 1.$$

This means that, in general,

$$g^{(m+1)}(z) = \frac{\mu}{\delta} g_A^{(m)}(z), \text{ for } m \geq 0 \text{ and } z \neq 1, \tag{6}$$

where the superscript denotes the mth derivative of the generating function. However, it is important to take note that the mean and variance of the distribution of M cannot be evaluated with this equivalence relationship in (6) since it is not valid for $z = 1$.

It is generally known [2] that setting $z = 0$ in the probability generating function yields

$$g^{(m)}(z = 0) = m! P(m).$$

Using this result in (6) above leads to the exact relationship

$$P(m+1) = \frac{\mu}{\delta} \frac{1}{m+1} P_A(m), \text{ for } m \geq 0. \tag{7}$$

From the set of differential equations in (5), the following higher order differential equations are derived by taking their mth derivatives.

$$\delta(z-1)g_I^{(m+1)}(z) = -(\gamma+\delta m)g_I^{(m)}(z) + \lambda g_A^{(m)}(z)$$
$$\delta(z-1)g_A^{(m+1)}(z) = \gamma g_I^{(m)}(z) + \big(\mu(z-1) - \lambda - \delta m\big)g_A^{(m)}(z) + \mu m g_A^{(m-1)}(z).$$

Since the stationary probabilities are of main interest, set $z=0$ and denote $g^{(m)}(z=0) = g^{(m)}$. Then the above set of equations can be simplified to

$$\begin{aligned} g_I^{(m+1)} &= (\frac{\gamma}{\delta} + n)g_I^{(m)} - \frac{\lambda}{\delta}g_A^{(m)} \\ g_A^{(m+1)} &= -\frac{\gamma}{\delta}g_I^{(m)} + \big(\frac{\mu+\lambda}{\delta} + m\big)g_A^{(m)} - \frac{\mu}{\delta}m g_A^{(m-1)}, \text{ for } m \geq 0. \end{aligned} \tag{8}$$

Using the identity $g(z) = g_A(z) + g_I(z)$ and (6) in the first equation of (8) yields

$$g^{(m+1)} - \frac{\delta}{\mu}g^{(m+2)} = (\frac{\gamma}{\delta} + m)(g^{(m)} - \frac{\delta}{\mu}g^{(m+1)}) - \frac{\lambda}{\mu}g^{(m+1)}$$
$$\iff$$
$$g^{(m+2)} = \big(\mu_1 + \gamma_1 + \lambda_1 + m\big)g^{(m+1)} - (\gamma_1 + m)\mu_1 g^{(m)}, \tag{9}$$

where $\mu_1 = \frac{\mu}{\delta}$, $\gamma_1 = \frac{\gamma}{\delta}$, and $\lambda_1 = \frac{\lambda}{\delta}$. This is a second order linear recursive equation with non-constant coefficients for the probability generating function of M, which is approximated using perturbation techniques in section 2.3. Take note that the unknown parameters μ_1, γ_1, λ_1 are dimensionless as they are quotients of two rates, which means the time dimension is cancelled out. Dividing by δ throughout the parameters is just a rescaling of time, thus having $\delta_1 = 1$, which does not change the underlying kinetic scheme.

An exact steady-state solution to the probability mass function of M has been derived by [3] which involves a confluent hypergeometric function of the first kind. Expressing the hypergeometric function in its integral form yields an integral expression that depicts a Poisson-beta mixture distribution (see appendix), where the mRNAs are poisson distributed with its mean being beta distributed. However, due to the presence of an integral, working with this distribution to infer the unknown parameters amounts to huge numerical difficulties. On the other hand, the problem can be simplified once the inherent properties of mRNA transcription are taken into account.

Consider now the case where both μ_1 and λ_1 are large but finite. These two parameters being large implies that the mean time in which an active gene synthesizes a mRNA as well as the mean time in which an active gene becomes inactive are much shorter than the mean time in which a mRNA degenerates. In other words, a gene goes into an active state and stays active for a relatively short period of time. But during this short active period, it synthesizes many mRNAs such that the number of mRNAs that degenerates during the active period is negligible relative to the number being synthesized. This is the so-called *burst-like* behaviour of mRNA transcription. Raj *et al.* [3] observed a burst-like behaviour in the synthesis of mRNAs during the short period of time when the gene was transcriptionally active, which corresponds to a large μ_1 and λ_1. Golding *et al.* [5] also reported burst-like transcription in their study of single-cell transcription in *Escherichia coli*. Furthermore, similar burst-like mRNA synthesis in eukaryotes was also reported by [6] and [7].

When these properties are taken into consideration, it becomes feasible to approximate the exact Poisson-beta solution by means of perturbation. The key is to identify a small, real-valued and dimensionless parameter ϵ so that the original problem can be expressed as a power series in terms of ϵ. The idea is that the original problem can be approximated by a truncated power series made up of an exact solution to a simpler problem plus some perturbative terms.

When ϵ is really small, then the perturbative terms are negligible and the exact solution to the simpler problem serves as a good approximation to the original problem. In the next section, perturbation techniques are used to approximate the exact Poisson-beta distribution.

2.3 Perturbation analysis

Perturbation techniques are used to find analytical approximations to the solution of (9). It is not the focus of this thesis to dwell deep into perturbation methods. In essence, the aim of using perturbation methods is to break a complex problem down into a simpler problem whose exact solution can be derived and other parts that are dependent on a small, real-valued parameter and dimensionless ϵ such that the behaviour of the complex problem can be observed as ϵ goes to zero. Intuitively, if ϵ is small enough, then the solution to the simpler problem serves as a good approximation to the complex problem. The interested reader is encouraged to consult [8], in particular Chapter 1, for a general understanding of perturbation methods. Hence, it is important to first identify the small, real-valued and dimensionless ϵ parameter in the problem. It is clear from (9) that the recursive equation depends on a large parameter μ_1. The problem can, therefore, be set in a perturbation framework by letting $\epsilon = \frac{1}{\mu_1}$. This works because μ_1 is assumed to be large and it is dimensionless, hence ϵ is small and dimensionless as well. In other words, the perturbation problem is

$$\begin{aligned} g^{(m+2)} &= \mu_1\big(1 + \frac{\gamma_1}{\mu_1} + \frac{\lambda_1}{\mu_1} + \frac{m}{\mu_1}\big)g^{(m+1)} - \mu_1(\gamma_1 + m)g^{(m)} \\ &\iff \\ \epsilon g^{(m+2)} &= \Big(\big(1 + \frac{\lambda_1}{\mu_1}\big) + \epsilon\big(\gamma_1 + m\big)\Big)g^{(m+1)} - (\gamma_1 + m)g^{(m)} \\ &\iff \\ \Big(\big(1 + \frac{\lambda_1}{\mu_1}\big)g^{(m+1)} - (\gamma_1 + m)g^{(m)}\Big) &- \epsilon\Big(g^{(m+2)} - (\gamma_1 + m)g^{(m+1)}\Big) = 0 \end{aligned} \tag{10}$$

Using conventional perturbation notation, the perturbation series of $g^{(m)}$ can be expressed as

$$g^{(m)} = g_0^{(m)} + \epsilon g_1^{(m)} + \epsilon^2 g_2^{(m)} + ..., \text{ for } m \geq 0,$$

where $\epsilon g_1^{(m)}$ is called the first order correction term and $\epsilon^2 g_2^{(m)}$ the second order correction term and so on. Substituting the perturbation series into (10) and grouping the coefficients of like powers of ϵ together results in

$$\begin{aligned} &\Big(\big(1 + \frac{\lambda_1}{\mu_1}\big)g_0^{(m+1)} - (\gamma_1 + m)g_0^{(m)}\Big) + \\ &\quad + \epsilon\Big(- g_0^{(m+2)} + \big(\gamma_1 + m\big)g_0^{(m+1)} + \big(1 + \frac{\lambda_1}{\mu_1}\big)g_1^{(m+1)} - (\gamma_1 + m)g_1^{(m)}\Big) + ... = 0. \end{aligned}$$

Since this power series in ϵ is constantly equals to zero, then the coefficients of all powers of ϵ must be equal to zero too. It is easy to see this by drawing equivalence to the commonly known power series $\sum_{n\geq 0} a_n x^n$, where a_n's are the coefficients of the series. It is known that this power series is guaranteed to converge for $x = 0$. If the power series is constantly equals to zero, then a_0 must equal to zero. Furthermore, its first derivative must also equal zero, since differentiating zero is still zero. Its first derivative then form a new power series that is again guaranteed to converge for $x = 0$. This means that $a_1 = 0$. Following the same line of reasoning, this implies that $a_j = 0$ for all $j \geq 0$. Hence, the leading order term of the perturbation series is

$$g_0^{(m+1)}\Big(1 + \frac{\lambda_1}{\mu_1}\Big) = (\gamma_1 + m)g_0^{(m)} \iff g_0^{(m+1)} = \Big(\frac{\mu_1}{\mu_1 + \lambda_1}\Big)(\gamma_1 + m)g_0^{(m)}. \tag{11}$$

It is shown (in the appendix) that (11) depicts the recursive relation of a negative binomial distribution parameterized by r and p, where

$$r = \gamma_1, \quad p = \frac{\mu_1}{\mu_1 + \lambda_1}.$$

It is useful to take note that the assumptions in the kinetic scheme imply that $r > 0$ and $0 < p < 1$. More discussion about the connection between the distributional parameters and the kinetic parameters can be found in 2.4. Continuing the perturbation analysis, the first order term is

$$g_0^{(m+2)} = (\gamma_1 + m) g_0^{(m+1)} + \big(1 + \frac{\lambda_1}{\mu_1}\big) g_1^{(m+1)} - (\gamma_1 + m) g_1^{(m)}. \tag{12}$$

It is now instructive to express both (11) and (12) in terms of the perturbation series of the probability mass function of M since it is of main interest in this section. Dividing the perturbation series of $g^{(m)}$ by $m!$ gives

$$P(m) = P_0(m) + \epsilon P_1(m) + \epsilon^2 P_2(m) + ..., \quad \text{for } m \geq 0,$$

which denotes the perturbation series of the probability mass function of M. Furthermore, let

- $P_A(m) = P_{0,A}(m) + \epsilon P_{1,A}(m) + \epsilon^2 P_{2,A}(m) + ...$;

- $P_I(m) = P_{0,I}(m) + \epsilon P_{1,I}(m) + \epsilon^2 P_{2,I}(m) + ...$

denote the perturbation series of the probability mass function of having m mRNAs in the active and inactive states, respectively. The mathematical derivation in what follows is rather long and technical. However, the reader should be able to follow the main results and the rest of this section by referring to the notation of the various perturbation series. Hence, for the sake of brevity, the results are shown immediately and the interested reader can locate the detailed mathematical derivation in the appendix. In short, (11) becomes

$$P_0(m+1) = \Big(\frac{\mu_1}{\mu_1 + \lambda_1}\Big) \frac{(\gamma_1 + m)}{m+1} P_0(m) = p \frac{(r+m)}{m+1} P_0(m).$$

This is the recursive relation of the negative binomial distribution depicted in terms of its probability mass function using the same r and p parameterization (see appendix).

Equation (12), on the other hand, becomes

$$P_{1,I}(m+1) = p\Big(\frac{r+m}{m+1}\Big) P_{1,I}(m) + (m+2)(p-1) P_0(m+2). \tag{13}$$

This is a recursive equation in $P_{1,I}(m)$, given $P_0(m+2)$. It is shown (in the appendix) that a general expression for this recursive relation can be found, which helps to yield an expression for $P_{1,I}(m)$. The general expression is

$$P_{1,I}(m+1) = p^{m+1} \binom{r+m}{m+1} a + \frac{1}{2}(m+2)(m+1)(p-1)\Big(\frac{2r+m+2}{r+m+1}\Big) P_0(m+2),$$

where $a = P_{1,I}(0)$. Since $P_0(m)$ is the probability mass function of a negative binomial distribution, its form is known. This means that solving for a leads immediately to a solution for $P_{1,I}(m)$, for $m \geq 1$. Then from there, the first order correction term of $P_0(m)$, that is $P_1(m)$, can also be solved for.

Multiplying both sides of the recursive formula above by $\frac{p-1}{rp}$ and summing over all m results in

$$\sum_{m=0}^{\infty}\frac{p-1}{rp}P_{1,I}(m+1)=\sum_{m=0}^{\infty}\Bigg[\frac{p-1}{r}p^m\binom{r+m}{m+1}a+ \\ +\frac{1}{2rp}(m+2)(m+1)(p-1)^2\Big(\frac{2r+m+2}{r+m+1}\Big)P_0(m+2)\Bigg]$$

$$\Longleftrightarrow$$

$$\frac{p-1}{rp}\sum_{m=0}^{\infty}P_{1,I}(m+1)=\frac{a(p-1)}{rp(1-p)^r}\sum_{m=0}^{\infty}P_0(m+1)+(p-1)^2\sum_{m=0}^{\infty}(m+1)P_0(m+1)+ \\ +\frac{(p-1)^2}{2r}\sum_{m=0}^{\infty}(m+1)mP_0(m+1)+\frac{(p-1)^2}{r}\sum_{m=0}^{\infty}(m+1)P_0(m+1), \tag{14}$$

where the recursive relation of the negative binomial distribution is again utilised in the right-hand side of the equation. The expression above can be further simplified but for the sake of brevity, the simplification mathematics are located in the appendix. Upon substituting all the simplifications into (14) and after some standard algebraic manipulations, a solution for a is found to be

$$a=P_{1,I}(0)=(1-p)^{r-1}rp\Big(\frac{p(r+1)(2-p)-2}{2}\Big). \tag{15}$$

Inserting (15) into the recursive formula for $P_{1,I}(m+1)$ and a final formula for $P_{1,I}(m)$, for $m\geq 0$, is obtained as follows:

$$P_{1,I}(m)=P_0(m)\Bigg[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)\Bigg].$$

Then, for $m\geq 0$, the first order correction to $P_0(m)$ is obtained using the relationship

$$\begin{aligned}P_1(m)&=P_{1,I}(m)+P_{1,A}(m)\quad[\text{ see appendix for derivation of } P_{1,A}(m)\ ]\\&=P_{1,I}(m)+(m+1)P_0(m+1)\\&=P_0(m)\Bigg[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)+p(r+m)\Bigg].\end{aligned}$$

Finally, the perturbation series of the probability mass function of M can now be expressed as

$$\begin{aligned}P(m)&=P_0(m)+\epsilon P_1(m)+\mathcal{O}(\epsilon^2)\\&=P_0(m)\Bigg[1+\epsilon\Bigg(\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}- \\ &\qquad -mrp(1-p)+p(r+m)\Bigg)\Bigg]+\mathcal{O}(\epsilon^2).\end{aligned}$$

This implies that, for $m\geq 0$,

$$P(m)\approx P_0(m)\Bigg[1+\epsilon\Bigg(\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}- \\ -mrp(1-p)+p(r+m)\Bigg)\Bigg]. \tag{16}$$

Furthermore, approximations of $P_A(m)$, respectively $P_I(m)$, are derived as follows:

$$\begin{aligned}P_A(m) &= P_{0,A}(m) + \epsilon P_{1,A}(m) + \epsilon^2 P_{2,A}(m) + O(\epsilon^3)\\ &\approx P_0(m)\Bigg(\epsilon p(r+m) + \epsilon^2 p(r+m)\Bigg[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) -\\ &\qquad - p(1-p)\frac{(m+1)(m+2)}{2} - (m+1)rp(1-p) + p(r+m+1)\Bigg]\Bigg);\end{aligned}$$

$$\begin{aligned}P_I(m) &= P_{0,I}(m) + \epsilon P_{1,I}(m) + O(\epsilon^2)\\ &\approx P_0(m)\left[1 + \epsilon\left(\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{m(m+1)}{2} - mrp(1-p)\right)\right].\end{aligned}$$

The solutions above have been checked to fulfill the balance equations in (1) at both the zeroth and first order. For the sake of brevity in the thesis proper, the calculations of the check are shown in the appendix.

