COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

DYNAMICAL MODELS IN GENE EXPRESSION

Dissertation Thesis

2018 **Mgr. Michal Hojčka**

COMENIUS UNIVERSITY IN BRATISLAVA

Faculty of Mathematics, Physics and Informatics

Department of Applied Mathematics and Statistics

DYNAMICAL MODELS IN GENE EXPRESSION

Dissertation Thesis

Mgr. Michal HOJČKA

9.1.9 Applied Mathematics

Applied Mathematics

Supervisor:

doc. Mgr. Pavol Bokes, PhD.

BRATISLAVA 2018

DYNAMICAL MODELS IN GENE EXPRESSION

Mgr. Michal Hojčka
E-mail: *michal.hojcka@fmph.uniba.sk*

doc. Mgr. Pavol Bokes, PhD.
E-mail: *pavol.bokes@fmph.uniba.sk*

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Mlynská dolina
842 48 Bratislava
Slovak Republic

Dissertation Thesis in 9.1.9 Applied Mathematics
Compilation date: April 25, 2018
Typeset in LaTeX

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

68167823

THESIS ASSIGNMENT

Name and Surname:	Mgr. Michal Hojčka
Study programme:	Applied Mathematics (Single degree study, Ph.D. III. deg., full time form)
Field of Study:	Applied Mathematics
Type of Thesis:	Dissertation thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Dynamical models in gene expression

Annotation: In this PhD thesis, we will use the methods of applied mathematical modelling of biological processes to study selected problems arising in the description of dynamical aspects of gene expression.

Tutor:	doc. Mgr. Pavol Bokes, PhD. (from 2015-12-17)
Department:	FMFI.KAMŠ - Department of Applied Mathematics and Statistics
Head of department:	prof. RNDr. Daniel Ševčovič, DrSc.
Assigned:	20.02.2014

Approved: 20.02.2014 prof. RNDr. Marek Fila, DrSc.
Guarantor of Study Programme

.....
Student

.....
Tutor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

68167823

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Mgr. Michal Hojčka
Študijný program: aplikovaná matematika (Jednoodborové štúdium, doktorandské III. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: dizertačná
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Dynamical models in gene expression
Dynamické modely v génovej expresii

Anotácia: V tejto PhD práci použijeme metódy aplikovaného matematického modelovania biologických procesov k štúdiu vybraných problémov vznikajúcich pri popise dynamických aspektov génovej expresie.

Školiteľ: doc. Mgr. Pavol Bokes, PhD. (od 17.12.2015)
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Spôsob sprístupnenia elektronickej verzie práce:
bez obmedzenia

Dátum zadania: 20.02.2014

Dátum schválenia: 20.02.2014

prof. RNDr. Marek Fila, DrSc.
garant študijného programu

.....
študent

.....
školiteľ

Abstract

We study the methods and approaches used in simulating systems of biochemical reactions. We present a stochastic model motivated by gene expression which includes production of protein molecules and their interactions with decoy binding sites. Then we formulate the associated Master equation. We focus on the distribution of free protein which cannot be expressed in a closed form. Therefore we present three different approaches to obtain it: employing singular perturbation reduction to obtain quasi-steady-state solution, simulating through stochastic algorithms and solving the associated system of ODEs. We also add large-system-size scaling to obtain statistical characteristics of free protein distribution like the Fano factor in a very simple form. We show that the Fano factor is greater than one for the intermediate levels of binding sites in contrast with Poissonian character (the Fano factor equals one) for no binding sites of their excess. In addition, we investigate the mRNA – microRNA system of reactions. Also here we derive quasi-steady-state solution and express the formula for the Fano factor in a closed form. It yields values below one for non-extreme levels of interaction strength. All results are supported and illustrated with the help of numerical simulations.

Keywords: gene expression • stochastic simulation • Master equation • singular perturbation

Acknowledgement

I would like to express my gratitude to my supervisor for help and time he gave me throughout the process of writing the thesis. I would also like to thank my family and friends for support.

Declaration on Word of Honour

I hereby declare this thesis was written on my own with using referred literature and help of my supervisor.

.....

Mgr. Michal Hojčka

UNIVERZITA KOMENSKÉHO V BRATISLAVE
Fakulta matematiky, fyziky a informatiky
Katedra aplikovanej matematiky a štatistiky

DYNAMICKÉ MODELY V GÉNOVEJ EXPRESII

Dizertačná práca

Mgr. Michal HOJČKA

9.1.9 aplikovaná matematika
Aplikovaná matematika

Školiteľ:
doc. Mgr. Pavol Bokes, PhD.

BRATISLAVA 2018

Abstrakt

V práci sa venujeme metódam a postupom používaných pri sledovaní systémov biochemických reakcií. Prezentujeme v nej stochastický model, ktorý je motivovaný génovou expresiou a zahrňuje tvorbu proteínov ako aj ich interakcie s falošnými väzobnými miestami. Uvádzame taktiež prislúchajúcu Master rovnicu. Snažíme sa vyjadriť pravdepodobnostné rozdelenie pre voľné proteíny, ktoré však nejde vyjadriť v uzavretom tvare. Preto použijeme tri rozličné spôsoby ako ho dostať: pomocou singulárnej perturbačnej redukcie nájsť riešenie v kvázistabilnom stave, cez stochastické simulačné algoritmy alebo vyriešením prislúchajúceho systému ODR. Pridaním podmienky pre veľký rozmer systému získame štatistické vlastnosti pre pravdepodobnostné rozdelenie voľných proteínov (ako napríklad Fano faktor) vo veľmi jednoduchej forme. Ukazuje sa, že pre strednú úroveň počtu väzobných miest je Fano faktor väčší ako jedna, na rozdiel od Poissonovského charakteru (čomu prislúcha Fano faktor rovný jednej) v prípadoch žiadného alebo extrémne veľkého počtu väzobných miest. Ďalej skúmame systém reakcií medzi mRNA a microRNA. Tiež vyjadríme riešenie v kvázistabilnom stave a nájdeme uzavretý výraz pre Fano faktor. Ten dosahuje hodnoty menšie ako jedna pre neextrémne úrovne interakčnej sily. Všetky výsledky dokumentujeme pomocou numerických simulácií.

Kľúčové slová: génová expresia • stochastické simulácie • Master rovnica • singulárna perturbácia

Contents

List of Figures

List of Tables

Introduction

According to [36], a gene is defined as a hereditary unit of DNA that is required to produce a functional product. This process of transforming the information from a gene to create further gene products is known as gene expression. The essential part in gene expression as well as in virtually every process on the cellular level is carried out by proteins, large biomolecular objects consisting mainly from the amino acids. In most mathematical models concerning gene expression, we focus on the first and the most important step of gene expression, the transcription. A special type of proteins, called transcription factors, possesses crucial functionality in this process as they activate transcription by binding to specific DNA sequences. The result of transcription is a primary RNA transcript; following the next steps of gene expression we ultimately obtain a functional protein. For further biological insight we refer to [28].

Biochemical reactions can be studied using a number of different mathematical formalisms and we can distinguish between the two main approaches. The first, deterministic, approach exploits deterministic ODE models to describe the dynamics of biochemical reactions. An alternative way to study biological systems is through stochastic models which consider each reaction as a single random event. The advantage of this approach is the fact that these models describe the behavior of the system well also at lower numbers of the involved species, as is the case for the number of proteins and other species present in the biological processes inside the cells such as gene expression [12, 48]. Therefore, deterministic modelling of such reactions can be quite inaccurate and we turn instead to stochastic methods [32]. As they work with discrete number of molecules, they can easily be simulated through stochastic simulation algorithms, in particular the Gillespie algorithm [19, 20].

Being a very timely topic, gene expression sparked a new wave of interest in Markovian models of chemical kinetics, e.g. [44]. In this thesis we present a simplified model in which we neglect the intermediary processes associated with

mRNA creation and focus solely on protein production. We assume that the protein is produced with a constant rate and that the rate of its decay is proportional to the number of proteins. We study the protein dynamics in presence of so-called decoy binding sites [52, 31] on the DNA. Our model further takes into account protein binding/unbinding reactions with these binding sites. Similar models have already been studied previously; in particular [18] investigated the model with protected complexes, i.e. the case when bound proteins were immune to degradation, showing that the steady-state distribution is Poissonian. Our model allows bounded proteins to degrade, which introduces additional noise into the model [5, 7]. For simplicity, we ignore effects of burst-like protein synthesis or transcriptional auto-regulation [5, 8, 46]. Unfortunately, as is often the case, the solution of free protein probability distribution cannot be obtained in a closed form. However, we can employ the fact that biochemical reactions often operate on different timescales [40, 47] to address the issue. History of applying these assumptions in stochastic modelling is rather new [9], but in recent years were thoroughly investigated in works such as [25, 23, 24]. Particularly, in the context of our model, the interactions between the protein and its binding sites occur on a substantially faster timescale than the production and decay of protein does [2]. Therefore we can successfully use singular perturbation methods [9, 38, 39] to obtain the quasi-steady-state solution to our problem.

Let us now go through the structure of the thesis. In the first chapter we summarise all the useful definitions from the fields of probability and differential equations, which we use in later parts of the thesis. In chapter two we focus on the theory regarding deterministic and stochastic modelling of biochemical reactions together with a couple of illustrative examples. Particularly, we present the example of deterministic approach together with the foundations of stochastic simulation of chemical reactions as well as numerical stochastic simulation algorithms used to simulate the system. The main results of the thesis follow afterward; most material of chapters three and four is also presented in the article [21]. In the third chapter, we derive the Master equation for our stochastic model and deploy a singular perturbation reduction to obtain quasi-steady-state approximations; this includes finding an equilibrium of binding/unbinding reaction, which is a specific case of a reversible bimolecular reaction studied by Laurenzi in [29]. We provide an alternative proof using mathematical induction. Using quasi-steady-state solution we are able to calculate the statistical properties and moments of the free protein distribution in the equilibrium. It follows that for large amounts of binding sites, the

free protein distribution is in fact Poissonian. Also, using three different methods we study the time evolution of the amount on both free and total protein and compare the results. In the fourth chapter we introduce the linear noise approximation, a tool with which the probability distribution can be obtained in the asymptotic case of large size of the system (as proposed in [49]). Using such an approximation, we derive the expression of Fano factor in a very simple form. Afterwards we investigate the quality of this approximation with respect to the results from the third chapter. In the last chapter we study another model also associated with gene expression, which considers the interactions between mRNA and microRNA molecules and the silencing effect on the population of mRNA. Information in this chapter are part of the article [6] which is currently submitted for publication.

CHAPTER 1

Preliminaries and Definitions

In this chapter we introduce a number of basic definitions and principles from the fields of probability and differential equations, which will be essential later in the thesis.

1.1 Probability

The case of stochastic simulation of chemical reaction is very closely connected to discrete probability distributions, as we simulate the numbers of molecules as they individually are. Therefore we work with integer amounts of molecules and use summation operations to obtain information about the distribution. In case of deterministic modelling or in specific asymptotic cases of stochastic models we deal mainly with the concentration of species and thus we also need to mention continuous probability distributions; for these we have to calculate integrals. Let us introduce some of the most common distributions and methods used and referred to later in the text. As the need to estimate parameters of the distribution from the sample observations will arise later in the text, we also mention the theory regarding the Maximum Likelihood Estimator (MLE). The information summarised here is based mostly on [10] and [22].

1.1.1 Discrete probability distributions

Let us consider a random process, such that its particular events occur with a known constant rate and independently of the time since the previous event. Examples of such processes can be customers arriving to the desk, dysfunctions of a server or in our case chemical reactions. The number of occurrences of a chosen phenomenon in a time interval is then based on *Poisson distribution* and such process if often

referred to as Poisson process. This distribution is parametrized by a rate parameter λ which also represents the average number of events in unit time interval. A random variable X, taking values in the set of non-negative integers, has a Poisson distribution if

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, \dots.$$

The first and second moments of X are $E(X) = Var(X) = \lambda$. In the text we often use another notation, often used in the field of mathematical biology, referring to the expected value of X as $\langle X \rangle$. The Poisson distribution arises in the non-bursting case of protein production.

Perhaps the most trivial case of probability distribution is the *Bernoulli distribution*, which is based on the Bernoulli trial, an experiment with only two possible outcomes

$$X = \begin{cases} 1 \text{ (success)}, & \text{with probability p}, \\ 0 \text{ (failure)}, & \text{with probability (1-p)}, \end{cases} \qquad 0 \leq p \leq 1.$$

Two widely used discrete probability distributions are based on Bernoulli trials.

At first, let us define a random variable, which tells us how many successes we get out of n identical independent Bernoulli trials with probability p. We say that such variable has *binomial distribution* with parameters p, n and probability mass function

$$P(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \dots, n.$$

The first two moments of binomial distribution are given by $\langle X \rangle = np$ and $Var(X) = np(1-p)$.

Secondly, let us define another random variable as the number of Bernoulli trials necessary to get a fixed number of successes. It is known as the *negative binomial distribution* and depends on two parameters, the probability of Bernoulli trials p and the number of successes r we require to happen. We can write

$$P(X = x|r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots.$$

Sometimes we can also use an alternative expression and model the number of failures Y before the r-th success instead. In this case we realize the relation $Y = X - r$ hold, obtaining

$$P(Y = y|r, p) = \binom{r+y-1}{y} p^r (1-p)^y, \qquad y = 0, 1, \dots.$$

The moments of the negative binomial distribution (case with the specified number of failures Y) are as follows: $\langle Y \rangle = r\frac{(1-p)}{p}$ and $Var(Y) = r\frac{(1-p)}{p^2}$. The special case of negative binomial distribution when $r = 1$ is called the *geometric distribution*. We meet this distributions often in cases when the production of some chemical species occurs in a bursting regime.

For the next distribution, suppose we have a box with N identical balls, except that M are red and $N - M$ are green. We blindly reach in and draw K balls at random, without replacement. We would like to express the probability that exactly x of the balls are red. If we denote the number of red balls in a sample of size K as X, then random variable X has a *hypergeometric distribution* given by

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \qquad x = 0, 1, \ldots, K.$$

The moments of hypergeometric functions can be obtained (after quite tedious calculations) as $\langle X \rangle = \frac{KM}{N}$ and $Var(X) = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right)$.

A very powerful tool used in the field of discrete probability is the method of *generating functions* [26]. The generic generating function for a sequence $\{g_0, g_1, \ldots\}$ has the form

$$G(s) = \sum_{n \geq 0} g_n s^n.$$

In case of *probability generating function (PGF)* we can replace g_k by $P(X = k)$ and write

$$G(s) = \sum_{n \geq 0} P(X = n) s^n.$$

If we know the functional form of generating function G, we can expand the formula using Taylor expansion in terms of s and obtain the probability $P(X = k)$ as the coefficient in front of s^k, i.e.

$$P(X = k) = \frac{1}{k!}\frac{\mathrm{d}^k G(s)}{\mathrm{d}s^k}\bigg|_{s=0}.$$

One of the applications of the PGF is the computation of various statistical moments of a given distribution, mainly the following:

$$\begin{aligned}
\mu'_r &= \langle X^r \rangle \qquad \text{(r-th uncorrected moment)}, \\
\mu_r &= \langle (X - \langle X \rangle)^r \rangle \qquad \text{(r-th central moment)}, \\
\mu_{(r)} &= \langle X!/(X - r)! \rangle \qquad \text{(r-th factorial moment)}.
\end{aligned}$$

From the computational point of view, most of the time it is easiest to calculate the factorial moments of distribution from the corresponding PGF using the formula

$$\mu_{(r)} = \left. \frac{\mathrm{d}^r G(s)}{\mathrm{d}s^r} \right|_{s=1}.$$

Most used statistical moments, the mean and the variance, can be expressed from the factorial moments as follows:

$$\begin{aligned} \langle X \rangle &= \mu_{(1)}, \\ Var(X) &= \mu_{(2)} + \mu_{(1)} - \mu_{(1)}^2. \end{aligned} \tag{1.1}$$

Probability generating functions are also a strong tool when dealing with the convolutions of random variables. For the sum of random variables X and Y we have

$$G_{X+Y}(s) = G_X(s) G_Y(s), \qquad \text{if X and Y are independent.} \tag{1.2}$$

Generating functions for the most used probability distributions are well-known; let us summarize them briefly:

- Poisson distribution: $G(s) = e^{\langle X \rangle (s-1)}$
- Bernoulli distribution: $G(s) = (1-p) + ps$
- Binomial distribution: $G(s) = ((1-p) + ps)^n$
- Negative binomial distribution: $G(s) = \left(\frac{ps}{1-(1-p)s} \right)^r$, $|s| < \frac{1}{1-p}$
- Hypergeometric distribution: $G(s) = \frac{\binom{N-M}{K} {}_2F_1(-K,-M;N-M-K+1;s)}{\binom{N}{K}}$

Results for binomial distribution can easily be verified combining Bernoulli distribution and (1.2). More information about the hypergeometric function ${}_2F_1$ from the last PGF together with hypergeometric functions in general form is outlined in Section 1.2.2.

1.1.2 Continuous probability distributions

The *gamma family of distributions* is a flexible family defined on $[0, \infty)$ and is closely related to gamma function, which is defined for a positive constant α as the integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

For the most values of α we cannot find closed form of gamma function, but it satisfies many useful relationships, in particular,

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha). \tag{1.3}$$

Combining (1.3) with the trivial case $\Gamma(1) = 1$, we get for any integer $n > 0$,

$$\Gamma(n) = (n-1)!.$$

Using (1.3) we obtain recursion relation which allows us to calculate value of any gamma function from knowing the value $\Gamma(c),\ \ 0 < c \leq 1$.

The full gamma family is characterized by two parameters, α and β; the probability density function (PDF) of gamma(α, β) distribution is given as

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{x/\beta},\ \ 0 < x < \infty,\ \ \alpha > 0,\ \ \beta > 0. \tag{1.4}$$

The parameter α is known as the shape parameter and β is often called the scale parameter of the distribution. The mean of the distribution (1.4) can be calculated as $\langle X \rangle = \alpha\beta$ and its variance as $Var(X) = \alpha\beta^2$.

Choosing particular values for α and β, we arrive at various special cases of the gamma family. If we fix $\alpha = 1$, we get the so-called *exponential distribution*. This distribution is closely related to Poisson processes as it represents the distribution of an expected time interval between two random events in such a process. The probability density function (PDF) of an exponential distribution has the form

$$f(X|\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & ,\ \ x \geq 0 \\ 0 & ,\ \ x < 0. \end{cases}$$

The parameter β can be represented as λ^{-1}, where λ is the rate of the corresponding Poisson process. It follows that the moments of an exponential distribution are $\langle X \rangle = \beta = \lambda^{-1}$ and $Var(X) = \beta^2 = \lambda^{-2}$.

Other special cases of the gamma family are the *chi-squared distribution* ($\alpha = p/2,\ \ p \in \mathbb{N}$, and $\beta = 2$) and *Erlang distribution* ($\alpha \in \mathbb{N}$).

The *normal distribution* (sometimes called also the Gaussian distribution) is a continuous probability distribution with two parameters μ and σ^2, which are also its mean and variance. The PDF of a normal distribution with mean μ and variance σ^2, denoted as $\mathcal{N}(\mu, \sigma^2)$, is given by the formula:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}.$$

One of the many situations where normal distribution comes into play is the random noise in the stochastic processes, Brownian motion, or diffusion.

We can describe the time evolution of the probability density function of various phenomena in chemical or physical mathematics by a partial differential equation, which is commonly known as the *Fokker-Planck equation*. In the field of stochastic processes, one of the most basic examples is the so-called *Ornstein-Uhlenbeck process*, which is defined as follows:

$$dX_t = -aX_t \mathrm{d}t + \sigma \mathrm{d}W_t,$$

where W_t is a standard Wiener process (see [37]) and $0 < a < 1$. The Fokker-Planck equation corresponding to this process has the form

$$\frac{\partial p(x,t)}{\partial t} = a\frac{\partial}{\partial x}(xp(x,t)) + \frac{\sigma^2}{2}\frac{\partial^2 p(x,t)}{\partial x^2}.$$

This Fokker-Planck equation is linear in p and autonomous, as a and σ does not change over the course of time and therefore we already know the stationary solution $\partial_t p = 0$ for the probability distribution, which is Gaussian and given by

$$p(x,t) = \sqrt{\frac{a}{\pi\sigma^2}} e^{-\frac{ax^2}{\sigma^2}}.$$

1.1.3 Maximum Likelihood Estimator

We use the method of maximum likelihood as it is one of the most popular techniques to obtain a parameter estimator from an observed random sample. Let us assume we observe an IID (independent and identically distributed) sample $X_1, X_2, \ldots, X_n$ from a population with PDF or PMF $f(x|\theta_1, \theta_2, \ldots, \theta_n)$. Then the corresponding likelihood function is defined by

$$L(\theta|\mathbf{x}) = L(\theta_1, \theta_2, \ldots, \theta_n | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, \ldots, \theta_n).$$

The log-likelihood function $l(\theta|\mathbf{x}) = (\ln L(\theta|\mathbf{x}))$ is often used instead. For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value in which $L(\theta|\mathbf{x})$ (or $l(\theta|\mathbf{x})$) attains its maximum as a function of θ with $\mathbf{x}$ kept fixed. A maximum likelihood estimator (MLE) of the parameter θ based on sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

Let us now derive the MLE for the sample from the Poisson distribution. For n independent observations drawn from the Poisson distribution, we can write likelihood function as the product of individual PMF's as

$$L(\lambda|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!},$$

and thus the log-likelihood function as

$$l(\lambda|x_1, x_2, \ldots, x_n) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!).$$

The first-order condition for the maximum is

$$\frac{\partial}{\partial\lambda} l(\lambda|x_1, x_2, \ldots, x_n) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0,$$

which implies a solution

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

It is nothing else than the basic sample mean. This made intuitive sense as expected value of a random Poisson variable equals to the value of λ parameter of Poisson distribution.

1.1.4 Statistical distance

When we try to approximate one probability distribution with another, important question is how to evaluate the quality of the approximation as some kind of 'distance' between these two distributions. There exist a variety of such functions. Some statistical distance measures are not metrics and they need not be symmetric. We are interested in measuring distance between discrete distributions and therefore we focus on formulas for such distributions. The continuous case can be derived analogously by changing the sums to integrals. Let us present a brief introduction to the most popular measures.

- Total variation distance: $\delta(P, Q) = \max_i(P(i) - Q(i))$.
- Kullback–Leibler divergence: $D_{KL}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$.
- Hellinger distance: $H(P, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{P(i)} - \sqrt{Q(i)}\right)^2}$.
- Rényi's divergence: $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \ln\left(\sum_i \frac{P(i)^\alpha}{Q(i)^{\alpha-1}}\right)$.
- Jensen–Shannon divergence: $JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(M\|Q)$, where $M = \frac{1}{2}(P + Q)$.
- Bhattacharyya distance: $D_B(P, Q) = -\ln\left(\sum_i \sqrt{P(i)Q(i)}\right)$.

In our thesis we use Bhattacharyya distance [4] as the measure of statistical distance for its symmetry and relative simplicity (but a better descriptive power than Total variation distance). As expected, $D_B = 0$ implies identical distributions and $D_B = 1$ implies mutually exclusive distributions.

1.1.5 Law of total variance

This law is also sometimes refereed to as Eve's law or variance decomposition formula. It states that if X and Y are random variables on the same probability space with the requirements of finite variance for Y, then the following formula holds:

$$Var(Y) = E\left(Var(Y|X)\right) + Var\left(E(Y|X)\right).$$

In other words, total variance can be expressed as the sum of the expectation of conditional variances and the variances of conditional expectations. Its proof uses another useful law, law of total expectation and both can be found in [50].

