TAMPERE UNIVERSITY OF TECHNOLOGY

JARNO MÄKELÄ

DYNAMICS OF STOCHASTIC SEQUENCE-LEVEL MODELS OF TRANSCRIPTION AND TRANSLATION IN PROKARYOTES

Master of Science Thesis

Subject Approved by Department
Council 09.03.2011
Examiners: Professor Olli Yli-Harja,
Assist. Prof. Andre Ribeiro

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Biotechnology
MÄKELÄ, JARNO: Dynamics of stochastic sequence-level models of transcription and translation in prokaryotes
Master of Science Thesis, 59 pages
June 2011
Major: Computational Systems Biology
Examiners: Professor Olli Yli-Harja and Assistant Professor Andre Ribeiro
Keywords: Stochastic model, transcription, translation, fluorescence measurements

In prokaryotes, transcription and translation are dynamically coupled, as the latter starts before the former is completed. Also, from one transcript, several translation events occur in parallel. To study how events in transcription elongation affect translation elongation and fluctuations in protein levels, we propose a delayed stochastic model of prokaryotic transcription and translation at the nucleotide and codon level that includes the promoter open complex formation and alternative pathways to elongation, namely pausing, arrests, editing, pyrophosphorolysis, RNA polymerase traffic, and premature termination. Stepwise translation can start after the ribosome binding site is formed and accounts for variable codon translation rates, ribosome traffic, back-translocation, drop-off, and trans-translation.

The recent development of measurement techniques in genetics promises better understanding of the functioning of biological systems. To attain the most out of these techniques, new methods are needed of interpreting the data, since most existent methods have been developed to analyze population level measurements, rather than extracting information from single cell dynamics. For example, one needs accurate estimation of the measurement noise from single cell measurements of gene expression. We use recently developed methods to measure gene expression *in vivo* in individual cells, at the single RNA and protein molecule levels. Such measurements of gene expression, attained in various conditions, as well as the proposed modeling strategy, are used to study and model the dynamics of gene expression at the single event level and to estimate noise sources in the processes.

First, the model is shown to accurately match the measurements of sequence-dependent translation elongation dynamics. Next, the degree of coupling between fluctuations in RNA and protein levels, and its dependence on the rates of transcription and translation initiation is characterized. Finally, sequence-specific transcriptional pauses are found to have an effect on protein noise levels. For parameter values within realistic intervals, transcription and translation are found to be tightly coupled in *Escherichia coli*, as the noise in protein levels is mostly determined by the underlying noise in RNA levels. Sequence-dependent events in transcription elongation, e.g. pauses, are found to cause tangible effects in the degree of fluctuations in protein levels, implying that these are evolvable.

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Biotekniikan koulutusohjelma
MÄKELÄ, JARNO: Stokastisten sekvenssitason mallien transkriptio- ja translaatiodynamiikka prokaryooteissa
Diplomityö, 59 sivua
Kesäkuu 2011
Pääaine: Laskennallinen systeemibiologia
Tarkastajat: Professori Olli Yli-Harja ja yliassistentti Andre Ribeiro
Avainsanat: Stokastinen malli, transkriptio, translaatio, fluoresenssimittaukset

Prokaryooteissa transkriptio ja translaatio tapahtuvat samanaikaisesti, jolloin jälkimmäinen voi alkaa ennen kuin edellinen on loppunut. Yhdestä RNA:sta voidaan myös tuottaa useita proteiineja samanaikaisesti. Tässä työssä tutkimme kuinka transkriptio vaikuttaa translaatioon ja proteiinitasojen heilahteluun. Rakensimme stokastisen mallin prokaryoottien transkriptiosta ja translaatiosta nukleotidin tarkkuudella, joka sisältää polymeraasi-promoottorikompleksin muodostumisreaktiot ja useita vaihtoehtoisia kilpailevia reaktioita prosessin sisällä, kuten hetkittäiset ja pidemmät RNAp:n pysähtymiset, virheiden korjaamiset, pyrofosforilaatiot, RNAp:n väliset ruuhkat ja ennenaikaiset irtoamiset DNA:sta. Portaittainen translaatio alkaa heti kun ribosomin sitoutumispaikka RNA:ssa on muodostettu ja ottaa huomioon kodonien vaihtelevat translaationopeudet, ribosomien väliset ruuhkat, erisuuntaiset translokaatiot, ribosomien irtoamiset RNA:sta ja trans-translaation.

Viimeaikaiset mittaustekniikan kehitysaskeleet genetiikassa auttavat meitä ymmärtämään biologisia mekanismeja paremmin. Hyödyntääksemme näitä mittaustuloksia täydellisesti, tarvitsemme uusia työkaluja datan käsittelyyn ja analysointiin, koska useimmat olemassa olevat metodit soveltuvat ainoastaan populaatiotason mittausdataan, eivätkä yksittäisistä soluista saataviin aikasarjoihin. Hyvänä esimerkkinä pelkästään mittauskohinan arviointi yksittäisten solujen kuvadatasta on haastavaa. Tässä työssä käytämme viime vuosina kehitettyjä mittaustapoja, joilla voidaan tutkia reaaliaikaista geeniekspressiota *in vivo* yksittäisissä soluissa yksittäisten RNA- ja proteiinimolekyylien tasolla. Sovellamme näitä geeniekspression mittaustapoja, kuten myös edellä mainittua lähestymistapaa mallinnuksessa, tutkiaksemme geeniekspression dynamiikkaa ja säätelymekanismeja.

Havaitsemme mallin sopivan tarkasti sekvenssipohjaisen translaation mittaustuloksiin. Tutkimme RNA- ja proteiinitasojen välisten heilahtelujen yhtenevyyden ja riippuvuuden transkription ja translaation aloitusnopeuksista. Lopuksi analysoimme sekvenssiriippuvaisten RNAp:n pysähtymiset ja niiden vaikutukset proteiinin tuottoon. Realistisilla parametreilla transkriptio ja translaatio ovat tarkasti synkronoituja *Escherichia coli*ssa ja proteiinitasojen heilahtelut määrittää enimmäkseen RNA-tasojen heilahtelu. Sekvenssiriippuvaiset transkription säätelymekanismit, kuten esimerkiksi RNAp:n pysähtymiset, vaikuttavat konkreettisesti proteiinitasojen heilahteluun.

PREFACE

This Master’s thesis is carried out in the Department of Signal Processing of the Faculty of Computing and Electrical Engineering at Tampere University of Technology.

First of all, I would like to thank my thesis supervisor, Andre Ribeiro, PhD, for his guidance throughout this work. I am grateful for always being available for discussion, introducing me with the topic, helping me during different phases of this work and providing me his invaluable comments to improve this work. My gratitude goes as well to professor Olli Yli-Harja who introduced me to the Computational Systems Biology research group, guided my academic studies and examined my master’s thesis.

My warm thanks also got to all my colleagues in the Laboratory of Biosystem Dynamics for their help and the inspiring work atmosphere they provided. I am especially grateful for the contribution and advices of Jason Lloyd-Price. Without such cooperation this thesis would not have been possible in its present form.

Finally, I would like to express my heartfelt thanks to my family. Without their help and encouragement this work would never have been completed.

Tampere, May 12, 2011

Jarno Mäkelä

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND NOTATION

aTc	Anhydrotetracycline
CME	Chemical master equation
CV^2	Variance over the mean squared
DNA	Deoxyribonucleic acid
Eσ	RNAp holoenzyme
E. coli	*Escherichia coli*
FISH	Fluorescence *in situ* hybridization
GFP	Green fluorescent protein
In vivo	Latin for “within the living”
In situ	Latin for “in position”
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LDM	Logarithmic direct method
mRNA	Messenger RNA
OD_{600}	Optical density at a wavelength of 600 nm
ODE	Ordinary differential equations
PCR	Polymerase chain reaction
Pro	Promoter region at DNA
RBS	Ribosome binding site
RFP	Red fluorescent protein
Rib	Ribosome

RNA	Ribonucleic acid
RNAp	RNA polymerase
RP_c	Closed complex of RNA polymerase
RP_o	Open complex of RNA polymerase
RRE	Reaction-rate equation
rRNA	Ribosomal ribonucleic RNA
SSA	Stochastic simulation algorithm
tmRNA	Transfer-messenger RNA
tRNA	Transfer RNA
YFP	Yellow fluorescent protein

1 INTRODUCTION

In prokaryotes, both transcription and translation are stochastic, multi-stepped processes that involve many components and chemical interactions. Several events in transcription and in translation [1-8] are probabilistic in nature, and their kinetics are sequence dependent. One example is sequence-dependent transcriptional pausing [1]. When they occur, these events can affect the degree of fluctuations of RNA and protein levels. Since noise in gene expression affects cellular phenotype, sequence dependent noise sources are subject to selection [9, 10] and are thus evolvable [7]. Recent evidence suggests that these noise sources may be a key for bacterial adaptability in unpredictable or fluctuating environmental conditions [11, 12].

To better understand the evolvability of bacteria, it is important to understand how fluctuations in RNA levels propagate to protein levels. Transcription and translation are coupled in prokaryotes, in that translation can initiate after the formation of the ribosome binding site region of the RNA, which occurs during the initial stages of transcription elongation. The extent to which sequence-dependent events in transcription elongation affect the noise in RNA, and consequently protein levels is largely unknown. Due to this, it is also not yet well understood how phenotypic diversity is regulated in monoclonal bacterial populations.

Two recent experiments have given a preliminary glimpse at the dynamics of production of individual proteins [13] and RNA molecules [14] *in vivo* in bacteria. However, as of yet, there is no experimental setting to simultaneously observe the production of both RNA and proteins at the molecular level. Further, in the aforementioned experiments [13, 14], the rate of gene expression was kept very weak, as otherwise the number of molecules would not be easily quantifiable. This implies that they cannot be used to study the effects of events such as the promoter open complex formation [15]. The present shortcomings of these techniques enhance the need for realistic models of gene expression in prokaryotes.

Several measurements have shed light on the dynamics of transcription and translation elongation [16, 17], and revealed the occurrence of several stochastic events during these processes, such as transcriptional pauses [2, 4]. The kinetics of RNA and protein degradation are also somewhat known [18]. These measurements allowed the recent development of realistic kinetic models of transcription at the nucleotide level [5, 19] and translation at the codon level [20]. These models were shown to match the measurements of RNA production at the molecule level [6, 21] and of translation elongation dynamics at the codon level [20]. In this regard, it was shown that measurements of sequence dependent translation rates of synonymous codons could be modeled with neither deterministic nor uniform stochastic models [20], thus the need for models with explicit translation elongation. Similarly, transcription elongation also needs to be modeled explicitly to accurately capture the fluctuations in RNA levels for fast transcription initiation rates [5, 19, 22].

Here, we propose a model of transcription and translation at the nucleotide and codon level for *Escherichia coli*. The model of transcription is the same as in [5], and includes the promoter occupancy time, transcriptional pausing, arrests, editing, premature termination, pyrophosphorolysis, and accounts for the RNAp footprint in the DNA template. The model of translation at the codon level proposed here is based on the codon-dependent translation model proposed in [20], which includes translation initiation, codon-specific translation rates and the stepwise translation elongation and activation. The model also accounts for the ribosome's footprint in the RNA template as well as the occupancy time of the ribosome binding site. Here, beside these features, we further include the processes of back-translocation, drop-off, and trans-translation. Finally, we include protein folding and activation, as well as degradation, modeled as first-order processes, so as to study fluctuations in the protein levels.

The dynamics of the model follows the Delayed Stochastic Simulation Algorithm [19, 23] and is simulated by a modified version of SGNSim [24]. While it's most relevant innovation is the coupling between realistic stochastic models of transcription and translation at the nucleotide and codon levels, which allows the study of previously unaddressed aspects of the dynamics of gene expression in prokaryotes, this introduces a level of complexity that required simulation capabilities that SGNSim did not possess. Namely, the simulator is required to create and destroy compartments at run time within the reaction vessel, where a separate set of reactions can occur.

We start by validating the dynamics of translation elongation in the model. Next, using realistic parameter values extracted from measurements, we address the following questions: how different are the distributions of time intervals between translation initiation events and between translation completion events, i.e., how stochastic is translation elongation? To what extent do fluctuations in temporal RNA levels propagate to temporal protein levels, and what physical parameters control this propagation of noise between the two? Finally, we investigate whether transcriptional pauses have a significant effect on the dynamics of protein levels.