2.4 mRNA transcriptional frequency and size

It is mentioned in section 2.2 that a gene is assumed to exhibit burst-like mRNA synthesis behaviour during the relatively short period when it is active. Let us call the event when a gene exhibits such behaviour a *burst*. Then it follows that the onset of a burst is the same as the gene being activated.

Let T be the inter-arrival time of two consecutive bursts from a gene. Then let X be the time the gene stays active and Y be the time the gene stays inactive. It follows that T is the sum of X and Y. Recall in section 2.1 that both the time the gene stays active and the time it stays inactive are exponentially distributed with parameters λ and γ respectively. After rescaling as in section 2.2, these parameters become λ_1 and γ_1. Hence,

$$X \sim Exp(\lambda_1), \qquad Y \sim Exp(\gamma_1).$$

Then assuming independence between X and Y, which is reasonable because knowing how long a gene stays active does not give information about how long the same gene stays inactive, the probability density function of T is

$$\begin{aligned}f_T(t) &= \int_0^t f_X(x)\cdot f_Y(t-x)dx\\ &= \frac{\lambda_1\gamma_1}{\lambda_1-\gamma_1}\big(e^{-\gamma_1 t} - e^{-\lambda_1 t}\big).\end{aligned}$$

This is recognised as the probability density function of a hypoexponential distribution with parameters λ_1 and γ_1. In other words, the inter-arrival time of two consecutive bursts follows a hypoexponential distribution and its mean is

$$E[T] = \frac{1}{\lambda_1} + \frac{1}{\gamma_1},$$

which is the sum of the mean time a gene stays active and the mean time it stays inactive. Using the r and p parametrization of the negative binomial distribution, the mean can be expressed as

$$E[T] = \epsilon\frac{p}{1-p} + \frac{1}{r},$$

where the first summand is derived using $\epsilon = \frac{1}{\mu_1}$ and $\frac{\mu_1}{\lambda_1} = \frac{p}{1-p}$. From this reparametrization, it can be seen that if ϵ is small, then $E[T]$ is dominated by the mean time the gene stays inactive. Hence, the frequency of bursts is predominantly controlled by how many times the gene switches from inactive to active during one time unit and is approximately r.

Another quantity of interest is the number of new mRNAs that are synthesised whenever the gene is active. In section 2.1, it is mentioned that mRNAs are synthesised at a constant rate of μ. Hence, it is reasonable to regard the mRNA synthesis process as a Poisson process where the event of interest is the arrival of a mRNA and the inter-arrival time of two consecutive mRNAs is exponentially distributed with rate μ. Then the number of mRNAs synthesised during an active period t follows a Poisson distribution with parameter μt. Using the rescaled parametrization μ_1 and letting Z denote the number of new mRNAs synthesised during an active period t, then

$$Z|X = t \sim Poi(\mu_1 t), \qquad X \sim Exp(\lambda_1).$$

The marginal probability mass function of Z can be derived as follows

$$\begin{aligned}
P(Z = m) &= \int_0^\infty f_{Z|X=t}(m) \cdot f_X(t) dt \\
&= \int_0^\infty e^{\mu_1 t} \frac{(\mu_1 t)^m}{m!} \cdot \lambda_1 e^{\lambda_1 t} dt \\
&= \lambda_1 \mu_1^m \Big(\frac{1}{\mu_1 + \lambda_1}\Big)^{m+1} \int_0^\infty \frac{1}{\Gamma(m+1)} t^{m+1-1} \big(\mu_1 + \lambda_1\big)^{m+1} e^{-(\mu_1+\lambda_1)t} dt \\
&= \Big(\frac{\mu_1}{\mu_1 + \lambda_1}\Big)^m \Big(\frac{\lambda_1}{\mu_1 + \lambda_1}\Big) \\
&= p^m (1-p).
\end{aligned}$$

The expression above is immediately identified as the probability mass function of geometric distribution with parameter $(1 - p)$. Using this parametrization, the geometric distribution should be read as the number of new mRNAs a gene manages to synthesise until the first time it turns inactive before it manages to synthesise a new mRNA. Then the mean number of new mRNAs synthesised is

$$E[Z] = \frac{p}{1-p}.$$

In other words, the average size of each burst is $\frac{p}{1-p}$. Relating back to the kinetic parameters, this is the same as $\frac{\mu_1}{\lambda_1}$, which is the quotient of the mean time in which the gene stays active and the mean time it takes to synthesize a mRNA. This makes perfect sense.

Since both $E[T]$ and $E[Z]$ are dependent on the parameters of the negative binomial distribution and consequently, the kinetic parameters, it would be interesting to take a deeper look at what extreme values of p and r entails biologically. For a small value of p, this means that burst size is small. Given the assumption of a large μ_1 holds, then this also implies that λ_1 is much larger than μ_1, which in turn means that the gene goes into the inactive state more frequently relative to how fast it synthesises mRNAs. In other words, it stays in the active state for only very brief periods of time and each time only manages to synthesise small quantities of mRNAs. Even without the involvement of ϵ, the expression for $E[T]$ says that the mean time between bursts is just the mean time the gene is inactive. And if γ_1 is small, then $E[T]$ becomes large, which in turn means that the gene is generally very quiet. Putting the pieces together, this behaviour constitutes a gene that switches rarely to the active state and at the same time synthesises few mRNAs during the brief moments it is active. While this may occur biologically, it is questionable if any data about this gene can be captured given finite sampling. On the other hand, if γ_1 is large as well, then $E[T]$ becomes really small, which essentially means

bursts happen back to back. A gene as such rapidly switches back and forth between active and inactive but only manages to synthesise few mRNAs each time it is active. It is plausible that finite sampling is still able to capture such behaviour. For a large value of p, burst size is large and the gene synthesise mRNAs faster than it goes into the inactive state. So even if it still stays in the active state for brief periods of time, it manages to synthesise large quantities of mRNAs every time it is active. If this is coupled with a large γ_1, then this is a gene that is very active and highly productive that will definitely be captured in finite sampling. In the case if γ_1 is small, this would mean a gene that is generally quiet but very productive whenever it is active. This scenario fits better with the description of a burst-like mRNA synthesis behaviour and also agrees well with empirical observations in biological experiments.

At this point, it is evident that meaningful biological quantities related to the mRNA synthesis and degeneration process of a gene can be expressed in terms of the parameters of the negative binomial distribution from the zeroth order. In the next section, an extensive investigation of the probability currents in the Markov Chain is conducted and reveals even more how the negative binomial distribution is able to provide meaningful statistical and biological insights.

2.5 Probability current

It has been mentioned in section 2.2 that the stationary distribution of the Markov Chain in Figure 3 is of main interest in this thesis. In order to have a deeper understanding of the characteristics of the stationary distribution, it is instructive to study and analyse the probability currents between subsets of states in the Markov Chain.

Consider a continuous-time Markov Chain with state space Ω that has transition rate from state i to j as q_{ij} and from state j to i as q_{ji}. Plus, this Markov Chain has a stationary distribution $\boldsymbol{\pi}$, that is, π_i is the long-run probability that the chain is in state i. Then the net probability current between a pair of states i and j is defined as

$$J_{i\to j} = \pi_i q_{ij} - \pi_j q_{ji}.$$

If $J_{i\to j}$ equals to zero for all i and j, then the Markov Chain is said to be in an *equilibrium* state. Otherwise, if there exists non-zero $J_{i\to j}$ between some pairs of i and j, then the chain is said to be in *steady* state, which is the case for the bivariate Markov Chain in this thesis.

In this thesis, the stationary distribution is approximated using perturbation analysis up to the first order in ϵ. It would have been ideal if the investigation of the probability currents includes the first order terms. This would, however, involve tedious mathematical computations that may not provide any useful interpretation. Hence, it is decided that this thesis investigates probability currents involving only zeroth order terms.

2.5.1 Horizontal flow

From the kinetic scheme in Figure 3, it follows that the horizontal probability currents occur between the active state and the inactive state of the gene at different values of m. In other words, this net probability current is a function of m.

Using conventional probability current notation and the rescaled kinetic parameters, let $J_{I\to A}(m)$ denote the net probability current from the inactive state to the active state, that is,

$$
\begin{aligned}
J_{I\to A}(m) &= \gamma_1 P_I(m) - \lambda_1 P_A(m) \\
&\approx \gamma_1 P_0(m) - \lambda_1 \epsilon p(r+m) P_0(m) \qquad [\text{Recall } \epsilon = \frac{1}{\mu_1}] \\
&= r P_0(m) - (1-p)(r+m) P_0(m) \\
&= P_0(m)\Big[r - (1-p)(r+m)\Big] \\
&= P_0(m)\Big[rp - m(1-p)\Big].
\end{aligned}
$$

It is not difficult to see that for small m, $J_{I\to A}(m)$ is positive. When this happens, it implies that there is net probability current flowing into the active state. Conversely, when m is large, $J_{I\to A}(m)$ turns negative, implying a net probability current flowing out of the active state. In other words, the number of mRNAs present in the cell acts as a regulatory mechanism that induces the gene to switch between the active and inactive states. Intuitively, when the cell is "low" on mRNAs, the gene wakes up more often to mRNA production in order to "top up" the number of mRNAs until it goes back to sleep again. Conversely, when the cell is "high" on mRNAs, the gene goes to sleep more often to facilitate the degeneration process.

More precisely, when m is smaller than $\frac{rp}{1-p}$, the factor $\big[rp - m(1-p)\big]$ is always positive. This, in turn, means that the net current's behaviour on this range of m is induced by that of the negative binomial probability mass function $P_0(m)$ up to its mean, which is $\frac{rp}{1-p}$. This further implies that the net current will typically increase first, peak at some point, then decreases to the zero mark at m equals to $\frac{rp}{1-p}$. In other words, the probability current flow from the inactive state to the active state is the same as that from the active state to the inactive state when the number of mRNAs is equals to the mean of the negative binomial distribution. Intuitively, at this m, the gene is at its most "comfortable" state since the system is in balance. However, $J_{I\to A}(m)$ can go into the negative region when m is larger than $\frac{rp}{1-p}$. But it does not go into negativity indefinitely because $P_0(m)$ goes quickly to zero as m becomes large. This implies that the net current decreases to a certain point before it starts to increase again until it eventually plateaus out to zero. It is important to take note that plateauing out to zero does not mean that it reaches long-term equilibrium. This is because since $P_0(m)$ exists in both $P_I(m)$ and $P_A(m)$, then $J_{I\to A}(m)$ essentially becomes zero minus zero as m becomes large. In essence, as m becomes large, the probability currents in both directions become increasingly weak to the extent they plateau out to zero. See Figure 4 for an intuitive understanding of how $J_{I\to A}(m)$ behaves for various values of $r > 1$. It is evident that for a moderate value of p, a larger r has the effect of pushing out the "wave" along the x-axis. This is because $P_0(m)$ is small for relatively small values of m, which is characteristic of a negative binomial distribution with a large r and moderate p. Statistically, if r is the predefined number of failures (see appendix for definition of negative binomial distribution) and is large, then the probability of getting only a small number successes with a moderate p should be small. So for a gene with a large r and moderate p, the probability of observing a small number of mRNAs is small. Hence, the beginning of the net current is also a zero minus zero case. From a burst frequency point of view, a large r means that the gene is quite active. Coupled with a moderate burst size, it is likely that it is seldom very low on mRNAs, which agrees well with the behaviour of its $P_0(m)$. In other words, this means that the gene has a higher threshold of m before it considers itself low on mRNAs and needs to be more active. This appropriately explains why the net current starts to ascend only at larger values of m. For a moderate value of r, a larger p also seems to have the effect of pushing out the wave along the x-axis. However, upon taking a closer look at the y-axes of the second column of Figure 4, it is evident that the range of the net current shrink tremendously as p gets larger. This means that the wave straightens out almost to a flat line as p becomes increasingly close to one. Again, this is characteristic of a negative binomial

distribution with a moderate r and a large p. Statistically, if the probability of success is so large, then the probability of ever observing even a moderate number of failures is very small. From a burst size point of view, a large p implies a large burst size, which means that this gene is very productive every time it is active. Hence, there is no necessity for a strong net current into the active state to top up the number of mRNAs.

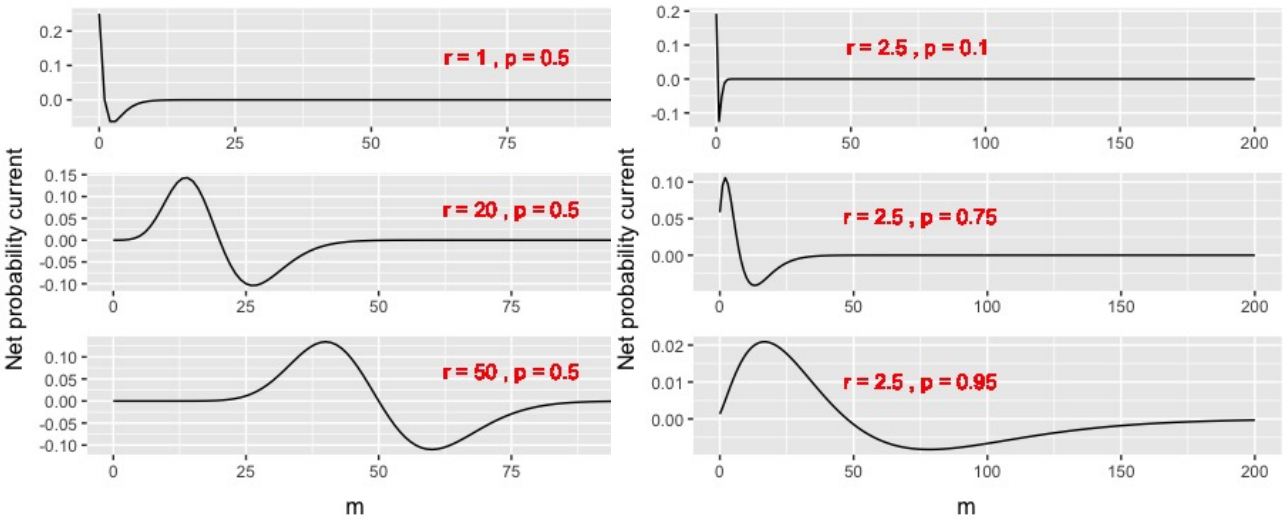

Figure 4: $J_{I\to A}(m)$ for different values of $r > 1$ and p.

For $0 < r \leq 1$, the maximum net current already occurs at m equals to zero for all values of p since the maximum value of the probability mass function of $P_0(m)$ occurs at this value of m. It then decreases until it hits the zero mark when m equals to $\frac{rp}{1-p}$. Similar to the $r > 1$ case above, it then goes into the negative region until a certain point before it starts to increase again, where it also eventually plateaus out to zero. Figure 5 shows how the current behaves for various combinations of $0 < r \leq 1$ and p. Values of r that are between zero and one corresponds to a quiet gene that has a long mean time between two consecutive bursts. Juxtaposing the top-right plot in Figure 4 and the top plot in Figure 5, it can be seen that they are very similar except for the magnitude of the y-axis. For a gene that has a larger burst frequency, that is $r = 2.5$, there is higher net probability current flow from the inactive to the active state as compared to the gene with $r = 0.1$.