1.2 Differential equations

Let us summarise some basic methods of solving ordinary differential equations (ODEs) and partial differential equations (PDEs), which will arise later in this thesis. Theory for the PDE part is based on [41] and in the hypergeometric ODE part we cite [22].

1.2.1 Quasilinear PDE

In this part we present a method to solve linear and quasilinear PDE as we use them in the next chapters.

We call a linear homogenous PDE an equation in the form

$$a_1(x)\frac{\partial u}{\partial x_1} + a_2(x)\frac{\partial u}{\partial x_2} + \ldots + a_n(x)\frac{\partial u}{\partial x_n} = 0, \tag{1.5}$$

where $a_1, a_2, \ldots, a_n : \ \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}$ are given continuous functions. The problem consists of finding a C^1 smooth function $u : \ \Omega \subset \mathbb{R}^n \mapsto \mathbb{R}$ such that in every point of Ω (1.5) holds. Finding a solution of (1.5) is based on the so-called characteristic system, which is a system of ordinary differential equations in $\mathbb{R}^n$ in the form

$$\begin{cases} \dot{x}(\tau) = \vec{a}(x(\tau)), \quad \tau \in \mathbb{R}, \\ x(0) = x^0 \in \mathbb{R}^n. \end{cases} \tag{1.6}$$

The function u is a solution of the system (1.5) if and only if u is constant on every characteristic, i.e. on every solution of characteristic system (1.6).
Now, let us consider a general quasilinear first-order equation, which has the form

$$a_1(x,u)\frac{\partial u}{\partial x_1} + a_2(x,u)\frac{\partial u}{\partial x_2} + \ldots + a_n(x,u)\frac{\partial u}{\partial x_n} = a_{n+1}(x,u), \tag{1.7}$$

with the equivalent conditions for a and u as in (1.5). The idea of solving quasilinear equation (1.7) is in the construction of an auxiliary linear homogenous equation in the form

$$a_1(x,u)\frac{\partial w}{\partial x_1} + a_2(x,u)\frac{\partial w}{\partial x_2} + \ldots + a_n(x,u)\frac{\partial w}{\partial x_n} + a_{n+1}(x,u)\frac{\partial w}{\partial u} = 0. \tag{1.8}$$

A link between equations (1.7) and (1.8) can be epitomised in the following conditions.

a) $w(x,u(x)) = const$ for each $x \in \Omega \subset \mathbb{R}^n$,

b) $\frac{\partial w}{\partial u}(x,u(x)) \neq 0$ for each $x \in \Omega \subset \mathbb{R}^n$.

If these two requirements are fulfilled, then u is a solution of equation (1.7) on the set Ω. Further information, such as proofs of the mentioned propositions, together with some other useful information about different types of PDEs and their solution methods can be found in [41]. For more extensive information about PDEs we refer to [16].

1.2.2 Hypergeometric functions and ODEs

A hypergeometric series can be formally defined as a power series

$$\beta_0 + \beta_1 z + \beta_2 z^2 + \ldots = \sum_{n\geq 0} b_n z^n,$$

in which the ratio of successive coefficients is a rational function of n. That is,

$$\frac{\beta_{n+1}}{\beta_n} = \frac{A(n)}{B(n)},$$

where $A(n)$ and $B(n)$ are polynomials in variable n.

Let us consider the most simple example, the series for the exponential function

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots = \sum_{n\geq 0} \frac{z^n}{n!}. \tag{1.9}$$

In series (1.9) we have chosen $\beta_n = \frac{1}{n!}$, which implies $\frac{\beta_{n+1}}{\beta_n} = \frac{1}{n+1}$ and thus $A(n) = 1$ and $B(n) = n + 1$. Series (1.9) can be denoted as hypergeometric function ${}_0F_0(z)$.

In generalised form, hypergeometric function ${}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z)$ has the form

$$ {}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z) = 1 + \frac{a_1 \cdots a_p}{b_1 \cdots b_q} \cdot \frac{z}{1!} + \frac{a_1(a_1+1) \cdots a_p(a_p+1)}{b_1(b_1+1) \cdots b_q(b_q+1)} \cdot \frac{z^2}{2!} + \ldots. \tag{1.10} $$

Let us define the notation for the rising factorial or Pochhammer symbol as

$$ \begin{aligned} &(a)_0 = 1, \\ &(a)_n = a(a+1)(a+2) \cdots (a+n-1), \qquad n \geq 1. \end{aligned} \tag{1.11} $$

Using (1.11) we can rewrite (1.10) as

$$ {}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z) = \sum_{n \geq 0} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n} \cdot \frac{z^n}{n!}. $$

The history of hypergeometric functions can be traced back to 1769 and the work of Euler, [14], where he mentioned the hypergeometric differential equation in the form

$$ x(1-x)\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + (c - (a+b+1)x)\frac{\mathrm{d}y}{\mathrm{d}x} - aby = 0, $$

which has three regular singular points: 0, 1 and ∞. A solution to this equation is the hypergeometric function ${}_2F_1(a, b; c; z)$, which was introduced by Gauss [17] and is often refereed to as the Gaussian hypergeometric function.

In the general form, ${}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z)$ is the regular solution of the differential equation in the form

$$ x(\theta + a_1)(\theta + a_2) \ldots (\theta + a_p)y = \theta(\theta + b_1 - 1)(\theta + b_2 - 1) \ldots (\theta + b_q - 1)y, \tag{1.12} $$

where θ stands for differential operator in the form $x\frac{\mathrm{d}}{\mathrm{d}x}$.

CHAPTER 2

Biochemical Reactions

In this chapter we present two main approaches to the study of biochemical reactions. The first, deterministic, approach works mainly with the concentrations of the chemical species in the system. The second, stochastic, approach works with the discrete number of molecules and seeks to find the probability distribution as a function of time for the given species.

2.1 Deterministic approach

Let us illustrate the basic methods and tools used in this section of mathematical biology on the one of the most basic enzymatic reactions, first mentioned by Michaelis and Menten in 1913 [33] and studied in many works including [27] and [11]. We also introduce here the idea of different timescales and quasi-steady-state approximation. Our main source for this part is [34], which provides an extensive introduction into this field.

The reaction of our interest involves substrate S, enzyme E; together they form a complex SE, which can be converted into the product P plus the enzyme E.

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} SE \xrightarrow{k_2} P + E. \tag{2.1}$$

A both-ways arrow symbolises the fact that the first reaction is reversible; the second reaction can only go forward, which is represented by a single arrow. Parameters k's are rate constants associated with the individual reactions. We call E a catalyst as it is conserved in the overall course of the reaction. Our goal is to write down differential equations describing the kinetics of system (2.1). We use the *Law of Mass Action*, which says that the rate of a given reaction is proportional to the product of the concentrations of the reactants. A conventional way how to refer to

a concentration is using []; for the sake of brevity let us use lowercase letters to denote the concentrations (we use lowercase letters to denote concentrations also in the further parts of this thesis):

$$s = [S], \qquad e = [E], \qquad c = [SE], \qquad p = [P].$$

Applying the Law of Mass Action to (2.1) we obtain the system of nonlinear reaction equations (one for each reactant):

$$\begin{aligned} \dot{s} &= -k_1 es + k_{-1} c, \\ \dot{e} &= -k_1 es + (k_{-1} + k_2) c, \\ \dot{c} &= k_1 es - (k_{-1} + k_2) c, \\ \dot{p} &= k_2 c, \end{aligned} \tag{2.2}$$

where $\dot{x}$ refers to the derivation with respect to time $\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)$ as usual. To complete the formulation of the problem, we have to provide initial conditions for the system. Let us say we start with non-zero amounts of enzyme E and substrate S only, i.e.

$$s(0) = s_0, \qquad e(0) = e_0, \qquad c(0) = 0, \qquad p(0) = 0. \tag{2.3}$$

The solution of (2.2) together with (2.3) gives us the concentrations as the functions of time. We have to take into account also obvious non-negativity condition for the concentrations.

The last equation is easy to solve: we can express the concentration of product as

$$p(t) = k_2 \int_0^t c(\tau) d\tau,$$

which implies that as soon as we know $c(t)$, we can compute also $p(t)$, and thus we need to solve just first three equations. As we mentioned earlier, enzyme E acts as catalyst in the system (2.1) and thus its concentration should be constant during the process. If we realize that E can exist either in free state (E) or in bounded state (SE), we can write the conservation law for its concentration as

$$\dot{e} + \dot{c} = 0, \qquad \text{i.e.} \qquad e + c = const = e_0.$$

Applying this, the system of ODEs is reduced to only two equations, let us take s and c as our independent variables and rewrite the system as

$$\begin{aligned} \dot{s} &= -k_1 e_0 s + (k_1 s + k_{-1}) c, \\ \dot{c} &= k_1 e_0 s - (k_1 s + k_{-1} + k_2) c, \end{aligned} \tag{2.4}$$

subject to the initial conditions

$$s(0) = s_0, \qquad c(0) = 0.$$

An usual approach to solving chemical reactions utilises the assumption that one 'fast' reaction is essentially always at equilibrium. This procedure is referred to as the pseudo- or quasi-steady-state approximation. In this case, we set $\mathrm{d}c/\mathrm{d}t \approx 0$; using this fact together with (2.4) yields

$$c(t) = \frac{e_0 s(t)}{s(t) + K_m}, \tag{2.5}$$

where

$$K_m = \frac{k_{-1} + k_2}{k_1}$$

is called the Michaelis constant. If we substitute (2.5) back to (2.4) we obtain

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{k_2 e_0 s}{s + K_m}.$$

Since the amount of enzyme is considered to be small compared to the amount of substrate we can assume $s_0/e_0 = \varepsilon \ll 1$ and obtain an implicit solution for the substrate:

$$s(t) + K_m \ln s(t) = s_0 + K_m \ln s_0. \tag{2.6}$$

We can easily check that (2.5) does not satisfy the condition $c(0) = 0$. We can distinguish between two timescales in the system, the initial timescale near $t = 0$ when the level of complex rises quickly and the longer timescale when the amount of complex is well approximated by (2.5) with $s(t)$ determined by (2.6). Therefore we consider it to be a reasonable approximation of the solution and mainly focus on situations in which we can use it.

2.2 Stochastic approach

Deterministic approach works with the concentration of individual chemical species, which assumes that system size is large. On the contrary, truth is that the number of proteins expressed from a single gene can often be quite small [53]; it would therefore be inaccurate to treat reactants as continuous variables as in deterministic approach. Answer to this problem is stochastic approach, in which each species is modeled as a discrete random variable and each reaction as a random event with a given probability of occurrence. It is not even uncommon to observe even less than 10 copies of protein species per *E. coli* cell [48], therefore we work with mean

number of proteins below 100 in this thesis. As the amount of species is represented with discrete variables it is possible to simulate these reactions through simulation algorithms (special versions of Monte Carlo simulations) to approximate probability distribution in a given time. These simulations were pioneered by Joseph L. Dool as early as around 1945, but only gained proper recognition after publishing of the paper by Gillespie [19] in 1976; therefore they are usually referred to as Gillespie algorithm. The same author next year (1977) in [20] used the algorithm to simulate systems of biochemical reactions. A nice summary with the focus on practical approach was given in [13], which is also one of the main sources for this section together with [49]. It contains exhaustive amount of information about chemical reactions, enzyme kinetics and related stochastic theory.

2.2.1 Simple case

First, let us consider the most simple case of reaction, degradation of chemical species A,

$$A \xrightarrow{k} \emptyset. \tag{2.7}$$

The symbol $\emptyset$ is often used to represent the chemical species which are outside of our interest. The letter k stands for the rate constant of the reaction and it is defined as following: probability, that a randomly chosen molecule of species A reacts during the time period $[t, t + \mathrm{d}t)$ equals $k\mathrm{d}t$. Therefore the probability that exactly one reaction happens in the whole amount of species A in time interval $[t, t+\mathrm{d}t)$ is $A(t)k\mathrm{d}t$, where $A(t)$ represents the number of molecules A in the system at time t. The probability that two or more reactions occur during the time period $[t, t+\mathrm{d}t)$ is $o(\mathrm{d}t^2)$ and thus can be disregarded as 0. A naive approach to numerically simulate this problem would be to choose sufficiently small time step Δt, generate a random number r, uniformly distributed in $(0, 1)$ and if $r < A(t)k\Delta t$, then reaction occurs and $A(t + \Delta t) = A(t) - 1$. It comes with obvious defects that in most of the time steps Δt no reaction occurs, so we generate a tremendous amount of random numbers for no reason at all.

A much better approach would be to always determine the time $t + \tau$, when the next reaction takes place. Let us first denote the probability that the next reaction occurs during time interval $[t + s, t + s + \mathrm{d}s)$ as $f(A(t), s)\mathrm{d}s$, ($\mathrm{d}s$ is infinitesimally small). Furthermore, let us define the probability that no reaction takes place in period $[t, t + s)$ as $g(A(t), s)$. These two expressions are connected to each other by

the following formula:

$$f(A(t), s)\mathrm{d}s = g(A(t), s)A(t + s)k\mathrm{d}s.$$

As no reaction takes place between t and $t + s$, it implies that $A(t + s) = A(t)$ and we can proceed with

$$f(A(t), s)\mathrm{d}s = g(A(t), s)A(t)k\mathrm{d}s. \tag{2.8}$$

Since we want to better understand the term $g(A(t), s)$, let us take $\sigma > 0$. As no reaction happens in $[t, t+\sigma+\mathrm{d}\sigma)$, we know that no reaction can happen in $[t, t+\sigma)$ and also in $[t + \sigma, t + \sigma + \mathrm{d}\sigma)$. We can write

$$g(A(t), \sigma + \mathrm{d}\sigma) = g(A(t), \sigma)(1 - A(t)k\mathrm{d}\sigma)$$

and with some rearrangements we obtain

$$\frac{g(A(t), \sigma + \mathrm{d}\sigma) - g(A(t), \sigma)}{\mathrm{d}\sigma} = -A(t)kg(A(t), \sigma).$$

Now if we take the limit of $\mathrm{d}\sigma \to 0$, we arrive at the following ordinary differential equation

$$\frac{\mathrm{d}g(A(t), \sigma)}{\mathrm{d}\sigma} = -A(t)kg(A(t), \sigma),$$

which is subject to the obvious initial condition $g(A(t), 0) = 1$. This ODE has a solution

$$g(A(t), \sigma) = e^{-A(t)k\sigma}.$$

We can substitute it into (2.8) and we obtain

$$f(A(t), s) = A(t)ke^{-A(t)k\sigma}\mathrm{d}s. \tag{2.9}$$

To change the problem of generating time $\tau \in (0, \infty)$, when the next reaction (2.7) occurs, into the problem of generating a uniformly distributed number in the interval $(0, 1)$, let us consider the function F, defined as

$$F(\tau) = e^{-A(t)k\tau}. \tag{2.10}$$

To prove that, the probability that $F(\tau) \in (a, b)$ has to be equal to the probability that $\tau \in (F^{-1}(b), F^{-1}(a))$. Using (2.9) and (2.10) we get

$$\int_{F^{-1}(b)}^{F^{-1}(a)} f(A(t), s)\mathrm{d}s = \int_{F^{-1}(b)}^{F^{-1}(a)} A(t)ke^{-A(t)ks}\mathrm{d}s$$

$$= -\int_{F^{-1}(b)}^{F^{-1}(a)} \frac{\mathrm{d}F}{\mathrm{d}s}\mathrm{d}s = -F(F^{-1}(a)) + F(F^{-1}(b)) = b - a$$

which proves the statement. It implies that we can generate the time step τ by putting $r = F(\tau)$, where r is random number, uniformly generated in $(0, 1)$. Using (2.10), we can write

$$r = e^{-A(t)k\tau}$$

and solving the equation for τ, we obtain

$$\tau = \frac{1}{A(t)k} \ln\left(\frac{1}{r}\right). \tag{2.11}$$

Stochastic simulation algorithm for the reaction (2.7) can be written down in the following steps we perform in each time t:

1) Generate a random number r uniformly distributed in $(0, 1)$.

2) Time, when the next reaction happens is given by $t + \tau$, where τ is computed using (2.11).

3) Number of molecules $A(t + \tau)$ is given by $A(t) - 1$.

Three realizations of the provided algorithm with initial conditions $A(0) = 20$ and $k = 0.1$ are illustrated in the Figure 2.1.

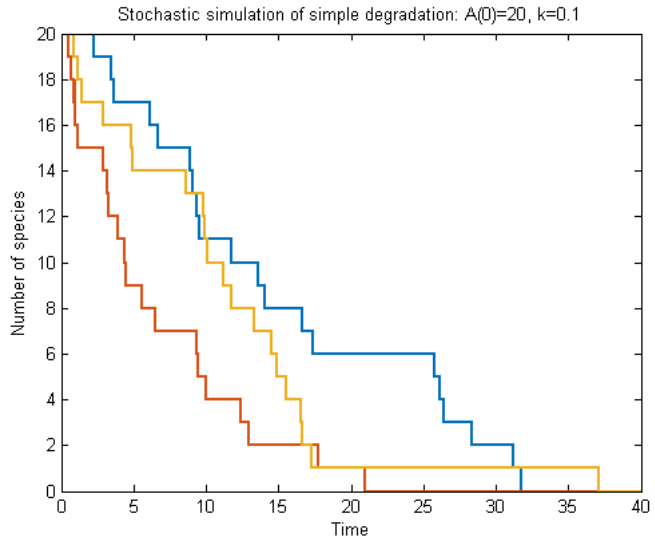

Figure 2.1: *Simulated trajectories of degradation using Gillespie algorithm.*

Let us now take a look on the probability distribution of chemical species A at the time t, let us denote $P(A(t) = n)$ as $P_n(t)$ for the sake of simplicity. There are

two possible ways, how $A(t+\mathrm{d}t)$ can be equal to n, either $A(t) = n$ and no reaction happened in the period $[t, t+\mathrm{d}t)$, or $A(t) = n+1$ and a degradation reaction occurred in the period $[t, t+\mathrm{d}t)$. The probability that two or more reactions occurred in the period $[t, t+\mathrm{d}t)$ can be neglected. We can express these observations in the form of equation as

$$P_n(t+\mathrm{d}t) = P_n(t) \cdot (1 - kn\mathrm{d}t) + P_{n+1}(t) \cdot k(n+1)\mathrm{d}t,$$

which can be rewritten as

$$\frac{P_n(t+\mathrm{d}t) - P_n(t)}{\mathrm{d}t} = P_{n+1}(t) \cdot k(n+1) - P_n(t) \cdot kn.$$

Sending $\mathrm{d}t \to 0$ we obtain a set of ODEs for the probabilities in the form

$$\dot{P}_n(t) = k(n+1)P_{n+1}(t) - knP_n(t). \tag{2.12}$$

The set of equations (2.12) is called a *chemical Master equation* or simply a *Master equation* of reaction (2.7). The term *Master equation* is very old as it first appeared in the paper [35] from 1940 in which it had a function of main equation from which all other results were derived.

Let us introduce a couple of techniques to obtain more information about the probability distribution of A given by P. If we introduce initial condition $A(0) = n_0$ (resp. $P_n = \delta_{n,n_0}$ where δ stands for Kronecker delta) to the system, we can set $P_n = 0$ for all $n > n_0$ as only degradation is possible. This will leave us with the marginal term in the Master equation in the form

$$\dot{P}_{n_0}(t) = -kn_0P_{n_0}(t),$$

which together with the initial conditions yields the result

$$P_{n_0}(t) = e^{-kn_0t}.$$

With known value of P_{n_0} it is possibly to explicitly express the value of P_{n_0-1} and we can continue with this algorithm all the way to P_0, until all P_n's are known. This naive way of solving can only be used if no production of A is present. As soon as we introduce it to the system, we can no longer find an upper bound for possible value of A and thus the Master equation becomes an infinite set of ODEs.

Another possible approach can be used if we are not necessarily interested in whole distribution of A, but we care for its statistical moments, in particular the

mean and the variance; let us denote them as $M(t)$ and $V(t)$ for this example. They can be calculated from the probabilities as follows:

$$M(t) = \sum_n n \cdot P_n, \quad V(t) = \sum_n (M(t) - n)^2 \cdot P_n, \tag{2.13}$$

where $\sum_n$ is the abbreviated symbol for summing over all possible values of n.

In order to derive $M(t)$ from (2.12), let us introduce the following widely used transformation of equations; we multiply both sides of an equation by n and afterward sum them with respect to all possible values of n to obtain

$$\sum_n n\dot{P}_n = k\sum_n n(n+1)P_{n+1} - k\sum_n n^2 P_n. \tag{2.14}$$

Another useful trick is to make use of the fact that the summing is performed through all values of n and thus we can shift the value of n inside the sum to $n+k$ without changing the value of the sum (i.e. $\sum_n n\dot{P}_{n+1} = \sum_n (n-1)\dot{P}_n$). Applying this together with (2.13) we can simplify (2.14) as follows:

$$\begin{aligned} \dot{M}(t) &= k\sum_n (n-1)nP_n - k\sum_n n^2 P_n \\ &= -k\sum_n nP_n = -k \cdot M(t). \end{aligned} \tag{2.15}$$

Using the separation-of-variables method together with the aforementioned initial condition $M(0) = n_0$, we can write the solution of (2.15) as

$$M(t) = n_0 e^{-kt}.$$

In order to obtain the value of $V(t)$, let us express the value of $\sum_n n^2 P_n$ with the help of the values of mean and variance as

$$\sum_n n^2 P_n = V(t) + M^2(t);$$

after applying the time-derivative we obtain

$$\sum_n n^2 \dot{P}_n = \dot{V}(t) + 2M(t)\dot{M}(t).$$

Now we can multiply both sides of the equation by n^2 and sum them through all values of n to rewrite (2.12) as

$$\sum_n n^2 \dot{P}_n = k\sum_n n^2(n+1)P_{n+1} - k\sum_n n^3 P_n,$$

$$\sum_n n^2 \dot{P}_n = k \sum_n (-2n^2 + n) P_n.$$

Using the formulas for mean and variance we can transform the equation to obtain non-homogenous linear ODE for the $V(t)$ in the form

$$\dot{V}(t) + 2M(t)\dot{M}(t) = -2kV(t) - 2kM^2(t) + kM(t),$$

$$\dot{V}(t) = -2kV(t) + kn_0 e^{-kt},$$

together with the initial condition $V(0) = 0$ (as $P_n = \delta_{n,n_0}$). An ODE in this form can be solved using the method of variation of constants to obtain

$$V(t) = n_0 e^{-kt} + Ce^{-2kt},$$

which together with the initial condition for V yields

$$V(t) = n_0 \left(e^{-kt} - e^{-2kt}\right).$$

However, we are often unable to use an analogous approach for more complex forms of the Master equation and have to settle for some approximated solution.

2.2.2 General case

Now, let us assume we have J different chemical compounds $X_1, X_2, \ldots, X_J$ in our system. A typical reaction in such system is given by a set of stoichiometric coefficients s_j, r_j as

$$s_1 X_1 + s_2 X_2 + \ldots s_J X_J \underset{k_-}{\overset{k_+}{\rightleftharpoons}} r_1 X_1 + r_2 X_2 + \ldots r_J X_J \tag{2.16}$$

In a stochastic formulation we assume all possible system states to be lattice points on a J-dimensional lattice. This set of accessible points is given by the coefficients s_i and r_i. For two points accessible from each other, we say they satisfy the equivalence relation called the stoichiometic compatibility [15]. A two-dimensional example is given in Figure 2.2.