We use recently developed methods to measure gene expression *in vivo* in individual cells, at the single RNA [14] and protein molecule levels [13]. Such measurements of gene expression, attained in various conditions, as well as the proposed modeling strategy, are used to study the dynamics of gene expression at the single event level and to estimate noise sources in the processes.

The results presented in this thesis are partly from a project done in collaboration with fellow research group members. The original scientific paper was published in the peer-reviewed journal "BMC Bioinformatics": Mäkelä et al., "Stochastic sequence-level model of coupled transcription and translation in prokaryotes," *BMC Bioinformatics* 2011, 12(1):121. The measurements of gene expression were carried out by our research group and are published in the peer-reviewed journal “BMC Molecular Biology”: Smolander et al., “Cell-to-cell diversity in protein levels of a gene driven by the tetracycline inducible $P_{\text{Ltet-o1}}$ promoter,” *BMC Molecular Biology* 2011, *in press.*

Chapter 2 has a description of the background of the stochastic modeling strategy, of the processes of transcription and translation, of the state-of-the-art models of these processes, and of fluorescence microscopy applications. These descriptions provide knowledge to understand the studies described in this thesis. Chapter 3 describes the materials and methods used to obtain the results reported in the thesis. Namely, it describes in detail the novel models of transcription and translation, the concept of time-averaging, the quantification of correlations and the measurements of gene expression. Chapter 4 contains the results from the modeling study and from the measurements of gene expression, as well as an explanation of the underlying mechanisms of these processes. In chapter 5, the conclusions and a discussion on future developments are presented.

2 BACKGROUND

2.1 Stochastic Simulation Algotrithm

2.1.1 Stochastic chemical kinetics

Unimolecular and bimolecular chemical reactions are instantaneous physical events that change the quantities of the chemical species involved. When considering the time evolution of the quantities of chemically reacting molecules, one must know position and velocities of individual molecules over time in order to accurately calculate the dynamics of the system. More complex processes cannot, usually, be accounted for explicitly, unless they are simplified to, for example, combinations or sequences of uni- and bimolecular events.

From the classical mechanics point-of-view, systems of chemical kinetics are considered as deterministic, since provided the complete description of the initial conditions it is possible to predict the evolution of the system. The deterministic approach has been the most common approach to model chemical kinetics, but when the number of molecules involved is small, the deterministic approach is not exact. One can easily find reasons why this approach is not accurate in such conditions [25]: For example, the molecule numbers is not continuous, changing only in discrete amounts. Also, it is not possible to determine, before hand, when a unimolecular event takes place, only the probability of occurrence can be estimated accurately. In many cases the deterministic approach produces realistic results, however, its accuracy should not be taken for granted, e.g. when modeling oscillating systems [26].

In the deterministic approach, the system's dynamics is described by continuous, coupled systems of ordinary differential equations (ODEs). The chemical kinetics is predicted assuming well-stirred, thermally equilibrated systems, and provided the number of molecules X_i of each chemical species S_i (i = 1,…, N), whose temporal evolution follows the set of ODEs of the form [26]:

$$\frac{dX_i}{dt} = f_i(X_i,...,X_N) \quad (i = 1,...,N) \tag{2.1}$$

where the functions f_i are inferred from the specifics of the various reactions. This set of equations forms the reaction-rate equation (RRE) [26].

The stochastic approach attempts to describe the temporal evolution of a well-stirred system of chemical interactions, accounting for the system's discreteness and inherent stochasticity. This approach consists of a single differential-difference equation, the chemical master equation (CME) [26]. Let us consider the following scenario: a well-stirred system of molecules of N chemical species existing in the quantities $\{S_1, . . . , S_N\}$, and interacting through M chemical reactions $\{R_1, . . . , R_M\}$.We assume that the system has a constant volume V and is in thermal but not in chemical equilibrium at constant temperature. We use $X_i(t)$ to denote the number of molecules of species S_i in the system at time t. From this initial condition, we can estimate the state vector $\mathbf{X}(t) \equiv (X_1(t), . . . , X_N(t))$, given that the system was in state $\mathbf{X}(t_0) = \mathbf{x}_0$ at initial time t_0 [26].

The aim of following individual molecules' positions and velocities can be discarded if the system is well stirred. This approximation derives from having elastic, non-reactive, collisions that distribute uniformly the molecules in V, as well as having the velocities of the molecules thermally randomized according to the Maxwell-Boltzmann velocities distribution [26]. Relevantly, one can ignore nonreactive molecular collisions that would consume most of the simulation time and consider only those collisions that change the population numbers of the chemical species.

Chemical reactions are the only cause of changes in numbers of the species in the system. Each reactive reaction R_μ is characterized by two quantities. The first is its state-change vector $\boldsymbol{v}_\mu \equiv (v_{1\mu} , . . . , v_{N\mu})$, where $v_{i\mu}$ is the change in the S_i molecular population caused by reaction R_μ, meaning that if the system is in state $\mathbf{x}$ and one R_μ reaction occurs, the system immediately changes to state $\mathbf{x} + \boldsymbol{v}\mu$. The other characterizing quantity for R_μ is its propensity function a_μ, which is defined as follows [26]:

$$a_\mu(\mathbf{x})dt \triangleq \text{the probability, given } \mathbf{X}(t) = \mathbf{x}\text{, that one } R_\mu \text{ reaction will occur somewhere inside V in the next infinitesimal time interval } [t, t + dt]. \tag{2.2}$$

This is the fundamental assumption behind stochastic chemical kinetics simulations because everything else follows from Probabilities theory. The propensities of the reactions fully characterize the system's state and temporal evolution. The calculation of the propensity for uni and bimolecular reactions is now exemplified.

For unimolecular reaction ($S_1 \rightarrow$ product) R_μ has some constant c_μ, defined by the underlying physics, such that $c_\mu\, dt$ is the probability that any particular molecule of this species will react in the next infinitesimal time interval dt. The laws of probabilities dictate that if there are x_1 S_1 molecules in the system, the probability that one of them will react according to R_μ in the next dt is $x_1 \cdot c_\mu\, dt$ [27]. Thus, the propensity function is $a_\mu(\mathbf{x}) = c_\mu x_1$. A bimolecular reaction involving two species (S1 + S2→ product) has a propensity function of the form $a_j(\mathbf{x}) = c_\mu x_1 x_2$. If it is a bimolecular reaction between two molecules of the same species (S1 + S1 → product), the propensity function is $a_\mu(\mathbf{x}) = (1/2)c_\mu x_1(x_1 - 1)$ [27].

There is a difference between unimolecular and bimolecular c_μ's [26]. The unimolecular c_μ is independent of the system volume V while the bimolecular c_μ is inversely proportional to V. This reflects the fact that increasing the volume V decreases the chances of collisions during dt. When comparing c_μ's with the reaction-rate constant k_μ of deterministic chemical kinetics it turns out that unimolecular c_μ is equal to k_μ while in bimolecular reactions, c_μ equals k_μ / V (or $2k_\mu$ / V if of the same species) [27]. This is consequence of the formulas of deterministic chemical kinetics being an approximation for large number of molecules of the formulas of stochastic chemical kinetics. Propensity functions are grounded in molecular physics and are thus more general than deterministic reactions of chemical kinetics [26].

Although the probabilistic nature of (2.2) prohibits an exact prediction of **X**(t), we can estimate the probability as follows [26]:

$$P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0) \triangleq \text{Prob}\left\{\mathbf{X}(\text{t})\text{=}\mathbf{x}\text{, given }\mathbf{X}(\text{t}_0)=\mathbf{x}_0\right\} \qquad (2.3)$$

Time-evolution equation for $P(\boldsymbol{x}, t \mid \mathbf{x}_0, t_0)$ can be derived by the laws of probability to (2.2). The result is the equivalent chemical master equation (CME) [28, 29]:

$$\frac{\partial P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)}{\partial t}=\sum_{\mu=1}^{M}\left[a_\mu(\mathbf{x}-\mathbf{v}_\mu)P(\mathbf{x}-\mathbf{v}_\mu,t\,|\,\mathbf{x}_0,t_0)-a_\mu(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)\right] \qquad (2.4)$$

The CME defines the function $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ completely via a set of coupled ODEs, with one equation for every possible combination of reactant molecules. The CME can be analytically solved for only a few simple cases.

Inferring anything regarding the evolution of average quantities, such as $\left\langle b(\mathbf{X}(t))\right\rangle \equiv \sum_x b(\mathbf{x})P(\mathbf{x},t\,|\,\mathbf{x}_0,t_0)$ is also difficult for bimolecular reactions [26]. By multiplying the CME (2.4) by **x** and then sum over all **x**, we get

$$\frac{d\left\langle \mathbf{X}(t)\right\rangle}{dt}=\sum_{\mu=1}^{M}\mathbf{v}_\mu\left\langle a_\mu(\mathbf{X}(t))\right\rangle \qquad (2.5)$$

Unimolecular reactions have linear propensity functions in the state variables, and we would have $\left\langle a_\mu(\mathbf{X}(t))\right\rangle = a_\mu(\left\langle \mathbf{X}(t)\right\rangle)$, which means that (2.5) would be a ODE for the first moment $\left\langle \mathbf{X}(t)\right\rangle$ [26]. Bimolecular reactions cause the right-hand side of (2.5) to have at least one quadratic moment of the form $\left\langle X_i(t)X_{i'}(t)\right\rangle$, and (2.5) would become a set of infinite number of open ended of equations for all moments in time. In a rare case when there are no fluctuations, if **X**(*t*) was a deterministic process, (2.5) would be reduced, named RRE (2.1). The RRE is valid if all fluctuations can be ignored, but this is rarely the case.

Because solving the CME (2.4) for the density function of **X**(*t*) is difficult, one way to calculate the dynamics of system is to construct a set of numerical realizations of **X**(*t*), i.e., simulated trajectories of **X**(*t*). This is not the same as solving the probability density function of **X**(*t*) but instead we get random samples of **X**(*t*). To simulate trajectories of **X**(*t*), instead of using the function $P(x, t \mid \mathbf{x}_0, t_0)$, we define a new probability function $p(\tau, \mu \mid \mathbf{x}, t)$, as follows [26]:

$$p\left(\tau,\mu \mid x,t\right)dt \triangleq \text{the probability, given } \mathbf{X}(t) = \mathbf{x}\text{, that the next reaction in the system will occur in the infinitesimal time interval } [t+\tau, t+\tau+d\tau)\text{, and will be an } R_\mu \text{ reaction.} \tag{2.6}$$

This function is the joint probability density function of the two random variables (time until the next reaction (τ) and index of the next reaction (μ)), given that the system is in state **x**. An exact formula for $p(\tau, \mu \mid x, t)$ can be derived from (2.2). The result of this procedure is [25, 27]

$$p(\tau,\mu \mid \mathbf{x},t) = a_\mu(\mathbf{x})\ \exp(-a_0(\mathbf{x})\tau), \tag{2.7}$$

where

$$a_0(\mathbf{x}) \triangleq \sum_{j=1}^{M} a_j(\mathbf{x}). \tag{2.8}$$

This equation is the mathematical basis for the stochastic simulation algorithm [26]. It defines that τ is an exponential random variable with mean (and standard deviation) of $1/a_0(\mathbf{x})$, while μ is an independent random variable with point probabilities $a_\mu(\mathbf{x})/a_0(\mathbf{x})$. There are several exact Monte Carlo algorithms for generating τ and μ according to their distributions. The simplest is the so-called direct method, which uses the standard inversion generating method of Monte Carlo theory [25, 26]: We draw two random numbers r_1 and r_2 from the uniform distribution in the unit interval, and take

$$\tau = \frac{1}{a_0(\mathbf{x})}\ln\left(\frac{1}{r_1}\right), \tag{2.9a}$$

$$\mu = \text{the smallest integer satisfying } \sum_{j=1}^{\mu} a_{j(\mathbf{x})} > r_2 a_0(\mathbf{x}). \tag{2.9b}$$

This and the procedure for simulating the trajectories, constitute the stochastic simulation algorithm (SSA) for constructing exact numerical realizations of the process **X**(*t*) [25, 27]:

1. Initialize time $t = t_0$ and the system's state $\mathbf{x} = \mathbf{x}_0$.
2. With the system in state **x** at time *t*, evaluate all $a_\mu(\mathbf{x})$ and their sum, $a_0(\mathbf{x})$.
3. Generate values for τ and μ from (2.9a) and (2.9b).
4. Execute the next reaction, replacing $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v}_\mu$.
5. Record (**x**, *t*). Return to Step 1, else end the simulation.