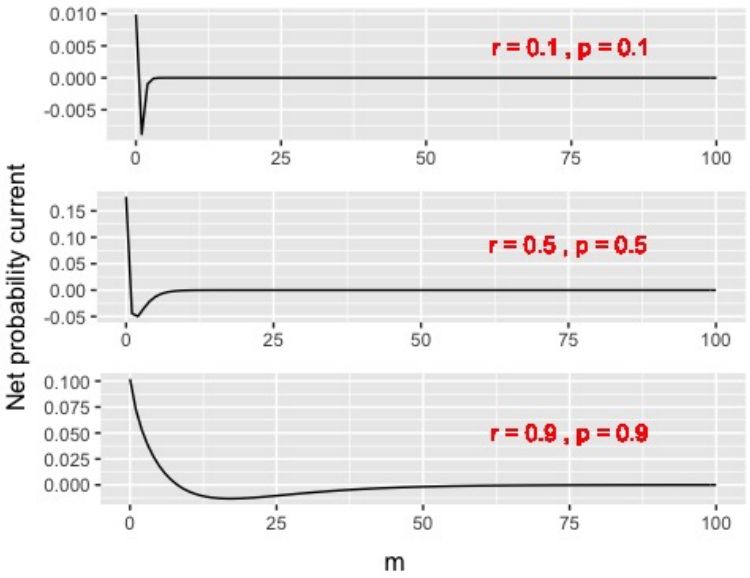

Figure 5: $J_{I\to A}(m)$ for different values of $0 < r \leq 1$ and p

The straightforward way to find the m values where the current peaks and troughs is to find the stationary points of $J_{I\to A}(m)$. However, since m is discrete, it does not make mathematical sense to derive the first derivative of $J_{I\to A}(m)$ and then solve for m. One way to locate these points is to construct ratios of two consecutive currents in the upward as well as downward directions, that is,

$$\frac{J_{I\to A}(m)}{J_{I\to A}(m+1)} \quad ; \quad \frac{J_{I\to A}(m)}{J_{I\to A}(m-1)}.$$

Let m_1^* denote the value of m where the current is at its maximum. Then it must hold that

$$\frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*+1)} > 1 \quad ; \quad \frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*-1)} > 1.$$

The mathematical calculations to solve these two inequalities are lengthy and technical, and hence are put into the appendix. In brief, the m_1^* that gives rise to maximum $J_{I\to A}(m)$ fulfills the following compound inequality

$$\frac{2rp-(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} < m_1^* < \frac{2rp+(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}, \tag{17}$$

for $r > \frac{2-p}{p}$ and $0 < p < 1$.

It can be seen from (17) that the difference between the bounds of the strict inequality is exactly one, which means that m_1^* can be uniquely determined since it is clamped between two real numbers. In other words, m_1^* can take either the ceiling of the lower bound or the floor of the upper bound. The case where the bounds are integers is considered in the appendix. It is now instructive to express (17) in terms of the mean and variance of the negative binomial distribution at zeroth order to facilitate interpretation of what this inequality entails. Let $E[M_0]$ and $Var(M_0)$ denote these two values respectively. Then (17) can be rewritten as

$$E[M_0] - \frac{1}{2} - \sqrt{Var(M_0)+\frac{1}{4}} < m_1^* < E[M_0] + \frac{1}{2} - \sqrt{Var(M_0)+\frac{1}{4}}. \tag{18}$$

It follows from (18) that m_1^* is approximately $E[M_0] - \sqrt{Var(M_0)}$ rounded to the nearest integer. In other words, the value of m which gives rise to maximum $J_{I\to A}(m)$ is approximately

equals to the mean of the negative binomial distribution at the zeroth order minus one standard deviation. Intuitively, this implies that when the number of mRNAs in the cell is one standard deviation away from $\frac{rp}{1-p}$ to the left, the inactive process starts kicking in at a faster rate than the active process to progressively slow down the mRNA production process in order for the number of mRNAs to reach $\frac{rp}{1-p}$. This is where the net current is zero.

It is instructive to take note that the case of $r > \frac{2-p}{p}$ essentially covers the scenarios where $r > 1$ for all values of p. For the case of $0 < r \leq 1$, it has been mentioned earlier that the maximum of $J_{I\to A}(m)$ occurs at m equals to zero.

Let now m_2^* denote the value of m where the current is at its minimum. Take note that m_2^* must be larger than $\frac{rp}{1-p}$ since the current has passed the zero mark. Constructing the ratios of two consecutive currents to find m_2^* requires careful consideration because now the currents are in the negative region. This means that

$$\frac{-J_{I\to A}(m_2^*)}{-J_{I\to A}(m_2^*+1)} > 1 \quad ; \quad \frac{-J_{I\to A}(m_2^*)}{-J_{I\to A}(m_2^*-1)} > 1,$$

which is the same as

$$\frac{J_{I\to A}(m_2^*)}{J_{I\to A}(m_2^*+1)} < 1 \quad ; \quad \frac{J_{I\to A}(m_2^*)}{J_{I\to A}(m_2^*-1)} < 1.$$

Similar to the mathematical calculations that resulted in (17), the ones to solve for these two inequalities are put into the appendix for the sake of brevity. The result is that the m_2^* that gives rise to minimum $J_{I\to A}(m)$ fulfills the following compound inequality

$$\frac{2rp-(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} < m_2^* < \frac{2rp+(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}. \tag{19}$$

Standard calculations show that $r > 0$ is the only constraint in order for (19) to hold as the left-hand side of (19) fulfills the general criteria of of $m \geq 0$ given any value of $0 < p < 1$. Hence, this constraint can be encompassed into the constraints for (17) and (19) still holds. Similar to (17), m_2^* is clamped between two real numbers and hence can be uniquely determined by either taking the ceiling of the lower bound or the floor of the upper bound. The case of integer bounds is considered in the appendix. Furthermore, (19) can be rewritten in terms of $E[M_0]$ and $Var(M_0)$ as such

$$E[M_0] - \frac{1}{2} + \sqrt{Var(M_0)+\frac{1}{4}} < m_2^* < E[M_0] + \frac{1}{2} + \sqrt{Var(M_0)+\frac{1}{4}}.$$

This gives the same interpretation as above that the value of m which gives rise to minimum $J_{I\to A}(m)$ is approximately equals to the mean of the negative binomial distribution at the zeroth order plus one standard deviation. In contrast to the maximum of $J_{I\to A}(m)$ above, when the number of mRNAs in the cell is one standard deviation away from $\frac{rp}{1-p}$ to the right, both the active and inactive processes start to weaken until $J_{I\to A}(m)$ gradually plateaus out to zero.

2.5.2 Vertical flow I

Let $J_{(A,m)\to(A,m+1)}(m)$ denote the net probability current when the gene is in the active state, then

$$\begin{aligned} J_{(A,m)\to(A,m+1)}(m) &= \mu_1 P_A(m) - (m+1)P_A(m+1) \\ &\approx p(r+m)P_0(m) \end{aligned}$$

Since only the zeroth order terms are of interest here, the second term in the first equality is omitted due to it being of order ϵ. It can be seen that $J_{(A,m)\to(A,m+1)}(m)$ is always positive at

the zeroth order. Intuitively, this implies that when the gene is active, there is consistently more mRNA production than degeneration independent of the value of m. However, the magnitude of the current changes as m increases as it follows the shape of the probability mass function $P_0(m)$. This means that there is one m where the current peaks, after which it progressively plateaus out to zero. Similar to the interpretation above, plateauing out to zero does not mean that $J_{(A,m)\to(A,m+1)}(m)$ reaches long-term equilibrium but rather the net current dies out as $P_0(m)$ approaches zero. See Figure 6 and Figure 7 for an intuitive understanding of how $J_{(A,m)\to(A,m+1)}(m)$ behaves for various values of r and p. Similar to Figure 4, for a fixed value of p, a larger r has the effect of pushing out the "wave" as well as increasing the magnitude of the peak. This agrees well with the functional form of $J_{(A,m)\to(A,m+1)}(m)$ since the mode occurs at the approximately the mean of $P_0(m)$ when r is large.

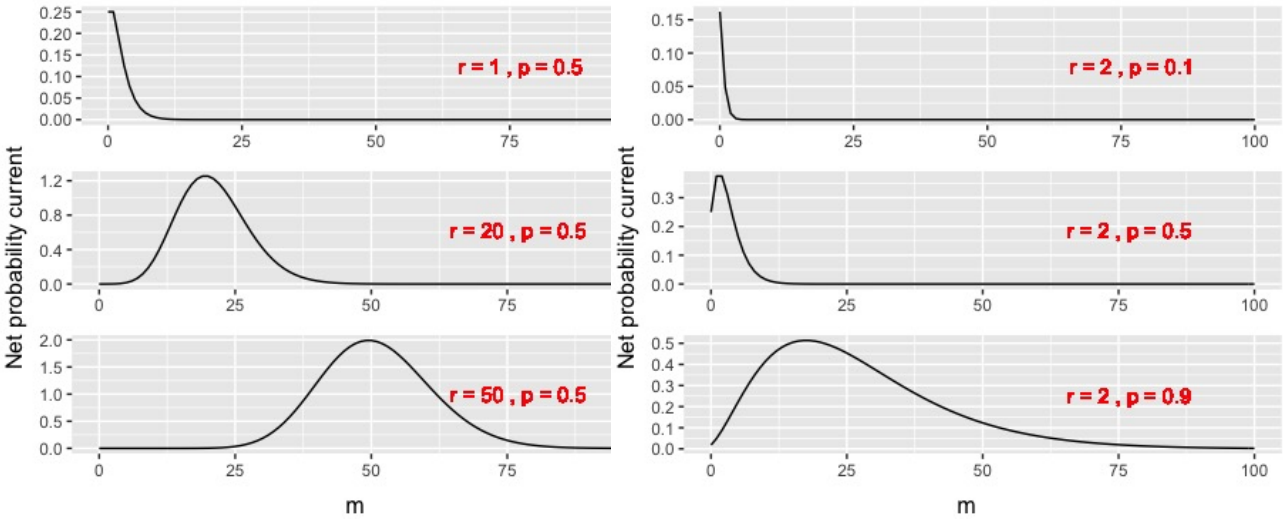

Figure 6: $J_{(A,m)\to(A,m+1)}(m)$ for different values of $r > 1$ and p.

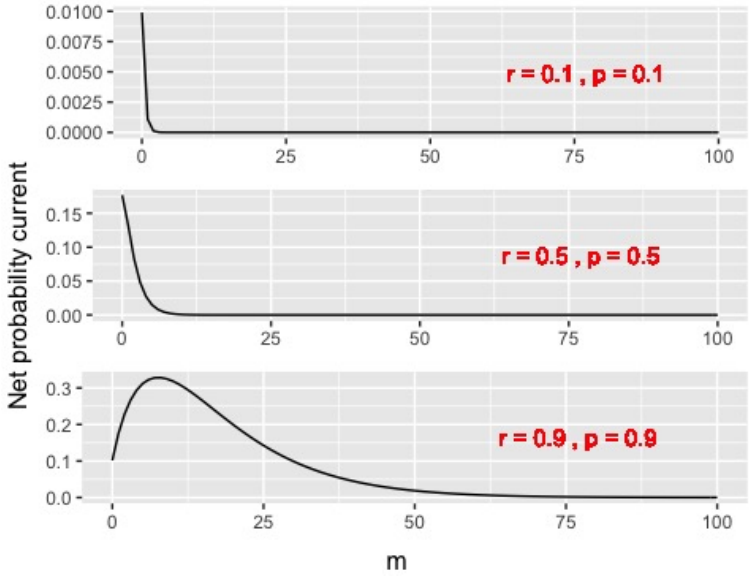

Figure 7: $J_{(A,m)\to(A,m+1)}(m)$ for different values of $0 < r \leq 1$ and p

Using the same technique as the horizontal flow, it is possible to locate the value of m where the current peaks. Let m^* denote this value of m where the current is at its maximum. Then

it must hold that

$$\frac{J_{(A,m)\to(A,m+1)}(m^*)}{J_{(A,m)\to(A,m+1)}(m^*+1)} > 1 \quad ; \quad \frac{J_{(A,m)\to(A,m+1)}(m^*)}{J_{(A,m)\to(A,m+1)}(m^*-1)} > 1.$$

From these two inequalities, it can be shown (in the appendix) that $m*$ fulfills the following compound inequality

$$\frac{rp}{1-p} - 1 < m^* < \frac{rp}{1-p}. \tag{20}$$

Standard computations reveal that $r \geq \frac{1-p}{p}$ and $0 < p < 1$ in order for (20) to hold. It can be seen that m^* is essentially the mean of the negative binomial distribution in the zeroth order. Similar to the compound inequalities for m_1^* and m_2^*, the difference between the upper and lower bound of (20) is always one. This means that the value of m^* can be uniquely determined if $\frac{rp}{1-p}$ is a real number. The case of integer bounds is similar to that of (17) and (19), which is mentioned in the appendix. From a biological perspective, this means that the optimal number of mRNAs in a cell when the gene is active is approximately the mean of the negative binomial distribution. Essentially, when there are not enough mRNAs, the gene's mRNA production process will work faster than the degeneration process to top up the number of mRNAs to the ideal level. This connects very well with the horizontal net current $J_{I\to A}(m)$ above where it has been shown that the mean of the negative binomial distribution is the value of m where the gene is at its most "comfortable" state.

2.5.3 Vertical flow II

Let $J_{(I,m)\to(I,m+1)}(m)$ denote the net probability current when the gene is in the inactive state, then

$$\begin{aligned} J_{(I,m)\to(I,m+1)}(m) &= 0 - (m+1)P_I(m+1) \\ &\approx -(m+1)P_0(m+1) \\ &= -p(r+m)P_0(m). \end{aligned}$$

As implied by the kinetic scheme in Figure 3, the probability current flow when the gene is inactive is one-directional because only mRNA degeneration occurs. Similar to $J_{(A,m)\to(A,m+1)}(m)$, it has always the same sign. Since its difference from $J_{(A,m)\to(A,m+1)}(m)$ is the sign, its behaviour is a mirror reflection of the latter. This means that it is the same m^* that gives rise to $J_{(I,m)\to(I,m+1)}(m)$'s minimum and $J_{(A,m)\to(A,m+1)}(m)$'s maximum. In other words, the m^* where the minimum of $J_{(I,m)\to(I,m+1)}(m)$ occurs fulfills the compound inequality in (20), which in turn means that this current shares the same optimal m as $J_{(A,m)\to(A,m+1)}(m)$. Intuitively, the degeneration process works to reduce the number of mRNAs back to the optimal level in the event the gene produces too many mRNAs when it is active. Again, this result connects very well with that of the horizontal net current $J_{I\to A}(m)$.

Concluding, the mean of the negative binomial distribution is an important quantity since it is the quantity where the gene is at its most comfortable state. As described at the beginning of this section, scRNA sequencing data is a matrix containing the gene expression counts of multiple genes in a sample of cells. This makes it possible to use standard inference methods such as maximum likelihood inference to estimate the r and p parameters. The ϵ parameter, while not taken into account here, can also be inferred if the perturbative solution that includes the first order term is used in the likelihood function.