Zero values of s_j and r_j are permitted. If $s_k = r_k \neq 0$, we call the corresponding X_k a catalyst. The reaction associated with the rate constant k_+ is called a forward reaction, and the reaction associated with rate constant k_- is referred to as a reverse reaction. The probability that a forward reaction occurs in $[t, t + \mathrm{d}t)$ is given as $f_+(t, \mathrm{d}t) = k_+ \prod_{i=1}^{J} ((X_i(t)))^{s_i} \mathrm{d}t$, where $((X))^s$ stands for $\frac{X!}{(X-s)!}$. Analogously, for a

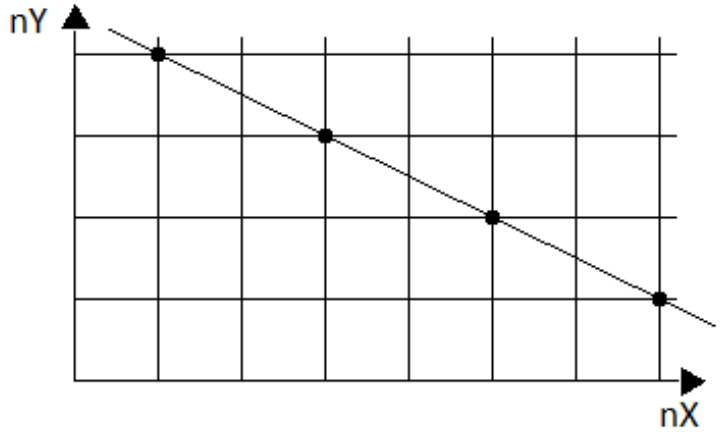

Figure 2.2: *Possible states of X and Y following reaction* $2X \rightleftharpoons Y$.

reverse reaction we have the probability $f_-(t, \mathrm{d}t) = k_- \prod_{i=1}^{J}((X_i(t)))^{r_i}\mathrm{d}t$. The probability that more than one reaction occurs in an infinitesimally small interval $\mathrm{d}t$ can be assumed to be zero; and therefore the probability that some reaction occurs in $[t, t + \mathrm{d}t)$ is given as $\alpha(t)\mathrm{d}t = f_-(t, \mathrm{d}t) + f_+(t, \mathrm{d}t)$. In case we have more reactions in the form (2.16), we sum up probabilities for all of them together. Using (2.11) we can define the time when the next reaction takes place in general case as

$$\tau = \frac{1}{\alpha(t)} \ln\left(\frac{1}{r}\right), \tag{2.17}$$

where $\alpha(t)$ is defined as the probability coefficient for $\mathrm{d}t$: the probability that any reaction occurs in $[t, t + \mathrm{d}t)$ is $\alpha(t)\mathrm{d}t$. A new problem which did not concern us in the simple case is now to determine which reaction takes place in time $t + \tau$. To determine that, we can generate another uniformly distributed random number r_2. If r_2 is uniformly distributed on $(0, 1)$, $r_2\alpha(t)$ is uniformly distributed on $(0, \alpha(t))$; as $\alpha(t) = \sum_{i=1}^{k} f_i$ (by f_i we mean both forward (f_i^+) and backward (f_i^-) reactions; and we removed $\mathrm{d}t$ from both sides of the equation), the probability that the value $\alpha(t)r_2$ will fall to the interval $\left(\sum_{i=1}^{j} f_i(t), \sum_{i=1}^{j+1} f_i(t)\right)$ is equal to $f_{j+1}(t)$. We can summarise the algorithm into four steps that we perform iteratively:

1) Generate two random numbers r_1, r_2 uniformly distributed in $(0, 1)$.

2) Time when the next reaction happens is given by $t + \tau$, where τ is given by (2.17) with r_1 in place of r.

2) Which reaction occurs in time $t + \tau$ is determined by r_2, according to the interval into which $r_2\alpha(t)$ falls.

4) Number of molecules for each species $X_i(t+\tau)$ is given by $X_i(t)+(s_i-r_i)$ for the forward reaction (or $-(s_i-r_i)$ for the reverse reaction).

Following the same rationale as for the simple case, we can also write down the Master equation for reaction (2.16) as

$$\begin{aligned}\dot{P}_{\mathbf{n}}(t) =& k_+ \left(\prod_j \mathbb{E}_j^{s_j - r_j} - 1 \right) \prod_j ((n_j))^{s_j} P_{\mathbf{n}}(t) \\ &+ k_- \left(\prod_j \mathbb{E}_j^{r_j - s_j} - 1 \right) \prod_j ((n_j))^{r_j} P_{\mathbf{n}}(t).\end{aligned} \tag{2.18}$$

We use the shift operator $\mathbb{E}$ from [49] to keep the equation more compact. It is defined (for given shift vector $\mathbf{i} = (i_1, i_2, \ldots, i_J)$) as follows:

$$\mathbb{E}_1^{i_1} \mathbb{E}_2^{i_2} \ldots \mathbb{E}_J^{i_J} f(\mathbf{n}) = f(\mathbf{n} + \mathbf{i}).$$

If more possible reactions occur in the system, the associated Master equation includes the sum of all such reaction terms. Therefore, it is often too complex for solving it explicitly or gaining information about the statistical attributes than for the simple case. In such situations we resort to combining numerical methods (stochastic simulation or the numerical integration of the Master equation). We will use this strategy in this thesis to study selected systems of reaction kinetics pertaining to cellular biology.

CHAPTER 3

A model for gene expression in the presence of decoy binding sites

In this chapter we describe a system of chemical reactions motivated by the dynamics of gene expression. We focus on the level of protein in the system which is subject to interactions with decoy binding sites. We are taking into account a simplified, non-bursting, regime of protein production. As we do not consider other binding sites in this work, we can omit 'decoy' from the notation and will refer to them simply as binding sites (or BS). Our main goal is to investigate the distributions of free (i.e. unbound) and total protein (both bound and free) in the system. In the beginning we write down associated Master equation. Then we focus on the distribution of total protein. The main part of our analysis consists of performing singular perturbation reduction to obtain quasi-steady-state solution for the free protein distribution. We prove the correctness of the solution by mathematical induction. Then we investigate the statistical characteristics of obtained distribution. Finally we introduce two additional methods to obtain free protein distribution and compare all of them using numerical simulations. Most of the chapter was also presented in paper [21].

Let us introduce the following notation for our variables:

X - total protein, X_f - free protein,

Y - all binding sites, Y_f - free binding sites,

C - complex (protein bound to the binding site).

We assume that three reversible reactions can take place:

1) Protein production/decay.

$$\emptyset \underset{\gamma}{\overset{k}{\rightleftharpoons}} X_f$$

2) Protein binding/unbinding reaction.

$$X_f + Y_f \underset{k_-}{\overset{k_+}{\rightleftharpoons}} C$$

3) Decay of the complex (a free binding site is vacated).

$$C \xrightarrow{\gamma} Y_f$$

We use upper-case letters in *italics* to represent a number of corresponding species throughout this thesis. We reserve the corresponding lower-case letter as a notation for a concentration of a given species. In order to avoid confusion with X, we use N instead of X_f as the number of free protein.

Although we have defined five different variables, the problem is in fact only two-dimensional. First, we assume that the total number of binding sites (Y) is constant. Let us express the number of the remaining species in terms of X and N (and constant Y). The number of complexes is the same as the number of bound protein. Therefore we get a relationship $C = X - N$. The number of free binding sites is equal to the number of all binding sites without the sites which are included in complexes, $Y_f = Y - C = Y - X + N$.

3.1 Master equation for the non-bursting case

Let us have a look into the system using stochastic simulations. We assume the case when one protein is created in each event of production. To make the notation clearer, we introduce the abbreviation of $P(X(t) = X, X_f(t) = N)$ as $P_{X,N}$. The Master equation corresponding to the system of reactions together with the non-bursting case of production and degradation of X (in X_f as well as in C) assumes the form

$$\begin{aligned}
\dot{P}_{X,N} = {} & kP_{X-1,N-1} - kP_{X,N} \\
& + \gamma(N+1)P_{X+1,N+1} - \gamma N P_{X,N} \\
& + k_+(N+1)(Y-X+N+1)P_{X,N+1} \\
& - k_+ N(Y-X+N)P_{X,N} \\
& + k_-(X-N+1)P_{X,N-1} - k_-(X-N)P_{X,N} \\
& + \gamma(X-N+1)P_{X+1,N} - \gamma(X-N)P_{X,N},
\end{aligned} \tag{3.1}$$

Let us illustrate one stochastic simulation of such system by the Gillespie algorithm ($Y = 10,\ k = 3,\ \gamma = 0.1,\ k_+ = 1,\ k_- = 10$), with no proteins at the

beginning: $P_{X,N}(0) = \delta_{X,0}\delta_{N,0}$. The time evolution of each species can be seen in Figure 3.1. It is easy to see that reactions affecting the amount of total protein in the system are slow in comparison with the free protein level changes; we exploit this observation later.

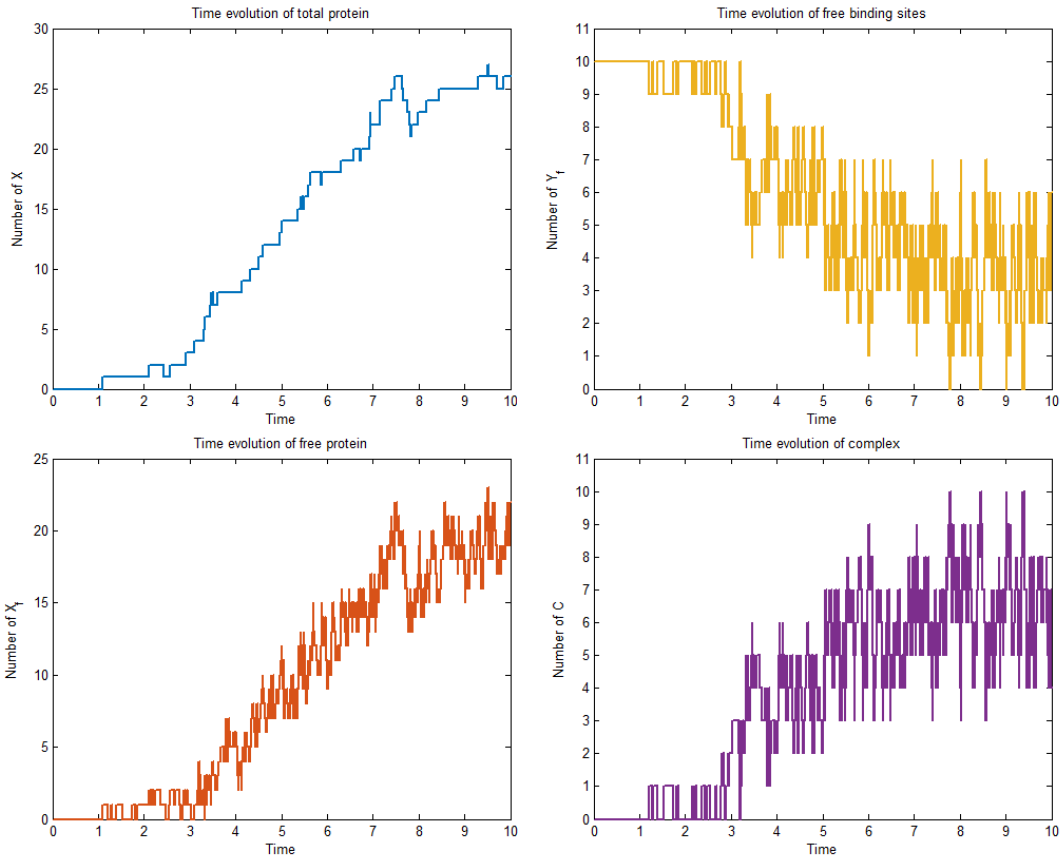

Figure 3.1: *Simulation of the system using Gillespie algorithm.*

3.2 Total protein distribution

In this section we use the equation (3.1) to obtain the distribution of all protein (variable X) in the system. In order to do so, let us remind that we obtain P_X by summing up all probabilities $P_{X,N}$ for a given X, i.e. $P_X = \sum_{N=0}^{\infty} P_{X,N}$. So, in order to get a differential equation for P_X, let us use a well-known trick and sum up both sides of the Master equation (3.1) with respect to N, from 0 to ∞. In order to keep it clear, we use the abbreviation $\sum_N$. As the sum goes through all integers, we can easily change all $N-1$ to N etc. It causes the rows of (3.1), corresponding to the binding/unbinding reaction, to cancel out, as the distribution of all protein in this

reaction does not change. We obtain:

$$\begin{aligned}\dot{P}_X &= k\sum_N \left(P_{X-1,N} - P_{X,N}\right) \\ &+ \gamma\sum_N N\left(P_{X+1,N} - P_{X,N}\right) \\ &+ \gamma\left(\sum_N \left((X-N+1)P_{X+1,N} - (X-N)P_{X,N}\right)\right),\end{aligned}$$

which simplifies to

$$\dot{P}_X = k\left(P_{X-1} - P_X\right) + \gamma\left((X+1)P_{X+1} - XP_X\right). \tag{3.2}$$

A system of differential equations in this form can be solved using the method of generating functions (see Section 1.1.1). In order to compute P_X in this way we multiply the equation by s^X and we sum them over X and we get

$$\begin{aligned}\frac{\partial}{\partial t}\sum_X s^X P(X,t) = k&\left(\sum_X s^X P_{X-1} - \sum_X s^X P_X\right) \\ &+ \gamma\left((X+1)\sum_X s^X P_{X+1} - X\sum_X s^X P_X\right).\end{aligned} \tag{3.3}$$

Now we use the definition of the generating function as $G(s,t) = \sum_x s^x P_x$. By differentiating this formula we can find an expression in terms of generating function also for other terms of the equation as $\frac{\partial G}{\partial s} = \sum_x x s^{x-1} P_x$. Knowing this we obtain from (3.3) the following partial differential equation:

$$\frac{\partial G}{\partial t} = k(s-1)G + \gamma(1-s)\frac{\partial G}{\partial s} = (s-1)\left(kG - \gamma\frac{\partial G}{\partial s}\right). \tag{3.4}$$

First, let us take a look at the result at the steady state ($t \to \infty$), in which case we simplify the equation to $kG = \gamma\frac{\partial G}{\partial s}$. This can easily be solved by the separation of variables and obtaining $G(s) = Ce^{\frac{k}{\gamma}s}$. Now let us recall our generating function definition, which at the steady state takes the form $G(s) = \sum_x s^x P(x,\infty)$, whereby $G(1) = 1$ follows from the normalization condition. Hence $C = e^{-\frac{k}{\gamma}}$ and we get

$$G(s,\infty) = e^{\frac{k}{\gamma}(s-1)}.$$

According to the definition of generating function, the value of P_x is equal to the value of coefficient multiplying s^x in the power series expansion of $G(s)$. Taylor-expanding the exponential gives us:

$$G(s) = e^{-\frac{k}{\gamma}}\sum_n \frac{\left(\frac{k}{\gamma}\right)^n s^n}{n!}.$$

Therefore

$$P_X = \frac{\left(\frac{k}{\gamma}\right)^X e^{-\frac{k}{\gamma}}}{X!},$$

which is the Poisson distribution parametrized by $\lambda = \frac{k}{\gamma}$. Using the fact that for the mean we have $\langle X \rangle = \lambda$ we can write

$$P_X = \frac{\langle X \rangle^X e^{-\langle X \rangle}}{X!}.$$

As the problem (3.4) is a linear partial differential equation of the type we considered in Section 1.2.1, we can find a solution for some feasible initial condition. Using the usual trick (1.7) we rewrite (3.4) into an auxiliary equation

$$\frac{\partial u}{\partial t} + \gamma(s-1)\frac{\partial u}{\partial s} + k(s-1)G\frac{\partial u}{\partial G} = 0.$$

The characteristic system of this equation has the form:

$$\begin{aligned} \dot{t} &= 1, & t(\tau) &= \tau + C_1 \\ \dot{s} &= \gamma(s-1), & s(\tau) &= C_2 e^{\gamma\tau} + 1 \\ \dot{G} &= k(s-1)G, & \frac{\mathrm{d}G}{G} &= k(C_2 e^{\gamma\tau})\mathrm{d}\tau. \end{aligned} \tag{3.5}$$

To solve the equation, we have to find the functions which are constant on the characteristics. Equations for t and s can be solved independently. Combining these two functions we gain the first constant expression, $t - \frac{\ln(s-1)}{\gamma}$. Solving the equation for G we obtain another constant expression in the form $C_3 = \frac{k(s-1)}{\gamma} - ln(G)$ or even more simple form $C_4 = \frac{ks}{\gamma} - ln(G)$. And thus the general solution of the auxiliary equation is

$$u(t, s, G(s,t)) = \Phi\left(t - \frac{\ln(s-1)}{\gamma},\ \frac{ks}{\gamma} - ln(G)\right).$$

Now, according to (1.8) let us take u in form $\varphi\left(t - \frac{\ln(s-1)}{\gamma}\right) + \frac{ks}{\gamma} - ln(G) = 0$. It follows, that we get the solution for G in the form $G(s,t) = \psi\left(t - \frac{\ln(s-1)}{\gamma}\right) \cdot e^{\frac{ks}{\gamma}}$.

To successfully solve the equation, we need to provide boundary conditions for the generating function G. Let us investigate the case when there is no protein at the beginning, thus $P(x,0) = \delta_0$, which gives us the equation $G(s,0) = 1$. We already know the solution for stationary case, which gives us another condition $G(s,\infty) = e^{\frac{k}{\gamma}(s-1)}$. Using the condition at time 0 we can calculate

$$\psi\left(-\frac{\ln(s-1)}{\gamma}\right) = e^{-\frac{ks}{\gamma}} \Rightarrow \psi(a) = e^{-\frac{k}{\gamma}\left(e^{-a\gamma}+1\right)}.$$

It implies that

$$G(s,t) = e^{-\frac{k}{\gamma}\left(e^{-\gamma t}(s-1)+1\right)+\frac{ks}{\gamma}} = e^{-\frac{k}{\gamma}(s-1)\left(e^{-\gamma t}-1\right)},$$

which is the same kind of generating function as the stationary state, just with a different value of the parameter in Poisson distribution, which now means that $\langle X\rangle = \frac{k}{\gamma}\cdot(1-e^{-\gamma t})$.

3.3 Singular perturbation reduction

As Master equation (3.1) does not have solution in the closed form (unless $Y=0$), we seek to determine a quasi-steady-state solution. To achieve that, we assume that binding/unbinding reactions are fast compared to production/degradation reactions (i.e. $k_- \gg \gamma$). In order to do that, let us nondimensionalise time by setting

$$t = \frac{\tau}{\gamma}.$$

Inserting the above into (3.1) yields

$$\begin{aligned}
\varepsilon\frac{d}{d\tau}P_{X,N} = {}& \varepsilon\frac{k}{\gamma}\left(P_{X-1,N-1} - P_{X,N}\right) \\
&+ \varepsilon\left((N+1)P_{X+1,N+1} - \gamma N P_{X,N}\right) \\
&+ \frac{1}{k_b}\big((N+1)(Y-X+N+1)P_{X,N+1} \\
&- N(Y-X+N)P_{X,N}\big) \\
&+ (X-N+1)P_{X,N-1} - (X-N)P_{X,N} \\
&+ \varepsilon\left((X-N+1)P_{X+1,N} - (X-N)P_{X,N}\right),
\end{aligned} \tag{3.6}$$

where

$$\varepsilon = \frac{\gamma}{k_-}, \quad k_b = \frac{k_-}{k_+}$$

are nondimensionalised parameters. By sending ε to zero, we obtain an equation for the leading-order approximation of $P_{X,N}$.

$$\begin{aligned}
0 = {}& \frac{1}{k_b}(N+1)(Y-X+N+1)P_{X,N+1} - \frac{1}{k_b}N(Y-X+N)P_{X,N} \\
&+ (X-N+1)P_{X,N-1} - (X-N)P_{X,N}.
\end{aligned}$$

As X stays the same in this approximation, we can treat it as constant (abbreviation P_N stands for $P_{X,N}$) and we get

$$
\begin{aligned}
&(N+1)(Y-X+N+1)P_{N+1} = \\
&\left(N(Y-X+N)+k_b(X-N)\right)P_N - k_b(X-N+1)P_{N-1},
\end{aligned}
\tag{3.7}
$$

whereby we require that boundary conditions $P_N(N<0)=0$ and $P_N(N>X \ \vee \ N<X-Y)=0$ are satisfied. This difference equation has a solution

$$
P_N = P_{X,N} = \frac{k_b^N C(X)}{N!(X-N)!(Y-X+N)!}, \tag{3.8}
$$

where $C(X)$ is a constant with respect to N, dependent only on the value of X. Let us prove statement (3.8) by mathematical induction with respect to N.

$$
\begin{aligned}
1^\circ: \quad & N = \max\{X-Y,0\} \\
& N=0: \quad P_0 = \frac{C(X)}{X!(Y-X)!} \\
& N=X-Y: \quad P_{X-Y} = \frac{k_b^{X-Y}C(X)}{(Y)!(X-Y)!}
\end{aligned}
\tag{3.9}
$$

This holds, as P_0 or P_{X-Y} depends only on X and the constant Y. therefore the validity of (3.9) can be ensured by an appropriate choice of $C(X)$.

$$
\begin{aligned}
2^\circ: \quad & N = \max\{X-Y,0\}+1 \\
& N=1: \quad P_1 = \frac{k_b X P_0}{Y-X+1} = \frac{k_b C(X)}{(X-1)!(Y-X+1)!} \\
& N=X-Y+1: \quad P_{X-Y+1} = \frac{k_b Y P_{X-Y}}{X-Y+1} = \frac{k_b^{X-Y+1}C(X)}{(Y-1)!(X-Y+1)!}
\end{aligned}
$$

which is easy to see by substituting $N=0$, or $N=X-Y$ into (3.7).

Induction step $(N-1,N)\to N+1$:

$$
\begin{aligned}
&(N+1)(Y-X+N+1)P_{N+1} = \\
&\quad = \left(N(Y-X+N)+k_b(X-N)\right)\frac{k_b^N C(X)}{N!(X-N)!(Y-X+N)!} - \\
&\quad - k_b(X-N+1)\frac{k_b^{N-1}C(X)}{(N-1)!(X-N+1)!(Y-X+N-1)!} \\
&P_{N+1} = \frac{k_b^{N+1}C(X)}{(N+1)!(X-N-1)!(Y-X+N+1)!}.
\end{aligned}
$$

This completes the proof by induction of the formula (3.8).

To explore further the term $C(X)$, we use the notation $P_X = \sum_{N=\max\{X-Y,0\}}^{X} P_{X,N}$ to expand (3.8) as

$$P_X = \sum_{i=\max\{0,X-Y\}}^{X} P_{X,i} = C(X) \sum_{i=\max\{0,X-Y\}}^{X} \frac{k_b^i}{((X-i)!(Y-X+i)!i!},$$

and thus we can express $C(X)$ as

$$C(X) = P_X \left(\sum_{i=\max\{0,X-Y\}}^{X} \frac{k_b^i}{(X-i)!(Y-X+i)!i!} \right)^{-1}.$$

Now, since we already know about the Poissonian character of steady-state solution for P_X, we are ready to calculate the number of free protein in quasi-steady state:

$$P_N = \sum_{X=N}^{N+Y} P_X \cdot \frac{k_b^N}{N!(X-N)!(Y-X+N)!} \cdot \left(\sum_{i=\max\{0,X-Y\}}^{X} \frac{k_b^i}{(X-i)!(Y-X+i)!i!} \right)^{-1}. \tag{3.10}$$

Let us now investigate how the value of $\varepsilon = \frac{\gamma}{k_-}$ influence the quality of this approximate solution. In order to do that, let us fix the mean of free protein $\langle X \rangle = 30$ and the dissociation constant $k_b = 10$; this will result in the same quasi-steady-state distribution regardless of the choice of ε. On the other hand, exact simulated distribution depends also on the value of $\varepsilon = \frac{\gamma}{k_-}$. If we divide both k_+ and k_- by 10, we keep the value of k_b, but decrease the value of ε by a factor of ten. We use the values $\varepsilon = 0.001, 0.01, 0.1, 1, 10$. As the distribution of total protein is Poissonian, we are also curious, how much our distribution differs from Poisson distribution. In order to do that we calculate MLE (maximum likelihood estimation) of Poisson distribution based on result of stochastic simulation (see Section 1.1.3). We calculate these distributions for different values of binding sites; $Y = 0, 10, 20, 30$. Results can be seen on Figures 3.2 and 3.3. The green histogram is distribution generated by the repeated simulations (10^5) of the Gillespie algorithm (see Section 2.2.2) for a sufficiently long time, blue line is the probability distribution of free protein in quasi-steady-state (3.10) and red line is the best-fit Poisson distribution calculated with MLE.