The **X**(t) trajectory produced by the SSA is a numerical solution of RRE (2.6). Notice that the time step τ in the SSA is exact and not a finite approximation to some infinitesimal dt, as the time step in any ODE numerical solver [26].

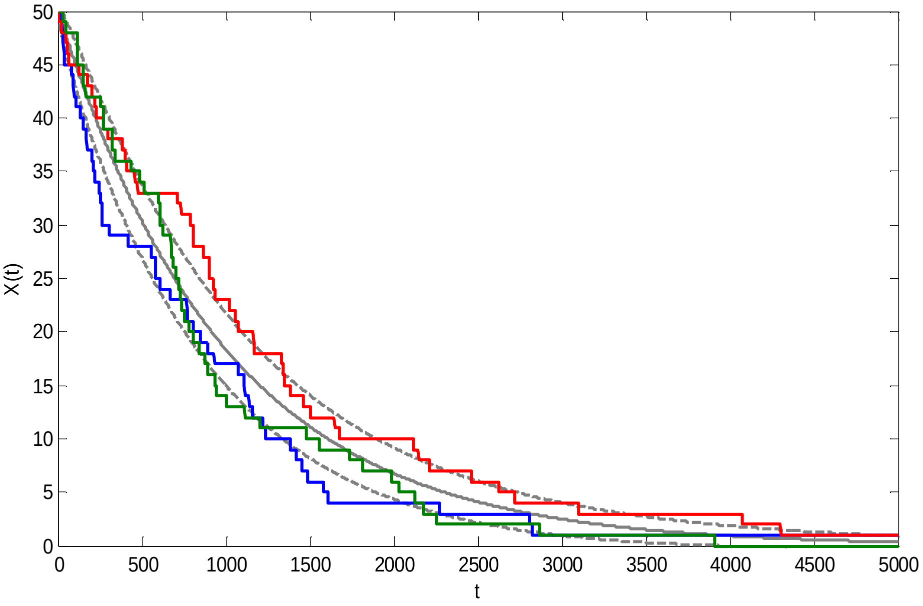

Figure 2.1. *Simulation of a 1*st*-order decay reaction. The continuous grey line is the solution of RRE. The two dashed lines are predicted CME solutions of one-standard-deviation from mean. The red, green and blue jagged lines are trajectories of three independent simulations with the SSA.*

To illustrate differences between deterministic solutions and stochastic simulations, Figure 2.1 shows the RRE as the mean behavior and stochastic trajectories from SSA as individual simulations. The CME solution provides the standard deviation of possible results in every point of time. If we simulate the SSA enough times and average the individual trajectories we attain a “mean dynamics” of the process.

Because the SSA and the CME are derived without the need for approximations (2.2), they are equivalent. When the CME is intractable, the SSA is thus usable. Being a numerical procedure, the SSA is simpler than most procedures used to solve numerically the RRE (2.6). The drawback is that the SSA is often very slow, essentially because it corresponds on simulating individual reaction events. The source for this slowness is the factor $1/a_0(\mathbf{x})$ in (2.9a), which will be small if any population is large, and if so, will force the time steps to be small as well [26].

2.1.2 Improvements to the Stochastic Simulation Algorithm

Given a large number of chemical species and reactions, the direct method of SSA [27] is slow and burdensome. Finding the next reaction is the most time consuming step of SSA. In previous formulations, the search procedure accumulates the sum of $a_i(x)$ for $i = 1, . . . , j$ by adding each propensity until the sum is larger than the product of $a_0\ (x)$

and a uniform random number [25]. The idea behind improved methods is to reduce the average number of operations to find the next reaction [26]. The average of this process is called search depth. The search depth of all current SSA algorithms is highly dependent on the biochemical system simulated. On each step of the direct method, the total propensity a_0 is first calculated by adding all of the propensities a_i together. Then, to select the next reaction, the propensities a_i are summed again. Thus, propensities are summed almost twice. The direct method uses two random numbers per iteration taking a time proportional to the number of reactions to update the propensities, to choose the reaction, and to calculate the time for the next reaction [27]. Attempts were made on reducing the search depth of each iteration, making possible to simulate greater number of reactions.

Gillespie also developed the first reaction method [27] which generates a putative time τ_i for each reaction to occur if no other reaction occurred first. μ is the reaction whose putative time is first, and τ is the putative time τ_μ. The probability distributions used to choose μ and τ are the same as in direct method. This algorithm uses M random numbers per iteration (where M is the number of reactions) and takes a time proportional to M to update the a_is, and takes a time proportional to M to identify the smallest τ_μ. However, if the number of possible reactions is large, this method is less efficient than the direct method.

The next reaction method [30] stores the next firing times of all reactions in an indexed binary tree priority queue (Figure 2.2), in which the firing time of each parent node is smaller than the firing times of its two daughter nodes. Thus, the time and index of the next occurring reaction are always available at the top node of the queue. The queue is updated when there are changes due to reactions and this simplifies the indexing scheme and the binary-tree structure of the queue. Another improvement of the next reaction method is the possibility to re-use random numbers, reducing the number of generation steps of random numbers to half of the direct method. Although the next-reaction method can be considerable faster than the direct method, it is challenging to code.

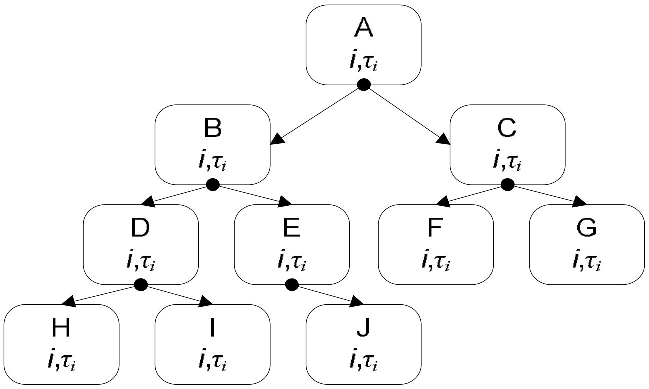

Figure 2.2. *Example of indexed priority queue. A tree structure of ordered pairs of the form (i, τ_i), where i is the number of a reaction and τ_i is the time when reaction i occurs, and i^{th} element is a pointer to the position in the tree that contains (i, τ_i). The tree structure has the property that each parent has a lower τ_i than its children.*

The advantages of next reaction method are: the method is exact (i.e. equivalent to the CME); it uses only a one random number per simulation event, and it takes a time proportional to the logarithm of the number of reactions, and not to the number of reactions [30]. Although the efficiency of updates in each step is proportional to log M, if some reactions are much faster than others, the next reaction method may be effectively proportional to log M', where $M' \ll M$. The next reaction method can also be extended to both Markov and non-Markov time-varying processes [30].

Another improvement proposed to the SSA is the logarithmic direct method [31]. The better efficiency of the logarithmic direct method (LDM) spawns from locating the next reaction via binary search, which has an average search depth proportional to the logarithm of the number of reactions and independent of the ordering of the reactions. LDM reduces the computation time and avoids the pre-simulation step of previous SSA formulations. The efficiency can be further improved by stating update stage through the use of sparse matrix techniques [31].

2.1.3 Delayed Stochastic Simulation Algorithm

The first stochastic models of gene expression assumed gene expression events to be instantaneous reactions, even though it sometimes takes a long time for a product to be released. [15] For example, transcription and translation consist of a number of chained reactions that happen in successive fashion. The time to execute this step is also dependent on the kinetics and number of chained reactions, meaning that the time of release of the products differs, according to some distribution. Measurements of transcription elongation times showed that the velocities of separate events followed a wide normal distribution [32]. A simple idea to account for this would be to transform the multi-step process into a single delayed process, removing the intermediate steps. This leads to a generalization of the commonly used elementary reactions, where products appear without delay, to a delayed reactions, where the initiating events are separated from the appearance of products by defined distributions of time intervals. However, this transforms the time-independent Markov process, in which the value a_i is calculated from the state, and τ_i is the sum of t and a random variable with exponential distribution and parameter a_i, into a non-Markov process where the distribution of τ_i depends on the history of states of the system [30].

In general, it is difficult to include non-Markov processes in stochastic simulations since the distribution of transition times to the next state includes the history of the system [30]. However, some non-Markov processes in chemical reaction simulations have useful properties that make the implementation easier. The history of the system consists of the series of discrete transitions and transition times. In this case, the propensity of the next reaction is usually the same as in the previous, and can thus just be repeated. This makes the implementation easier as it allows converting identical state transitions into a single step, provided that one stores the entire state history. Second, the entire history may not be needed. For any reaction μ, one only needs to store the part of the history that affects reaction μ, which in a chain of similar, consecutive reactions, implies significant reduction in stored steps [30]. The following reaction describes a general multi-step reaction:

$$\mathrm{A}+\mathrm{B}\xrightarrow{k_1}\mathrm{S}_0\xrightarrow{k}\mathrm{S}_1\xrightarrow{k}\mathrm{S}_2 \quad \cdots \quad \mathrm{S}_{n-1}\xrightarrow{k}\mathrm{S}_n\xrightarrow{k_2}\mathrm{C}+\mathrm{D} \tag{2.10}$$

Consider the following delayed process: a molecule type S_0 is produced and then undergoes an n-step process. The resulting molecule, S_n, affects the system dynamics. The first and last reactions differ, but all n intermediate equations are identical. Assuming time independence (first-order reactions are not affected by change in volume), one can simplify the n-step exponential process into a single step gamma process [30]. The number of processes to consider is equal to the number of distinct molecules produced within the process, which is much smaller than the number of states. The assumptions made here can be used to describe many common biological processes [15].

The implementation of non-Markov processes is described by Gibson [30]. Delayed processes are stored into a waitlist that keeps track of all reactions happening at each moment in the simulation. The procedure is as follows: rather than storing the state and time directly, the waitlist is a list of the time of occurrence of L processed events. At each step of the SSA, the time for occurrence of the next event in the waitlist is compared to the time for the next event in the SSA. If the first is smaller than the latter, the event stored in the waitlist is executed. Else, a SSA step takes place. Since SSA simulates a memoryless process, after an event in the waitlist is executed, a new random number is generated to determine the time for the next event. The algorithm proceeds as follows [19]:

1. Set $t = 0$, t_{stop} = stop time, set initial number of molecules and reactions, and create empty waitlist L.
2. Generate an SSA step for reacting events to get the next reacting event R_1 and the corresponding time t_1.
3. Compare t_1 with the least time in L, t_{min} . If $t_1 < t_{min}$ or L is empty, set: $t = t_1$. Update the number of molecules by performing R_1, adding delayed products (if existing) and the time delay that they have to stay in L from the appropriate distribution.
4. If L is not empty and if $t_1 \geq t_{min}$, set $t = t_{min}$. Update the number of molecules and waitlist L, by releasing the first element in L.
5. If $t < t_{stop}$, go to step 2; otherwise stop.

Recent studies on gene expression in prokaryotes [33] and eukaryotes [34] used non-delayed reaction models to match the experimental measurements. It is of relevance to note that these studies focused on steady-state dynamics of gene expression, where delayed and non-delayed dynamics would agree with each other after a transient. The delays cause some effects to be stronger in the dynamics of gene expression than one would assume from deterministic or non-delayed stochastic models. In particular, when modeling gene expression within more complex systems of reactions, such as given feedback loops, delayed reaction models are needed. The role of delays [35] and fluctuations [36, 37] in the dynamics gene expression has been the interest of a number of studies.