3 Application

In this section, the model presented in section 2 is applied to a set of single-cell RNA sequencing data consisting of mRNA counts in 939 pyramidal cells taken from the hippocampus CA1 of the mouse cerebral cortex. Each cell has corresponding numbers of mRNAs synthesized by 19,972 genes. This data set is curated by Ziesel et al [17]. Maximum likelihood estimation is used to infer the parameters of the model. From now on, the first order model denotes the one that includes the first order ϵ term while the zeroth order model denotes the one that only has the negative binomial distribution.

3.1 Exploring pyramidal cells from mouse cerebral cortex

It is commonly known that single-cell RNA sequencing data has a large number of zeros counts. The same can be said for this data set, as shown in Figure 8 below, where an overwhelmingly large number of genes have very few mRNAs detected in most of the 939 cells. In fact, a total of 763 genes do not have any counts at all in any of the 939 cells. Figure 9 shows a collage of histograms of the number of mRNAs detected in the cells for some selected genes. It is evident that besides having large number of zero counts, the distribution of mRNAs is often skewed with a long right tail.

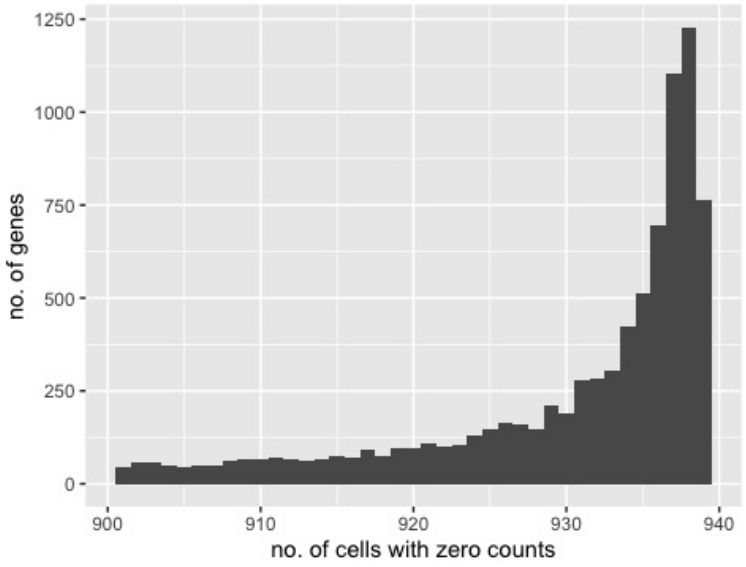

Figure 8: Zoomed-in view of histogram showing the number of zero-count cells among the genes.

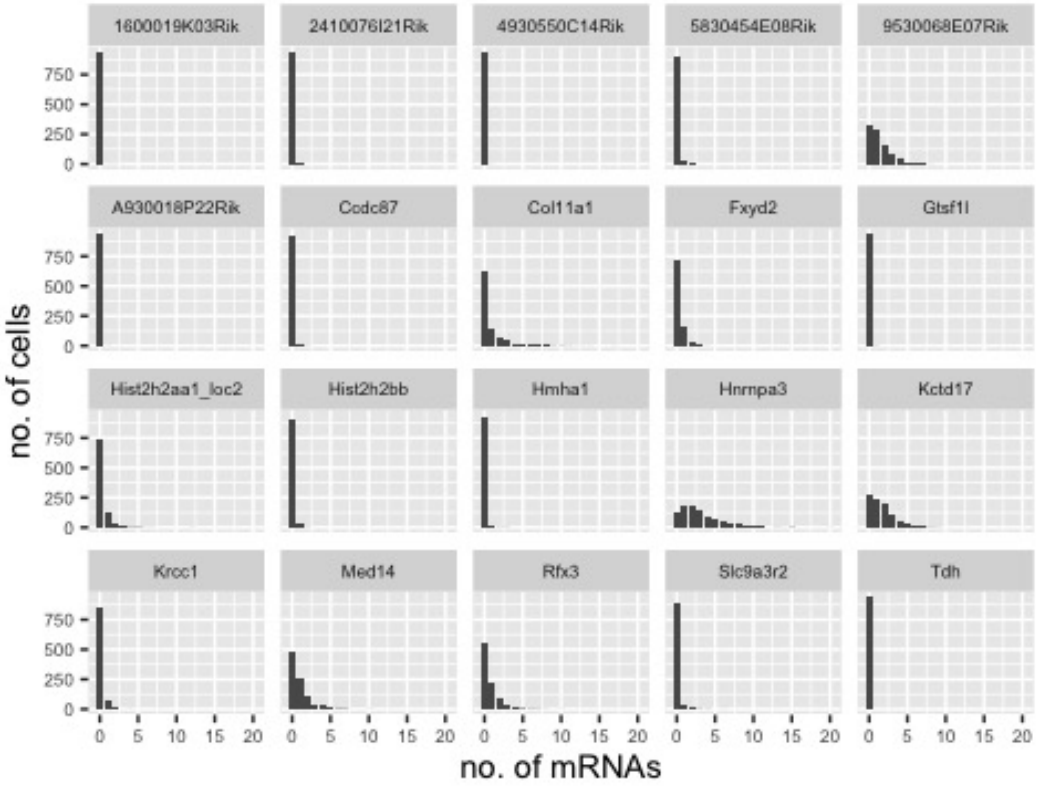

Figure 9: Histograms of mRNAs of randomly selected genes. The name of each panel is the name of the gene to which the histogram corresponds to.

Furthermore, it is generally acknowledged that single-cell RNA sequencing data is overdispersed, although the degree of dispersion can vary across datasets. This is one common reason why a negative binomial model, instead of Poisson, is often used to model this kind of data. However, equidispersed and underdispersed are not unheard of either. Figure 10 shows the sample variance versus sample mean of 19,209 genes on logarithmic 10 scale. It can be seen that most of the genes have sample variance larger than sample mean, but there is a sizable number of them whose sample variance is less than or equals to the sample mean. As shown in the next section, these equidispersed and underdispersed genes do not have maximum likelihood estimates and hence are removed.

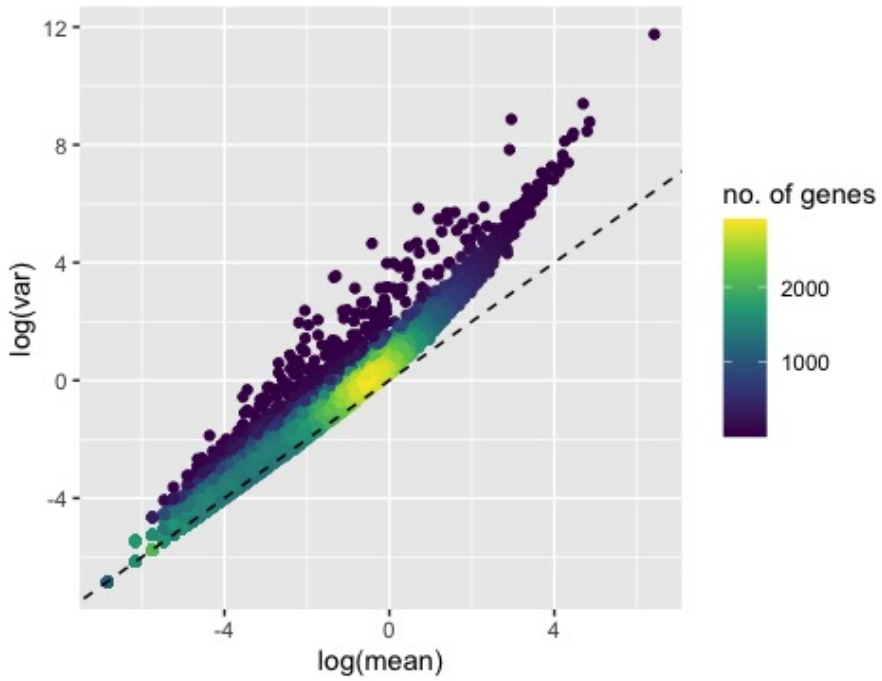

Figure 10: Density plot of sample variance versus sample mean for 19,209 genes on logarithmic 10 scale. 763 genes are removed due to zero count in all 939 cells. Dash line is $y = x$.

3.2 Determining $\hat{r}_{ML}$, $\hat{p}_{ML}$ and $\hat{\epsilon}_{ML}$

As mentioned in section 3.1, it is not possible to do any parameter inference for genes that do not have any counts at all. Hence these genes are removed from further analysis. Furthermore, when both parameters in the negative binomial distribution are unknown, maximum likelihood inference is possible only if sample variance is larger than sample mean [18]. In other words, genes whose sample variance is smaller than or equals to the sample mean are also discarded from further analysis. After removal of all these genes, there are 15,115 genes left in the data set.

Assuming the first order model, the log-likelihood function takes the form of

$$l(\boldsymbol{\theta}; \boldsymbol{m}) = \sum_{k=1}^{n} log\Big[\Gamma(m_k + r)\Big] - \sum_{k=1}^{n} log\Big[m_k!\Big] - nlog\Big[\Gamma(r)\Big] + nrlog(1-p) + log(p)\sum_{k=1}^{n} m_k + \\ + \sum_{k=1}^{n} log\left[1 + \epsilon\Big(\frac{rp}{1-p}\frac{p(r+1)(2-p)-2}{2} - p(1-p)\frac{m_k(m_k+1)}{2} - m_k rp(1-p) + p(r+m_k)\Big)\right],$$

where $\boldsymbol{\theta} = (r = \gamma_1, p = \frac{\mu_1}{\mu_1+\lambda_1}, \epsilon)^T$. Its partial derivatives with respect to r, p and ϵ are

$$\frac{\partial}{\partial r} l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \frac{\Gamma'(m_k + r)}{\Gamma(m_k + r)} - n\frac{\Gamma'(r)}{\Gamma(r)} + nlog(1-p) + \sum_{k=1}^{n} \frac{\epsilon \frac{\partial}{\partial r} h(r, p, m_k)}{1 + \epsilon h(r, p, m_k)};$$

$$\frac{\partial}{\partial p} l(\boldsymbol{\theta}) = -\frac{nr}{1-p} + \frac{\sum_{k=1}^{n} m_k}{p} + \sum_{k=1}^{n} \frac{\epsilon \frac{\partial}{\partial p} h(r, p, m_k)}{1 + \epsilon h(r, p, m_k)}; \tag{21}$$

$$\frac{\partial}{\partial \epsilon} l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \frac{h(r, p, m_k)}{1 + \epsilon h(r, p, m_k)},$$

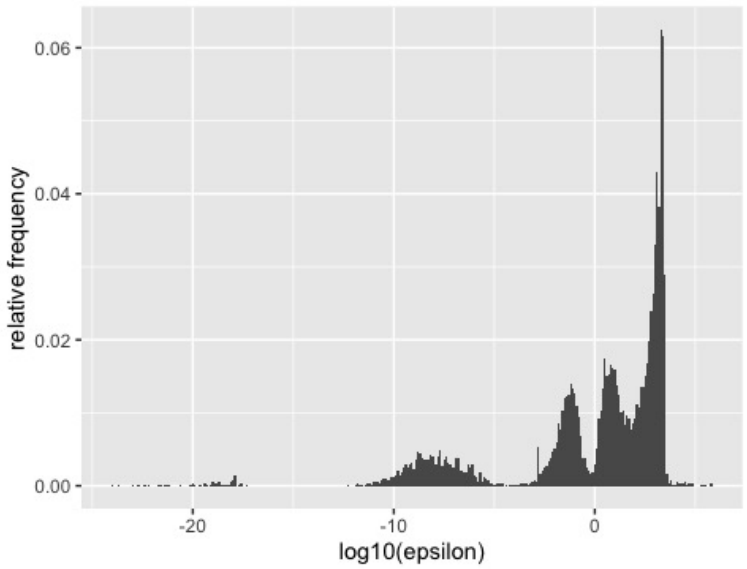

Figure 11: Histogram showing the spread of $\hat{\epsilon}_{ML}$ on logarithmic 10 scale. 5,231 values are removed because they are exactly zero on the linear scale.

where

$$h(r, p, m_k) = \frac{rp}{1-p}\frac{p(r+1)(2-p)-2}{2} - p(1-p)\frac{m_k(m_k+1)}{2} - m_k rp(1-p) + p(r+m_k).$$

There are no analytical solutions to these three parameters and hence, they have to be estimated numerically. Since the set of equations in (21) is solved by numerical approximation, there is a need to make choices about initial values. For the negative binomial parameters r and p, method of moments estimates are used as initial values. In other words, each gene has a different pair of r and p initial values. As for ϵ, since it is expected to be small, its initial value is a random real number sampled from $[0, 0.005]$ and is the same for every gene.

The Nelder-Mead method is used to solved (21). More discussion on numerical methods is found in section 4. While the Nelder-Mead method is generally used for unconstrained optimisation problems, modifications have been made that enable constraints to be placed on the unknown parameters. In this thesis, since the r and p parameters must satisfy the properties of the negative binomial distribution, they have to be positive and p has to be between 0 and 1. Furthermore, according to the assumption of ϵ, it has to be positive too as μ_1 must be positive. All computations are done in `R` [19] and Nelder-Mead optimisation of the log-likelihood function is done using the `Nelder_Mead` function in `R` package `lme4` [20].

The result is that all of the ML estimates of r and p stay well within the constraints, which means none of them hit the boundaries of the constraints placed on them. The ML estimates of ϵ is problematic, on the other hand, as more than 43% of them are larger than one, which essentially violates the model assumption of a small ϵ. Figure 11 shows the spread of $\hat{\epsilon}_{ML}$ minus those that are exactly zero. The 43% corresponds to the conjoined masses to the right of the zero mark . Furthermore, approximately 35% of them have values that are exactly zero, which is the lower bound of ϵ. This can mean that these ϵ's may have wanted to cross into the negative region but hit the wall instead at the boundary of the constraint. Indeed, a cross-check with the unconstrained maximum likelihood estimates of ϵ shows that over 99% of them are negative, albeit quite a number of them are close to zero. See Figure 12 that shows the spread of negative $\hat{\epsilon}_{ML}$ that are between 0 and -1.

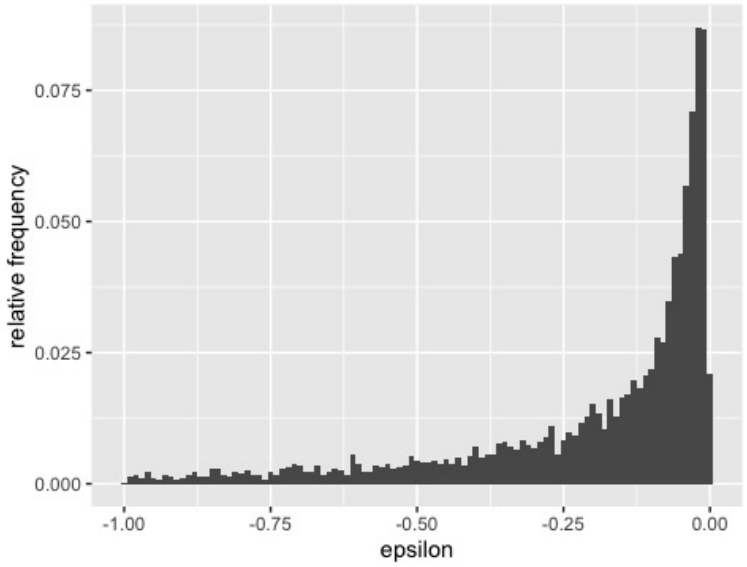

Figure 12: Histogram showing the spread of negative $\hat{\epsilon}_{ML}$ that are between 0 and -1.