Case $Y = 0$ is special as no binding/unbinding reactions can occur. Therefore the choice of ε has no effect on the distribution and both Gillespie and QSS results

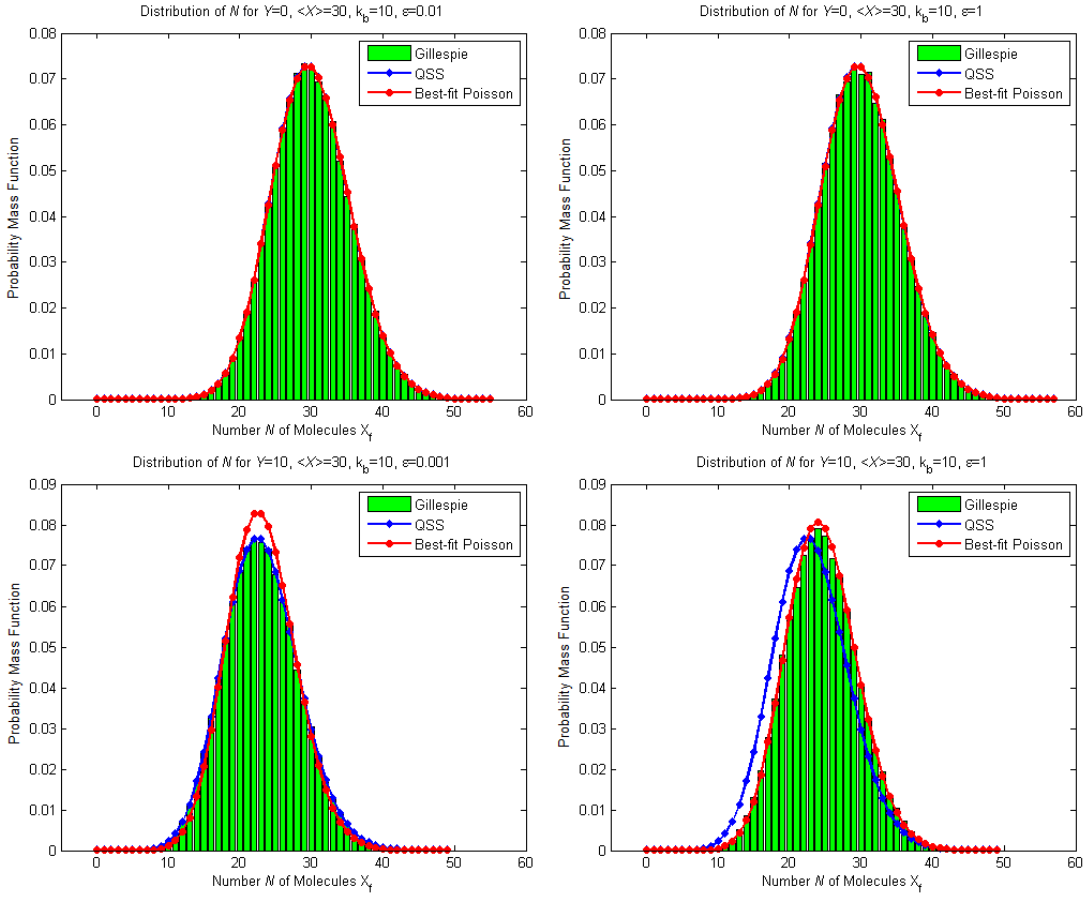

Figure 3.2: *Quality of quasi-steady-state solution for* $Y = 0$ *and* $Y = 10$.

yield Poisson distribution. Other observations are also expected. It is possible to see even with a bare eye that with increasing value of ε the QSS distribution fits the simulated distribution less well. To evaluate the goodness of the fit, we calculate the Bhattacharyya distance (see Section 1.1.4) between these two distributions and put it together in Table 3.1. In Table 3.2 we also present the distance between simulated distribution and MLE of Poisson distribution. As we add further noise into the Poisson process of total protein production/decay, the Gillespie distribution is broader than the Poisson distribution. But, as we increase ε, the number of binding/unbinding reactions decreases and we once again converge to the Poisson distribution.

3.3.1 Moments of quasi-steady-state probability distribution

In this section we focus on the analysis of the free protein probability distribution in the quasi-steady-state. We focus on four basic statistical characteristics, the mean μ, the variance σ^2, the Fano factor ($F = \frac{\sigma^2}{\mu}$) and the squared coefficient

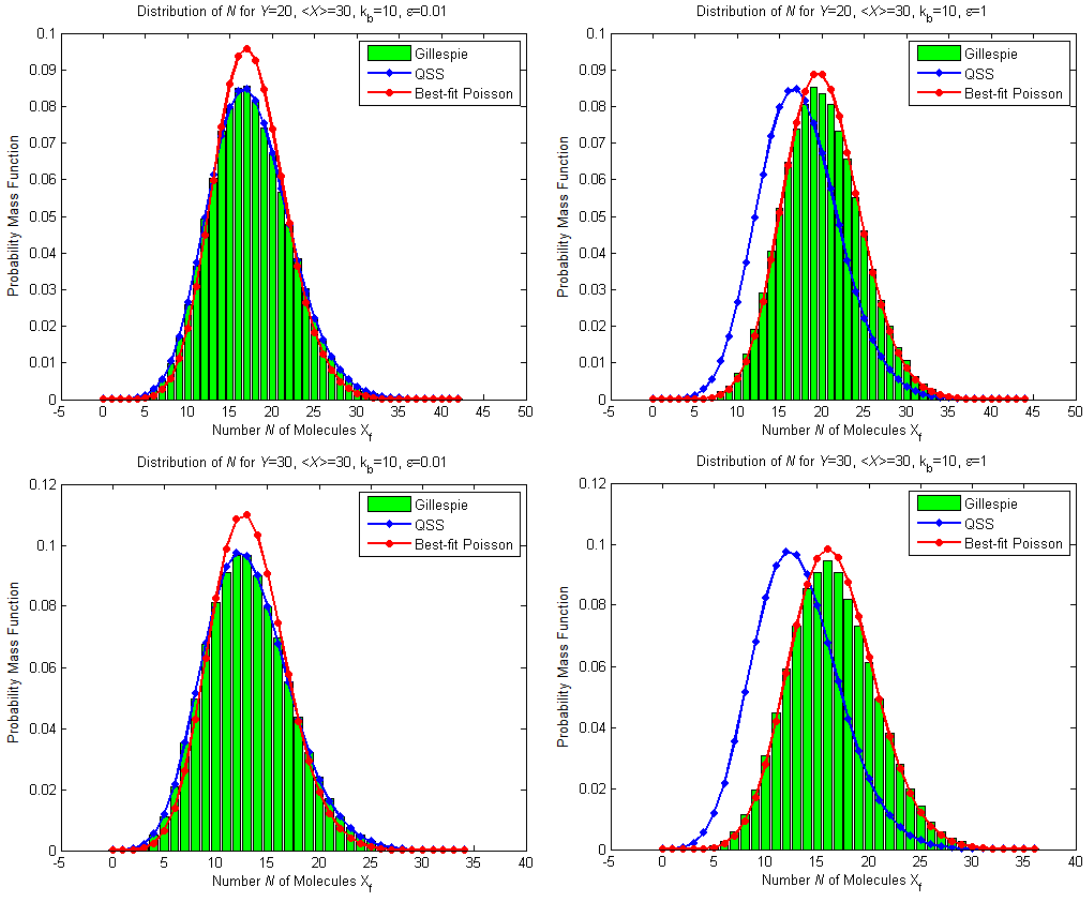

Figure 3.3: *Quality of quasi-steady-state solution for* $Y = 20$ *and* $Y = 30$.

$Y \setminus \varepsilon$	0.001	0.01	0.1	1	10
0	$8.15 \cdot 10^{-5}$	$6.36 \cdot 10^{-5}$	$7.93 \cdot 10^{-5}$	$5.86 \cdot 10^{-5}$	$4.52 \cdot 10^{-5}$
10	$5.04 \cdot 10^{-5}$	$7.87 \cdot 10^{-5}$	$2.27 \cdot 10^{-4}$	0.011	0.111
20	$5.30 \cdot 10^{-5}$	$5.13 \cdot 10^{-5}$	$7.80 \cdot 10^{-4}$	0.042	0.398
30	$6.39 \cdot 10^{-5}$	$7.47 \cdot 10^{-5}$	$1.74 \cdot 10^{-3}$	0.085	0.780

Table 3.1: *Statistical distance between simulated distribution and quasi-steady-state approximation.*

of variation ($CV^2 = \frac{\sigma^2}{\mu^2}$). We vary the total number of binding sites. We study the changes in the characteristics of the free protein distribution for the selected choice of the mean of total protein production $\langle X \rangle$ and the dissociation constant k_b. These values are fixed: $\gamma = 0.1,\ k_- = 10$; thus $\varepsilon = 0.01$ stays constant. As we showed in previous section, for such value of ε the quasi-steady-state results provide a good approximation of the final distribution. Different choices of k and k_+ are used to modify $\langle X \rangle$ and k_b. In order to calculate these statistical characteristics, we use quasi-steady-state approximation, as running Gillespie algorithm would be extremely time-consuming mainly for small values of ε due to simulation

$Y \setminus \varepsilon$	0.001	0.01	0.1	1	10
0	$7.80 \cdot 10^{-5}$	$6.33 \cdot 10^{-5}$	$7.92 \cdot 10^{-5}$	$5.67 \cdot 10^{-5}$	$4.41 \cdot 10^{-5}$
10	$1.60 \cdot 10^{-3}$	$1.77 \cdot 10^{-3}$	$1.36 \cdot 10^{-3}$	$2.76 \cdot 10^{-4}$	$7.02 \cdot 10^{-5}$
20	$3.74 \cdot 10^{-3}$	$3.90 \cdot 10^{-3}$	$3.02 \cdot 10^{-3}$	$6.41 \cdot 10^{-4}$	$6.32 \cdot 10^{-5}$
30	$4.37 \cdot 10^{-3}$	$4.44 \cdot 10^{-3}$	$3.17 \cdot 10^{-3}$	$7.56 \cdot 10^{-4}$	$8.03 \cdot 10^{-5}$

Table 3.2: *Statistical distance between simulated distribution and best-fit Poisson distribution.*

of significant increase in number of binding/unbinding reactions. In order to show that these results are appropriate we run Gillespie algorithm just for a few cases and plot them together with QSS results. Results are presented in Figures 3.4-3.5.

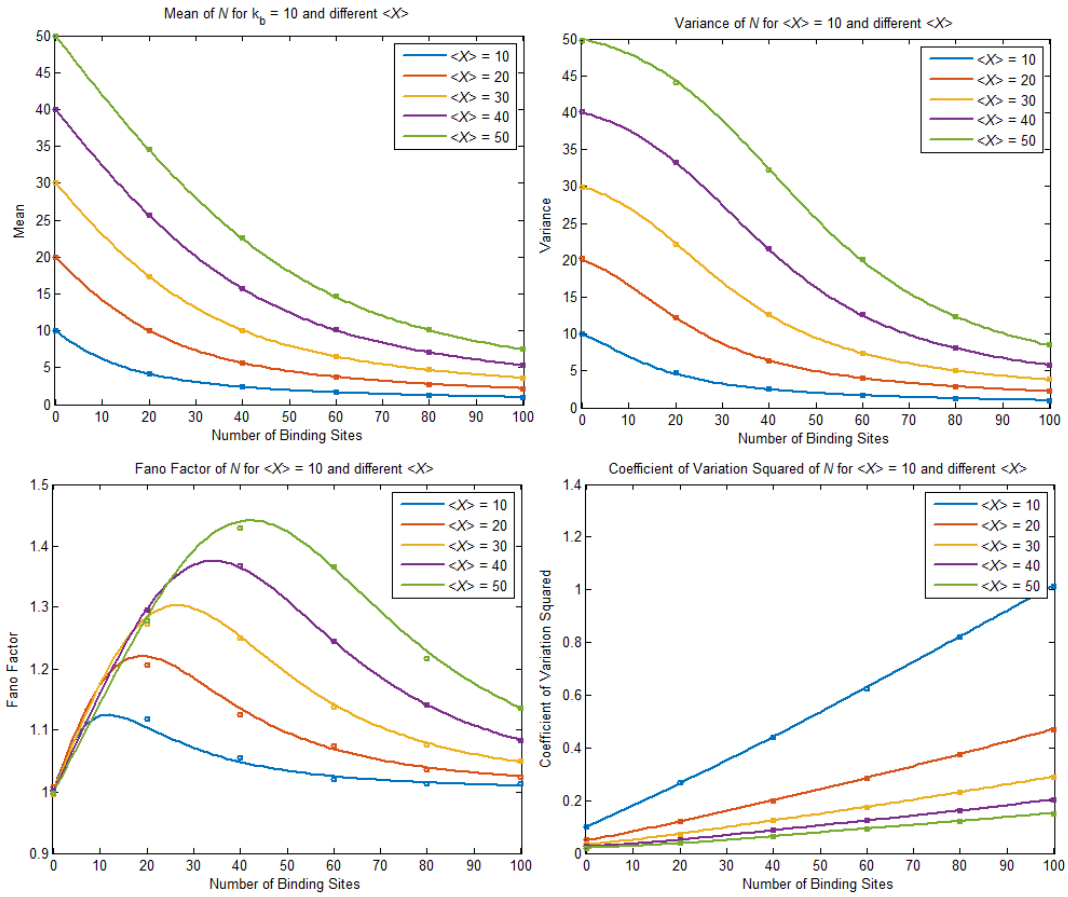

Figure 3.4: *Moments of the free protein probability distribution for different* $\langle X \rangle$.

The case $Y = 0$ gives the Poisson distribution, so we expect $\mu = \sigma^2$ and thus $F = 1$. Very large dissociation constants imply that proteins bind weakly to the binding sites and thus the system keeps the characteristics similar to the Poisson distribution for large numbers of binding sites. As the coefficient of variation squared increases very fast with adding binding sites, we use a logarithmic scale in the last panel.

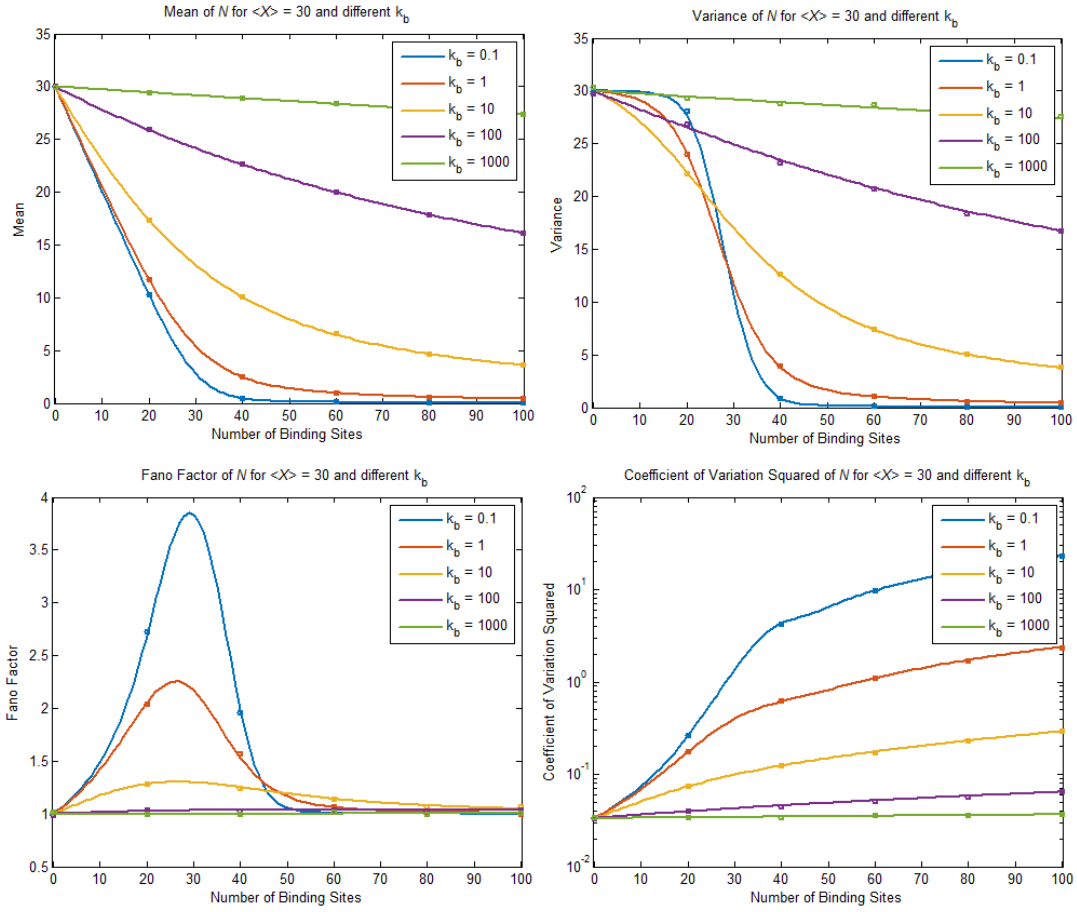

Figure 3.5: *Moments of the free protein probability distribution for different* k_b.

From the pictures it is obvious that for large values of Y, the mean and variance tend to 0, the Fano factor tends to 1 and CV^2 diverges to infinity. That could imply an approximate Poisson distribution. Let us investigate this hypothesis a little more.

3.3.2 Large Y regime

In the case, when $Y \gg X$ (which also implies $Y \gg N$), we can use the approximation for the factorial

$$Y! = \underbrace{Y(Y-1)\cdot \ldots \cdot (Y-X+N+1)}_{\approx Y^{X-N}}(Y-X+N)!$$

which can can rearranged to obtain

$$(Y-X+N)! \approx Y!\, Y^{N-X}. \tag{3.11}$$

Substituting (3.11) into (3.10) and using binomial theorem and some basic manipulation we obtain a simplified equation for the probability distribution of the free

protein:

$$
\begin{aligned}
P_N &= \sum_{X=N}^{N+Y} P_X \cdot \frac{k_b^N\, Y^{X-N}}{N!(X-N)!\,Y!} \cdot \left(\sum_{i=0}^{X} \frac{k_b^i\, Y^{X-i}}{(X-i)!\,Y!i!} \right)^{-1} \\
&= \sum_{X=N}^{N+Y} P_X \cdot \frac{k_b^N\, Y^{X-N}}{N!(X-N)!\,Y!} \cdot \left(\frac{Y^X}{X!} \sum_{i=0}^{X} \left(\frac{k_b}{Y} \right)^i \binom{X}{X-i} \right)^{-1} \\
&= \sum_{X=N}^{N+Y} P_X \cdot \binom{X}{N} \left(\frac{k_b}{Y} \right)^N \left(\frac{Y}{k_b+Y} \right)^X \\
&= \sum_{X=N}^{N+Y} P_X \cdot \binom{X}{N} \left(\frac{k_b}{k_b+Y} \right)^N \left(\frac{Y}{k_b+Y} \right)^{X-N}.
\end{aligned}
\tag{3.12}
$$

This formula can be also rewritten as $\sum P_X \cdot P_{N|X}$, where P_X has a Poisson distribution with parameter $\langle X \rangle$ and we see that $P_{N|X}$ has a binomial distribution with probability of Bernoulli trial given as $\frac{k_b}{k_b+Y}$ and number of trials set as X. Therefore we can express the free protein distribution as $N = \sum_{i=0}^{X} \xi_i$, where $P(\xi_i = 1) = \frac{k_b}{k_b+Y}$. Let us perform now some algebraic manipulation on the probability generating functions:

$$
\begin{aligned}
G_N(s) &= \sum_{k\geq 0} P(N=k) s^k \\
&= \sum_{k\geq 0} \sum_{j\geq 0} P(X=j) P\left(\sum_{i=0}^{j} \xi_i = k \right) s^k \\
&= \sum_{j\geq 0} P(X=j) \left(G_\xi(s) \right)^j \\
&= G_X\left(G_\xi(s) \right).
\end{aligned}
\tag{3.13}
$$

In the second step we used the rule for calculating probability generating function of the sum of independent random variables (1.2). It is clear now that probability generating function for free protein is given as a composition of generating functions for G_X (Poisson distribution) and G_ξ (Bernoulli distribution). Using the formulae for the generating functions of these distributions (see Section 1.1.1) and substituting them into (3.13) we obtain

$$
\begin{aligned}
G_N(s) &= \exp\left(\langle X \rangle \left(G_\xi(s) - 1 \right) \right) \\
&= \exp\left(\langle X \rangle \left(\frac{Y + k_b s}{Y + k_b} - 1 \right) \right) \\
&= \exp\left(\left(\frac{k_b \langle X \rangle}{k_b + Y} \right) (s-1) \right),
\end{aligned}
\tag{3.14}
$$

which again indicates the Poisson distribution, but with a different parameter

$$\langle N \rangle = Var(N) = \frac{k_b \langle X \rangle}{k_b + Y}.$$

Computing the probability distribution in quasi-steady state for large values of binding sites is infeasible as the formula works with the terms in the order $Y!$, therefore we use Gillespie algorithm (with 10^5 repetitions) to estimate the mean and the variance of the distribution. The results are provided in table 3.4. Other parameters used are ($\langle X \rangle = 10,\ \ k_b = 10$). As the distribution is Poisson, the Fano factor tends to 1 and $CV^2 = \frac{1}{\langle N \rangle}$, which tends to infinity.

y	theoretical	$\langle N \rangle$	$Var(N)$
100	0.9091	1.0027	1.0124
500	0.1961	0.1999	0.2009
1000	0.099	0.0999	0.0998
5000	0.02	0.0201	0.02
10000	0.01	0.01	0.01

Table 3.3: *Quality of approximated estimation for large Y.*

3.4 Numerical simulations

In this section, we introduce and compare three different ways how to obtain the number of free protein using numerical simulations.

1) Using Gillespie algorithm. In this case we simulate the reaction as it really happens through time.

2) Using explicit formula for free protein count in quasi-steady state. In this case we assume that $k_- \gg \gamma$.

3) Solving the system of ODEs given by the Master equation. As the problem contains infinite number of equations, we have to set maximal value of the number of all protein at which the system is truncated.

We used these parameters in all three cases: $Y = 10,\ k = 3,\ k_+ = 1,\ k_- = 10$ and $\gamma = 0.1$. As the initial condition of the system we use the no-protein case, i.e. $P_{X,N} = \delta_{X,N,(0,0)}$.

3.4.1 Gillespie algorithm

Applying the algorithm described in Section 2.2.2 to the current problem we obtain the following step-by-step algorithm.

1) Calculate α, such that probability that any reaction will occur in interval $[t, t+\mathrm{d}t]$ is equal to $\alpha \cdot \mathrm{d}t$. This is given by

$$\alpha = k + N\gamma + k_+ N(Y - X + N) + Ck_- + C\gamma.$$

2) Choose two random numbers $r_1,\ r_2$ from the uniform distribution on the interval $[0, 1]$.