Several studies [38, 39] proposed delayed SSA algorithms that allow explicit delays in protein production. Another work [40] proposed the use of delayed SSA for reactions with a single delayed event. This algorithm considered two types of reactions with a delayed event. In non-consuming reactions, the products of an unfinished reaction cannot participate in new reactions, while in consuming ones, the reactants change immediately. The algorithm proposed [19] differs from the previous as it is able to handle multiple delayed events in a single reaction and that each of these multiple delays can follow a distribution (i.e. are random variables). The algorithm divides the delayed reactions into reacting events and generating events (appearance of products, possibly delayed). Non-delayed generating events are carried out at the same time as the corresponding reacting event. Delayed generating events are stored in a waitlist which is sorted by time of occurrence (also stored in the waitlist). The algorithm benefit is to have more than one generating event per reacting event, meaning several products appearing with different delays from one reaction event. As mentioned, generating a new reaction event while discarding the one previously generated due to the occurrence of the delayed event, does not introduce errors as the process is memoryless (obeys Poisson statistics) [15].

2.2 Biology and modeling of gene expression

Transcription and translation are the means of reading-out or expressing the genetic information by cells. The progression of genetic information in presented in Figure 2.3. The genetic information is stored in a sequence that contains four different bases (A, T, G, C). The first step in expressing the genetic information is to copy a particular portion of the DNA nucleotide sequence, the gene, into a RNA nucleotide sequence. The RNA differs from DNA in that nucleotides in RNA are ribonucleotides (rather than deoxyribonucleotides) and the RNA contains the base uracil (U) instead of the thymine (T) in the DNA. Another difference is that while the DNA always stored in cells as a double-stranded helix, RNA is single stranded. Therefore, RNA has the possibility of folding into complex three-dimensional shapes, with precise structural and catalytic functions [41].

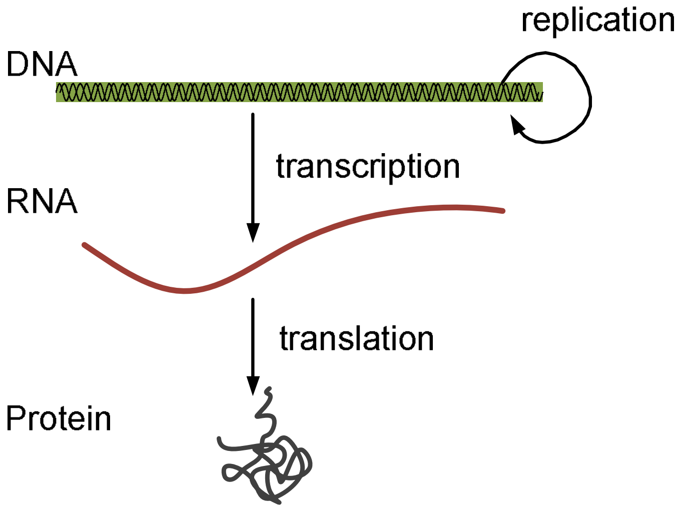

Figure 2.3. *The progression of genetic information from DNA to RNA and from RNA to protein occurs in all living cells. The former is called transcription and latter is translation. Replication is the process of copying the genetic information to next generations.*

Once mRNA contains the needed genetic information, its sequence is used to synthesize a protein. Instead of copying the information into a message, as it happens in transcription, the information is translated into a different form, an amino acid sequence. While the nucleotide sequence codes information in only four different nucleotides, proteins consist of sequences composed of twenty different amino acids. This implies that there is no one-to-one correspondence between these messages. Each group of three consecutive nucleotides in RNA, named codon, each codon specifying one amino acid (or stops the translation process) [41]. The transcription and translation are dynamically coupled in prokaryotes as shown in Figure 2.4.

2.2.1 Transcription in prokaryotes

RNA polymerases initiate RNA synthesis at sites in the DNA called promoters. These sites are defined by both genetic and biochemical criteria [42]. The DNA sequence of individual bacterial promoters determines the strength of promoter (defined as the average number of initiations events per unit of time). The strength of promoters varies over a wide range. Sequence homologies and the location of promoter mutations have shown that two separate regions within the bacterial promoter participate in the RNA polymerase initiation reaction. The two regions are located at, respectively, approximately 35 and 10 base pairs upstream the start point of RNA synthesis [42].

The control of transcription initiation involves several enzymes, which help recognizing the promoter, and various types of activator- and repressor molecules that control RNA chain initiation frequencies which have a dynamic range of about 10^4. Thus, some genes are transcribed every few seconds while the others only once in a generation [43]. The regulatory proteins recognize short specific sequences of double-helical DNA. Although each of these proteins has unique features, most bind DNA as homodimers or heterodimers and recognize structural motifs in DNA. The motifs near the promoter are called "operator" sites. The amino acid sequence of the repressor molecule defines the folding structure of the protein, which in turn determines the particular DNA sequence that the gene regulatory protein recognizes. Heterodimerization increases the range of DNA sequences that can be recognized. The repressor molecules in some cases block the access of RNA polymerase, but in others the repressor may just affect the confirmation of the bind RNA polymerase complex or looping of DNA to prevent binding, as in the ara and gal operons [42].

The main enzyme involved in transcription is the RNA polymerase (RNAp) which is a complex structure consisting of several subunits (β', β, σ and α) [42]. The structure and sequence of RNAp has been determined by various methods, such as crystallization [42]. The core enzyme ($\beta\beta'\alpha_2$) termed E contains all necessary enzymatic components required for the synthesis of RNA chains. However, the combination of E and σ subunits forms the so-called holoenzyme (Eσ), which can bind specifically to promoter sites and initiate the RNA chain elongation correctly. Every molecule of RNAp contains exactly one σ subunit (sigma factor). Sigma factor is a family of several factors, each used for specific purposes, e.g., σ^{70} is the "housekeeping" or primary sigma factor, while σ^{32} is the heat shock sigma factor, which is turned on when the cell is exposed to heat.

The investigation of transcription initiation has relied for many years on a rather simple model involving three overall steps: (1) binding, (2) isomerization and (3) promoter clearance [42]. There are various versions of this scheme vary from one promoter to the next, that may include additional steps and equilibrium reactions, but all are in accordance with the following scheme:

$$\mathrm{R}+\mathrm{P}\underset{K_B}{\rightleftharpoons}\mathrm{RP_c}\xrightarrow[k_f]{}\mathrm{RP_o}\rightarrow\rightarrow\cdots\rightarrow\mathrm{RNA} \tag{2.11}$$

This scheme was suggested by Zillig and coworkers [42, 44, 45] and involves the initial binding of RNA polymerase to the promoter with a equilibrium binding constant, K_B, to form an inactive complex called the closed complex (RP_c). The closed complex subsequently isomerizes with rate constant k_f to form the transcriptionally active open complex (RP_o) and dissociates the sigma factor. Before transcription initiation, the RNAp holoenzyme adheres only weakly to bacterial DNA when the two collide, and slides rapidly along the DNA molecule until it dissociates again [42]. Reaching the promoter region, the RNAp holoenzyme recognizes the promoter site by making specific contacts with the bases that are exposed on the outside of the helix. After the binding, the RNAp opens up the double helix to expose the nucleotides on each strand. The limited opening of the DNA helix does not require energy of ATP hydrolysis [41]. Instead, the polymerase undergoes reversible structural change that is more favorable than initial state. With the unwound DNA, one of the exposed strands acts as a template for complementary base-pairing. The initial RNA synthesis in this step is relatively inefficient and involves abortive initiation [46]. After the ten first nucleotides are assembled, the RNAp holoenzyme breaks its interactions with the promoter and releases the sigma factor [41].

The elongation continues until reaching the terminator site, where the polymerase halts and releases both the elongated RNA chain and the DNA template. In prokaryotes, the termination signal consists usually of a string of A-T nucleotide pairs preceding a two-fold symmetric DNA sequence that causes the transcribed RNA to fold into a "hairpin" structure and to be released from the RNAp, which causes the dissociation of the RNAp from the DNA [41, 47]. After the RNAp has been released, it can associate with a free sigma factor to form a new holoenzyme and begin the process of transcription again [41].

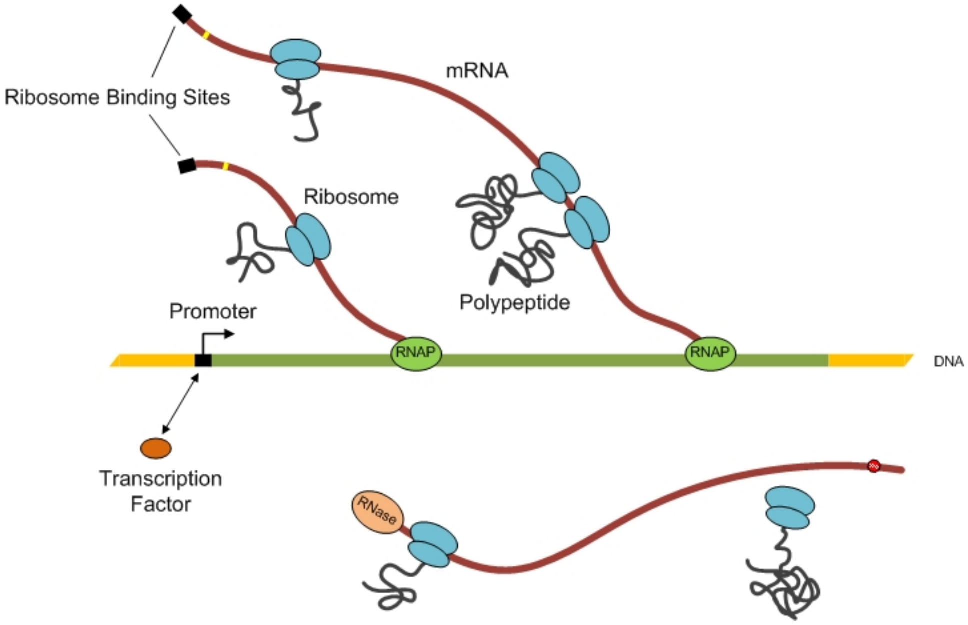

Figure 2.4. *Overview of coupled transcription and translation in prokaryotes. The two processes are coupled in space and time, causing the information propagation to not be preceded by intermediate steps. Notice the possibility of traffic between ribosomes and between RNAps.*

2.2.2 Translation in prokaryotes

Translation in prokaryotes can be divided into three main phases: initiation, elongation and termination. It begins with the binding of the ribosome complex to the mRNA strand. During elongation, the amino acids, determined by the RNA sequence, are added to the elongating peptide chain. Termination is the final step, as specific release factors detach the peptide and the RNA chain from the ribosome [41].

The translation of an mRNA begins with the codon AUG, and a special initiator tRNA carrying formylmethionine is required to start translation. Each bacterial mRNA contains a specific ribosome binding site (RBS; also called as Shine-Dalgarno sequence) that is located a few nucleotides upstream from AUG [41]. This nucleotide sequence forms base pairs with the 16S rRNA of the ribosome to direct the ribosome to correct initiation site. The binding of the ribosome to the ribosome binding site starts with the binding of the 30S ribosomal subunit to the nascent mRNA. After that, fMet-tRNA binds to the P-site forming a 30S complex. The 50S ribosome subunit attaches to it, forming the 70S initiation complex [41]. *E. coli* has specific translation factors for initiation (IF1, IF2 and IF3). As the ribosome binding is possible as soon as the RBS is revealed by the RNAp, bacterial mRNAs are often polycistronic i.e. they can encode different proteins from a single mRNA.

Translation elongation is efficient and accurate in prokaryotes because of the existence of specific translation factors, namely, EF-G and EF-Tu, which assist the ribosome during each cycle by coupling the GTP hydrolysis with the transitions between the ribosomal states [41]. Translation elongation occurs through successive translocation-and-pause cycles [3]. Translocation includes three steps, followed by a pause, during which the bond between amino acids is formed. EF-Tu assists the incoming aminoacyl-tRNA and checks whether the tRNA-amino acid match is correct. If the codon-anticodon match is correct, the ribosome triggers the hydrolysis of GTP, whereupon tRNA donates its amino acid to protein synthesis [41].

The genetic code contains two mechanisms for redundancy: some tRNAs can be charged with the same amino acid, and a single tRNA can recognize more than one codon due to a “wobble” effect in position three of the anti-codon [41]. The net effect is that multiple codons code for the same amino acid. These codons are called synonymous codons. Synonymous codons read by the same tRNA have been shown to translate at significantly different rates [17, 48], implying that translation rates are per-codon dependent, rather than per-tRNA or per-amino acid dependent. Only a few of these translation rates have been measured directly [17] but indirect assessment is available [20].