Hence, disregarding genes that have either large $\hat{\epsilon}_{ML}$ or exactly zero $\hat{\epsilon}_{ML}$, there are 3,352 genes, which is about 22% of the total number of genes, that have $\hat{\epsilon}_{ML}$ values that reasonably fulfill the assumption of the model. Their values have a distribution that corresponds to the two masses on the left of Figure 11. A sanity check of the $\hat{r}_{ML}$ and $\hat{p}_{ML}$ values of these 3,352 genes shows reasonably acceptable values with approximately 76% of them having $\hat{r}_{ML}$ ranging between 0.50 and 3.00 while $\hat{p}_{ML}$ ranging between 0.25 and 0.75. This means that most of them have an estimated mean burst frequency ranging between 0.50 and 3.00 and estimated mean burst size ranging between of 0.33 and 3.00. Figure 13 shows all 3,352 value pairs of estimated mean burst frequency and estimated mean burst size on logarithmic 10 scale. It can be seen that there are no genes with the combination of low burst frequency and low burst size, which makes sense because it is likely that there may not be enough mRNAs captured for these genes due to finite sampling. It is now instructive to examine how different burst frequencies and burst sizes manifest in the spread of the mRNAs. See Figure 14 for such a comparison.

The upper-left panel in Figure 14 features a gene that is randomly chosen from those in the yellow blob in Figure 13. Contrasting this with a gene that has a higher burst frequency and larger burst size like `Actb` that is featured in the upper-right panel, it is evident that the cells contain larger numbers of mRNAs synthesised by the latter. This is reasonable since in addition to the gene being frequently active, it also synthesizes many mRNAs each time it switches to active. This implies that `Actb` "tops up" the number of mRNAs much faster than `Nr2f1` . The two panels in the second row features genes that have the exact opposite burst frequency and burst size characteristics. It can be seen, however, that a common feature between these two genes is the disproportionately large number of zero-count cells. For `Madcam1`, this can be explained by its low burst frequency, meaning that it is inactive most of the time. For `Ldlrad4`, the small burst size implies that even though it is a very active gene, it does not do much synthesis work when it is active. The striking difference between them is the length of the tail. For a gene like `Madcam1`, on the rare occasion that it switches to active, it synthesizes a large number of mRNAs, which explain the exceptionally long tail. But this also means that if the sample size is not large enough, then there is a high chance that no mRNAs are captured for this gene at all.

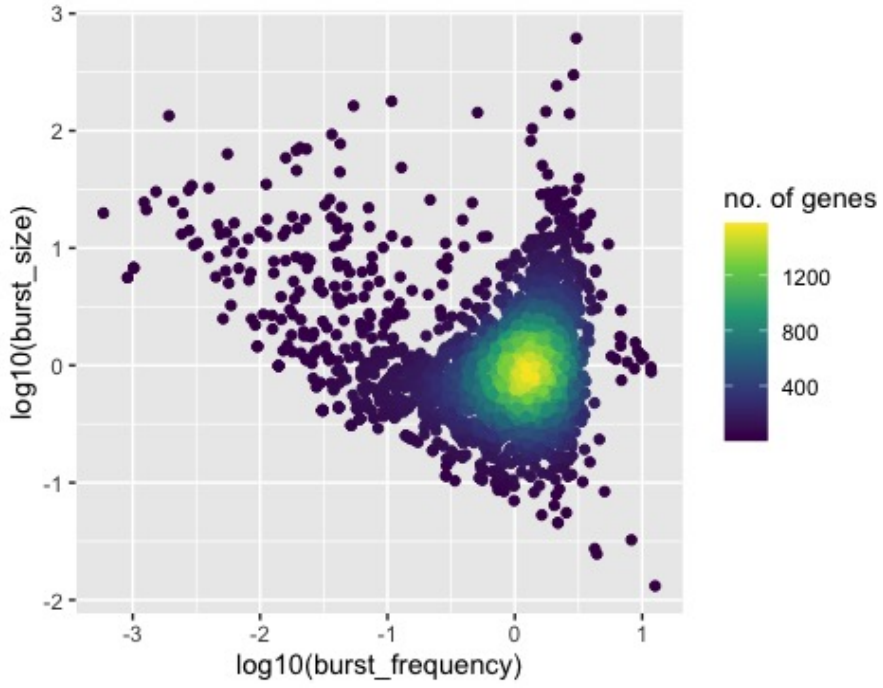

Figure 13: Estimated mean frequency and mean burst size of the 3,352 genes on logarithmic 10 scale.

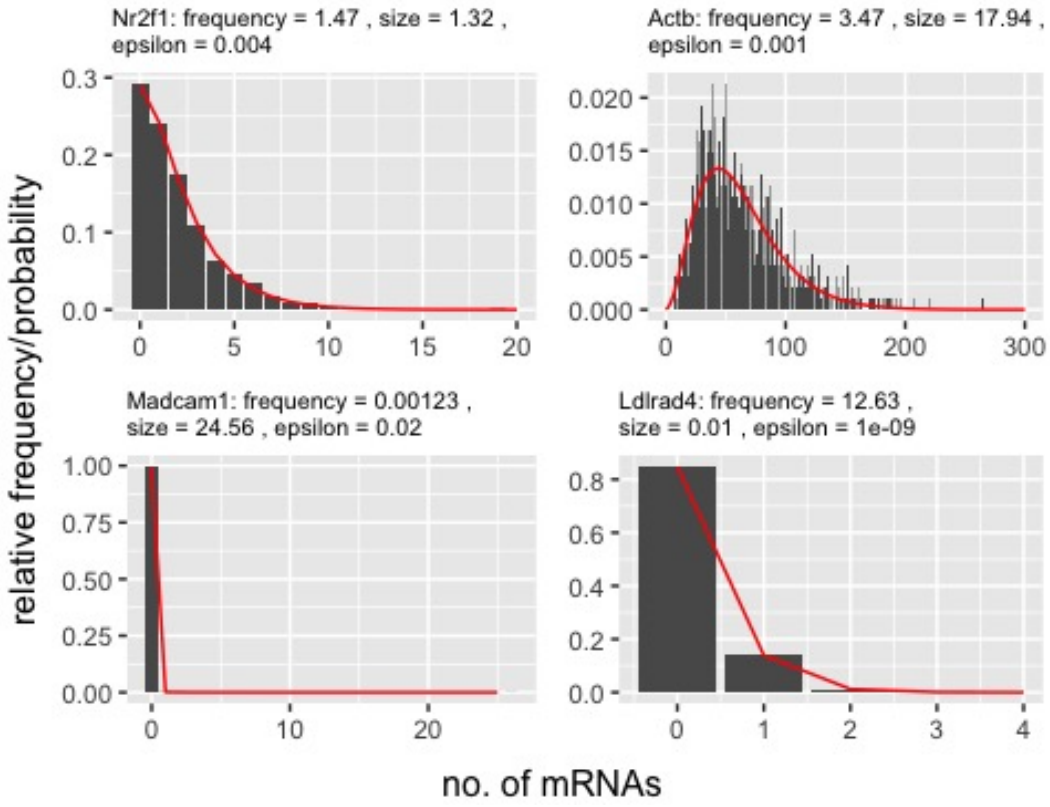

Figure 14: Histograms showing the spread of mRNAs (grey) for genes with different burst frequency and burst size characteristics, each superimposed with the fitted negative binomial distribution (red) using the individual gene's estimated r and p parameters.

Finally, as an informal check that the model assumption of a small ϵ holds, the log-likelihood function values of the zeroth order model are juxtaposed against those of the first order model.

The idea is if just the negative binomial model alone, that is the zeroth order model, is a good enough approximation for the exact distribution, then its log-likelihood value should be at least as good as that of the first order. As shown in the left panel of Figure 15, most of the log-likelihood values pairs lie very close to the straight line, which indicates that adding a small correction to the zeroth order model does not help to materially increase the likelihood of observing the given data. In fact, only 1,413 genes have their first order log-likelihood values larger than their zeroth order ones. In addition, the difference between the two values is very small. This is evident in the right panel of Figure 15, which shows that almost all of them are less than 10^{-5}. Hence, the negative binomial model alone can be said to be a good enough approximation to the exact distribution of M for these 3,352 genes. It is instructive, however, to take note that this is only a qualitative comparison. A more mathematically rigorous comparison is to first ascertain that the maximum likelihood optimisation indeed locates the global maximums of both log-likelihood functions, otherwise this can well be a comparison of two local maximums. This problem is exacerbated by the fact that the first order model is more non-linear than the zeroth order model as can be seen in (16), which means the chances of hitting a local maximum are higher. Also, as shown in the left panel of Figure 15, a small number of genes have zeroth order log-likelihood values that are much higher than their first order ones. Since ϵ is small, it quite surprising to observe the occurrence of such large differences. It is believed that this is due to numerical issues in the maximum likelihood optimisation of the first order model, which is discussed in more details in section 4.

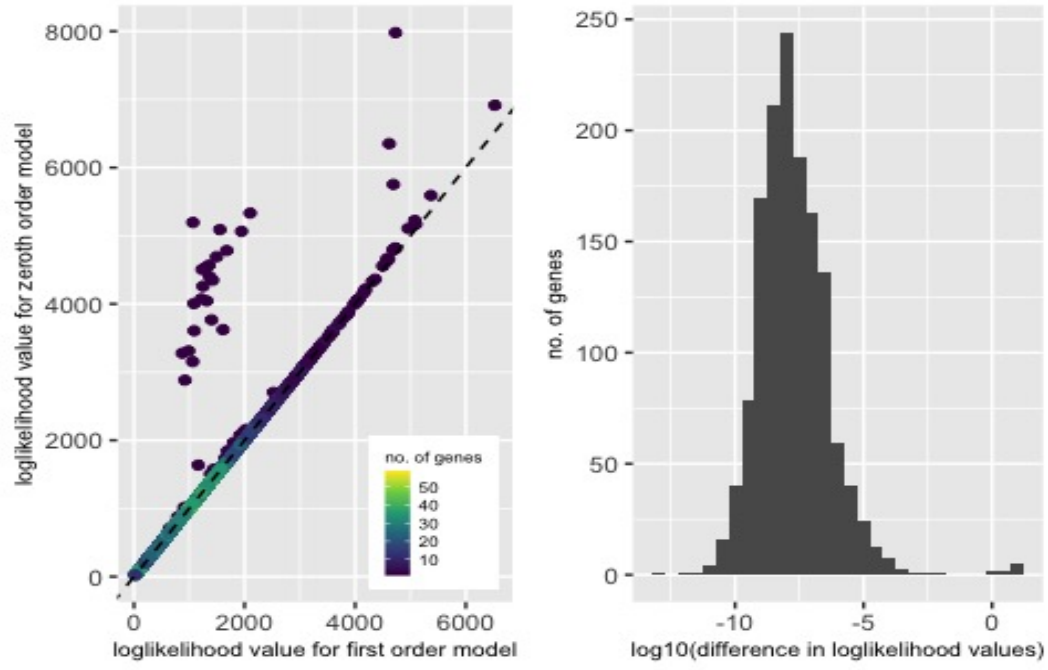

Figure 15: Left: density plot of log-likelihood function values of the first order model against the zeroth order mode; Right: spread of difference in log-likelihood values among genes whose first order log-likelihoods are larger than zeroth order log-likelihoods.

4 Discussion

In this thesis, the mRNA synthesis and degeneration process of a gene in a cell is modelled with a bivariate Markov Chain. A kinetic scheme is drawn up to represent this Markov Chain. Under the assumption that a gene synthesizes mRNAs in a burst-like manner when it is active, perturbation techniques are then used to derive an approximation of the steady-state stationary distribution of the number of mRNAs, whose exact distribution is a Poisson-beta mixture

distribution. The negative binomial distribution emerges as the solution to the unperturbed problem. In other words, the negative binomial distribution can be a good approximation to the stationary distribution. A number of literature that model the synthesis of gene products [3] [21] [22] have also arrived at the negative binomial distribution as the limiting distribution of the number of mRNAs (and proteins). The work in this thesis provides an alternative mathematical explanation. Furthermore, this thesis derives the distribution of the inter-arrival time of bursts as well as that of the number of new mRNAs synthesised at each burst. It is shown that the means of these two distributions are related to the parameters of the negative binomial distribution. Besides, an in-depth investigation of the probability currents shows further relevance of the negative binomial distribution. All these help to connect the negative binomial distribution to interesting quantities that are related to mRNA transcription and adds depth to a cell's gene expression profile. Maximum likelihood inference using the first order model is applied on a set of scRNA sequencing data that is collected from pyramidal cells of the hippocampus CA1 of the mouse cerebral cortex. A total of 3,352 genes return parameter values that reasonably fulfill the model assumptions and hence can be used to profile the pyramidal cells.

As seen in section 3, there are some problems in maximum likelihood fitting when the first order model is used. But this is within expectation since $P(m)$ in the first order model may not even be a proper probability mass function when it turns negative for certain combinations of r, p, ϵ and m. Indeed, as shown in section 3, the inclusion of the epsilon parameter in maximum likelihood fitting has shown to render the model invalid for a large number of genes in the dataset. First, because μ_1 is positive, this implies that ϵ has to be positive. But there is actually no such constraint in pertubation analysis. Hence, in cases where a negative $\hat{\epsilon}_{ML}$ is obtained, the gene is discarded for further analysis as it does not fulfill the model assumption, even though the fit is good. A case in point is gene `Mrpl43` shown in Figure 16. The first order model (green) with a negative $\hat{\epsilon}_{ML}$ actually fits the observed data quite well. Another example that exemplifies problem with the first order model is illustrated using gene `Nr2f1` in Figure 17. It can be seen that there are (r, p, ϵ) values that are quite far from the Nelder-Mead estimates but are close in the log-likelihood values. For example, $r = 1.51$, $p = 0.891$ and $\epsilon = 0.460$ returns a log-likelihood value of -1,745, which is lower than the log-likelihood that is computed using the Nelder-Mead estimates. Hence the first order model is not unimodal and may not even be continuous. Even with ϵ just slightly larger than zero, the model can become unphysical as $P(m)$ turns negative for some values of m and log-likelihood becomes infinite.