3) Use r_1 to calculate the time when the next reaction occurs, $\tau = \frac{1}{\alpha}\ln(\frac{1}{r_1})$.

4) Use r_2 to calculate which reaction occurs at the time $t+\tau$. If $0 \leq \alpha r_2 < k$, then free protein production occurred; if $k \leq \alpha r_2 < k + N\gamma$, then free protein decay occurred; if $k + N\gamma \leq \alpha r_2 < k + N\gamma + k_+ N(Y - X + N)$, then binding reaction occurred; if $k + N\gamma + k_+ N(Y - X + N) \leq \alpha r_2 < k + N\gamma + k_+ N(Y - X + N) + Ck_-$, then unbinding reaction occurred and if $k + N\gamma + k_+ N(Y - X + N) + Ck_- \leq \alpha r_2 < \alpha$, then bounded protein decay just occurred.

5) Repeat steps $(2-4)$ and write down the values of N (free protein) and X (total protein) in given time points.

6) Let the simulation run 10^5 times to obtain the probability distribution of these variables.

3.4.2 Quasi-steady-state approximation

In this section we use the result obtained in Section 3.2, that under the given initial condition, total protein distribution P_X follows the Poisson distribution with the mean $\langle X \rangle = \frac{k}{\gamma} \cdot (1 - e^{-\gamma t})$. Together with (3.10) we get the formula for P_N:

$$P_N = C(X) \frac{k_b^N}{((X-N)!(Y-X+N)!N!},$$

$$\text{where } C(X) = P_X \left(\sum_{i=\max\{0, X-Y\}}^{X} \frac{k_b^i}{((X-i)!(Y-X+i)!i!} \right)^{-1},$$

let us remind the substitution for dissociation rate constant $k_b = \frac{k_-}{k_+}$. Problem with this approach is that in order to get desired P_N, we need to multiply very huge number $C(X)$ with very small fraction. As the fraction also contains member $(X-N)!$, we have to set upper level for X and N in order to compute non-zero values of $P_{X,N}$. We use upper bound $X = 100$ in order to be consistent with the ODE approach. All protein distribution calculations are based on the results from Section 3.3.

In Figure 3.6 we display the time evolution of the four statistical characteristics of both free and total protein levels as calculated with the Quasi-steady-state approach.

3.4.3 Differential equations

In order to turn Master equation (3.1) into a finite system of ODEs, we set all probabilities, where X (the number of total protein) is greater than 100 to zero. We can also use the fact that the amount of free protein cannot be greater than the amount of total protein and set all $P_{X,N}$ such that $N > X$ to zero. Second restricting condition is based on the number of binding sites Y. As number of bound protein $(X-N)$ is Y in maximal case, we can set all $P_{X,N}$, such that $N < X - Y$, to zero. Therefore the number of non-zero equations is actually less than $10 \cdot 101$ instead of $101 \cdot 101$. From the general from of Master equation we can write them as

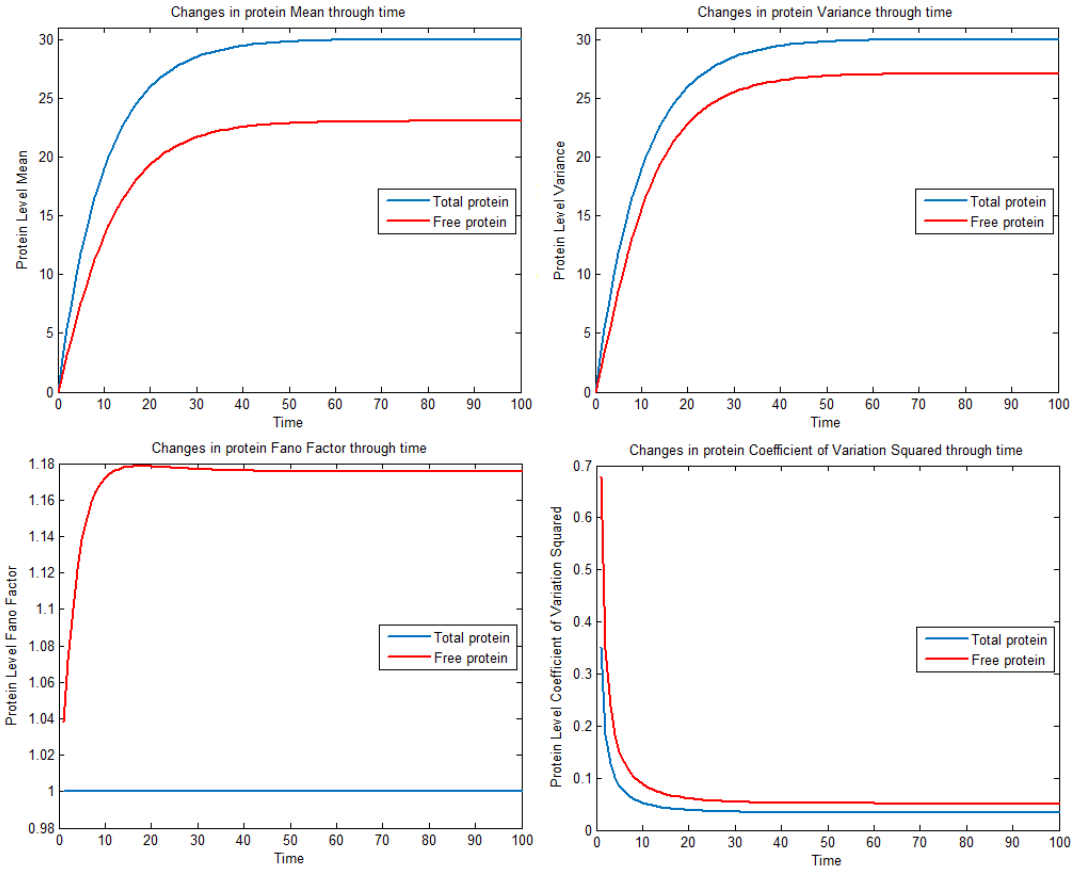

Figure 3.6: *Time evolution of statistical characteristics.*

$$
\begin{aligned}
\dot{P}_{0,0} &= -kP_{0,0} + \gamma P_{1,1} + \gamma P_{1,0},\\
\dot{P}_{1,0} &= -kP_{1,0} + \gamma P_{2,1} + k_+ Y P_{1,1} - k_- P_{1,0} + 2\gamma P_{2,0},\\
\dot{P}_{1,1} &= kP_{0,0} - kP_{1,1} + 2\gamma P_{2,2} - \gamma P_{1,1} - k_+ Y P_{1,1} + k_- P_{1,0} - k_- P_{1,0} + \gamma P_{2,1}, \qquad (3.15)\\
&\vdots\\
\dot{P}_{100,100} &= kP_{99,99} - kP_{100,100} - 100\gamma P_{100,100} - 100k_+ Y P_{100,100} + k_- P_{100,99}.
\end{aligned}
$$

In order to solve it, we can write this system of ODEs in the matrix form as $\dot{P} = A.P$, where P is vector of all relevant non-zero probabilities and A is a transition matrix. From the form of (3.15) we see that A is matrix with constant terms, therefore we obtain the linear system of ODEs. As the system is linear, its Jacobian matrix stays constant during the whole process. Its eigenvalues range from -1075 to a little bit under zero, with eleven eigenvalues smaller than -1000. Therefore we can say that no solution component is unstable (eigenvalue with large positive real value) and some components are very stable (eigenvalue with large negative

real value). This situation is not good for the propagation of error in a numerical scheme and we refer to such system of ODEs as stiff. This is a common situation with Master equations. For more information about the stiff ODEs, we recommend looking at [42]. In order to obtain solution of this system we use MATLAB ordinary differential equation solver *ode15s*, which is a variable-step, variable-order solver based on the numerical differentiation formulas of orders one to five capable of solving stiff ODEs (see [43] for reference).

3.4.4 Comparison

The output of the numerical simulations is the probability surface $P_{X,N}$ in the given time points. In Table 3.4 we observe the differences in the probability surfaces obtained by different methods. Difference between the two methods is again defined as the Bhattacharyya distance (see Section 1.1.4) applied to the differences of the two probability surfaces. In Figure 3.7 we provide examples of heatmaps of probability distributions in the chosen moments in time.

$Time$	Gill vs. ODE	Gill vs. Quasi	ODE vs. Quasi
0	0	0	0
1	$1.88 \cdot 10^{-3}$	$7.86 \cdot 10^{-4}$	$2.69 \cdot 10^{-3}$
2	$4.74 \cdot 10^{-4}$	$5.34 \cdot 10^{-4}$	$6.80 \cdot 10^{-4}$
3	$2.80 \cdot 10^{-4}$	$5.20 \cdot 10^{-4}$	$2.16 \cdot 10^{-4}$
4	$3.84 \cdot 10^{-4}$	$4.43 \cdot 10^{-4}$	$1.99 \cdot 10^{-4}$
5	$1.61 \cdot 10^{-3}$	$4.81 \cdot 10^{-4}$	$1.28 \cdot 10^{-3}$
10	$1.37 \cdot 10^{-3}$	$5.34 \cdot 10^{-4}$	$9.88 \cdot 10^{-4}$
20	$5.37 \cdot 10^{-4}$	$5.58 \cdot 10^{-4}$	$3.01 \cdot 10^{-5}$
50	$5.72 \cdot 10^{-4}$	$5.67 \cdot 10^{-4}$	$3.18 \cdot 10^{-5}$
100	$6.20 \cdot 10^{-4}$	$6.44 \cdot 10^{-4}$	$2.40 \cdot 10^{-5}$

Table 3.4: *Statistical distance between simulated probability surfaces obtained by different numerical methods.*

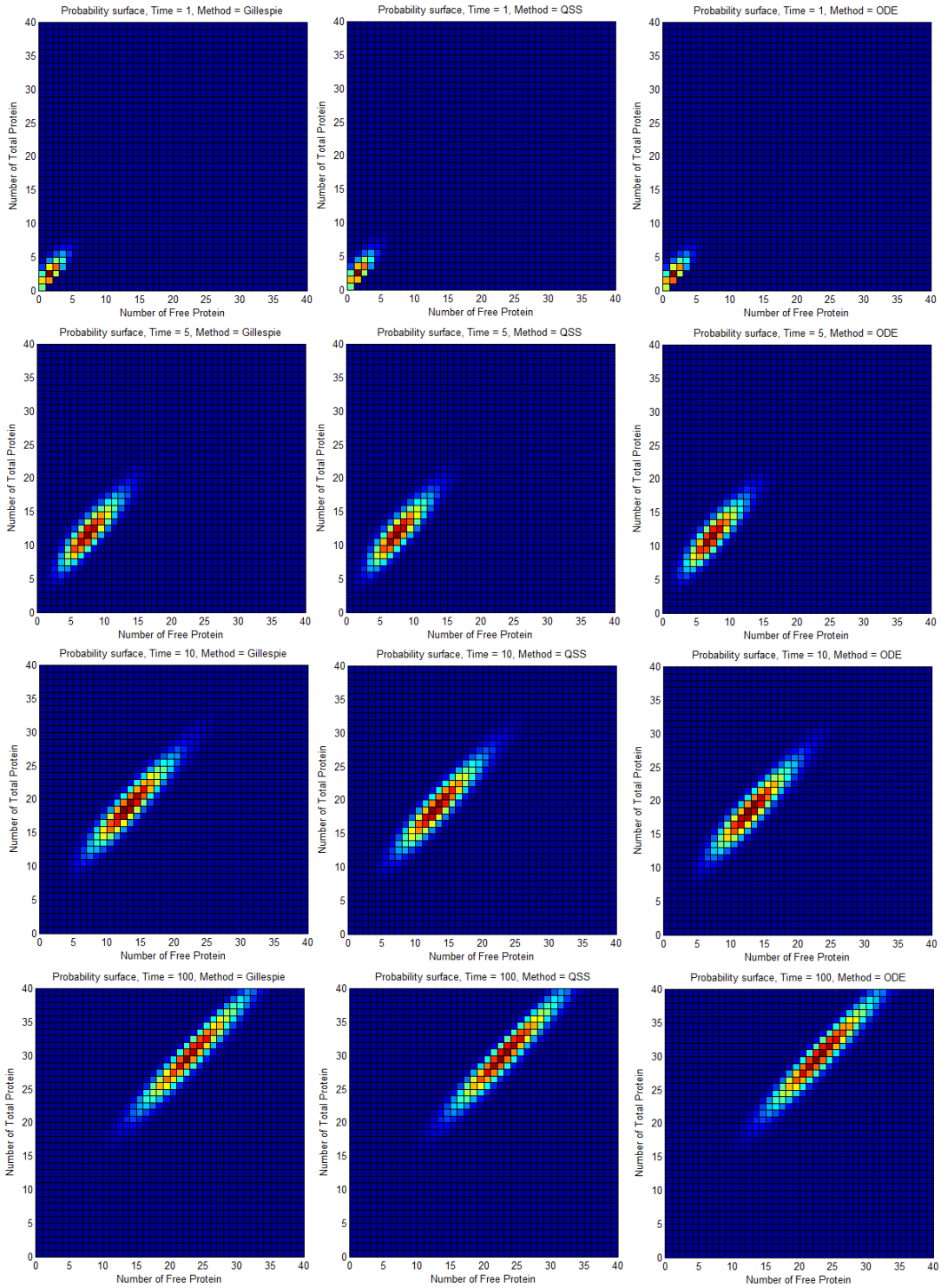

Figure 3.7: *Time evolution of probability surfaces for different methods.*

We formulated a simplified gene expression model and presented the associated Master equation. Then we obtained the distribution of total protein and after employing singular perturbation reduction we also derived quasi-steady-state solution for the free protein distribution. Then we compared it with the distribution obtained using two other methods: simulated by Gillespie algorithm and from a system of ODEs. With the help of numerical simulations we studied its statistical moments; we also compared the distributions obtained from the three methods. We reported a very good agreement between the distributions; thus we justified the use of the quasi-steady-state solution as a very good approximation for the free protein distribution.

CHAPTER 4

Small noise approximation

In this chapter we try to further simplify the distribution of free protein species (in system of reactions presented in Chapter 3) and its Fano factor for some reasonable specific case. In the previous part we already obtained free protein probability distribution (in quasi-steady state) in the closed form; but in this chapter we aim to obtain a closed-form formula for statistical moments. In order to do that we perform an expansion of the Master equation in a linear-noise scenario in the similar manner as presented in Chapter X in [49]. We divide this procedure into three main stages. First we derive deterministic mean of the distribution, secondly we include noise from binding/unbinding reactions and finally we add noise from total protein number fluctuation. We also justify our results and compare them to formulae from Chapter 3 using numerical simulations. This chapter is also part of paper [21].

We focus on the limit case when the size Ω of the system is large enough ($\Omega \gg 1$). Under these assumption we write down our variables in terms of concentrations rather than total numbers:

$X = \Omega \cdot x$ - total protein, $Y = \Omega \cdot y$ - total binding sites,

$N = \Omega \cdot n$ - free protein, $Y_f = \Omega \cdot y_f$ - free binding sites,

$C = \Omega \cdot c$ - complex.

We would like to underscore the fact that lowercase letters here denote the concentration of a particular reactant denoted by corresponding uppercase letter. We identify the system size Ω with the dissociation constant $k_b = k_-/k_+$ in our model. This is a standard approach and it guarantee that the binding and unbinding reaction rates are of the same order: the probability of a binding reaction occurrence depends on $N \cdot Y_f \cdot k_+ = n \cdot y_f \cdot \Omega^2 \cdot k_+$ and the probability of an unbinding reaction occurrence depends on $C \cdot k_- = c \cdot \Omega \cdot k_-$.

We focus here on the Fano factor as the characteristic of free protein distribution noise; in our case this noise can have two sources: the first one is the reversible association, and the second one is protein production/degradation. As we already mentioned, the timescales of these two processes are diametrically different; thus, we can treat them separately. For the reversible association we first investigate the deterministic approach to the reactions to obtain expected values of the reactants. Afterward we focus on stochastic component to inspect the noise present in the system. For protein production/degradation we already know about the Poissonian character of the process, therefore the main challenge here is to combine the two individual results here in the correct way.

4.1 Constant X (total protein count)

In this part, let us concentrate on the reversible association $\mathrm{X_f} + \mathrm{Y_f} \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{C}$. In the course of this reaction, we can treat the values of x (total protein concentration) and y (total binding sites concentration) as constants which will allow us to express other reactants in terms of these values. As we already mentioned previously, we were not able to express the distribution of free protein in the closed form, therefore we seek to approximate it as well as we can.

4.1.1 Deterministic case

If we assume large protein numbers, we can obtain the mean of the distribution by finding the stationary state of deterministic reaction kinetics, which we reviewed in Section 2.1. The concentrations x and y are treated as constants here. The variables n, y_f and c are dependent on time. Let us summarise the information we have available. Firstly, we know that the reversible association reaction $X_f + Y_f \underset{k_-}{\overset{k_+}{\rightleftharpoons}} C$ gives us the condition $n \cdot y_f = c$ in the stationary state. As we mentioned above, all the concentrations (n of free protein, y_f of free binding site, and c of their complex) are measured in units of the dissociation constant. Secondly, we can transform already mentioned conservation laws to the concentration terms and obtain $n + c = x$ for the total protein and $y_f + c = y$ for the total binding-site concentration. These equations together form the system of three equations with three unknown variables, therefore it is easy to get the solution. If we combine them together and express c from them, we obtain the quadratic equation

$$\bar{c}^2 + \bar{c}(-x - y - 1) + xy = 0.$$

It has a single non-negative solution, which we refer to as $\bar{n}$, $\bar{y}_f$, $\bar{c}$, and has the form

$$\begin{aligned} \bar{c}(x,y) &= \frac{x+y+1-\sqrt{x^2+y^2+1+2x+2y-2xy}}{2}, \\ \bar{n}(x,y) &= x-\bar{c}, \\ \bar{y}_f(x,y) &= y-\bar{c}. \end{aligned} \tag{4.1}$$

For the sake of greater parsimony, we omit the explicit notation of dependence of $\bar{c},\ \bar{n},\ \bar{y}_f$ on the total concentration of x and y in all non-ambiguous cases further down in the rest of the chapter.

4.1.2 Stochastic component

We assume that the number of total protein X remains constant, focusing only on the binding/unbinding reactions:

$$\mathrm{X_f} + \mathrm{Y_f} \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{C}.$$

In this setup, the problem is only one-dimensional. We choose C as our main variable for computation here, as it turns out that the calculations are the least complicated in that case. Other two dependent variables can be calculated in the straightforward manner as $N = X - C$ for free protein number and $Y_f = Y - C$ for free binding sites number. We can write down the Master equation in the similar manner as in 3.1 (just in terms of different variables) and obtain

$$\dot{P_C} = k_- \left((C+1)P_{C+1} - CP_C\right) + k_+ \left((N+1)(Y_f+1)P_{C-1} - N\,Y_f P_C\right),$$

where P_C denotes probability mass function of C. In order to simplify this expression, let us recall the shift-operator notation from [49] ($\mathbb{E}^i f(n) = f(n+i)$) to transform the Master equation into nicer compact form. Also as we investigate the distribution in the steady state, we set the time-derivative on the left side of the equation to zero. Applying this yields an equation in the form

$$0 = (\mathbb{E}-1)k_- C P_C + (\mathbb{E}^{-1}-1)k_+ N\,Y_f P_C, \tag{4.2}$$

As the equation is written in terms of C, the shifting is also executed with respect to C. Let us emphasize the fact that applying $\mathbb{E}^{-1}$ to N will cause it to shift into $N+1$ as N depends on C through the formula $N = X - C$ and the same applies for Y_f.

Inserting the scaling between reactants numbers and concentrations ($N = \Omega \cdot n$, $Y_f = \Omega \cdot y_f$, $C = \Omega \cdot c$ and $k_b = \Omega$) into (4.2) yields

$$0 = (\mathbb{E}-1)cP_c + (\mathbb{E}^{-1}-1)n y_f P_c, \tag{4.3}$$

Let us now investigate the behavior of the shift operator for a moment. As the shifting is executed with respect to C, arbitrary function $f(\tilde{C})$ will be changed to $f(\tilde{C}+1)$ after applying the operator. If we now rewrite the same step in terms of variable c (concentration of C) we go from $f(\Omega \cdot \tilde{c})$ to $f(\Omega \cdot \tilde{c}+1)$ or alternatively we can substitute and rewrite it as $g(\tilde{c}) \mapsto g(\tilde{c}+\Omega^{-1})$. Let us now recall the fact that Ω is big, thus Ω^{-1} is small increment in terms of c. Therefore it is possible to estimate $g(\tilde{c}+\Omega^{-1})$ using the Taylor expansion near $\tilde{c}$ as

$$g(\tilde{c}+\Omega^{-1}) = g(\tilde{c}) + \Omega^{-1}\frac{\partial g}{\partial c}(\tilde{c}) + \frac{\Omega^{-2}}{2}\frac{\partial^2 g}{\partial c^2}(\tilde{c}) + \dots.$$

If we use this fact and apply the same rationale to $\mathbb{E}^{-1}$ we can formally expand the shift operators as follows:

$$\begin{aligned} \mathbb{E} &= e^{\partial_C} = e^{\Omega^{-1}\partial_c} = 1 + \Omega^{-1}\partial_c + \frac{\Omega^{-2}}{2}\partial_c^2 + \dots \\ \mathbb{E}^{-1} &= e^{-\partial_C} = e^{-\Omega^{-1}\partial_c} = 1 - \Omega^{-1}\partial_c + \frac{\Omega^{-2}}{2}\partial_c^2 - \dots \end{aligned} \tag{4.4}$$

Inserting (4.4) into (4.3) we can write

$$0 = \left(\Omega^{-1}\partial_c + \frac{\Omega^{-2}}{2}\partial_c^2 + \dots\right) cP_c + (-\Omega^{-1}\partial_c + \frac{\Omega^{-2}}{2}\partial_c^2 - \dots) ny_f P_c.$$

It is possible now to multiply the equation by Ω. To deal with the infinite number of terms, let us recall the fact, that $\Omega^{-1} \ll 1$ and thus we can neglect all terms of order lower than Ω^{-1} to obtain

$$\partial_c\left[(c-ny_f)P_c + \frac{\Omega^{-1}}{2}\partial_c\left((c+ny_f)P_c\right)\right] = 0. \tag{4.5}$$

If we now integrate this equation with respect to c (with zero-flux conditions) we come to

$$(c-ny_f)P_c + \frac{\Omega^{-1}}{2}\partial_c\left((c+ny_f)P_c\right) = 0.$$

Let us denote the terms which appear in the above equation as $A = c - ny_f$ and $B = c + ny_f$. Here comes into play the assumption of small (or linear) noise, hence the name small-noise (or linear-noise) approximation. Using small-noise approximation (variance of c is of order Ω^{-1}) in this case is based on Taylor-expanding A and B around the deterministic mean value. Let us remind that since $\Omega^{-1} \ll 1$ we neglect all terms smaller than Ω^{-1} which means all terms smaller than Ω^{-1} in A

and all terms smaller than Ω^0 in B. Let us calculate:

$$\begin{aligned} A(c) &= c - (x-c)(y-c) \\ A'(c) &= 1 + x + y - 2c = 1 + n + y_f \\ A(c) &\simeq A(\bar{c}) + A'(\bar{c})(c-\bar{c}) = A'(\bar{c})(c-\bar{c}) \\ &= (1 + \bar{n} + \bar{y}_f)(c-\bar{c}) \\ B(c) &\simeq B(\bar{c}) \\ &= \bar{c} + \bar{n}\bar{y}_f = 2\bar{n}\bar{y}_f. \end{aligned} \tag{4.6}$$

We used the facts that $\bar{c} = \bar{n}\bar{y}_f$, $n = x-c$ and $y_f = y-c$. Substituting approximations (4.6) into (4.5) yields

$$(1 + \bar{n} + \bar{y}_f)(c-\bar{c})P_c + \Omega^{-1}\bar{n}\bar{y}_f \partial_c P_c = 0$$

or after rearranging we come to

$$\partial_c P_c = -\frac{1 + \bar{n} + \bar{y}_f}{\Omega^{-1}\bar{n}\bar{y}_f} \cdot (c-\bar{c})P_c. \tag{4.7}$$

This equation can be easily solved by the separation of variables c and P_c, which yields

$$P_c \propto \exp\left(\frac{1}{2} \cdot \frac{(1 + \bar{n} + \bar{y}_f)}{\Omega^{-1}\bar{n}\bar{y}_f} \cdot (c-\bar{c})^2\right).$$

This is the inner part of formula for normal distribution with mean $\bar{c}$ and variance $\frac{\Omega^{-1}\bar{n}\bar{y}_f}{1+\bar{n}+\bar{y}_f}$, the term before the exponential can be chosen so as to make the distribution integrate to one. We have thus found an approximate distribution of c; to put it symbolically, we can say that

$$c \sim \mathcal{N}\left(\bar{c}, \frac{\Omega^{-1}\bar{n}\bar{y}_f}{1 + \bar{n} + \bar{y}_f}\right) \tag{4.8}$$

as $\Omega \to \infty$. As n and y_f can be computed directly from c as $x - c$, resp. $y - c$, they follow the analogous distributions as c with the same variance and only different mean.