Translation elongation, while efficient and rapid, it contains equilibrium reactions and error-correction that can inhibit forward translocation. For example, back-translocation generally occurs when the tRNA has not yet locked into the peptide chain, causing the ribosome to move backwards on the mRNA template to the position of the previous codon. While the occurrence of back-translocation has been observed and can be promoted by certain antibiotics [49, 50], its exact causes remain somewhat unknown. Nevertheless, the kinetic rates for translocation and back-translocation have been measured under various conditions [49]. Alternatively, when the process becomes inefficient, there is the possibility of the ribosomes dissociating from the RNA prior to completion. The overall rate of dissociations has been measured in under various conditions [51]. It seems that the error in translation most affecting the fitness of bacteria under normal laboratory growth conditions is the drop-off event [51].

Trans-translation is the process by which the ribosome is released from the RNA template after stalling, which can occur for various reasons, such as incorporation of an incorrect codon, premature mRNA degradation, or spontaneous frameshifting [52]. Trans-translation is executed by the tmRNA that together with SmpB and EF-Tu, binds to the A-site of the ribosome and releases it from the mRNA [52]. Once the ribosome is released, the mRNA is usually degraded.

Translation elongation continues until reaching a stop codon. These are not recognized by a tRNA and do not specify an amino acid, but instead signal release factors to bind the ribosome. In *E. coli* there are three release factors: RF1 or RF2 binds and releases the ribosome together with RF3 [41]. These factors force the peptidyl transferase in the ribosome to add a water molecule instead of an amino acid to the peptidyl-tRNA. This frees the carboxyl end of the polypeptide from the tRNA and, thus, releases the whole peptide chain into the cytoplasm [41].

2.2.3 Modeling gene expression in prokaryotes

Modeling gene expression is a means to examine the dynamics of transcription and translation to estimate and profile the dynamical role of steps and parameters. The problem in modeling is the decision of what features to include in model and what is not relevant and can be excluded. Many models claim to account for the critical steps in gene expression, yet we have no certainty what those steps are, i.e., what sets of chemical transitions should be integrated into reaction channels and what concentrations should be absorbed into kinetic rate constants. Most relevantly, the relevance of the various steps may differ from gene to gene, between different environmental conditions, and even with the dynamic state of the gene network. Experimental results show that many genes can produce very different fluctuations in RNA and protein numbers depending on the environment and conditions of experiment [36, 53, 54]. This poses great difficulty in designing a general model of gene expression.

Most models [21, 55-57] are based on the premise that gene expression dynamics follows a cascade: gene activation determines the mRNA concentration, which in turn determines the protein concentration. This premise can be extended into other processes as well: any process that indirectly affects the concentrations of proteins and RNA or the dynamics of gene expression at any level ought to be included into realistic models. Most models focus on gene activation, RNA and protein numbers, discarding other possible processes as affecting effective rate constants.

The gene activation can be described as a random process, in which genes spontaneously switch between on and off states at a certain rate. However, depending on the growth rate of the cell, bacteria can have several copies of genes in partially replicated chromosomes [58], or in multi-copy plasmids. Transcription and translation are considered to be Poisson processes where the production probabilities per unit time are proportional to the number of activated genes and mRNAs, respectively [59]. This assumption does not account possible interfering events e.g. RNAp usage, elongation pauses [16, 47], premature terminations [60, 61] and amino acid starvations. Finally, mRNAs and proteins are often described as having exponentially distributed lifetimes, assuming that each degradation event is independent and memoryless [59]. This may not hold if the degradation rate depends on competition between ribosomes and RNases, for example [62].

The first stochastic models of genetic circuits assumed gene expression as an instantaneous process. The models' dynamics were simulated with the SSA. However, transcription and translation take time and these durations depend on the gene length, thus, vary from gene to gene. Recent models introduced delays in the reactions such as transcription and translation, so as to allow RNAs and proteins to appear with delays, following expression [19, 23, 38]. While non-delayed models can match gene expression fluctuations [34] under certain conditions, gene regulation network models with complex dynamics, e.g. feedback loops, require delayed reactions to reproduce the dynamics [38, 63]. Several reactions in gene expression, such as transcription, translation, post-translational modifications, and folding, are time consuming [64].

Here we present the first multi-delayed stochastic model whose dynamics was compared to, and found to match, measurements of gene expression at the single event level [13]. The model comprises transcription, translation, repression of transcription and RNA and protein degradation [21].

$$\mathrm{Pro} + \mathrm{RNAp} \xrightarrow{\mathrm{k_t}} \mathrm{Pro}(\tau_1) + \mathrm{RBS}(\tau_1) + \mathrm{RNAp}(\tau_2) + \mathrm{R}(\tau_2), \qquad (2.12)$$

$$\mathrm{Rib} + \mathrm{RBS} \xrightarrow{\mathrm{k_{tr}}} \mathrm{RBS}(\tau_3) + \mathrm{Rib}(\tau_4) + \mathrm{P}(\tau_5), \qquad (2.13)$$

$$\mathrm{RBS} \xrightarrow{d_{\mathrm{RBS}}} \varnothing, \qquad (2.14)$$

$$\mathrm{Pro} + \mathrm{Rep} \xrightarrow{k_{\mathrm{rep}}} \mathrm{Pro{\bullet}Rep}, \qquad (2.15)$$

$$\mathrm{ProRep} \xrightarrow{k_{\mathrm{unrep}}} \mathrm{Pro} + \mathrm{Rep}. \qquad (2.16)$$

This model (reactions (2.12) to (2.16)) is based on a previous one proposed in [57]. Reactions (2.12) and (2.13) describe prokaryotic transcription and translation, respectively. Pro represents the promoter region of the gene, RNAp is an RNA polymerase, Rib is a ribosome, and R is a transcribed RNA molecule. The RBS (ribosome binding site) is the initial sequence of the RNA to which the ribosomes bind to and initiate translation. In prokaryotes, translation can occur as soon as the RBS emerges from the RNA exit channel of RNAp (delay of τ_1 seconds). This delay determines the time to products appearing in the cell. Note that τ can be a random variable following a distribution, thus vary from one reaction event to the next. Reaction (2.14) is the degradation of mRNA (more specifically, of the RBS, which prevents new translation events of that mRNA). Reaction (2.15) models transcription repression by a repressor molecule (Rep) binding into operator site next to the promoter, while (2.16) models the unbinding of the repressor from the operator. Only when the operator is free from repression, can transcription initiate. With low induction the model produces Poisson distribution of RNA numbers and geometric distribution of protein numbers from a single RNA, as shown in Figure 2.5.

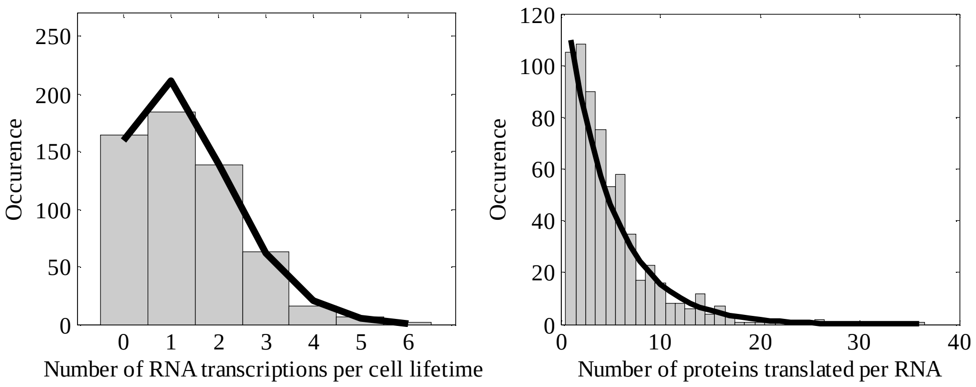

Figure 2.5. *(Left) Histogram (gray bars) of the number of mRNAs per cell cycle. The data fit well to a Poisson distribution (solid line) with an average of 1.2 RNAs per cell cycle. (Right) Distribution of the number of protein molecules from a single RNA, which follows a geometric distribution (solid line).*

2.3 Single nucleotide models

Although translation-transcription coupling has been known for decades, direct impacts of the coupling has only been described in the phenomena of transcription attenuation and polarity [60]. The regulatory mechanisms involved in these processes take place during transcription and translation elongation processes rather than, e.g. transcription initiation, thus representing a case of regulation in an independent layer of control.

In most modeling strategies proposed so far, elongation has been modeled as a simple delayed or non-delayed single step event, thus not allowing the modeling possible regulatory mechanism of RNA and protein numbers dynamics in this stage. However, elongation dynamics is far more complex than these models assumed [2, 16, 47, 60]. Recent studies have thus proposed the explicit modeling of transcription elongation as a chain of consecutive reactions, one for each nucleotide [5, 62].

Transcription elongation includes many competitive reaction channels to regulate the RNAp movement on the template [2, 16]. Transcriptional regulation during elongation can be accounted for if there are explicit single-nucleotide addition reactions in the model. The regulative reaction pathways can be made available to the RNAp at each template position, and the kinetic competition between the alternative reaction channels and normal elongation, not only determine greatly, for example, the stochasticity of the process, but by changing reaction rates, it results in completely different transcriptional outcomes at the operon level (e.g., premature terminations and productive transcriptions). Forward transcription elongation, which includes nucleotide activation and addition, is the dominating reaction channel, but, especially in specific sequences, or during transient concentrations fluctuations, it may enhance the activation of different pathways, e.g. sequence-dependent pausing [16]. This concept of kinetic competition at the nucleotide level makes the dynamics of transcription elongation much more diverse, leading to complex patterns of noise in gene expression.

One of the first stochastic models of transcription at the nucleotide level was proposed in [62]. The phage-λ lysis-lysogeny decision circuit was the model system, and the novel modeling strategy of its dynamics was able to explain the experimental results, which other models (deterministic in nature) could not. The movement of the RNAP along the DNA was modeled as a sequence of independent one-nucleotide reaction steps. It was assumed that each step forward has constant probability of occurring per unit time. This analysis shows how stochastic molecular level fluctuations can be exploited by the regulatory circuit to produce different phenotypes from the same genotype in monoclonal cell populations.

Another model with multi-stepped transcription elongation [22] introduced ubiquitous pauses due to backtracking of the RNA polymerase. Such pauses led to a non-Poisson distributed, broad and heavy-tailed distribution of transcription elongation times. It also enhances bursts in mRNA production, when the time intervals between RNAps are shorter than in Poisson statistics. Results suggest that transcriptional pausing may lead to a range of variability in transcription rates between consecutive events, with a non-negligible effect on noise in mRNA levels as well as in cell-to-cell variability in RNA numbers.

Recently, a model including the elongation at the nucleotide level was proposed [5] that, in addition to alternative regulation pathways presented in previous models; it additionally includes e.g. RNA polymerase arrests [2] and promoter complex formation [65]. The study of the dynamics of this model showed that the occurrence of pausing and other chemical pathways in step-wise elongation can increase collisions between preceding RNAp molecules and amplify bursting. The proposed delayed stochastic model of transcription at the nucleotide level incorporates the promoter occupancy time, pausing, arrests, misincorporation and editing, pyrophosphorolysis, premature termination [5], and accounts for the footprint of an RNAP when bound to the DNA template [2, 66].

The dynamics of the single nucleotide model did not fully match the dynamics of single-step delayed models, indicating that the noise in elongation affects the time interval distributions [5]. However, the difference between the delayed and detailed models was not only due to traffic between polymerases. Competitive events in elongation such as pauses and arrests have a role in shaping the time-interval distribution even when not sufficiently frequent or long to cause collisions between RNA polymerases. Beforehand, it was assumed that single-step multi-delayed models of transcription were accurate as long as the level of expression is low, i.e. no collisions between the elongating RNA polymerases [38].

The measured distribution of intervals between completion events, at the single protein level [13] or at the single RNA level [14] were matched with both the delayed and detailed models. In fully induced genes, the time for the promoter complex formation [65] was found to be the rate limiting step determining the rate of collisions between elongating RNAps. Pauses and arrests allow the RNA completions to be separated by a time interval smaller than the duration of the open complex formation, producing "bursty" transcription dynamics. Simulations of the model proposed in [5] showed that the delays in elongation processes are not well fit by uniform distributions. For example, a pause-prone sequence is likely to cause a broader distribution of elongation delays than otherwise.