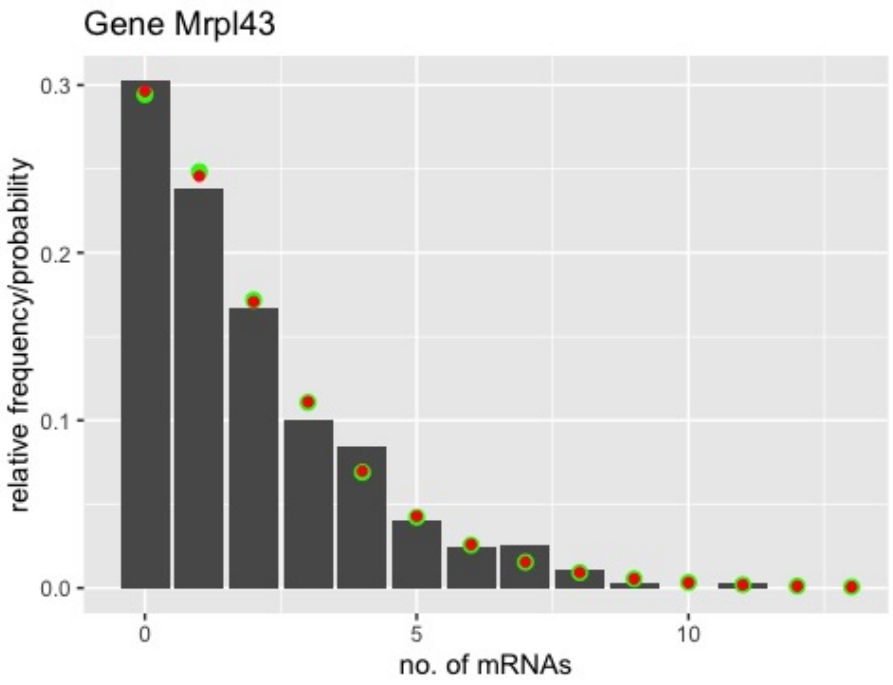

Figure 16: Spread of mRNAs. Grey: Observed counts. Red: Fitted distribution using zeroth order model, i.e. $\epsilon = 0$, where $\hat{r}_{ML} = 1.48$ and $\hat{p}_{ML} = 0.561$. Numerical method is Nelder-Mead. Green: Fitted distribution using first order model, where $\hat{r}_{ML} = 1.61$, $\hat{p}_{ML} = 0.521$ and $\hat{\epsilon}_{ML} = -0.044$. Numerical method is Newton-Raphson.

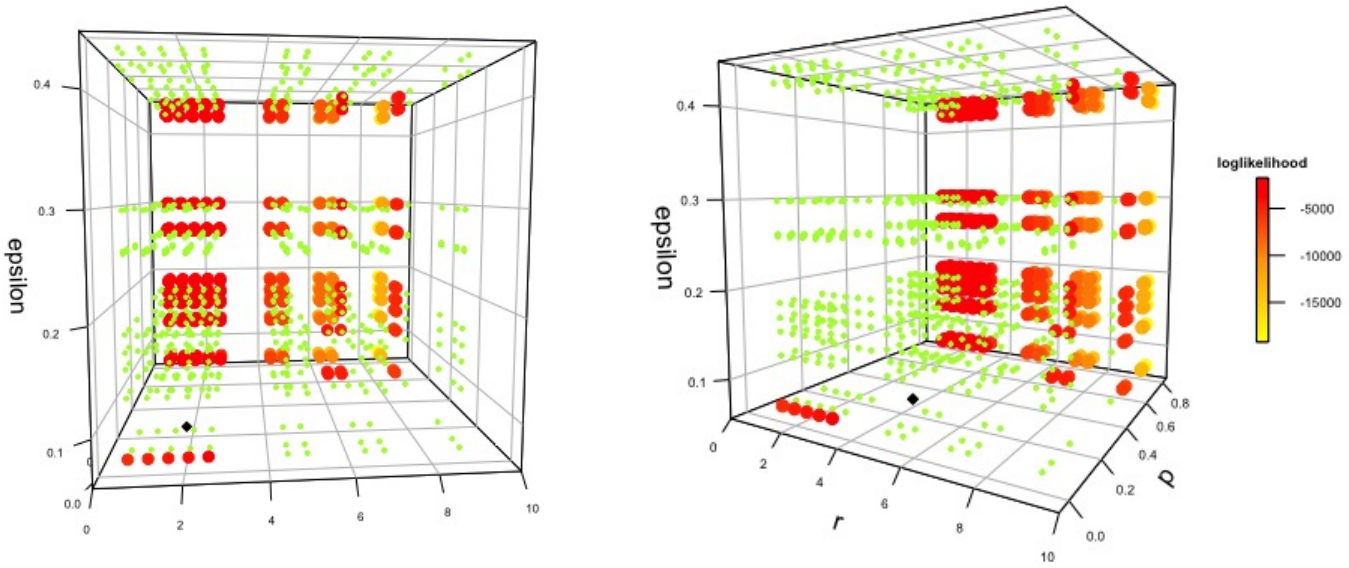

Figure 17: 4-dimensional view of the log-likelihood function values for gene Nr2f1 for 1,000 randomly selected values of r, p and ϵ. The black diamond locates the coordinates of the Nelder-Mead maximum likelihood estimates ($\hat{r}_{ML} = 1.47$, $\hat{p}_{ML} = 0.569$, $\hat{\epsilon}_{ML} = 0.0040$), which has a log-likelihood value of -1,755. Green points locate the coordinates for which the log-likelihood is undefined.

Second, different numerical methods return quite different results on the first order model. The optimisation method used in this thesis for the maximum likelihood optimisation problem is the Nelder-Mead method, which does not require derivatives. While it is commonly used

in unconstrained problem, there is also a box constraint version where both lower and upper bounds on the parameter values can be imposed. This method works well when used to estimate parameters in the zeroth order model without the need to impose constraints. However, problems arise when it is used to estimate parameters in the first order model. Quite a sizable number of $\hat{\epsilon}_{ML}$ hit the lower bound of zero while some others return unphysical results. The $\hat{\epsilon}_{ML}$ for gene `Mrpl43`, for example, is 29,243 while $\hat{r}_{ML}$ is 5.10 and $\hat{p}_{ML}$ is 0.660. With such a large $\hat{\epsilon}_{ML}$, the first order model is no longer valid since $P(m)$ is no longer a probability function. On the other hand, the estimates in the zeroth order model are 1.48 and 0.561 for $\hat{r}_{ML}$ and $\hat{p}_{ML}$ respectively. As shown in Figure 16, this model fits the observed data reasonably well. A deeper investigation into why the algorithm does not end up with a small $\hat{\epsilon}_{ML}$ would be instructive but is likely to fall into the realm of numerical analysis, which is beyond the scope of this thesis. Besides Nelder-Mead, this thesis has also considered the gradient descent method, which is a commonly used method in optimisation problems. However, the first order model again presents some difficulties for this method.

The presence of a presumably small parameter ϵ in the denominator of the third partial derivative in (21) can become problematic. This is because if $h(r, p, m_k)$ becomes large, which is highly possible since it is a quadratic function in m_k, then division by a small ϵ can cause some of the summands to become really large, which in turn causes the gradient to become very large. This can throw the descent process off-track and return spurious results, as attested by running this method on some genes (code is available). While the choice of learning rate can help to mitigate this problem, this involves making yet another choice of parameter value. Finally, the Newton-Raphson root search method is another viable method and also tested on some genes, in particular those that end up with large $\hat{\epsilon}_{ML}$ (code is available) with the Nelder-Mead method. As shown in Figure 16, the numerical solution returned by the Newton-Raphson method is a good fit for the observed data albeit with a negative $\hat{\epsilon}_{ML}$. It has also returned unphysical results for many other genes, mainly due to negative $\hat{\epsilon}_{ML}$. When tested on the zeroth order model, however, an overwhelming majority of the estimates from this method and the Nelder-Mead method are consistent with one another. Unphysical results aside, the main reason it is not chosen as the optimisation method is the lengthy computation time because the Hessian needs to be computed for so many genes. All in all, while the assumption of a large μ_1 has made it possible to approximate the exact Poisson-beta mixture distribution by means of perturbation techniques, its inclusion in the optimisation problem causes numerical instability and results in unphysical values. More work needs to be done to better ascertain if this is a numerical issue or wrong assumptions of the kinetic parameters in the model.

Finally, up to this point, there is no mention of dependence between a gene's mRNA synthesis and degeneration process and the type of cell that it resides in. In this thesis, it has been assumed that cell type is known and maximum likelihood inference is done on cells that of the same type. In biology literature [17] [23] [24] [25] , the same gene that resides in different cell types is believed to behave differently. In other words, the gene expression profiles of cells are not homogeneous. However, it is rare that the gene expression data available for analysis contains information about cell types. Hence, cell type is a latent variable that is not directly observed. Inclusion of cell type into parameter inference is definitely a possible extension of this thesis. This is a general idea of how it can be taken into consideration: Let U denote the discrete latent variable of cell type with a finite set of J types. Let $\boldsymbol{m} = \{m_g\}$ denote a vector of mRNA counts from $g = 1, 2, ..., G$ genes. In other words, $\boldsymbol{m}$ is the number of mRNA counts of G genes in a cell. Then given J cell types, $\boldsymbol{m}$ comes from the following joint probability mass function

$$P(\boldsymbol{m}|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j P(\boldsymbol{m}|\boldsymbol{\theta}_j),$$

where $\boldsymbol{\theta}_j$ holds the parameters of the joint distribution of $(m_1, m_2, ..., m_G)$ condition on cell

type j and π_j is the mixing coefficient (which can be fixed or inferred). $\boldsymbol{\theta}$ denotes all unknown parameters in the model. Intuitively, π_j is the probability that a randomly selected cell is of type j and all J of them need to sum to one. If the components in the conditional joint distribution are independent of one another and assuming the negative binomial model for the mRNAs, then $P(\boldsymbol{m}|\boldsymbol{\theta}_j)$ is a product of G negative binomial distributions and $\boldsymbol{\theta}_j = \{r_{gj}, p_{gj}\}$. Then given an observed set of N independent distributed cells, the likelihood is given by

$$P(\boldsymbol{m}, \boldsymbol{u}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{j=1}^{J} \Big[\pi_j P(\boldsymbol{m}_n|\boldsymbol{\theta}_j)\Big]^{u_{nj}},$$

where $\boldsymbol{u}$ denotes the indicator vector of $\{u_{nj}\}$ such that it is one if $\boldsymbol{m}_n$ is of cell type j and zero otherwise. Markov Chain Monte Carlo methods such as Gibbs sampling can be used to draw a sample from the posterior $P(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{m})$.

5 Appendix

1. ***Definition of Negative Binomial Distribution, NB***

   Let M be a discrete, integer-valued and non-negative random variable that denotes the number of successes observed in a sequence of Bernoulli experiments until a predefined number of failures are attained, that is,

   $$M \sim NB(r, p).$$

   The predefined number of failures is r and p is the probability of success in each experiment. Take note that $r > 0$ and $p \in [0, 1]$. The probability mass function of M is

   $$P(M = m) = \binom{m + r - 1}{m} p^m (1 - p)^r, \quad \text{for } m \geq 0.$$

   While r is defined here as integer-valued, it can be extended to real-valued which comes from the negative binomial distribution being a Poisson distribution with gamma distributed Poisson rate with parameters r and $(1 - p)/p$. Take note that the negative binomial distribution implemented in the `R` package `stats` has the opposite interpretation, that is, it describes the number of failures observed in a sequence of Bernoulli experiments until a predefined number of successes are attained.

   It is also known that the following recurrence relation holds for the negative binomial distribution

   $$(m + 1)P(m + 1) = p(r + m)P(m), \text{ for } m \geq 0.$$

   Let $g(z)$ denote the probability generating function of M. Then using the commonly known formula of $\frac{g^{(m)}(0)}{m!} = P(m)$, where $g^{(m)}(z)$ denotes the mth derivative of $g(z)$, the recurrence relation above can be expressed in terms of the probability generating function as

   $$\begin{aligned} (m + 1)\frac{g^{(m+1)}(0)}{(m + 1)!} &= p(r + m)\frac{g^{(m)}(0)}{(m)!} \\ &\iff \\ g^{(m+1)}(0) &= p(r + m)g^{(m)}(0). \end{aligned}$$

2. Proof of long-run probabilities P_I and P_A.

   Recall in section 2.2 that

   $$\begin{aligned} P_I &= \frac{\lambda}{\lambda + \gamma}; \\ P_A &= \frac{\gamma}{\lambda + \gamma}. \end{aligned}$$

This can be shown by summing over all m in the first equation in (1) as such

$$\sum_{m=0}^{\infty}(\gamma+m\delta)P_I(m)=\sum_{m=0}^{\infty}\Big((m+1)\delta P_I(m+1)+\lambda P_A(m)\Big)$$
$$\Longleftrightarrow$$
$$\gamma\sum_{m=0}^{\infty}P_I(m)+\delta\sum_{m=0}^{\infty}mP_I(m)=\delta\sum_{m=0}^{\infty}(m+1)P_I(m+1)+\lambda\sum_{m=0}^{\infty}\Big(P(m)-P_I(m)\Big)$$
$$\Longleftrightarrow$$
$$\gamma\sum_{m=0}^{\infty}P_I(m)+\delta\sum_{m=0}^{\infty}(m+1)P_I(m+1)=\delta\sum_{m=0}^{\infty}(m+1)P_I(m+1)+\lambda-\lambda\sum_{m=0}^{\infty}P_I(m)$$
$$\Longleftrightarrow$$
$$\sum_{m=0}^{\infty}P_I(m)=P_I=\frac{\lambda}{\lambda+\gamma}.$$

And the proof for P_A follows as

$$\sum_{m=0}^{\infty}P(m)=\sum_{m=0}^{\infty}P_I(m)+\sum_{m=0}^{\infty}P_A(m)$$
$$\Longleftrightarrow$$
$$1=\frac{\lambda}{\lambda+\gamma}+\sum_{m=0}^{\infty}P_A(m)$$
$$\Longleftrightarrow$$
$$\sum_{m=0}^{\infty}P_A(m)=P_A=\frac{\gamma}{\lambda+\gamma}.$$

3. Derivation of higher order differential equations from (5)

$$\delta\frac{\partial^m}{\partial z^m}(z-1)g_I'(z)=-\gamma g_I^{(m)}(z)+\lambda g_A^{(m)}(z)$$
$$\delta\frac{\partial^m}{\partial z^m}(z-1)g_A'(z)=\gamma g_I^{(m)}(z)-\lambda g_A^{(m)}(z)+\mu\frac{\partial^m}{\partial z^m}(z-1)g_A(z)$$
$$\Longleftrightarrow$$
$$\delta\Big(mg_I^{(m)}(z)+(z-1)g_I^{(m+1)}(z)\Big)=-\gamma g_I^{(m)}(z)+\lambda g_A^{(m)}(z)$$
$$\delta\Big(mg_A^{(m)}(z)+(z-1)g_A^{(m+1)}(z)\Big)=\gamma g_I^{(m)}(z)-\lambda g_A^{(m)}(z)+\mu\Big(mg_A^{(m-1)}(z)+(z-1)g_A^{(m)}(z)\Big)$$

4. Poisson-beta distribution

Using equation 1 in the supplementary information of [3], the exact solution to the steady-state marginal stationary distribution of the number of mRNAs M is

$$P(m)=\frac{\Gamma(\gamma_1+m)}{\Gamma(m+1)\Gamma(\gamma_1+\lambda_1+m)}\frac{\Gamma(\gamma_1+\lambda_1)}{\Gamma(\gamma_1)}\mu_1^m{}_1F_1(\gamma_1+m,\gamma_1+\lambda_1+m,-\mu_1).$$

This can be rewritten by using the integral representation of the confluent hypergeometric function of the first kind, that is,

$$ {}_1F_1(\gamma_1+m,\gamma_1+\lambda_1+m,-\mu_1)=\frac{\Gamma(\gamma_1+\lambda_1+m)}{\Gamma(\lambda_1)\Gamma(\gamma_1+m)}\int_0^1 e^{-\mu_1 t}t^{\gamma_1+m-1}(1-t)^{\lambda_1-1}\,dt. $$