4.2 Fluctuating X (total protein count)

In the second part, instead of taking X to be a constant, we assume that the total protein count fluctuates due to creation of new protein molecules and decay of old ones, which is denoted by a reversible pair of chemical reactions

$$\emptyset \underset{k}{\overset{\gamma}{\rightleftharpoons}} \mathrm{X}.$$

We already know that these processes operate on a much slower timescale than the association/dissociation reactions. Furthermore, we have already found the steady-state distribution of X (from Section 3.2); the result is that it is Poissonian with $\langle X \rangle = \mathrm{Var}(X) = \frac{k}{\gamma}$. As $\Omega \to \infty$, we can approximate the Poissonian distribution of X with a Gaussian distribution as $\mathcal{N}\left(\langle X \rangle, \langle X \rangle\right)$ As we use the system-size scaling $\frac{k}{\gamma} = \langle x \rangle \cdot \Omega$, we can estimate the distribution of $\langle x \rangle$ by a small-noise Gaussian distribution

$$x = \frac{X}{\Omega} \sim \mathcal{N}\left(\langle x \rangle, \Omega^{-1} \cdot \langle x \rangle\right). \tag{4.9}$$

In order to calculate the statistics of n (free protein concentration) we have to combine the results (4.8) and (4.9) given that n is expressed in terms of slowly fluctuating x (total protein concentration). The free protein concentration also naturally depends on y (total concentration of binding sites), but as it is a constant through the whole process, we can neglect it from our notation for the sake of simplicity. We can use the law of total variance, described in Section 1.1.5, in order to solve this problem. In our case, we can express the total variance as

$$\mathrm{Var}(n) = E\left(\mathrm{Var}(n|x)\right) + \mathrm{Var}\left(E(n|x)\right). \tag{4.10}$$

We have already obtained in (4.8) the solution for $E(n|x)$ and $\mathrm{Var}(n|x)$; we utilize the fact that x fluctuates much more slowly that n and thus we can obtain results for n subject to constant x. Using the large Ω approximation, we can write

$$\begin{aligned} E(n|x) &= \bar{n} = \bar{n}(x) \\ \mathrm{Var}(n|x) &= \frac{\Omega^{-1}\bar{n}(x)\bar{y}_f(x)}{1 + \bar{n}(x) + \bar{y}_f(x)}. \end{aligned} \tag{4.11}$$

Since the variance of x is of order Ω^{-1} and we neglect all terms of order higher than Ω^{-1}, we can use the approximations $\bar{n}(x) \simeq \bar{n}(\langle x \rangle)$ and $\bar{y}_f(x) \simeq \bar{y}_f(\langle x \rangle)$ in the formula for the conditional variance, which yields

$$E(\mathrm{Var}(n|x)) \simeq \mathrm{Var}(n|\langle x \rangle) = \frac{\Omega^{-1} \cdot \bar{n}\bar{y}_f}{1 + \bar{n} + \bar{y}_f}, \tag{4.12}$$

evaluated at $\langle x \rangle$ (the mean of total protein concentration) and y (constant concentration of binding sites). The final term left to calculate is $\mathrm{Var}(\bar{n}(x))$. If we used the approximation $\bar{n}(x) = \bar{n}(\langle x \rangle)$, we would end with zero, which would incorrectly neglect all the variance. Therefore we have to also include terms of next order by using the linear variance approximation (first-order Taylor expansion), i.e. by writing

$$\bar{n}(x) \simeq \bar{n}(\langle x \rangle) + \frac{d\bar{n}}{dx} \cdot (x - \langle x \rangle).$$

As the first part of expression is constant we can express the variance of $\bar{n}(x)$ from this equation as

$$\begin{aligned} \mathrm{Var}(\bar{n}(x)) &\simeq \left(\frac{d\bar{n}}{dx}\right)^2_{|x=\langle x\rangle} \cdot \mathrm{Var}(x) \\ &= \left(\frac{d\bar{n}}{dx}\right)^2_{|x=\langle x\rangle} \cdot \Omega^{-1}\langle x\rangle \end{aligned} \tag{4.13}$$

We used here known Poissonian character of x. Hence, the last item left to calculate is the derivation of $\bar{n}$ with respect to x. In order to find it, let us start with taking the equations that define the dependence of $\bar{n}$, among others, on x:

$$\bar{n}\bar{y}_f = \bar{c}, \;\; \bar{n} + \bar{c} = x, \;\; \bar{y}_f + \bar{c} = y.$$

To obtain the term $\frac{d\bar{n}}{dx}$, let us now differentiate these equations with respect to variable x; this yields

$$\frac{d\bar{n}}{dx} \cdot \bar{y}_f + \frac{d\bar{y}_f}{dx} \cdot \bar{n} = \frac{d\bar{c}}{dx}, \tag{4.14}$$

$$\frac{d\bar{n}}{dx} + \frac{d\bar{c}}{dx} = 1, \tag{4.15}$$

$$\frac{d\bar{y}_f}{dx} + \frac{d\bar{c}}{dx} = 0. \tag{4.16}$$

Substituting (4.16) into (4.14), we can eliminate $\frac{d\bar{y}_f}{dx}$ to obtain

$$\frac{d\bar{n}}{dx} \cdot \bar{y}_f = \frac{d\bar{c}}{dx} \cdot (1 + \bar{n}).$$

Now we can use equation (4.15) to substitute $\frac{d\bar{c}}{dx}$ with $1 - \frac{d\bar{n}}{dx}$ to get

$$\frac{d\bar{n}}{dx} \cdot \bar{y}_f = \left(1 - \frac{d\bar{n}}{dx}\right) \cdot (1 + \bar{n}).$$

From this form it is trivial to express desired derivative as

$$\frac{d\bar{n}}{dx} = \frac{1 + \bar{n}}{1 + \bar{n} + \bar{y}_f}$$

and we can substitute it back to (4.13) and write down the formula for variance of expected value from the law of total variance as

$$\mathrm{Var}(\bar{n}(x)) \simeq \left(\frac{1 + \bar{n}}{1 + \bar{n} + \bar{y}_f}\right)^2 \cdot \Omega^{-1}\langle x\rangle. \tag{4.17}$$

All partial terms were calculated and we are now in a position to express the (unconditioned) moments of n. The mean is obtained by substituting $\langle x\rangle$ into the

deterministic results (4.1), and the variance is determined by substituting the partial results (4.12) and (4.17) into the formula for total variance (variance decomposition theorem) (4.10). For the mean we obtained the formula

$$\begin{aligned} E(n) &= \bar{n}(\langle x \rangle) \\ &= \frac{\langle x \rangle - y - 1 + \sqrt{\langle x \rangle^2 + y^2 + 1 + 2\langle x \rangle + 2y - 2\langle x \rangle y}}{2}. \end{aligned} \tag{4.18}$$

The variance can be expressed in the form

$$\mathrm{Var}(n) = \Omega^{-1} \left(\frac{\bar{n}\bar{y}_f}{1 + \bar{n} + \bar{y}_f} + \left(\frac{1 + \bar{n}}{1 + \bar{n} + \bar{y}_f} \right)^2 \langle x \rangle \right). \tag{4.19}$$

Combining these two results allows us to obtain the quantity of our main focus, the Fano factor, as

$$F = \frac{\mathrm{Var}(N)}{E(N)} = \frac{\mathrm{Var}(\Omega \cdot n)}{E(\Omega \cdot n)} = \Omega \cdot \frac{\mathrm{Var}(n)}{E(n)}. \tag{4.20}$$

Substituting (4.18) and (4.19) into (4.20) we obtain

$$F = \frac{\bar{y}_f}{1 + \bar{n} + \bar{y}_f} + \left(\frac{1 + \bar{x}}{1 + \bar{n} + \bar{y}_f} \right)^2 \frac{\langle x \rangle}{\bar{n}}.$$

We can now use the facts that $\bar{n}\bar{y}_f = \bar{c}$ and $\langle x \rangle = \bar{n} + \bar{c}$ to obtain

$$F = \frac{\bar{y}_f(1 + \bar{n} + \bar{y}_f) + (1 + \bar{n})^2(1 + \bar{y}_f)}{(1 + \bar{n} + \bar{y}_f)^2}.$$

Let us now perform a couple of simplifying steps in order to get the expression of the Fano factor in as simple a form as possible:

$$\begin{aligned} F &= \frac{(1 + \bar{n} + \bar{y}_f)^2 - (1 + \bar{n})(1 + \bar{n} + \bar{y}_f) + (1 + \bar{n})^2(1 + \bar{y}_f)}{(1 + \bar{n} + \bar{y}_f)^2} \\ &= 1 + \frac{-(1 + \bar{n})\bar{y}_f + (1 + \bar{n})^2\bar{y}_f}{(1 + \bar{n} + \bar{y}_f)^2} \\ &= 1 + \frac{\bar{n}\bar{y}_f(1 + \bar{n})}{(1 + \bar{n} + \bar{y}_f)^2}. \end{aligned} \tag{4.21}$$

Here the term 1 can be interpreted as the Poissonian noise, whose source is the production and degradation of new protein, and the residual fraction as an additional non-Poissonian noise present due to the interaction with binding sites.

4.3 Numerical simulations

In the current section we perform numerical simulations in order to obtain visualization of our computed results and to confirm the validity of the approximation

scheme presented above. We divide these simulations into two main parts: behavior of the Fano factor based on small-noise approximation (system-size approach) for different values of $\langle x \rangle$ and y, and judging the quality of this approximation with respect to results for the Fano factor obtained in previous chapter.

4.3.1 Fano factor based on system-size approach

As the basis for our numerical simulations, we have the following expression for the Fano factor:

$$F = 1 + \frac{\bar{n}\bar{y}_f(1+\bar{n})}{(1+\bar{n}+\bar{y}_f)^2}.$$

All the terms inside the formula depends internally on two characteristics of the system, total protein mean $\langle x \rangle$ and total number of binding sites y. Formulas how to calculate $\bar{n}$ and $\bar{y}_f$ have been derived in the previous section as

$$\bar{n} = \frac{\langle x \rangle - y - 1 + \sqrt{\langle x \rangle^2 + y^2 + 1 + 2\langle x \rangle + 2y - 2\langle x \rangle y}}{2}$$

and

$$\bar{y}_f = \frac{y - \langle x \rangle - 1 + \sqrt{\langle x \rangle^2 + y^2 + 1 + 2\langle x \rangle + 2y - 2\langle x \rangle y}}{2}.$$

In the first simulation we investigate the behavior of Fano factor with respect to different values of $\langle x \rangle$. In the first figure (Figure 4.1) we plot the dependence of Fano factor on the number of binding sites. We calculate, plot and compare the Fano factor for different values of $\langle x \rangle$. All values are meant to represent concentrations; therefore to obtain the corresponding number of molecules we have to multiply them by Ω. We clearly see that for larger values of $\langle x \rangle$ we are able to reach larger values of F as the possibility for binding/unbinding reactions increase and with them comes additional noise.

An interesting point to observe in the graphs is the slope near $y = 0$. We see that for small values of $\langle x \rangle$ the slope increases with increasing $\langle x \rangle$, but for larger values of $\langle x \rangle$ the slope starts to decrease. In order to investigate this phenomenon into further depth, let us calculate Taylor expansion near $F(x, 0)$:

$$F(x, \Delta y) = 1 + \Delta y \cdot \frac{x}{(1+x)^2}.$$

We can find the maximum of this value by differentiating

$$\frac{\partial F(x, \Delta y)}{\partial x} = \Delta y \cdot \frac{1-x}{(1+x)^3},$$

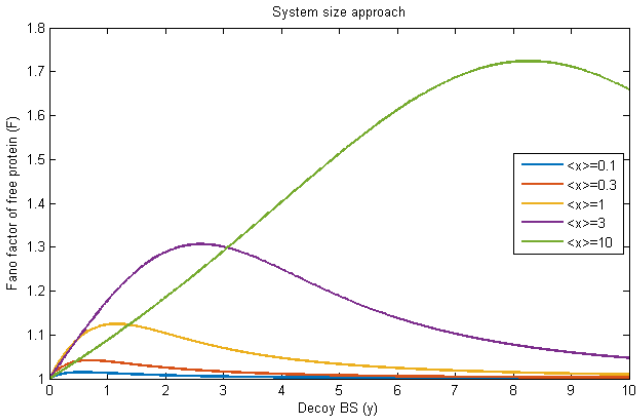

Figure 4.1: *Fano factor for large system size with y as an independent variable.*

which confirms that the maximal slope is obtained for $\langle x \rangle = 1$.

In the second picture displayed on Figure 4.2 we take the ratio $y/\langle x \rangle$ (number of BS divided by mean number of total protein) as the independent variable and plot Fano factor again for several different choices of $\langle x \rangle$.

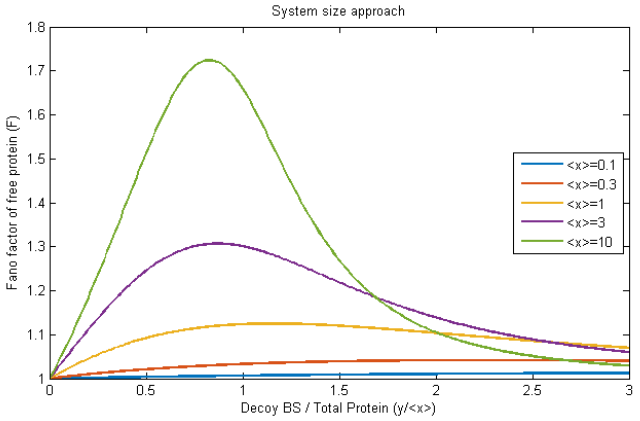

Figure 4.2: *Fano factor for large system size with $y/\langle x \rangle$ as an independent variable.*

We can see that for larger values of $\langle x \rangle$ the maximum of Fano factor is achieved near $y/\langle x \rangle = 1$. In order to investigate this phenomenon further, let us express the Fano factor in terms of new variables $a = \frac{y}{\langle x \rangle}$ and $b = \frac{1}{\langle x \rangle}$. This substitution yields

$$\begin{aligned} F(a,b) &= 1 + \frac{1}{2} \cdot \frac{a(1-a-b+\sqrt{\star})}{\star}, \\ \star &= a^2 + b^2 + 1 + 2b + 2a - 2ab. \end{aligned} \tag{4.22}$$

The problem of finding the maximum of Fano factor with respect to a is equivalent to finding the solution of $\frac{\partial F(a,b)}{\partial a} = 0$, with $F(a,b)$ from (4.22). This yields a very

complex implicit function. As we are interested in cases with large values of $\langle x \rangle$, we want to investigate its solution for small values of b. As we are unable to express the value of b in terms of a from the implicit function in reasonable way, we have to settle for numerical and graphical solution for this equation, which is provided in Figure 4.3. This verifies our hypothesis that for large $\langle x \rangle$ the maximal Fano factor is achieved near the point where $y = \langle x \rangle$.

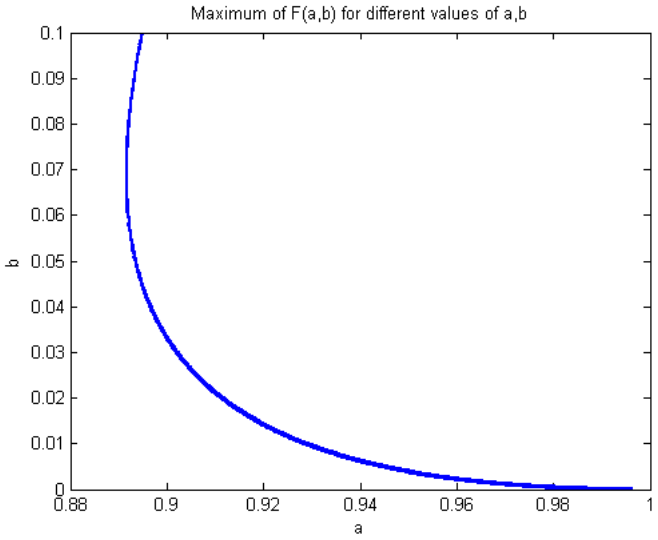

Figure 4.3: *Maximum of Fano factor; graphical solution.*

4.3.2 Quality of the system-size approach

In this section we investigate the quality of the linear noise approximation. In order to do so we compare the Fano factor calculated by the system size approach with the results of the quasi-steady state analysis and check whether they are consistent.

Let us recall the results for the free protein distribution based on quasi-steady-state approach:

$$P_N = \sum_{X=N}^{N+Y} P_X \cdot \frac{k_b^N}{N!(X-N)!(Y-X+N)!} \cdot \cdot \left(\sum_{i=\max\{0,X-Y\}}^{X} \frac{k_b^i}{(X-i)!(Y-X+i)!i!} \right)^{-1} . \tag{4.23}$$

Using this formula we can obtain probability distribution of free protein and therefore express its Fano factor. But this approach is not straightforward and it brings forward new issues. The formula (4.23) contains also the term $N!$ and terms of

similar order. As factorial function rises extremely steeply, mathematical software have problems calculating these values for large N. For example, MATLAB cannot calculate factorials for numbers bigger than 170. In order to get the results for big values of N, we have to rewrite the sum in a way that we get around these problems and to avoid multiplying zero by infinity. For this purpose, (3.10) can be rewritten as

$$P_N = \sum_{X=N}^{N+Y} P_X \cdot \left(\sum_{i=\max\{0,X-Y\}}^{X} k_b^{i-N} \cdot \frac{N!(X-N)!(Y-X+N)!}{i!(X-i)!(Y-X+i)!} \right)^{-1} . \qquad (4.24)$$

In the formula in this form we do not have to calculate $N!$ but instead express the term $\frac{N!}{i!}$ as the falling factorial: $N^{\underline{N-i}} = N \cdot (N-1) \cdot \ldots \cdot (i+1) \cdot i$, if $N > i$ or $\frac{1}{i^{\underline{i-N}}} = \frac{1}{i \cdot (i-1) \cdot \ldots \cdot (N+1) \cdot N}$, if $N < i$. Other parts of the fraction can be expressed in an analogous manner.

Now we can proceed with the calculations: we use 1, 5, 10 and 100 as the value for Ω and compare the quasi-steady-state expression of Fano factor with the linear-noise-approximation expression for Fano factor (4.21) for values of BS concentration y between 0 and 10. For the small values of Ω we expect larger difference between the interpretations of Fano factor, for the case $\Omega = 100$ we expect good fit. In the case where $\Omega = 1$ we can only use integer values of y, for $\Omega = 5$ we can use multiples of 0.2 for y, in other cases we use multiples of 0.1 for y in order to plot the graphs. We use the same setup for the different values of $\langle x \rangle$, in particular $0.1, 0.5, 1$ and 2. Resulting graphs are plotted on Figure 4.4 and confirm our assumptions. Also we see that the differences between approximations increase with $\langle x \rangle$, which is no surprise as Fano factor achieves higher values in such cases.

Furthermore, in Table 4.1 we present the sum of squared residuals in the points $y = 1, 2, \ldots, 10$. (If $y = 0$, then $F = 1$ always, so we can omit this point.)

Ω	$\langle x \rangle = 0.1$	$\langle x \rangle = 0.5$	$\langle x \rangle = 1$	$\langle x \rangle = 2$
1	$1.8 \cdot 10^{-5}$	$2.2 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	0.0032
5	$5.6 \cdot 10^{-7}$	$8.9 \cdot 10^{-6}$	$3.3 \cdot 10^{-5}$	$1.4 \cdot 10^{-4}$
10	$1.4 \cdot 10^{-7}$	$2.2 \cdot 10^{-6}$	$8.4 \cdot 10^{-6}$	$3.6 \cdot 10^{-5}$
100	$1.3 \cdot 10^{-9}$	$2.2 \cdot 10^{-8}$	$8.5 \cdot 10^{-8}$	$3.6 \cdot 10^{-7}$

Table 4.1: *Sum of squared residuals from LNA expression of Fano factor.*

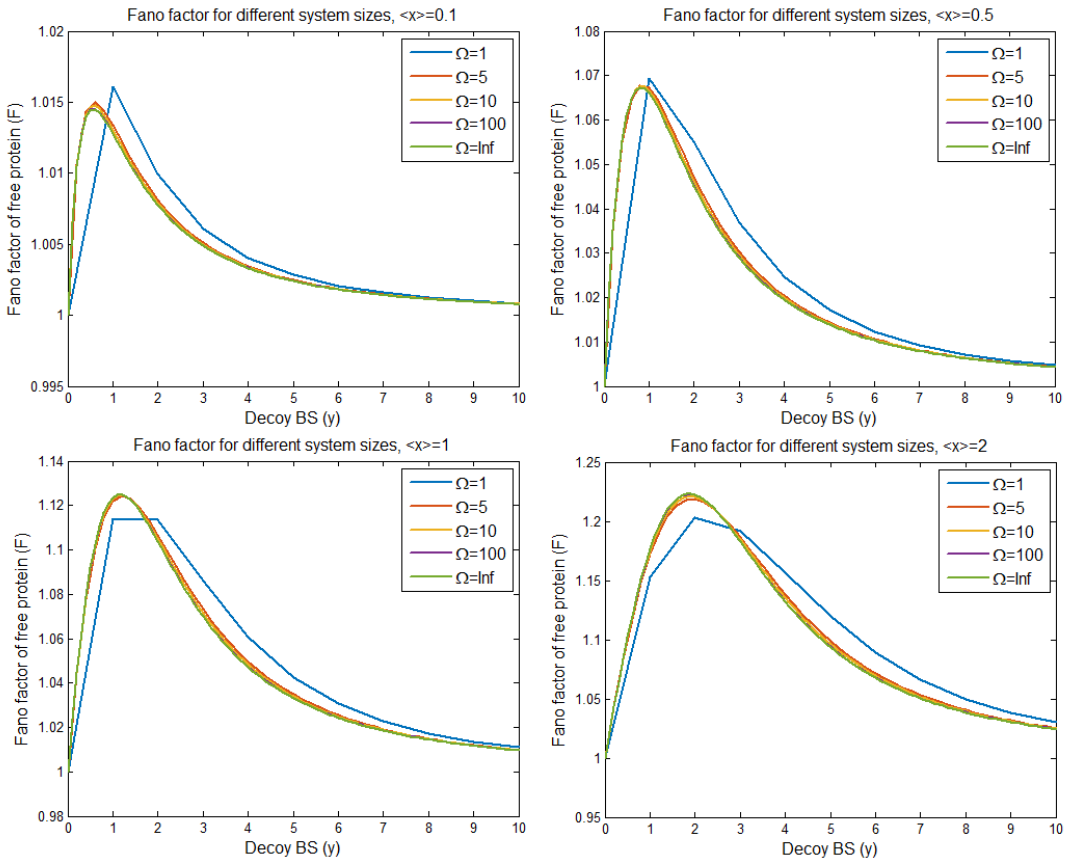

Figure 4.4: *Fano factor for different system sizes.*

In this chapter we considered the case of large system size using the dissociation constant as the measure of this size. We isolated two sources of the noise in the system and combined the results together to obtain a tractable expression for the Fano factor of the free protein distribution which differs from the Poissonian case. We also performed numerical simulations which showed that results for large system size are consistent with the quasi-steady-state results from Chapter 3.