The modeling of translation in nucleotide (or codon) level has also been proposed. Arkin and co-workers proposed a translation model consisting of *n* steps and competitive degradation control [62]. Statistics of intersteps times were described by the exponential probability function and ribosome queuing was involved as is observed in experiments [67, 68]. However, several experiments indicate that not all codons are translated at the same speed [17, 48]. Ribosome traffic was found to be dependent on sequence rather than gene length or other parameters. This modeling strategy was used in a recent study of translation efficiency and traffic [20]. They reproduced the observations from *in vivo* experiments for incorporation of radioactivity in different strains with slow-to-translate codons [17] into a protein, which could not be modeled with neither as deterministic nor by a uniform stochastic modeling strategy [20].

However, the translation speed has been measured only for a few codons under specific circumstances but indirect assessment is available for the overall translational efficiency [20]. Several studies propose that translation efficiency determines the transcription process outcome [68] and the speed of transcription elongation [69]. There are results suggesting the existence of direct translation-transcription coupling between the RNAp and ribosome that explain why mRNAs being translated cannot by terminated by Rho termination factor [70]. Uncoupling of transcription and translation at the end of operons enables transcription termination.

2.4 Fluorescence microscopy

Genes' expression levels have been measured using a variety of techniques such as Northern blotting, quantitative PCR and sequencing. While these techniques allow high throughput measurements of gene expression at the whole genome scale, they have drawbacks in studying gene expression dynamics at the molecular level. The data is noisy, the measurements are necessarily from populations of cells and *in vivo* measurements are not possible. In measurements of dynamics of gene expression, fluorescence microscopy has been for a long time the state-of-the-art method of measuring gene expression in individual cells. The limitations of light microscopy are well-known and only a few methods have been able to improve it. The diffraction barrier is still present and, while there are methods to circumvent this restriction, the physical limitation that the resolution is half of the wavelength of the light in measurement cannot be overcome.

The image quality can be assessed in two terms: contrast and resolution. Resolution is the physical concept that can be described, measured and manipulated according to rules derived from optics. On the other hand, contrast, the difference in visual properties that makes an object (or its representation in an image) distinguishable from other objects and the background, is not quantifiable. Contrast is limited by the noise in the measurements and is, usually, independent of the resolution available.

A confocal microscope has slightly higher resolution than a wide field microscope, but this is not its biggest advantage. The actual benefit of using confocal microscopy comes from its higher contrast, especially given thick specimens. It is based on restricting the volume under observation i.e. keeping overlying or nearby light sources from contributing to the detected signal. The disadvantage for this is slow as the instrument must observe only one point at a time (e.g. scanning laser confocal) or a group of separated points (e.g. the spinning-disc confocal). For now, we study the applicability of microscopy to measurements of gene expression in individual cells.

The most common way to study specific genes under microscope is to use green fluorescent protein (GFP) obtained from the jellyfish *Aequorea Victoria* [71]. GFP emits green light when excited with light of proper wavelength and only requires the presence of oxygen to maturate i.e. no external compounds are needed for organisms to express observable GFP [71]. The GFP gene may be inserted to and expressed in a wide range of organisms, e.g., mammals [72], fishes [73], yeasts [74], and a broad variety of bacteria [13, 14]. GFP normally doesn't affect the growth of the host and does not interact with other proteins. The optimal reporter for studying real-time gene expression in individual cells should possess an instability allowing monitoring of rates of expression. A major drawback of GFP is that, once formed, it is very stable [75], which in turn renders the protein less valuable for studies of transient gene expression. Andersen and co-workers constructed new variant GFP genes with reduced half-life compared to wild-type [75].

Recent advancements in fluorescence microscopy methods include measuring the real-time production of single protein molecules in individual *E. coli* cells [13]. A fusion protein of a fast-maturing yellow fluorescent protein (YFP) and a membrane-targeting peptide was used to monitor appearance of individual fluorescent molecules inside the cell under a repressed condition. This method included photobleaching the appeared molecules after taking image to retain the optimal conditions for detection i.e. to prevent the clustering of spots. The proteins were produced in bursts from a stochastically transcribed mRNA molecule with intervals large enough to track the proteins into individual mRNAs. They observed the distributions of protein production and found that protein copy numbers in the bursts follow a geometric distribution.

Real-time dynamics of RNA production is much harder to measure in the cell than fluorescent proteins and the most used constructs consist of tagging RNA with fluorescent probes. Singer and co-workers visualized localization of the mRNA in living yeast cells using green fluorescent protein (GFP) fused to the RNA-binding protein MS2 to bind a reporter mRNA containing MS2-binding sites [76, 77]. Golding modified this system to measure RNA production in bacteria [14]. The mechanism is illustrated in Figure 2.6. They found that transcription occurs in bursts, that the burst sizes are geometrically distributed and that the intervals between bursts are exponentially distributed.

Another method used is based on fluorescent-protein complementation regulated by the interaction of a split RNA-binding protein with its corresponding RNA binding protein [78]. Valencia-Burton and colleagues dissected the RNA binding protein in two fragments, and each fragment is fused to split fragments of green fluorescent protein (GFP). Binding of the fragments into RNA resulted in the expression of the assembled GFP in bacteria.

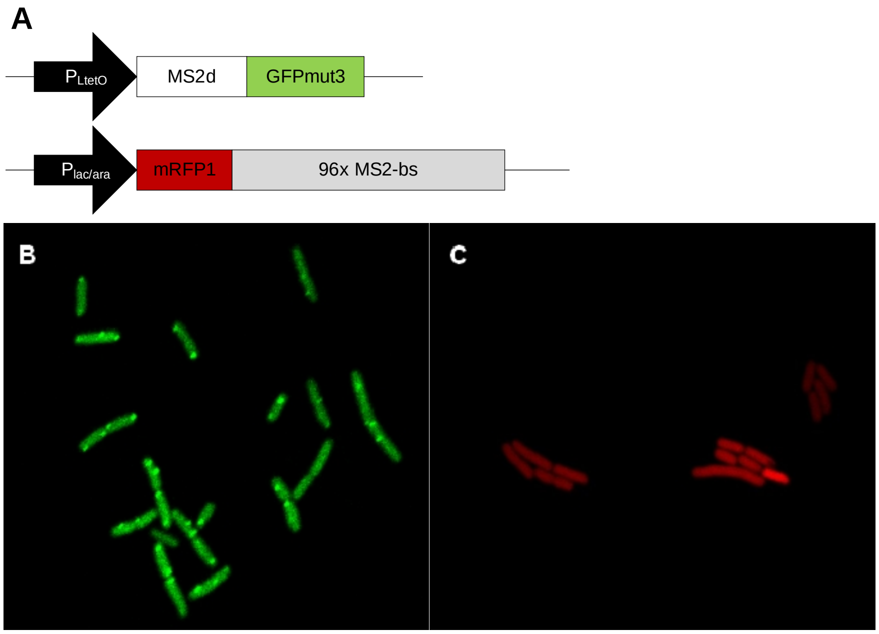

Figure 2.6. *(A) Genetic components of the detection system used in measurements of E. coli [14]. The tagging protein consists of a MS2 coat protein fused to GFP. Protein production is regulated by the P_{LtetO} promoter [43]. The RNA target consists of the coding region for mRFP1, a monomeric red fluorescence protein [79], followed by a tandem repeats of 96 MS2 binding sites. RNA target production is controlled by $P_{lac/ara}$ promoter [80]). This construct is on an F plasmid, with a single copy per bacterial chromosome. (B) Detection of mRNA in living cells from a microscope image of cells expressing the RNA target and tagging protein. The bright spots inside the cells are mRNAs tagged with approximately 96 fused GFP proteins. The green background represents freely diffusing fusion proteins. (C) Detection of protein in living cells. This microscope image shows the expression of red fluorescent protein (RFP) inside the bacteria.*

Real-time measurements of gene expression are applicable only to certain conditions. Constant imaging bleaches the fluorescent proteins and causes photo-toxicity in cells. Instead of using real-time measurements, one can observe the state of the population in a single time moment. It is possible to infer some properties of the dynamics from population level measurements. Xie and co-workers studied naturally occurring mRNA and protein numbers in individual cells by fixing the cells and using fluorescence in situ hybridization (FISH) with single-molecule sensitivity [81]. They confirmed the validity of transcript measurements with RNA-seq and measured of proteins levels with an YFP fusion library with single-molecule accuracy, providing quantitative analyses of both abundance and noise in the proteome and transcriptome for *E. coli*. They found no momentary correlation between mRNA and protein levels that causes the disconnection between proteome and transcriptome analyses of a single cell. However, this may be explained by the difference in mRNA and protein lifetimes.

3 MATERIALS AND METHODS

3.1 Model of transcription at a nucleotide level

We model the dynamics of gene expression as in [23]. This model was shown [21] to match the dynamics of RNA and protein production at the single molecule level [13]. The dynamics of the system of chemical reactions is driven by the delayed stochastic simulation algorithm (delayed SSA [19]) so as to include events whose time of completion once initiated is non negligible, in that it affects the dynamics of production of RNA and protein molecules. Specifically, several steps in gene expression, such as the promoter open complex formation, are time consuming [64]. In order to include these events when simulating gene expression, the delayed SSA was proposed [19].

All simulations are executed by an extended version of SGNSim [24] to allow multiple coupled chain elongation processes to run in parallel on each elongating RNA strand. The extension consists in providing the simulator with the ability to introduce new chemical reactions at run time (i.e., those corresponding to the translation of each individual RNA strand).

The delayed stochastic model of transcription at the nucleotide level [5] includes the promoter occupancy time, pausing, arrests, editing, premature terminations, pyrophosphorolysis, and accounts for the RNAp footprint in the DNA template [2]. Additional reactions model the stepwise forward movement and activation of the RNAp, pausing and unpausing of the RNAp due to collisions with adjacent RNAps, release of the promoter when the RNAp begins elongation, and error correction. A state diagram of one nucleotide in transcription elongation is shown in Figure 3.1.

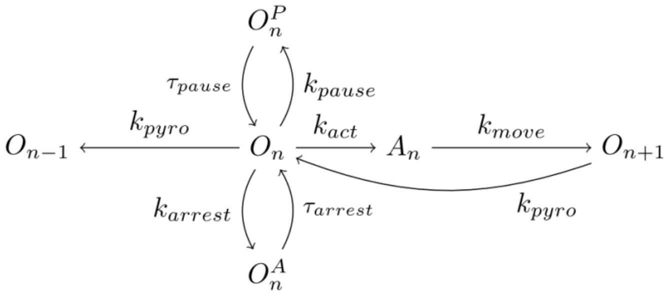

Figure 3.1. *State diagram of one nucleotide in transcription elongation.*

The reactions, stochastic rate constants and time delays, are shown in Table 3.1, and described in detail in [5]. Here, Pro stands for the promoter region, RNAp for the RNA polymerase, and RNAp•Pro for the promoter region occupied by an RNAp. A_n, O_n and

U_n stand for the *n*th nucleotide when activated, occupied, and unoccupied, respectively. Ranges of nucleotides are denoted such as $U_{[start,end]}$, denoting a stretch of unoccupied nucleotides from indexes *start* to *end*. O_{n_p}, $O_{n_{ar}}$ and $O_{n_{correcting}}$ are used to represent a paused, arrested, or error correcting RNAp at position *n*. On the template, each RNAp occupies ($2\Delta_{RNAp}+1$) nucleotides, where $\Delta_{RNAp} = 12$. These nucleotides cannot be occupied by any other RNAp at the same time. U_n^R denotes transcribed ribonucleotides which are free (i.e., not under the RNAp's footprint). These transcribed ribonucleotides are created in a separate part of the simulation (denoted by the R superscript pointing to RNA compartment), one separate set per RNA strand, so that we can simulate the translation of all individual RNA molecules independently and simultaneously.

We use a delayed reaction event to model the first step in transcription; the promoter closed and open complex formation (3.1). These processes could instead be modeled by a set of non-delayed, consecutive reactions [65]. We use a delayed reaction as it was shown to accurately model the dynamics of this process [19, 21, 23]. The duration of this step likely varies from one event to the next, but while values for the mean duration are known, as of yet, there are no exact measurements of the standard deviation. Nevertheless, it is likely small compared to the mean, given the very small standard deviations of promoter activity [80]. For these reasons, we set the promoter delay, $\tau_{oc,}$ as a random variable, following a normal distribution with a mean of 40 s and a standard deviation of 4 s, whose value is randomly drawn each time a transcription event occurs.