Then the exact solution becomes

$$ \begin{aligned} P(m) &= \int_0^1 \frac{\mu_1^m}{m!}e^{-\mu_1 t}\frac{\Gamma(\gamma_1+\lambda_1)}{\Gamma(\gamma_1)\Gamma(\lambda_1)}t^{\gamma_1+m-1}(1-t)^{\lambda_1-1}\,dt \\ &= \int_0^1 \frac{(\mu_1 t)^m}{m!}e^{-\mu_1 t}\frac{\Gamma(\gamma_1+\lambda_1)}{\Gamma(\gamma_1)\Gamma(\lambda_1)}t^{\gamma_1-1}(1-t)^{\lambda_1-1}\,dt \\ &= \Big[\text{change of variables: } x=\mu_1 t\Big] \\ &= \int_0^{\mu_1} \frac{x^m}{m!}e^{-x}\frac{1}{\mu_1}\frac{\Gamma(\gamma_1+\lambda_1)}{\Gamma(\gamma_1)\Gamma(\lambda_1)}\Big(\frac{x}{\mu_1}\Big)^{\gamma_1-1}\Big(1-\frac{x}{\mu_1}\Big)^{\lambda_1-1}\,dx \\ &= \int_0^{\mu_1} \frac{x^m}{m!}e^{-x}\frac{\Gamma(\gamma_1+\lambda_1)}{\Gamma(\gamma_1)\Gamma(\lambda_1)}\frac{x^{\gamma_1-1}(\mu_1-x)^{\lambda_1-1}}{\mu_1^{\gamma_1+\lambda_1-1}}\,dx. \end{aligned} $$

By the law of total probability, this is the unconditional distribution of M, where

$$ M|\mu_1 X=x\sim Po(x) \qquad \text{with} \qquad \mu_1 X\sim Beta(\gamma_1,\lambda_1). $$

5. Rewriting (11) and (12)

   Matching the corresponding terms between the perturbation series of g^m and $P(m)$, (11) can be expressed as

$$ P_0(m+1)=\Big(\frac{\mu_1}{\mu_1+\lambda_1}\Big)\frac{(\gamma_1+m)}{m+1}P_0(m)=p\frac{(r+m)}{m+1}P_0(m). $$

   Since $P_0(m)$ is the probability mass function of the negative binomial distribution, then it follows that

   - $\sum_{m=0}^{\infty}P_0(m)=1$
   - $\sum_{m=0}^{\infty}P_i(m)=0$, for $i\geq 1$

     This is because

$$ 1=\sum_{m=0}^{\infty}P(m)=\sum_{m=0}^{\infty}\Big(P_0(m)+\epsilon P_1(m)+\epsilon^2 P_2(m)+...\Big). $$

     Since $\sum_{m=0}^{\infty}P_0(m)=1$, it follows that $\sum_{m=0}^{\infty}P_i(m)=0$ for $i\geq 1$.

   Using the same parametrization of r and p, (12) can similarly be rewritten as

$$ (m+2)P_0(m+2)=\big(r+m\big)P_0(m+1)+\frac{1}{p}P_1(m+1)-\frac{r+m}{m+1}P_1(m). \tag{22} $$

   It turns out that (22) can be further simplified. The first manipulation is to use $\mu_1=\frac{\mu}{\delta}$ and $\epsilon=\frac{1}{\mu_1}$, after which (7) is rewritten as

$$ \epsilon P(m+1)=\frac{1}{m+1}P_A(m). $$

The second manipulation is to substitute the perturbation series of $P_I(m)$ and $P_A(m)$ into the rewritten (7), which results in

$$\epsilon P_0(m+1) + \epsilon^2 P_1(m+1) + ... = \frac{1}{m+1}\Big(P_{0,A}(m) + \epsilon P_{1,A}(m) + \epsilon^2 P_{2,A}(m) + ...\Big).$$

Expanding the equation and grouping the coefficients of like powers of ϵ together yields

$$\begin{aligned}\frac{1}{m+1}P_{0,A}(m) + \epsilon\Big(- P_0(m+1) + \frac{1}{m+1}P_{1,A}(m) + \Big) \\ + \epsilon^2\Big(- P_1(m+1) + \frac{1}{m+1}P_{2,A}(m)\Big) + ... = 0.\end{aligned}$$

Using the same argument in the thesis proper about the coefficients of a constantly zero perturbation series, the following results are obtained:

$$\begin{aligned}P_{0,A}(m) &= 0; \\ P_0(m+1) &= \frac{1}{m+1}P_{1,A}(m); \\ P_1(m+1) &= \frac{1}{m+1}P_{2,A}(m).\end{aligned} \tag{23}$$

These are important results because this means that the zeroth order of the stationary distribution of M, which is a negative binomial distribution, is determined by the zeroth order term from the inactive status of the gene. That is

$$P_0(m) = P_{0,A}(m) + P_{0,I}(m) = P_{0,I}(m).$$

Furthermore, the entire first order correction from the active status of the gene can be determined by the zeroth order negative binomial distribution. But this also means that

$$P_{0,I}(m+1) = \frac{1}{m+1}P_{1,A}(m).$$

Inserting the second result of (23) as well as the identity $P_i(m) = P_{i,I}(m) + P_{i,A}(m)$, for $i \geq 0$, into (22) gives the following:

$$\begin{aligned}(m+2)P_0(m+2) = \big(r+m\big)P_0(m+1) + \frac{1}{p}P_{1,A}(m+1) + \frac{1}{p}P_{1,I}(m+1) - \\ - \frac{r+m}{m+1}P_{1,A}(m) - \frac{r+m}{m+1}P_{1,I}(m)\end{aligned}$$

$$\iff$$

$$\begin{aligned}(m+2)P_0(m+2) = \big(r+m\big)P_0(m+1) + \frac{m+2}{p}P_0(m+2) + \frac{1}{p}P_{1,I}(m+1) - \\ - \big(r+m\big)P_0(m+1) - \frac{r+m}{m+1}P_{1,I}(m)\end{aligned}$$

$$\iff$$

$$P_{1,I}(m+1) = p\Big(\frac{r+m}{m+1}\Big)P_{1,I}(m) + (m+2)(p-1)P_0(m+2),$$

which is (13).

6. Deriving a general expression for the recursive relation in (13)

   In order to derive a general expression from this recursive relation to solve for $P_{1,I}(m)$, define $P_{1,I}(0) = a$. Then for

- $m = 0$

$$P_{1,I}(1) = p\Big(\frac{r}{1}\Big)a + 2(p-1)P_0(2)$$

- $m = 1$

$$\begin{aligned} P_{1,I}(2) &= p\Big(\frac{r+1}{2}\Big)P_{1,I}(1) + 3(p-1)P_0(3) \\ &= p^2\Big(\frac{r+1}{2}\Big)\Big(\frac{r}{1}\Big)a + 3\Big(\frac{r+1}{r+2}\Big)(p-1)P_0(3) + 3(p-1)P_0(3) \\ &= p^2\Big(\frac{r+1}{2}\Big)\Big(\frac{r}{1}\Big)a + 3(p-1)\Big(\frac{2r+3}{r+2}\Big)P_0(3). \end{aligned}$$

The second last equality is derived from (11), that is, the recurrence relation of a negative binomial distribution.

- $m = 2$

$$\begin{aligned} P_{1,I}(3) &= p\Big(\frac{r+2}{3}\Big)P_{1,I}(2) + 4(p-1)P_0(4) \\ &= p^3\Big(\frac{r+2}{3}\Big)\Big(\frac{r+1}{2}\Big)\Big(\frac{r}{1}\Big)a + 4(p-1)\frac{2r+3}{r+3}P_0(4) + 4(p-1)P_0(4) \\ &= p^3\Big(\frac{r+2}{3}\Big)\Big(\frac{r+1}{2}\Big)\Big(\frac{r}{1}\Big)a + 4(p-1)\Big(\frac{3r+6}{r+3}\Big)P_0(4). \end{aligned}$$

It follows by mathematical induction that, in general,

$$\begin{aligned} P_{1,I}(m+1) &= p^{m+1}\Big(\frac{r+m}{m+1}\Big)\Big(\frac{r+m-1}{m}\Big)\cdots\Big(\frac{r}{1}\Big) + \\ &\qquad + (m+2)(p-1)\Big(\frac{(m+1)r+1+2+3+\ldots+(m+1)}{r+m+1}\Big)P_0(m+2) \\ &= p^{m+1}\binom{r+m}{m+1}a + (m+2)(p-1)\Big(\frac{(m+1)r+\frac{(m+1)(m+2)}{2}}{r+m+1}\Big)P_0(m+2) \\ &= p^{m+1}\binom{r+m}{m+1}a + \frac{1}{2}(m+2)(m+1)(p-1)\Big(\frac{2r+m+2}{r+m+1}\Big)P_0(m+2). \end{aligned}$$

7. Simplification of (14)

To simplify the right hand side, take note that since $P_0(m)$ is the probability mass function of a negative binomial distribution with parameters r and p, then it follows that the summations in the second and fourth summands can be simplified as

$$\frac{rp}{1-p} = \sum_{m=0}^{\infty} mP_0(m) = \sum_{m=0}^{\infty}(m+1)P_0(m+1).$$

The second equality holds since the summand at $m = 0$ does not contribute to the sum. Furthermore, the third summand can be simplified as follows:

$$\begin{aligned} \frac{(p-1)^2}{2r}\sum_{m=0}^{\infty}(m+1)mP_0(m+1) &= \frac{(1-p)^2}{2r}\sum_{m=0}^{\infty}(m+1)m\binom{r+m}{m+1}(1-p)^r p^{m+1} \\ &= \frac{(1-p)^2}{2r}\sum_{m=1}^{\infty}\frac{(r+m)!}{(m-1)!(r+1)!}(1-p)^{r+2}p^{m-1}\frac{r(r+1)p^2}{(1-p)^2} \\ &= \Big[\text{Let } n = m-1\Big] \\ &= \frac{(1-p)^2}{2r}\frac{r(r+1)p^2}{(1-p)^2}\sum_{n=0}^{\infty}\binom{r+2+n-1}{n}(1-p)^{r+2}p^n \\ &= \frac{p^2(r+1)}{2}. \end{aligned}$$

The second last equality follows as the summands in the summation is the probability mass function of a negative binomial distribution with parameters $r+2$ and p. Hence, the sum is equals to one.

To simplify the left hand side, recall that $\sum_{m=0}^{\infty} P_i(m) = 0$ for $i \geq 1$. This means that $\sum_{m=0}^{\infty} P_1(m) = 0$. Then using $P_i(m) = P_{i,I}(m) + P_{i,A}(m)$ and the result $P_0(m+1) = \frac{1}{m+1} P_{1,A}(m)$ gives

$$\begin{aligned}
&\sum_{m=0}^{\infty} P_1(m) = \sum_{m=0}^{\infty} \Big(P_{1,I}(m) + P_{1,A}(m) \Big) = 0 \\
&\iff \\
&\sum_{m=0}^{\infty} P_{1,I}(m) = -\sum_{m=0}^{\infty} P_{1,A}(m) \\
&\iff \\
&\sum_{m=0}^{\infty} P_{1,I}(m) = -\sum_{m=0}^{\infty} (m+1) P_0(m+1) \\
&\iff \\
&\sum_{m=0}^{\infty} P_{1,I}(m) = \frac{rp}{p-1}.
\end{aligned}$$

Using this result, the left-hand side becomes

$$\frac{rp}{p-1} = \sum_{n=0}^{\infty} P_{1,I}(n) = P_{1,I}(0) + \sum_{n=0}^{\infty} P_{1,I}(n+1) \iff \sum_{n=0}^{\infty} P_{1,I}(n+1) = \frac{rp}{p-1} - a.$$

Substituting the above simplifications into (14) results in

$$\begin{aligned}
\frac{p-1}{rp}\Big(\frac{rp}{p-1} - a\Big) = \frac{a(p-1)}{rp(1-p)^r}\Big(1 - P_0(0)\Big) + (p-1)^2 \frac{rp}{1-p} + \frac{p^2(r+1)}{2} + \\
+ \frac{(p-1)^2}{r}\frac{rp}{1-p},
\end{aligned}$$

which upon rearranging of terms gives the solution to a.

8. Do the stationary probabilities fulfill the balance equations?

   Recall (1):

$$\begin{aligned}
(\gamma + m\delta) P_I(m) &= (m+1)\delta P_I(m+1) + \lambda P_A(m); \\
(\lambda + \mu + m\delta) P_A(m) &= (m+1)\delta P_A(m+1) + \mu P_A(m-1) + \gamma P_I(m).
\end{aligned}$$

   Zeroth order

   First equation:

$$\begin{aligned}
(\gamma + m\delta) P_0(m) + O(\epsilon) &= (m+1)\delta P_0(m+1) + \lambda \epsilon p (r+m) P_0(m) + O(\epsilon) \\
\delta r P_0(m) + m\delta P_0(m) + O(\epsilon) &= \delta p (r+m) P_0(m) + \delta(1-p)(r+m) P_0(m) + O(\epsilon) \\
\delta r P_0(m) + m\delta P_0(m) + O(\epsilon) &= \delta p (r+m) P_0(m) + \delta (r+m) P_0(m) - \delta p (r+m) P_0(m) + O(\epsilon) \\
\delta r P_0(m) + m\delta P_0(m) + O(\epsilon) &= \delta (r+m) P_0(m) + O(\epsilon). \text{ [balanced]}
\end{aligned}$$

Second equation:

$$\lambda\epsilon p(r+m)P_0(m) + \mu\epsilon p(r+m)P_0(m) + m\delta\epsilon p(r+m)P_0(m) = \\ = \delta\epsilon p^2(r+m+1)(r+m)P_0(m) + \mu\epsilon p(r+m-1)P_0(m-1) + \gamma P_0(m)$$

$$\delta(1-p)(r+m)P_0(m) + \delta p(r+m)P_0(m) + O(\epsilon) = \delta m P_0(m) + \delta r P_0(m) + O(\epsilon)$$

$$\delta(r+m)P_0(m) + O(\epsilon) = \delta(r+m)P_0(m) + O(\epsilon). \quad [\text{ balanced }]$$

First order

First equation:

$$(\gamma + m\delta)\epsilon P_{1,I}(m) + O(\epsilon^2) = (m+1)\delta\epsilon P_{1,I}(m+1) + \lambda\epsilon P_{1,A}(m) + \lambda\epsilon^2 P_{2,A}(m) + O(\epsilon^3).$$

LHS:

$$(\gamma + m\delta)\epsilon P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{m(m+1)}{2} - mrp(1-p)\Big] + O(\epsilon^2) = \\ = \delta(r+m)\epsilon P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{m(m+1)}{2} - mrp(1-p)\Big] + O(\epsilon^2).$$

RHS:

$$(m+1)\delta\epsilon P_0(m+1)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{(m+1)(m+2)}{2} - \\ - (m+1)rp(1-p)\Big] + \lambda\epsilon p(r+m)P_0(m) + \lambda\epsilon^2(m+1)P_1(m+1) + O(\epsilon^3) = \\ = \delta\epsilon p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{(m+1)(m+2)}{2} - \\ - (m+1)rp(1-p)\Big] + O(\epsilon^0) + \lambda\epsilon^2(m+1)P_0(m+1)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - \\ - p(1-p)\frac{(m+1)(m+2)}{2} - (m+1)rp(1-p) + p(r+m+1)\Big] + O(\epsilon^3) = \\ = \delta\epsilon p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{(m+1)(m+2)}{2} - \\ - (m+1)rp(1-p)\Big] + \delta\epsilon(1-p)(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - \\ - p(1-p)\frac{(m+1)(m+2)}{2} - (m+1)rp(1-p) + p(r+m+1)\Big] + O(\epsilon^2) = \\ = \delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{(m+1)(m+2)}{2} - \\ - (m+1)rp(1-p) + p(r+m+1)\Big] - \delta\epsilon p^2(r+m+1)(r+m)P_0(m) + O(\epsilon^2) = \\ = \delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{m(m+1)}{2} - mrp(1-p) - \\ - p(1-p)(r+m+1) + p(r+m+1)\Big] - \delta\epsilon p^2(r+m+1)(r+m)P_0(m) + O(\epsilon^2) = \\ = \delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big) - p(1-p)\frac{m(m+1)}{2} - mrp(1-p)\Big] + \\ + \delta\epsilon p^2(r+m+1)(r+m)P_0(m) - \delta\epsilon p^2(r+m+1)(r+m)P_0(m) + O(\epsilon^2). \quad [\text{ balanced }]$$