CHAPTER 5

Distribution of mRNA – microRNA system

A microRNA is a small, non-coding RNA and contains about 22 nucleotides (abbreviated form miRNA is sometimes used instead of microRNA). It can be found mainly in some viruses, plants and animals. MicroRNA was discovered for the first time in 1993 on the lin-4 gene, which was repressing another, lin-14 gene [30, 51]. This process of gene repression is a member of broad class known as RNA silencing. For more information about the biological background, we refer to [3]. The content of this chapter is also the part of paper [6], which is currently submitted for publication.

First, we introduce our chemical reaction system, formulate the Master equation, and use generating functions to transform the Master equation into a partial differential equation of the second order (Section 5.1). In a specific parametric regime, this partial differential is reduced to an ordinary differential equation (Section 5.2), which is solved using the hypergeometric functions (Section 5.3). This solution is used to construct non-trivial approximations to the probability mass function and moments of the probability distribution of the chemical reaction system (Section 5.4). We identify and discuss certain conditions in which the probability mass function or the moments assume relatively simple algebraic forms (Section 5.5). Finally, we compare our results with stochastic simulations and discuss the numerical observations (Section 5.6)

5.1 The model and its Master equation

We consider two reactants in our system: X (mRNA) and Y (microRNA). In order to describe their behavior, let us consider a discrete stochastic chemical kinetics system composed of two species which are subject to two slow reactions

$$\emptyset \xrightarrow{\kappa} \mathrm{X} + \mathrm{Y}, \quad \mathrm{X} \xrightarrow{1} \emptyset, \tag{5.1}$$

and three fast reactions

$$\mathrm{X} + \mathrm{Y} \xrightarrow{\frac{\alpha q}{\varepsilon}} \mathrm{Y}, \quad \mathrm{X} + \mathrm{Y} \xrightarrow{\frac{\alpha(1-q)}{\varepsilon}} \emptyset, \quad \mathrm{Y} \xrightarrow{\frac{1}{\varepsilon}} \emptyset. \tag{5.2}$$

The reaction rates have been normalized in a way such that X decay rate equals unity. Then κ is the normalized production rate, α is the normalized interaction strength between X and Y and $\frac{1}{\varepsilon}$ is Y decay rate; in other words $\varepsilon \ll 1$ is the half life ratio of Y to X. The last parameter $0 \leq q \leq 1$ gives the probability that Y survives a bimolecular reaction with X.

Since the production of Y is slow but its decay is fast, the probability of observing a nonzero amount of Y will be $O(\varepsilon)$ small. Each time a new pair of X and Y molecules is produced, a rapid corrective phase driven by the fast reactions ensues, during which the newly produced Y can degrade multiple copies of X molecules before it is itself degraded. On the slow timescale of production and decay of X, this corrective phase manifests as an instantaneous jump in the copy number of X from state m to one of the states $0, 1, \ldots, m+1$. We display simulated trajectory of this system for a few choices of parameters in Figure 5.1. We may notice that in all three plots there is only one case when the number of microRNA species is greater than one (we used value $\varepsilon = 10^{-4}$). Other observation worth noticing is the fact that there are no jumps in mRNA number on the third plot (case $q = 0$) in contrast with the first plot (e.g. around $t = 2.6$) and the second plot (e.g. around $t = 1.6$). This behavior is expected from the form of (5.2) and we focus on it in Section 5.5.

The Master equation for the system of reactions has the form

$$\begin{aligned}
\dot{P}_{M,N} =& \kappa \left(\mathbb{E}_M^{-1}\mathbb{E}_N^{-1} - 1\right) P_{M,N} + \\
&+ \left(\mathbb{E}_M - 1\right) M P_{M,N} + \\
&+ \frac{\alpha q}{\varepsilon} \left(\mathbb{E}_M - 1\right) MN P_{M,N} + \\
&+ \frac{\alpha(1-q)}{\varepsilon} \left(\mathbb{E}_M \mathbb{E}_N - 1\right) MN P_{M,N} + \\
&+ \frac{1}{\varepsilon} \left(\mathbb{E}_N - 1\right) N P_{M,N}.
\end{aligned} \tag{5.3}$$

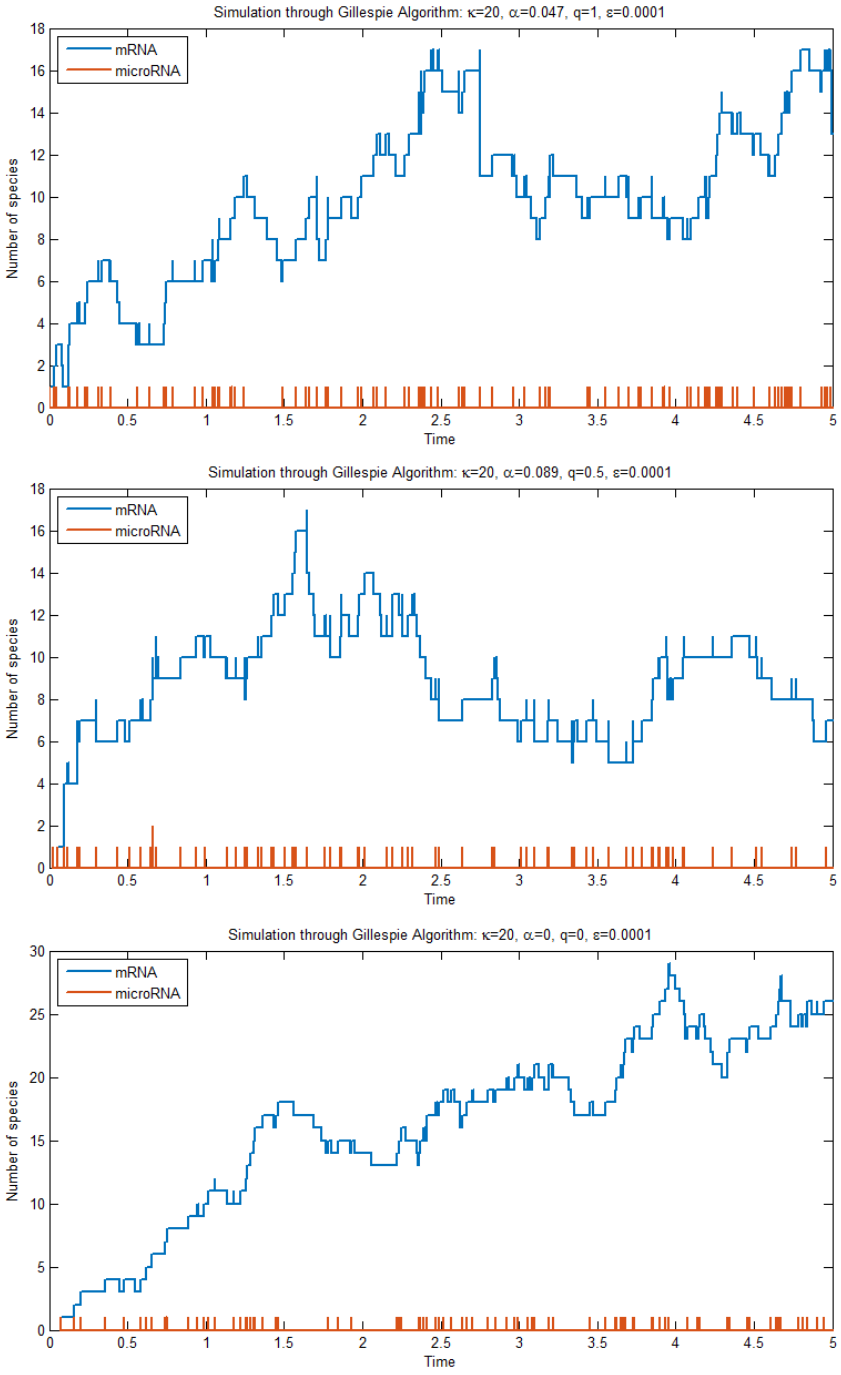

Figure 5.1: *Trajectories of mRNA and microRNA simulated by Gillespie algorithm.*

Grouping the terms of same order in (5.3) together and evaluating the Master equation in the steady state yields

$$\begin{aligned}\varepsilon\left(\kappa\left(\mathbb{E}_M^{-1}\mathbb{E}_N^{-1}-1\right)P_{M,N}+\left(\mathbb{E}_M-1\right)MP_{M,N}\right)+\left(\mathbb{E}_N-1\right)NP_{M,N}+\\ +\alpha\left(\left(q+(1-q)\mathbb{E}_N\right)\mathbb{E}_M-1\right)MNP_{M,N}=0.\end{aligned} \tag{5.4}$$

As we observe two species, we use unit shift operators with respect to each variable in the Master equation: $\mathbb{E}_M f(M,N)=f(M+1,N), \quad \mathbb{E}_N f(M,N)=f(M,N+1)$. Let us define the generating function for two-dimensional sequence $G(x,y)$ by

$$G(x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}P_{m,n}x^m y^n. \tag{5.5}$$

In order to transform (5.4) into the generating function form we need to multiply both sides of equation by $x^M y^N$ and subsequently sum them with respect to both M and N. Through this transformation $P_{M,N}$ maps into $G(x,y)$. Another term which appears in the Master equation is $MP_{M,N}$. After the transformation we obtain

$$\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}P_{M,N}Mx^M y^N=x\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}P_{M,N}Mx^{M-1}y^N=x\frac{\partial G(x,y)}{\partial x}.$$

It follows that multiplying by M in space of bivariate probability is analogous to applying operator $x\frac{\partial}{\partial x}$ in space of generating functions. Using the same rationale multiplying by N transforms into applying $y\frac{\partial}{\partial y}$. Let us now focus on shift operators in Master equation. Let us calculate

$$\begin{aligned}&\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}\mathbb{E}_M(P_{M,N})x^M y^N=\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}P_{M+1,N}x^M y^N=\\ &=\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}P_{M,N}x^{M-1}y^N=\frac{1}{x}\sum_{M=0}^{\infty}\sum_{N=0}^{\infty}P_{M,N}x^M y^N=\frac{1}{x}G(x,y),\end{aligned} \tag{5.6}$$

which implies that applying $\mathbb{E}_N$ in probability space transforms into multiplying by $\frac{1}{x}$ in generating function space. Analogously we can write remaining transformation rules as $\mathbb{E}_N\to\frac{1}{y},\ \ \mathbb{E}_M^{-1}\to x,\ \ \mathbb{E}_N^{-1}\to y$.

Applying the transformation rules mentioned above on (5.4), we find that the generating function of the steady-state probability distribution $P_{M,N}$ satisfies a second-order partial differential equation

$$\begin{aligned}\varepsilon\left(\kappa\left(xy-1\right)G+\left(\frac{1}{x}-1\right)x\frac{\partial G}{\partial x}\right)+\left(\frac{1}{y}-1\right)y\frac{\partial G}{\partial y}+\\ +\alpha\left(\left(q+(1-q)\frac{1}{y}\right)\frac{1}{x}-1\right)xy\frac{\partial^2 G}{\partial x\partial y}=0,\end{aligned} \tag{5.7}$$

which can be simplified and rewritten as

$$\alpha\,(1-q+qy-xy)\,\frac{\partial G}{\partial x\partial y}+(1-y)\frac{\partial G}{\partial y}+\varepsilon\left(\kappa(xy-1)G+(1-x)\frac{\partial G}{\partial x}\right)=0. \quad (5.8)$$

Abbreviated form G stands for $G(x,y)$. The requirement that the probabilities $P_{M,N}$ sum to one implies that the generating function satisfies the normalization condition $G(1,1)=1$, which can be used as an additional condition for (5.8).

5.2 Reduction to a 2nd-order ODE

In this moment we can incorporate the fact that the probabilities of observing nonzero amounts of Y are small, and we rescale the probability by

$$P_{M,N}=\varepsilon^N Q_{M,N}, \quad (5.9)$$

which in terms of the generating function translates to

$$G(x,y)=F(x,\varepsilon y),\quad \text{where}\quad F(x,\omega)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}Q_{m,n}x^m\omega^n. \quad (5.10)$$

As ε tends to 0, we can use Taylor expansion to represent F and $\partial F/\partial\omega$ by their leading order terms. Evaluating F and $\partial F/\partial\omega$ at $\omega=0$ yields a function depending only on x, we denote it $f(x)$ and $g(x)$ respectively, i.e.

$$F(x,0;\varepsilon)=f(x)+O(\varepsilon),\quad \frac{\partial F}{\partial\omega}(x,0;\varepsilon)=g(x)+O(\varepsilon). \quad (5.11)$$

In order to obtain value of $P_{M,N}$ we need to find the coefficient before term x^My^N. It is clear that probability greater than $O(\varepsilon)$ can exist only for $N=0$. We can extract the value of probability by repeatedly differentiating the equality $G(x,y)=F(x,\varepsilon y)=f(x)+O(\varepsilon)$ with respect to x and y and then setting $x=y=0$, which yields

$$P_{M,N}=\frac{\delta_{N,0}}{M!}\left.\frac{\mathrm{d}^M f}{\mathrm{d}x^M}\right|_{x=0}+O(\varepsilon). \quad (5.12)$$

Thus, assuming that (5.9) holds (thus for $N=0$ we can use $P_{M,N}=Q_{M,N}$), the probability distribution is completely determined, at the leading order, by the function $f(x)$. Using this information our task now reduced to determining the value of $f(x)$.

Substituting $y=\omega/\varepsilon$ into (5.8), we obtain a partial differential equation for the rescaled generating function $F(x,\omega)$, which reads

$$\begin{aligned}\kappa x\omega F-\alpha(x-q)\omega\frac{\partial^2 F}{\partial x\partial\omega}-\omega\frac{\partial F}{\partial\omega}+\\ +\varepsilon\left(-\kappa F+(1-x)\frac{\partial F}{\partial x}+\alpha(1-q)\frac{\partial F}{\partial x\partial\omega}+\frac{\partial F}{\partial\omega}\right)=0.\end{aligned} \quad (5.13)$$

We can obtain the first part of information by combining (5.13) with (5.11) and collecting $O(1)$ terms. As all $O(1)$ terms contain variable ω, we can divide the resulting equation by ω and afterward we can take $\omega \to 0$ to obtain the limiting behaviour which can be expressed in terms of functions $f(x)$ and $g(x)$ defined on the right-hand sides of (5.11). This yields

$$\kappa x f - \alpha(x-q)\frac{\mathrm{d}g}{\mathrm{d}x} - g = 0. \tag{5.14}$$

The second part of information can be obtained by letting $\omega = 0$ in (5.13) and collecting $O(\varepsilon)$ terms, which yields

$$-\kappa f - (x-1)\frac{\mathrm{d}f}{\mathrm{d}x} + \alpha(1-q)\frac{\mathrm{d}g}{\mathrm{d}x} + g = 0. \tag{5.15}$$

At this stage we would like to use this system of two equations to eliminate one of the unknown functions f and g. In order to eliminate g, we can add equations (5.14) and (5.15) up, and then divide the result by $x-1$, obtaining

$$\kappa f - \frac{\mathrm{d}f}{\mathrm{d}x} - \alpha\frac{\mathrm{d}g}{\mathrm{d}x} = 0. \tag{5.16}$$

In order to eliminate $\mathrm{d}g/\mathrm{d}x$ we add up the $(1-q)$-multiple of (5.14) and the $(x-q)$-multiple of (5.15) before dividing the result by $x-1$, whereby we obtain

$$-\kappa q f - (x-q)\frac{\mathrm{d}f}{\mathrm{d}x} + g = 0. \tag{5.17}$$

In order to eliminate g completely, we should change g in (5.17) to $\mathrm{d}g/\mathrm{d}x$, as this term appears also in (5.16). Differentiating (5.17), we get

$$-\kappa q\frac{\mathrm{d}f}{\mathrm{d}x} - \frac{\mathrm{d}}{\mathrm{d}x}\left((x-q)\frac{\mathrm{d}f}{\mathrm{d}x}\right) + \frac{\mathrm{d}g}{\mathrm{d}x} = 0. \tag{5.18}$$

Adding up $1/\alpha$-multiple of (5.16) and 1-multiple of (5.18) to eliminate $\mathrm{d}g/\mathrm{d}x$, we arrive at a ordinary differential equation of the second order for f, which reads

$$\frac{\mathrm{d}}{\mathrm{d}x}\left((x-q)\frac{\mathrm{d}f}{\mathrm{d}x}\right) + \left(\kappa q + \frac{1}{\alpha}\right)\frac{\mathrm{d}f}{\mathrm{d}x} - \frac{\kappa}{\alpha}f = 0. \tag{5.19}$$

In the next part we seek to solve this ODE.

5.3 Solving the 2nd-order ODE

Let us recall the generalized hypergeometric differential equation (1.12) from the Section 1.2.2. Now we check if we can transform equation (5.19) into such a form.

We are looking for the numbers p and q, which determine the order of hypergeometric function ${}_pF_q$ (not to be confused with the reaction rate parameter q from (5.2)). From the absence of term $x\frac{\mathrm{d}f}{\mathrm{d}x}$ in (5.19) we can deduce that only feasible value of p is 0. And as the equation is a second-order ODE, we reduce the set of possible values of q to 1.

Substituting the values $p = 0,\ q = 1$ into (1.12) yields that hypergeometric function ${}_0F_1(;\beta;x)$ is a solution to equation

$$x\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} + \beta\frac{\mathrm{d}f}{\mathrm{d}x} - f = 0. \tag{5.20}$$

Let us transform our 2nd-order ODE (5.19) obtained from the study of the stochastic model into the canonical form (5.20). Introducing the substitution $\frac{\kappa}{\alpha}(x - q) = y$ into (5.19) and, using the implied fact that

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\kappa}{\alpha}\frac{\mathrm{d}f}{\mathrm{d}y},$$

we can rewrite (5.19) as

$$\frac{\kappa}{\alpha}\frac{\mathrm{d}}{\mathrm{d}y}\left(y\frac{\mathrm{d}f}{\mathrm{d}y}\right) + \frac{\kappa}{\alpha}\left(\kappa q + \frac{1}{\alpha}\right)\frac{\mathrm{d}f}{\mathrm{d}y} - \frac{\kappa}{\alpha}f = 0.$$

After differentiating the first term and dividing by $\frac{\kappa}{\alpha}$ we come to equation

$$y\frac{\mathrm{d}^2 f}{\mathrm{d}y^2} + \left(\kappa q + \frac{1}{\alpha} + 1\right)\frac{\mathrm{d}f}{\mathrm{d}y} - f = 0,$$

which has the form of (5.20); therefore, we have found solutions in the form

$$f(x) = c_0 \cdot {}_0F_1(;\kappa q + \frac{1}{\alpha} + 1; \frac{\kappa}{\alpha}(x - q)), \tag{5.21}$$

in which c_0 is an arbitrary constant. The "doubly confluent" hypergeometric function ${}_0F_1$ is defined by the convergent series

$${}_0F_1(a, z) = \sum_{m=0}^{\infty}\frac{z^m}{(a)_m m!}. \tag{5.22}$$

Imposing the normalisation condition for the probability to sum up to unity, which in terms of our generating functions translates to $f(1) = 1$, we can determine the prefactor c_0 in (5.21) and obtain

$$f(x) = \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(x - q)\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1 - q)\right)}. \tag{5.23}$$

Basic properties of the doubly confluent hypergeometric function can be established using its power-series representation (5.22). Differentiating (5.22) with respect to z yields

$$\frac{\mathrm{d}}{\mathrm{d}z}{}_0F_1(a,z) = \sum_m \frac{mz^{m-1}}{(a)_m m!} = \sum_m \frac{mz^m}{(a)_{m+1} m!} = \frac{{}_0F_1(a+1,z)}{a}. \tag{5.24}$$

This procedure can be generalized for multiple differentiating and we obtain

$$\frac{\mathrm{d}^m}{\mathrm{d}z^m}{}_0F_1(a,z) = \frac{{}_0F_1(a+m,z)}{(a)_m}. \tag{5.25}$$

Comparing (5.22) with the power-series expansions of the normal and modified Bessel functions [1], we obtain

$${}_0F_1(c,z) = \Gamma(c) z^{\frac{1-c}{2}} I_{c-1}(2\sqrt{z}), \quad {}_0F_1(c,-z) = \Gamma(c) z^{\frac{1-c}{2}} J_{c-1}(2\sqrt{z}), \quad z>0, \tag{5.26}$$

where $\Gamma(z)$ is the gamma function, $J_\nu(z)$ is the Bessel function, and $I_\nu(z)$ is the modified Bessel function of order ν, more precisely

$$J_\nu(z) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{2m+\nu},$$

$$I_\nu(z) = i^{-\nu} J_\nu(ix) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{2m+v}.$$

5.4 Results

Applying the formula for repeated differentiation (5.25) on (5.23) yields

$$\frac{\mathrm{d}^m f(x)}{\mathrm{d}x^m} = \frac{\left(\frac{\kappa}{\alpha}\right)^m}{\left(\kappa q + \frac{1}{\alpha} + 1\right)_m} \times \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1 + m, \frac{\kappa}{\alpha}(x-q)\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1-q)\right)}, \tag{5.27}$$

which in combination with (5.12) and the properties of generating function provides an approximation

$$P_{M,N} = \frac{\delta_{N,0}\left(\frac{\kappa}{\alpha}\right)^M}{M!\left(\kappa q + \frac{1}{\alpha} + 1\right)_M} \times \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1 + M, -\frac{\kappa q}{\alpha}\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1-q)\right)} + O(\varepsilon) \tag{5.28}$$

for the desired probability distribution. In order to find mean and variance of distribution (5.28), let us calculate its factorial moments (theory in Section 1.1.1). Evaluating the derivatives of $f(x)$ at $x=1$, we obtain

$$\begin{aligned} \mu_{(M)} &= \langle X(X-1)\cdot\ldots\cdot(X-M+1)\rangle = \left.\frac{\mathrm{d}^M f(x)}{\mathrm{d}x^M}\right|_{x=1} + O(\varepsilon) \\ &= \frac{\left(\frac{\kappa}{\alpha}\right)^M}{\left(\kappa q + \frac{1}{\alpha} + 1\right)_M} \times \frac{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1 + M, \frac{\kappa}{\alpha}(1-q)\right)}{{}_0F_1\left(\kappa q + \frac{1}{\alpha} + 1, \frac{\kappa}{\alpha}(1-q)\right)} + O(\varepsilon). \end{aligned} \tag{5.29}$$

At the same time as noting that the mean $\langle X \rangle$ trivially coincides with the first factorial moment $\mu_{(1)}$, using (1.1) we also point out that the other characteristic of interest here, the Fano factor, can be expressed in terms of the first two factorial moments as

$$\mathrm{F} = 1 + \frac{\mu_{(2)}}{\mu_{(1)}} - \mu_{(1)}. \tag{5.30}$$

We expressly mention, without carrying out the somewhat tedious calculation, that the probability distribution (5.28) and the moments can be written in terms of Bessel's functions via (5.26).