Once the first nucleotide is occupied via reaction (3.2), stepwise elongation can begin (3.3). Also, as soon as the promoter is released, a new transcription initiation event can occur. Following each elongation step (3.3), an activation step occurs (3.4), which is necessary for the RNAp to move along the template to the next nucleotide. The following events compete with stepwise elongation: pausing (3.5) and (3.7), released via (3.5) or (3.6), arrests and their release (3.8), editing (3.9), premature terminations (3.10), and pyrophosphorolysis (3.11).

At the end of the elongation process, the RNAp is released (3.12). mRNA degradation is modeled, for simplicity, as a first order reaction (3.13). When (3.13) occurs, the first few ribonucleotides of the RNA are immediately removed from the system, preventing any new translation event [82]. Thus, we model the degradation process such that it begins in the vicinity of the RBS and then gradually cuts the mRNA as it is being released from the ribosomes. This allows the translating ribosomes to complete protein production before the whole mRNA is degraded. When the final ribosome unbinds from the RNA, the rest of the RNA strand, denoted by R in reaction (3.13), is destroyed.

Table 3.1. *Reactions in modeling transcription. Chemical reactions, rate constants (in s^{-1}), and time delays (in s) used to model transcription initiation, elongation, and termination. Parameter values were obtained from measurements in E. coli, mainly for LacZ. References are reported in the column Ref.*

Event	**Reaction**	**Rate constant**	**Ref.**
Initiation and promoter complex formation (3.1)	$\mathrm{Pro} + \mathrm{RNAp} \xrightarrow{k_{init}} \mathrm{RNAp{\bullet}Pro}(\tau_{oc})$	$k_{init} = 0.015$ $\tau_{oc} = 40 \pm 4$	[21]
Promoter clearance (3.2)	$\mathrm{RNAp{\bullet}Pro} + \mathrm{U}_{[1,\Delta_{\mathrm{RNAp}}+1]} \xrightarrow{k_m} \mathrm{O}_1 + \mathrm{Pro}$	$k_m = 114$	[69]
Elongation (3.3)	$\mathrm{A}_n + \mathrm{U}_{n+\Delta_{\mathrm{RNAp}}+1} + \mathrm{N}^{\mathrm{R}}_{n-\Delta_{\mathrm{RNAp}}} \xrightarrow{k_m}$ $\mathrm{O}_{n+1} + \mathrm{U}_{n-\Delta_{\mathrm{RNAp}}} + \mathrm{U}^{\mathrm{R}}_{n-\Delta_{\mathrm{RNAp}}}$	$k_m = 114$	[69]
Activation (3.4)	$\mathrm{O}_n \xrightarrow{k_a} \mathrm{A}_n + \mathrm{N}^{\mathrm{R}}_n$	$k_a = 114$, n > 10, $k_a = 30$, n ≤ 10	[69]
Pausing (3.5)	$\mathrm{O}_n \underset{1/\tau_p}{\overset{k_p}{\rightleftarrows}} \mathrm{O}_{n_{\mathrm{p}}}$	$k_p = 0.55$ $\tau_p = 3$	[2]
Pause release due to collision (3.6)	$\mathrm{O}_{n_{\mathrm{p}}} + \mathrm{A}_{n\text{-}2\Delta_{\mathrm{RNAp}}-1} \xrightarrow{0.8k_m} \mathrm{O}_n + \mathrm{A}_{n\text{-}2\Delta_{\mathrm{RNAp}}-1}$	$k_m = 114$	[83]
Pause induced by collision (3.7)	$\mathrm{O}_{n_{\mathrm{p}}} + \mathrm{A}_{n\text{-}2\Delta_{\mathrm{RNAp}}-1} \xrightarrow{0.2k_m} \mathrm{O}_{n_{\mathrm{p}}} + \mathrm{O}_{n\text{-}2\Delta_{\mathrm{RNAp}}-1_{\mathrm{p}}}$	$k_m = 114$	[83]
Arrests (3.8)	$\mathrm{O}_n \underset{1/\tau_{ar}}{\overset{k_{ar}}{\rightleftarrows}} \mathrm{O}_{n_{\mathrm{ar}}}$	$k_{ar} = 0.00028$ $\tau_{ar} = 100$	[5]
Editing (3.9)	$\mathrm{O}_n \underset{1/\tau_c}{\overset{k_{ec}}{\rightleftarrows}} \mathrm{O}_{n_{\mathrm{correcting}}}$	$k_{ec} = 0.008$ $\tau_c = 5$	[2]
Premature termination (3.10)	$\mathrm{O}_n \xrightarrow{k_{pre}} \mathrm{RNAp} + \mathrm{U}_{[n\text{-}\Delta_{\mathrm{RNAp}},\, n+\Delta_{\mathrm{RNAp}}]}$	$k_{pre} = 0.00019$	[84]
Pyrophosphorolysis (3.11)	$\mathrm{O}_n + \mathrm{U}_{n-\Delta_{\mathrm{RNAp}}-1} + \mathrm{N}^{\mathrm{R}}_{n-1} + \mathrm{U}^{\mathrm{R}}_{n-\Delta_{\mathrm{RNAp}}-1} \xrightarrow{k_{pyro}}$ $\mathrm{O}_{n-1} + \mathrm{U}_{n+\Delta_{\mathrm{RNAp}}-1} + \mathrm{N}^{\mathrm{R}}_{n-\Delta_{\mathrm{RNAp}}-1}$	$k_{pyro} = 0.75$	[85]
Completion (12)	$\mathrm{A}_{n_{last}} \xrightarrow{k_f} \mathrm{RNAp} + \mathrm{U}_{[n_{last},\, n_{last}\text{-}\Delta_{\mathrm{RNAp}}]}$	$k_f = 2$	[86]
mRNA degradation (3.13)	$\mathrm{R} \xrightarrow{k_{dr}} \varnothing$	$k_{dr} = 0.011$	[13]

If the model of RNA degradation was such that some of the ribosomes on the RNA template fell off when degradation begins (i.e. due to endonucleatic cleavage of the RNA chain at a random position [82]), one consequence would be the reduction of the mean protein burst size as these RNAs would contribute far fewer proteins than if the ribosomes were allowed to finish translating. This would likely result in a reduction of protein noise levels. Alternatively, the ribosome occupancy of the ribosome binding site might determine mRNA longevity [68]. In this case, for the same mean burst size, the noise is expected to increase since large bursts will get larger and small bursts will get smaller, likely increasing protein noise levels. We opted not to include these additions to the degradation model since they are not yet well characterized [82].

Finally, we note that in present model we do not add an explicit reaction for abortive initiation of transcription [46]. This could be done by adding a reaction (3.14) which would compete with reaction (3.2). Its rate, k_{ab}, would be set so as to match the fraction of abortive initiations after the formation of the promoter open complex [46]:

$$\text{RNAp}\bullet\text{Pro} \xrightarrow{k_{ab}} \text{Pro} + \text{RNAp} \qquad (3.14)$$

For simplicity, we opted not to include this reaction in the simulations, and instead set a value for the rate of transcription initiation that matches realistic rates of RNA production. From the point of view of RNA production, since (3.14) competes with reaction (3.2), it would be dynamically equivalent to decrease the rate of transcription initiation in (3.2) to account for the fraction of abortive initiations. The model of transcription and the reaction rates in Table 3.1 are described in greater detail in [5]. Parameter values were obtained from measurements in *E. coli*.

3.2 Model of translation at a codon level

The stochastic model of translation at the codon level includes initiation (3.15) and stepwise translocation (codon incorporation) (3.16-3.18) followed by activation (3.19). Reactions competing with translocation are back-translocation (3.20), drop-off (3.21), and trans-translation (3.22). The process ends with elongation completion (3.23), followed by protein folding and activation (3.24). Protein degradation (3.25) is included to allow us to study fluctuations in protein levels at steady state. All reactions and rate constants are presented in Table 3.2.

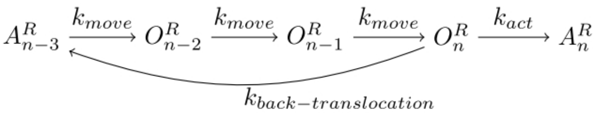

Figure 3.2. *State diagram of one codon in translation elongation*

Here, Rib denotes a free ribosome complex in the cellular medium, while Rib^R denotes a ribosome bound to a specific RNA strand. Similar to Δ_{RNAp}, Δ_{Rib} denotes the ribosome's footprint in the RNA template. Each ribosome occupies ($2\Delta_{\text{Rib}}+1$) ribonucleotides, where $\Delta_{\text{Rib}} = 15$ [20]. U_n^R, O_n^R and A_n^R are the ribonucleic equivalents of U_n, O_n and A_n. U_n^R denotes an unoccupied ribonucleotide, while O_n^R denotes that a translating ribosome is currently positioned at ribonucleotide *n*. Similarly, $\mathrm{A}_n^{\mathrm{R}}$ denotes that a ribosome has created peptide bond for the peptide coded by the codon at position [*n*-2,*n*], where *n* is a multiple of 3 (*n* = 3, 6, 9, …). Since different codons are translated at different rates, the activation reaction has a codon-specific rate [17]. Specific rates were set for four codons, while the remaining ones fall into three different classes [20], A, B and C, whose rates are denoted $k_{trans\{A,B,C\}}$.

Translation has three main phases: initiation, elongation and termination. It begins with the binding of the ribosome complex to the mRNA strand. During elongation, the amino acids, determined by the RNA sequence, are added to the elongating peptide chain. Termination is the final step, as specific release factors detach the peptide and the RNA chain from the ribosome. The specific translation factors of *E. coli* for each phase are not explicitly modeled, as they exist in abundance under normal conditions. The binding of the ribosome to the RBS of the RNA is modeled as a single step reaction (3.15). The next ribosome can only to bind after the preceding one has moved away from the RBS. This implies that the initiation of two consecutive translation events is separated by a non-negligible time interval.

Translation elongation occurs through successive translocation-and-pause cycles [3]. Translocation includes three steps (3.16-3.18), after which there is a pause (3.19), during which the bond between amino acids is formed. The time that (3.19) takes to occur accounts for this pause, which is much longer than the time for (3.16-3.18) to occur [3].

Synonymous codons read by the same tRNA have been shown to translate at significantly different rates [17], implying that our model must incorporate per-codon translation rates for reaction (3.19), rather than per-tRNA or per-amino acid rates. Only a few of these translation rates have been measured directly [17] but indirect assessment is available [20]. In our case, we assume normal cellular conditions, including an abundance of charged tRNA, implying that we do not need to model the tRNA explicitly. Since each codon is translated at a different rate, the codon frequency also needs to be accounted for explicitly [48]. In the model, the sequence can either be randomly generated or selected from a known gene. In the former case, the sequence is randomly generated according to the known statistical frequency of each codon in *E. coli*.

The competing reactions of stepwise translation elongation are back-translocation (3.20), drop-off (3.21) [51] and trans-translation (3.22), which are explicitly modeled. Back-translocation generally occurs when the tRNA has not yet locked into the peptide chain, causing the ribosome to move backwards on the mRNA template to the position of the previous codon. While the occurrence of back-translocation has been observed and can be promoted by certain antibiotics [49, 50], its exact causes remain somewhat unknown. Nevertheless, the kinetic rates for translocation and back-translocation have been measured under various conditions [49]. Alternatively, the ribosomes can randomly dissociate from the RNA, in a process called drop-off, modeled by reaction (3.21). The overall rate of drop-off has been measured in [51], from which we have inferred a per-codon rate.

In the model, stalling followed by trans-translation, which can occur for a variety of reasons, such as the incorporation of an incorrect codon, premature mRNA degradation, or spontaneous frameshifting [52], can occur spontaneously with a given probability at any codon via reaction (3.22). When this reaction occurs, the RNA strand is immediately destroyed in the simulation, and all translating ribosomes are released back into the cellular medium, denoted in reaction (3.22) by $[\mathrm{Rib}^{\mathrm{R}}]\mathrm{Rib}$, where $[\mathrm{Rib}^{\mathrm{R}}]$ denotes the number of ribosomes bound to the RNA at that moment.