Second equation:

$$(\lambda+\mu+m\delta)\Big[\epsilon P_{1,A}(m)+\epsilon^2 P_{2,A}(m)\Big]+O(\epsilon^3)=(m+1)\delta\epsilon P_{1,A}(m+1)+\mu\Big[\epsilon P_{1,A}(m+1)+ \\ +\epsilon^2 P_{2,A}(m)\Big]+\gamma\epsilon P_{1,I}(m)+O(\epsilon^3).$$

LHS:

$$\lambda\epsilon p(r+m)P_0(m)+\mu\epsilon p(r+m)P_0(m)+m\delta\epsilon p(r+m)P_0(m)+ \\ +\lambda\epsilon^2 p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{(m+1)(m+2)}{2}- \\ -(m+1)rp(1-p)+p(r+m+1)\Big]+\mu\epsilon^2 p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)- \\ -p(1-p)\frac{(m+1)(m+2)}{2}-(m+1)rp(1-p)+p(r+m+1)\Big]+ \\ +m\delta\epsilon^2 p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{(m+1)(m+2)}{2}- \\ -(m+1)rp(1-p)+p(r+m+1)\Big]+O(\epsilon^3)= \\ =O(\epsilon^0)+O(\epsilon^0)+\delta\epsilon pm(r+m)P_0(m)+\delta\epsilon(1-p)(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)- \\ -p(1-p)\frac{(m+1)(m+2)}{2}-(m+1)rp(1-p)+p(r+m+1)\Big]+ \\ +\delta\epsilon p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{(m+1)(m+2)}{2}- \\ -(m+1)rp(1-p)+p(r+m+1)\Big]+O(\epsilon^2)= \\ =\delta\epsilon pm(r+m)P_0(m)+\delta\epsilon p(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)- \\ -p(1-p)\frac{(m+1)(m+2)}{2}-(m+1)rp(1-p)+p(r+m+1)\Big]+O(\epsilon^2).$$

RHS:

$$
\begin{aligned}
&(m+1)\delta\epsilon p(r+m+1)P_0(m+1)+\mu\epsilon p(r+m-1)P_0(m-1)+\\
&\quad+\mu\epsilon^2P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)+p(r+m)\Big]+\\
&\quad+\delta r\epsilon P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)\Big]+O(\epsilon^3)=\\
&=\delta\epsilon p^2(r+m+1)(r+m)P_0(m)+O(\epsilon^0)+\delta\epsilon mP_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-\\
&\quad-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)+p(r+m)\Big]+\delta r\epsilon P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-\\
&\quad-p(1-p)\frac{m(m+1)}{2}-mrp(1-p)\Big]+O(\epsilon^2)=\\
&=\delta\epsilon p^2(r+m+1)(r+m)P_0(m)+\delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-\\
&\quad-p(1-p)\frac{(m+1)(m+2)}{2}-(m+1)rp(1-p)+p(r+m+1)\Big]-\delta\epsilon mpP_0(m)+\\
&\quad+\delta\epsilon m(m+1)p(1-p)P_0(m)+\delta\epsilon mrp(1-p)P_0(m)-\delta\epsilon rp(r+m+1)P_0(m)+\\
&\quad+\delta\epsilon(m+1)rp(1-p)P_0(m)+\delta\epsilon r^2p(1-p)+O(\epsilon^2)=\\
&=\delta\epsilon p^2(r+m+1)(r+m)P_0(m)+\delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-\\
&\quad-p(1-p)\frac{(m+1)(m+2)}{2}-(m+1)rp(1-p)+p(r+m+1)\Big]-\\
&\quad-\delta\epsilon p^2(r+m+1)(r+m)P_0(m)+\delta\epsilon pm((r+m)P_0(m)-\delta\epsilon rp(r+m+1)P_0(m)+\\
&\quad+\delta\epsilon rp(m+1)P_0(m)+\delta\epsilon r^2pP_0(m)+O(\epsilon^2)=\\
&=\delta\epsilon(r+m)P_0(m)\Big[\frac{rp}{1-p}\Big(\frac{p(r+1)(2-p)-2}{2}\Big)-p(1-p)\frac{(m+1)(m+2)}{2}-\\
&\quad-(m+1)rp(1-p)+p(r+m+1)\Big]+\delta\epsilon pm((r+m)P_0(m)+O(\epsilon^2).\ [\text{ balanced }]
\end{aligned}
$$

9. Deriving the compound inequality (17) for m_1^*

From the first inequality,

$$
\begin{aligned}
\frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*+1)}&=\frac{\delta P_0(m_1^*)\Big[rp-m_1^*(1-p)\Big]}{\delta P_0(m_1^*+1)\Big[rp-(m_1^*+1)(1-p)\Big]}\\
&=\frac{(m_1^*+1)\Big[rp-m_1^*(1-p)\Big]}{p(r+m_1^*)\Big[rp-(m_1^*+1)(1-p)\Big]}>1.
\end{aligned}
$$

After some extensive algebra and simplification yields the following inequality

$$
|m_1^*-\frac{2rp-(1-p)}{2(1-p)}|<\frac{\sqrt{4rp+(1-p)^2}}{2(1-p)},
$$

which is the same as

$$
\frac{2rp-(1-p)}{2(1-p)}-\frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}<m_1^*<\frac{2rp-(1-p)}{2(1-p)}+\frac{\sqrt{4rp+(1-p)^2}}{2(1-p)},
$$

Recall earlier that it is assumed that both μ_1 and λ_1 are always positive, then $0<p<1$ must hold since $p=\frac{\mu_1}{\mu_1+\lambda_1}$. Consequently, there is no issue of divergence caused by

division with zero in the denominator. Furthermore, the following needs to hold in order for $m_1^* \geq 0$,

$$\frac{2rp-(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} > 0,$$

which, in turn, implies that $r > \frac{2-p}{p}$. This agrees well with the constraint that $p > 0$. In addition to the check above, the compound inequality needs be checked that both summands are smaller than $\frac{rp}{1-p}$ since the maximum current must happen before it hits zero. Upon checking, only the left summand is smaller than $\frac{rp}{1-p}$. Hence, for $\frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*+1)} > 1$,

$$m_1^* > \frac{2rp-(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}. \tag{24}$$

The upper bound of m_1^* must then come from the second current ratio inequality, which is

$$\begin{aligned}\frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*-1)} &= \frac{\delta P_0(m)\Big[rp-m(1-p)\Big]}{\delta P_0(m_1^*-1)\Big[rp-(m_1^*-1)(1-p)\Big]} \\ &= \frac{p(r+m-1)\Big]\Big[rp-m(1-p)\Big]}{m\Big[rp-(m_1^*-1)(1-p)\Big]} > 1.\end{aligned}$$

The following inequality is derived

$$|m_1^* - \frac{2rp+(1-p)}{2(1-p)}| > \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)},$$

which is the same as

$$m_1^* > \frac{2rp+(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} \quad \text{or} \quad m_1^* < \frac{2rp+(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}. \tag{25}$$

A meticulous check shows that the left summand in (25) is larger than $\frac{rp}{1-p}$ while the right summand is smaller than $\frac{rp}{1-p}$ and definitely positive when $r > \frac{2-p}{p}$. Besides, it has the correct inequality sign that is needed to constrain m_1^* within an interval. Finally, it is larger than the summand in (24). Hence, for $\frac{J_{I\to A}(m_1^*)}{J_{I\to A}(m_1^*-1)} > 1$, the right inequality in (25) holds.

Summarizing, for $r > \frac{2-p}{p}$ and $0 < p < 1$, the m_1^* that gives rise to maximum $J_{I\to A}(m)$ fulfills the following compound inequality

$$\frac{2rp-(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} < m_1^* < \frac{2rp+(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}.$$

10. Deriving the compound inequality (19) for m_2^*

The expressions for these ratios are the same as the ones for positive currents except for the reversal of the inequality sign. Hence, they are not repeated for the sake of brevity. The following inequality is yielded from the first inequality

$$|m_2^* - \Big(\frac{2rp-(1-p)}{2(1-p)}\Big)| > \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)},$$

which is the same as

$$m_2^* > \frac{2rp-(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} \quad \text{or} \quad m_2^* < \frac{2rp-(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}. \tag{26}$$

Recall from the earlier checks that the left summand in (26) is larger than $\frac{rp}{1-p}$. Hence, for $\frac{J_{I\to A}(m_2^*)}{J_{I\to A}(m_2^*+1)} < 1$, the left inequality in (26) is the valid one. Similar to the maximum current case, the upper bound for m_2^* must then come from the second current ration inequality, that is, $\frac{J_{I\to A}(m_2^*)}{J_{I\to A}(m_2^*-1)} < 1$. From this inequality, the following result is obtained

$$\frac{2rp+(1-p)}{2(1-p)} - \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} < m_2^* < \frac{2rp+(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}.$$

Again, from earlier checks, the right summand is the one that is larger than $\frac{rp}{1-p}$. Besides, it has the correct inequality sign that is needed to constrain m_2^* within an interval. Hence, for $\frac{J_{I\to A}(m_2^*)}{J_{I\to A}(m_2^*-1)} < 1$,

$$m_2^* < \frac{2rp+(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}.$$

Summarizing, the m_2^* that gives rise to minimum $J_{I\to A}(m)$ fulfills the following compound inequality

$$\frac{2rp-(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)} < m_2^* < \frac{2rp+(1-p)}{2(1-p)} + \frac{\sqrt{4rp+(1-p)^2}}{2(1-p)}.$$

11. A note about the bounds of (17) and (19)

    In the unlikely event that the bounds are integers, then there are no unique values for m_1^* and m_2^*. This implies that it will not be possible to know the exact value of m that give rise to the peak or the trough of the net probability current. However, as an approximation, one can consider the value of either the lower and upper bound with immaterial difference.

12. Deriving the compound inequality (20) for m^*

    From the first inequality,

$$\begin{aligned}\frac{J_{(A,m)\to(A,m+1)}(m^*)}{J_{(A,m)\to(A,m+1)}(m^*+1)} &= \frac{p(r+m^*)P_0(m^*)}{p(r+m^*+1)P_0(m^*+1)} \\ &= \frac{m^*+1}{p(r+m^*+1)},\end{aligned}$$

    which yields the following result

$$m^* > \frac{rp}{1-p} - 1.$$

    From the second inequality,

$$\begin{aligned}\frac{J_{(A,m)\to(A,m+1)}(m^*)}{J_{(A,m)\to(A,m+1)}(m^*+1)} &= \frac{p(r+m^*)P_0(m^*)}{p(r+m^*-1)P_0(m^*-1)} \\ &= \frac{p(r+m^*)}{m^*},\end{aligned}$$

    which yields the following result

$$m^* < \frac{rp}{1-p}.$$

    Combining the two results yields the following compound inequality

$$\frac{rp}{1-p} - 1 < m^* < \frac{rp}{1-p}.$$

6 References

[1] Peccoud, J. & Ycart., B. (1995). Markovian Modeling of Gene-Product Synthesis. Theoretical Population Biology, 48(2), 222 - 234.

[2] Gut, A. (2009). An Intermediate Course in Probability. Springer New York, NY.

[3] Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biol, 4, 1 - 13.

[4] Raj, A., & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell, 135(2), 216-226.

[5] Golding, I. Paulsson, J., Zawilski, S.M. & Cox, E.C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell, 123(6), 1025 - 1036.

[6] Blake, W.J., Kaern, M., Cantor, C. & Collins, J. (2003). Noise in eukaryotic gene expression. Nature, 422, 633 - 637.

[7] Raser, J.M. & O'Shea, E.K. (2004). Control of Stochasticity in Eukaryotic Gene Expression. SCIENCE, 304(5678), 1811 - 1814.

[8] Holmes, M.H. (2013). Introduction to Perturbation Methods. Springer New York, NY.

[9] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139?140. https://doi.org/10.1093/bioinformatics/btp616

[10] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol, 11, R106. https://doi.org/10.1186/gb-2010-11-10-r106

[11] Di, Y., Schafer, D., Cumbie, J. & Chang, J. (2011). The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology, 10(1). https://doi.org/10.2202/1544-6115.1637

[12] Vu, T. N., Wills, Q. F., Kalari, K. R., Niu, N., Wang, L., Rantalainen & M., Pawitan, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics (Oxford, England), 32(14), 2128-2135.

[13] Shi, W., Fornes, O. & Wasserman, W. W. (2019). Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants. Bioinformatics (Oxford, England), 35(15), 2610-2617.

[14] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. (2009). Stochasticity of gene products from transcriptional pulsing. Phys. Rev. E, 79, 031911.

[15] Paulsson, J. (2005). Review: Models of stochastic gene expression. Physics of Life Reviews, 2 (2), 157 - 175.

[16] Friedman, N., Cai, L. & Xie, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Physical review letters, 97(16), 168302.

[17] Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., & Linnarsson, S. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (New York, N.Y.), 347(6226), 1138-1142. https://doi.org/10.1126/science.aaa1934

[18] Aragón, J., Eberly, D., & Eberly, S. (1992). Existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution. Statistics & Probability Letters, 15(5), 375 - 379

[19] R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[20] Douglas, B., Martin, M., Ben, B. & Steve, W. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48

[21] Vahid Shahrezaei & Peter S. Swain. (2008). Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences, 105(45), 17256 - 17261

[22] Bokes, P., King, J. R., Wood, A. T., & Loose, M. (2012). Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. Journal of mathematical biology, 64(5), 829-854.

[23] Goldman, Samantha L., MacKay, M., Afshinnekoo, E., Melnick, Ari M., Wu, S., and Mason, Christopher E. (2019). The Impact of Heterogeneity on Single-Cell Sequencing. Frontiers in Genetics, 10. https://www.frontiersin.org/article/10.3389/fgene.2019.00008

[24] Harris, Kenneth D., Hochgerner, H., Skene, Nathan G., Magno, L., Katona, L., Bengtsson Gonzales, C., Somogyi, P., Kessaris, N., Linnarsson, S., & Hjerling-Leffler, J. (2018). Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. PLoS Biol, 16(6), 1-37. https://doi.org/10.1371/journal.pbio.2006387

[25] Ralston, A. & Shaw, K. (2008) Gene expression regulates cell differentiation. Nature Education 1(1):127

[26] Grün, D., Kester, L. & van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nature Methods, 11, 637-640. https://doi.org/10.1038/nmeth.2930

[27] Soneson, C. & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91. https://doi.org/10.1186/1471-2105-14-91

[28] Tang, X., Huang, Y., Lei, J., Luo, H., & Zhu, X. (2019). The single-cell sequencing: new developments and medical applications. Cell & Bioscience, 9, 53. https://doi.org/10.1186/s13578-019-0314-y