5.5 Commentary on special cases

In this section we go through a few special cases of the system based on values of α and q and try to simplify the (leading-order approximations of) probability mass function (5.28). Interestingly, these simplifications can be related back to the properties of the chemical system and its Master equation.

The most trivial case $\alpha = 0$ describes the system in which the interaction effects of Y on X are neglected. After each producing reaction Y almost immediately decays without effect. Thus we can approximate the system with a simple immigration-death process implying that steady-state distribution of X is Poisson with mean κ. Substituting $\alpha \to 0$ into (5.19) we obtain simplified ODR $\kappa f = \frac{\mathrm{d}f}{\mathrm{d}x}$ with the trivial solution $f = e^{\kappa(x-1)}$. This confirms our assumption as it is PGF of the Poisson distribution.

Other special cases can be obtained for boundary values of q, i.e. $q = 0$ and $q = 1$. Given that the hypergeometric function ${}_0F_1(a, z)$ is equal to one at $z = 0$, certain algebraic simplifications are available in the (leading-order approximations of) probability mass function (5.28) if $q = 0$ and the moments (5.29) if $q = 1$.

If $q = 0$, the probability mass function (5.28) becomes, up to the normalisation constant, a rational function of the parameters κ and α. Biologically, $q = 0$ means that one molecule Y cannot degrade more than one X molecule during its lifetime. When a pair of molecules X and Y is created, the probability that X survives until the corresponding Y molecule degrades is given as

$$\frac{(N+1)/\varepsilon}{\alpha(M+1)/\varepsilon + (N+1)/\varepsilon} = \frac{1}{\alpha(M+1)+1}.$$

The complementary probability pertains to the possibility that X is eliminated by its twin molecule Y, whereby Y is also destroyed owing to q being set to 0.

Consequently, no jumps in the copy number of X can occur in the limit of short Y lifetimes ($\varepsilon \to 0$). Therefore the limiting process on the slow timescale will be a one-step random walk with stationary Master equation

$$0 = \left\{ (\mathbb{E}_M - 1)M + (\mathbb{E}_M^{-1} - 1)\frac{\kappa}{\alpha(M+1)+1} \right\} P_M,$$

which can be transformed into the difference equation form

$$P_{M+1} = \frac{1}{M+1}\left(M + \frac{\kappa}{\alpha(M+1)+1} \right) P_M + \frac{1}{M+1}\left(\frac{\kappa}{\alpha M + 1} \right) P_{M-1}. \qquad (5.31)$$

Solving (5.31) we obtain

$$P_M = \frac{\left(\frac{\kappa}{\alpha}\right)^M}{M!\left(\kappa q + \frac{1}{\alpha} + 1\right)_M} P_0,$$

which is a result that agrees with (5.28) for $q = 0$.

If $q = 1$, the moments (5.29) are given by rational functions of the model parameters; our results for the factorial moments of first and second order can then be equivalently obtained from a set of algebraic moment equations, which turn out to be closed in this particular case [45].

5.6 Numerical examples

The Fano factor of species X, as given by (5.29)–(5.30), exhibits a non-monotonous response to an increase in the strength α of its interaction with species Y (Figure 5.2). The Poissonian character and thus $F = 1$ for the case $\alpha = 0$ can be also observed in the figure.

Interesting result for the Fano factor is that the Fano factor always eventually climbs up as interaction strengthens; this holds even in the case when Y always degrades in its interaction with X ($q = 0$). However, the increase in the Fano factor that occurs in the $q = 0$ case is substantially slower than the increase observed when the probability q of Y surviving the interaction is non-zero.

The circles in Figure 5.2 show individual values of Fano factor estimated by stochastic Gillespie simulation of the chemical system (5.1)–(5.2) with a finite value of $\varepsilon = 0.01$ (our analytic results being valid in the limit of $\varepsilon \to 0$). Each dot is calculated by averaging over 10^5 observations obtained by simulating the process until it reached stationarity. We observe a good agreement between the numerical and analytic results.

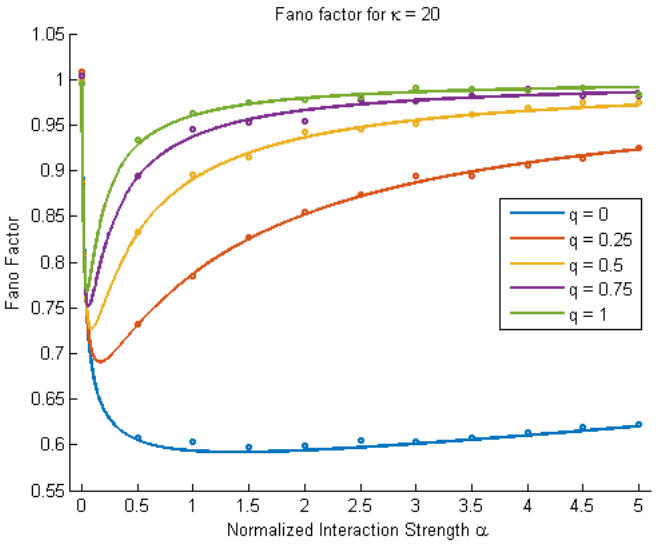

Figure 5.2: *Fano Factor of X as a function of the interaction strength with Y.*

In further analysis, we focus on the distribution of M for a particular chosen set of parameters. We continue to use $\kappa = 20$ for the normalised production rate, select three values, 1, 0.5 and 0 of the probability q of Y-survival, and calculate the values of α at which the Fano factor is minimal and thus the distribution is most 'extreme'. These values are 0.047, 0.089 and 1.385, respectively. For the three parameter sets we use (5.28) to determine the analytic distribution of species X copy number (Figure 5.3, which is displayed as blue circles joined by lines). Additionally, we construct empirical histograms of the exact process by performing 10^5 independent Gillespie simulations of the chemical system (5.1)–(5.2) with $\varepsilon = 0.1$ (Figure 5.3, panels on the left, green bars) and also with $\varepsilon = 0.01$ (Figure 5.3, panels on the right, green bars). For contrast with referential Poissonian statistics we include a best-fit Poisson distribution (using MLE) in each of the panels (Figure 5.3, represented by red circles joined by lines).

From all panels it is clear that the species X copy number distributions are narrower than Poissonian distributions, verifying previous results of Fano factor being less than one. Contrasting the cases with $\varepsilon = 0.1$ on the left with $\varepsilon = 0.01$ on the right we observe that, as can reasonably be expected, decreasing ε leads to a better agreement with our analytic results which were derived assuming $\varepsilon \to 0$. It is clear that $\varepsilon = 0.1$ is not small enough to obtain good fit; on the other hand, simulated distribution with $\varepsilon = 0.01$ are almost indistinguishable from the analytic values using bare eye. Overall, we see a good agreement between theory and simulations.

By Figure 5.2, the Fano factor is close to one if α is very small or, contrastingly,

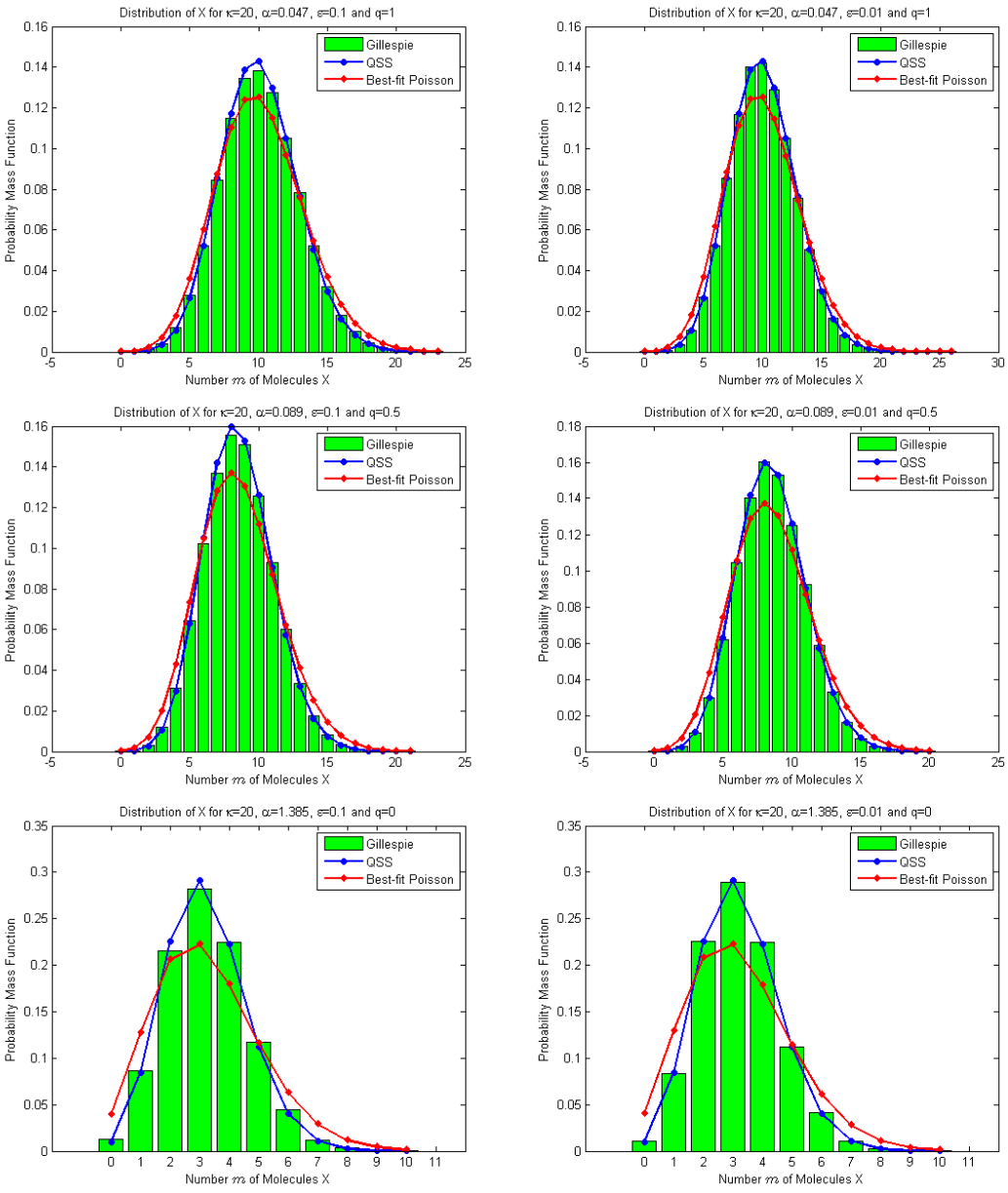

Figure 5.3: *Species X copy-number distributions (analytic and numerical results).*

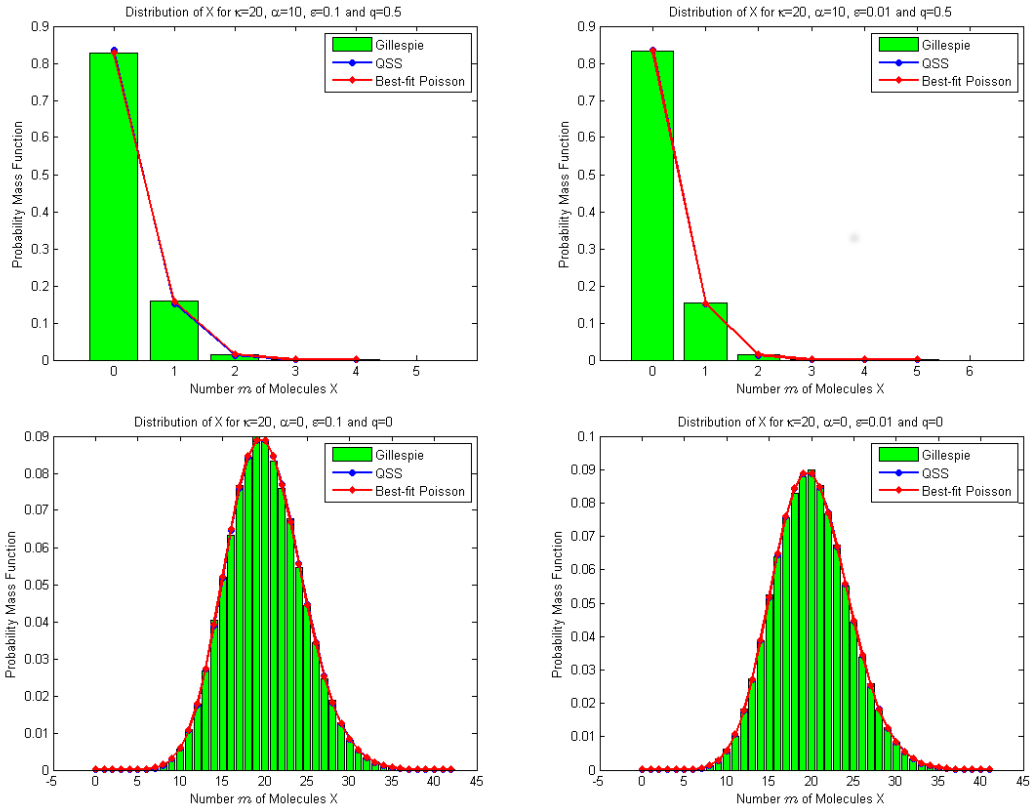

Figure 5.4: *Species X copy-number distributions (analytic and numerical results).*

it is very large. Closeness of the Fano factor to one suggests that the underlying distribution will be Poissonian if $\alpha \ll 1$ or $\alpha \gg 1$ as was also discussed in Section 5.5. Examples shown in Figure 5.4 confirm such supposition with the first row of panels showing examples in which α is very large (10) and the second row of panels showing examples in which α is set to zero. In either situation, the quasi-steady-state description yields a comparably accurate approximation to stochastic simulation results as the best-fit Poisson distribution.

Conclusion

In the first two chapters we summarised the theory regarding biochemical reactions as well as relevant information about probability and ordinary and partial differential equations. We introduced two approaches to reaction modelling: deterministic and stochastic; we chose the stochastic as the main approach. In the next part we introduced a simplified gene expression model in the presence of (decoy) binding sites. We presented its Master equation, which does not have closed-form solution. We derived the distribution of total protein and then we employed singular-perturbation reduction techniques to obtain a quasi-steady-state approximation. Using this approximation we were able to obtain explicit formula for the free protein distribution. In addition to quasi-steady-state approximation, we introduced and compared two other methods to obtain free protein distribution. First one was the stochastic simulation through Gillespie algorithm and the second one numerical solving of the stiff system of ODEs. Comparing with other methods, we justified the correctness of quasi-steady-state formula. Then we employed this formula to observe statistical moments for a wide range of input parameters. We focused on the Fano factor, which yielded substantially different results from Poissonian case. In the fourth chapter we extended our model with the assumption of large system size, using the dissociation constant as its measure. With linear noise approximation we obtained simple expression for the Fano factor of free protein distribution. With the help of numerical simulation we showed the consistency with results from previous chapter. In the final chapter we applied similar methods on mRNA – microRNA system of reactions. We obtained explicit formula for mRNA distribution and compared it with numerical simulations. Finally we studied the distribution for many input parameters and demonstrated the differences between the Fano factor and the benchmark Poissonian case. Although we applied our methodologies on relatively simple models, we expect that it can be helpful to employ analogous approaches in other stochastic models of gene expression or more general biological systems.

Bibliography

[1] Abramowitz, M., Stegun, I.: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Washington, D.C., 1972.

[2] Alon, U.: *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC, 2007.

[3] Bartel, D.P.: *MicroRNAs: genomics, biogenesis, mechanism, and function*, Cell 116 (2): 281-297, 2004.

[4] Bhattacharyya, A.: *On a measure of divergence between two statistical populations defined by their probability distributions*, Bulletin of the Calcutta Mathematical Society 35: 99–109, 1943.

[5] Bokes, P., Singh, A.: *Protein copy number distributions for a self-regulating gene in the presence of decoy binding sites* PLoS ONE Vol. 10, No. 3: 1-19, 2015.

[6] Bokes, P., Hojčka, M., Singh, A., : *Buffering gene expression noise by microRNA based regulation*, manuscript submitted for publication

[7] Burger, A., Walczak, A.M., Wolynes, P.G.: *Abduction and asylum in the lives of transcription factors*, P. Natl. Acad. Sci. USA Vol. 107: 4016-4021, 2010.

[8] Burger, A., Walczak, A.M., Wolynes, P.G.: *Influence of decoys on the noise and dynamics of gene expression*, Physical Review E Vol. 86(4): 041920, 2012.

[9] Cao, Y., Gillespie, D.T., Petzold, L.R.: *The slow-scale stochastic simulation algorithm*, Journal of Chemical Physics, 122(1), 014116, 2005.

[10] Casella, G., Berger, R. L.: *Statistical inference: Second edition*, Thomson Learning, Ann Arbor, 2002.

[11] Chen, W.W., Neipel, M., Sorger, P.K.: *Classic and contemporary approaches to modeling biochemical reactions*, Genes Dev 24 (17): 1861–1875, 2010.

[12] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: *Stochastic gene expression in a single cell*, Science Vol. 297: 1183-1186, 2002.

[13] Erban, R., Chapman, S.J., Maini, P.K.: *A practical guide to stochastic simulations of reaction-diffusion processes*, arXiv:0704.1908, 2007.

[14] Euler, L.: *Institutiones Calculi Integralis, vol. 1 of Opera Omnia Series*, 1769.

[15] Feinberg, M.: *Lectures onChemical Reaction Networks*, Lectures delivered at the Mathematics Research Center, University of Wisconsin Madison, 1979.

[16] Fritz, J.: *Partial differential equations (4th ed.)*, Springer, 1991.

[17] Gauss, C.F.: *Disquisitiones Generales Circa Seriem Infinitam, vol. 3*, Werke, 1813.

[18] Ghaemi, R., Del Vecchio, D.: *Stochastic analysis of retroactivity in transcriptional networks through singular perturbation*, Americal Control Conference, 2012: 2731-6.

[19] Gillespie, D.T.: *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions*, Journal of Computational Physics 22 (4): 403–434, 1976.

[20] Gillespie, D.T.: *Exact Stochastic Simulation of Coupled Chemical Reaction*, The Journal of Physical Chemistry 81 (25): 2340–2361, 1977.

[21] Hojčka, M., Bokes, P.: *Non-monotonicity of Fano factor in a stochastic model for protein expression with sequesterisation at decoy binding sites*, Biomath 6, 1710217, 2017.

[22] Johnson, N., Kotz, S., Kemp, A.: *Univariate Discrete Distributions, 3rd ed.*, Wiley-Interscience, 2005.

[23] Kang, H-W., Kurtz, T.G.: *Separation of time-scales and model reduction for stochastic reaction networks*, The Annals of Applied Probability Vol. 23, No. 2, 529-583, 2013.

[24] Kang, H-W., Kurtz, T.G., Popovic, L.: *Central limit theorems and diffusion approximations for multiscale Markov chain models*, The Annals of Applied Probability Vol. 24, No. 2, 721-759, 2014.

[25] Kim, J.K., Sontag, E.D.: *Reduction of multiscale stochastic biochemical reaction networks using exact moment derivation*, PLOS Computational Biology Vol.13(6): e1005571, 2017.

[26] Knuth, D.E., Graham R.L., Patashnik O.: *Concrete mathematics*, Addison Wesley, 1989.

[27] Koshland, D.E., Némethy, G., Filmer, D.: *Comparison of experimental binding data and theoretical models in proteins containing subunits*, Biochemistry 1966 Jan; 5(1):365-85, 1966.

[28] Latchman, D.S.:*Transcription factors: an overview*, The International Journal of Biochemistry & Cell Biology 29 (12): 1305–12, 1997.

[29] Laurenzi, I.J.: *An analytical solution of the stochastic master equation for the reversible bimolecular reaction kinetics*, The Journal of Chemical Physics 08/2000; 113(8): 3315-3322.

[30] Lee, R.C., Feinbaum, R.L., Ambros, V.: *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*, Cell 75: 843-854, 1993.

[31] Lee, T.H., Maheshri, N.: *A regulatory role for repeated decoy transcription factor binding sites in target gene expression*, Mol. Syst. Biol. Vol. 8: 576, 2012.

[32] McAdams, H.H., Arkin, A.: *Stochastic mechanisms in gene expression*, P. Natl. Acad. Sci. USA Vol. 94: 814-819, 1997.

[33] Michaelis, L., Menten, M.L.: *Die Kinetik der Invertinwirkung*, Biochem Z 49: 333–369, 1913.

[34] Murray, J.D.: *Mathematical biology: An introduction*, Springer, 2002.

[35] Nordsieck, A., Lamb, W.E., Uhlenbeck, G.E.: *On the theory of cosmic-ray showers in the furry model and the fluctuation problem*, Physica 7: 344-360, 1940.

[36] Nussbaum, R.L., McInnes, R.R.; Willard, H.: *Thompson & Thompson Genetics in Medicine (8 ed.)*, Elsevier, 2016.

[37] Øksendal, B.: *Stochastic Differential Equations*, Springer, 2000.

[38] Feigelman, J., Marr, C., Popovic, N.: *A Case Study on the Use of Scale Separation-Based Analytic Propagators for Parameter Inference in Stochastic Gene Regulation*, J. Coupled Syst. Multiscale Dyn. 3(2): 164-173, 2015.

[39] Popovic, N., Marr, C., Swain, P.S.: *A geometric analysis of fast-slow models for stochastic gene expression*, Journal of Mathematical Biology 72(1): 87-122, 2016.

[40] Schauer, M., Heinrich, R.:*Quasi-steady-state approximation in the Mathematical Modeling of Biochemical Reaction Networks*, Mathematical Biosciences Vol. 65: 155-170, 1983.

[41] Ševčovič, D.: *Parciálne deferenciálne rovnice a ich aplikácie*, IRIS, 2008.

[42] Shampine, L. F., Gear, C. W.:*A user's view of solving stiff ordinary differential equations*, SIAM Review 21 (1), 1–17, 1979.

[43] Shampine, L. F., Reichelt, M. W.: *The MATLAB ODE Suite*, SIAM Journal on Scientific Computing, Vol. 18, 1–22, 1997.

[44] Shahrezaei, V., Swain, P.S.: *Analytical distributions for stochastic gene expression*, P. Natl. Acad. Sci. USA Vol. 105: 17256–17261, 2008.

[45] Singh, A., Hespanha, J.P., *Moment closure techniques for stochastic models in population biology*, Proc. Amer. Control Conf., 4730-4735, 2006.

[46] Soltani, M., et al.: *Nonspecific transcription factor binding can reduce noise in the expression of downstream proteins*, Physical Biology 12(2015), 055002.

[47] Stiefenhofer, M.: *Quasi-steady-state approximation for chemical reactional networks*, Journal of Mathematical Biology Vol. 36: 593-609, 1998.

[48] Taniguchi, Y., et al.: *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells*, Science Vol. 329: 533-538, 2010. doi.org/10.2142/biophys.51.136

[49] Van Kampen N.G.: *Stochastic processes in physics and chemistry*, Elsevier Science B.V., 1992.

[50] Weiss, N.A.: *A Course in Probability*, Addison–Wesley, 2005.

[51] Wightman, B., Ha, I., Ruvkun, G.: *Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans*, Cell 75: 855-862, 1993.

[52] Wunderlich, Z., Mirny, L.A.: *Different gene regulation strategies revealed by analysis of binding motifs*, Trends Genet. 25(10): 434-440, 2009.

[53] Yu, J., Xiao, J., Ren, X., Lao, K., Xie, X.S.: *Probing gene expression in live cells, one protein molecule at a time*, Science Vol. 311: 1600-1603, 2006. doi.org/10.1126/science.1119623