Translation elongation continues until the stop codon is reached (3.23). The exact steps in termination are not modeled explicitly in the model. Its kinetic rate is higher than initiation, preventing queuing near the stop codon [20]. Reaction (3.23) is followed by folding and activation (3.24), modeled as a first order process for simplicity [21]. The rate of this reaction is set to model the maturation time of GFP, as most measurements of protein expression at the single cell level use this protein. $\mathrm{P}_{\mathrm{prem}}$ denotes the unfolded protein, while P denotes the complete activated protein, which can then degrade via reaction (3.25).

Table 3.2. *Reactions in modeling translation. Chemical reactions and rate constants (in s^{-1}) used to model translation initiation, elongation, and termination, as well as protein folding and activation, and protein degradation. Parameter values were obtained from measurements in E. coli, mainly for LacZ. References are reported in the Ref.*

Event	**Reaction**	**Rate constant**	**Ref.**
Initiation (3.15)	$\mathrm{Rib}+\mathrm{U}^{\mathrm{R}}_{[1,\Delta_{\mathrm{Rib}}+1]} \xrightarrow{k_{trans\_init}} \mathrm{O}^{\mathrm{R}}_{1}+\mathrm{Rib}^{\mathrm{R}}$	k_{trans_init} = 0.33	[20]
Stepwise translocation (3.16-3.18)	$\mathrm{A}^{\mathrm{R}}_{n-3}+\mathrm{U}^{\mathrm{R}}_{[n+\Delta_{\mathrm{Rib}}-3,n+\Delta_{\mathrm{Rib}}-1]} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n-2}$ $\mathrm{O}^{\mathrm{R}}_{n-2} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n-1}$ $\mathrm{O}^{\mathrm{R}}_{n-1} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n}+\mathrm{U}^{\mathrm{R}}_{[n-\Delta_{\mathrm{Rib}}-2,n-\Delta_{\mathrm{Rib}}]}$	k_{tm} = 1000	[3]
Activation (3.19)	$\mathrm{O}^{\mathrm{R}}_{n} \xrightarrow{k_{trans\{A,B,C\}}} \mathrm{A}^{\mathrm{R}}_{n}$	k_{transA} = 35, k_{transB} = 8, k_{transC} = 4.5	[20]
Back-translocation (3.20)	$\mathrm{O}^{\mathrm{R}}_{n}+\mathrm{U}^{\mathrm{R}}_{[n-\Delta_{\mathrm{Rib}}-2,n-\Delta_{\mathrm{Rib}}]} \xrightarrow{k_{bt}} \mathrm{A}^{\mathrm{R}}_{n-3}+\mathrm{U}^{\mathrm{R}}_{[n+\Delta_{\mathrm{Rib}}-3,n+\Delta_{\mathrm{Rib}}-1]}$	k_{bt} = 1.5	[41]
Drop-off (3.21)	$\mathrm{O}^{\mathrm{R}}_{n} \xrightarrow{k_{drop}} \mathrm{Rib}+\mathrm{U}^{\mathrm{R}}_{[n\text{-}\Delta_{\mathrm{Rib}},n+\Delta_{\mathrm{Rib}}]}$	k_{drop} = 0.000114	[61]
Trans-translation (3.22)	$\mathrm{R} \xrightarrow{k_{tt}} [\mathrm{Rib}^{\mathrm{R}}]\mathrm{Rib}$	k_{tt} = 0.000052	[87]
Elongation completion (3.23)	$\mathrm{A}^{\mathrm{R}}_{n_{last}} \xrightarrow{k_{trans\_f}} \mathrm{Rib}+\mathrm{U}^{\mathrm{R}}_{[n_{last},n_{last}\text{-}\Delta_{\mathrm{Rib}}]}+\mathrm{P}_{\mathrm{prem}}$	k_{trans_f} = 2	[20]
Folding and activation (3.24)	$\mathrm{P}_{\mathrm{prem}} \xrightarrow{k_{fold}} \mathrm{P}$	k_{fold} = 0.0024	[88]
Protein degradation (3.25)	$\mathrm{P} \xrightarrow{k_{dec}} \varnothing$	k_{dec} = 0.0017	[88]

Given the above, we note that the dynamics of transcription and translation are sequence dependent in the present model in the following ways. First, the model allows the insertion of, e.g., arrests or sequence specific pauses at a specific nucleotide (exemplified in the last section of the results section). In general, since the rates of all possible events are defined uniquely for each nucleotide, any event may be set to have a distinct propensity at a specific nucleotide rather than a constant rate for all nucleotides. Translation elongation is, in the same manner, sequence dependent, with the additional feature that the rates of elongation in this case are always codon dependent.

The chemical reactions and rate constants (in s^{-1}) used to model translation initiation, elongation, and termination, as well as protein folding and activation and protein degradation are in Table 3.2. Parameter values were obtained from measurements in *E. coli*, mainly for *LacZ.*

3.3 Correlation and time-averaging of noise

Protein levels do not respond instantaneously to changes in the number of mRNA molecules in the system since new proteins take time to synthesize after a new mRNA is produced, and excess proteins take time to degrade after an mRNA has been degraded. Since the processes of creation and degradation in proteins take longer than in RNA, the fluctuations in protein levels result from a time averaging of the fluctuations in mRNA levels [8]. The degree to which fluctuations propagate from RNA to protein levels depends on various parameters, the most relevant being the ratio between the degradation rates of the proteins and RNAs. Changing this ratio is likely to affect the degree of correlation between the RNA and protein time series.

The effect of the time-averaging phenomena on protein numbers is quantifiable as follows: $0 < \tau_1/(\tau_1 + \tau_2) < 1$, where τ_1 and τ_2 are the average lifetimes of mRNA and proteins, respectively. The noise (σ^2/μ^2) in protein levels due to fluctuations in mRNA levels, given a simple birth and death process following Poisson statistics, can thus be approximated as following [8]:

$$\frac{\sigma_2^2}{\langle n_2 \rangle^2} \approx \frac{1}{\langle n_2 \rangle} + \frac{1}{\langle n_1 \rangle} \frac{\tau_1}{\tau_2 + \tau_1}, \tag{3.26}$$

where $<n_1>$ and $<n_2>$ are the average number of mRNA and proteins, respectively. The first term on the right-hand-side includes the contributions from the small-number Poisson fluctuations of probabilistic individual birth and death events of proteins. This noise does not necessarily have to be Poissonian [8]. The second term accounts for the contributions from random changes in the rate of protein synthesis caused by fluctuations in mRNA numbers.

There is another noise source, namely, arising from the stochasticity in the gene activation dynamics. mRNA numbers adjust quickly to changes in gene activity, while proteins adjust slowly to changes in mRNA level. If we assume Poisson statistics in all steps a more accurate estimation of noise in protein numbers is given by [59]:

$$\frac{\sigma_3^2}{\langle n_3\rangle^2} \approx \frac{1}{\langle n_3\rangle} + \frac{1}{\langle n_2\rangle}\frac{\tau_2}{\tau_3+\tau_2} + \frac{1}{\langle n_1\rangle}\frac{\tau_2}{\tau_3+\tau_2}\frac{\tau_1}{\tau_3+\tau_1}\frac{\tau_1+\tau_3+\tau_1\tau_3/\tau_2}{\tau_1+\tau_2}, \tag{3.27}$$

where <n$_1$>, <n$_2$> and <n$_3$> are the average number of available genes, mRNA and proteins, respectively. τ_1, τ_2 and τ_3 are the average lifetimes of “gene availability for transcription”, mRNA and proteins, respectively. The first and second noise terms are the same as in the previous equation. The additional noise term accounts for random changes in gene availability, where the first factor is a measure of stationary small-number gene fluctuations which can be defined as a binomial variable in the case of single gene expression [59].

To assess the extent to which fluctuations in RNA levels are propagated to protein levels, we compute the normalized discrete cross-correlation [89] between the time series of RNA and protein numbers. The normalized cross-correlation function *r* for *m* pairs of time series (*x* and *y*) of discrete signals of length *n* is given by:

$$\mathrm{r}\left[\tau\right] = \frac{\sum_{l=1}^{N}\sum_{k=1}^{n-\tau}\left(x_l[k]-m_{x1,\dots,N[1,\dots,n-\tau]}\right)\left(y_l[k+\tau]-m_{y1,\dots,N[1+\tau,\dots,n]}\right)}{\left(\left(\mathrm{n}\text{-}\tau\right)N-1\right)s_{x1,\dots,N[1,\dots,n-\tau]}s_{y1,\dots,N[1+\tau,\dots,n]}} \tag{3.28}$$

where $\tau\in\{0,\dots,\mathrm{n}-1\}$ is the lag, and m_{w} and s_{w} are the sample mean and sample standard deviation of *w*, respectively, defined by:

$$\mathrm{m}_{\mathrm{w}_{1\dots\mathrm{N}}[i\dots j]} \doteq \frac{1}{(j-i+1)N}\sum_{l=1}^{N}\sum_{k=i}^{j} w_l[k] \tag{3.29}$$

$$s_{\mathrm{w}_{1\dots\mathrm{N}}[i\dots j]} \doteq \sqrt{\frac{1}{(j-i+1)N-1}\sum_{l=1}^{N}\sum_{k=i}^{j}\left(w_l[k]-m_{\mathrm{w}_{1\dots N}[i\dots j]}\right)^2} \tag{3.30}$$

3.4 Measurements of *in vivo* transcription and translation in *E. coli*

3.4.1 Bacterial strains and plasmids for measurements of protein levels

In this study we used a new bacterial strain constructed by Shannon Healy and Olli-Pekka Smolander using an intermediate lifetime green fluorescent protein, GFP(AAV) [75] which was placed under the control of the P$_{\mathrm{LtetO\text{-}1}}$ promoter [80] and inserted into the *E. coli* genome at the *galK* locus by homologous recombination of λRED. This strain was used to measure production from a single gene instead of from multi copy plasmids.

The method of RNA detection and quantification *in vivo* in *E. coli* cells DH5α-PRO uses the ability of the coat protein of bacteriophage MS2 to tightly bind specific RNA sequences [90]. Detection of single RNA transcripts with 96 tandem repeats of the MS2 binding sites in *E. coli* is possible by using dimeric MS2 fused to GFP (MS2d-GFP fusion protein) as a detection tag [14]. The method uses two genetic constructs. The first is a medium-copy vector that expresses the MS2d-GFP fusion protein, whose promoter ($P_{LtetO\text{-}1}$) is regulated by tetracycline repressor. The second is a single copy F-based vector, with a $P_{lac/ara}$ promoter controlling production of the transcript target, i.e. mRFP1 followed by a 96 MS2 binding site. Constructs were generously provided by I. Golding (University of Illinois). To detect the individual RNA molecules from P_{tet} promoter, an F-based single copy plasmid vector with P_{tet} and transcript target was created by Meenakshisundaram Kandhavelu. Together with this F-based plasmid, a medium copy number plasmid expressing MS2d-GFP fusion protein was inserted into the same cells.

3.4.2 Cell culturing and microscopy of proteins and mRNA molecules

For measuring mRNA molecules, $P_{LtetO\text{-}1}$ with mRFP1-MS2-96bs cells were grown in Miller LB medium, supplemented with antibiotics at 37 ^{o}C with shaking (250 RPM), diluted into fresh medium to reach a final optical density of OD_{600} of 0.3-0.5. The cells were incubated with the inducer IPTG (1 mM) for 60 min to reach a full induction of MS2-GFP, to produce detectable amount of protein tags for RNA. Various concentration of aTc (0, 0.1, 0.5, 1, 2 ng/ml) (IBA GmbH, Göttingen, Germany) were used to induce the promoter expressing the target RNA. Finally, the cells were incubated at 37 ^{o}C with shaking (250 RPM) for 60 min. After induction, a few microliters of culture were placed between a cover-slip and a thin slab of LB/1% agarose and imaged with microscope.

Protein level measurements were conducted using $P_{LtetO\text{-}1}$ cells with the same cell preparation protocol as in mRNA molecule measurements but the first induction step of protein tags with IPTG was not used. In both cases, cells were visualized by fluorescence microscopy, using a Nikon Eclipse (TE2000-U, Nikon, Tokyo, Japan) inverted C1 confocal laser-scanning system with a 100x Apo TIRF (1.49 NA, oil) objective. GFP fluorescence is measured using a 488 nm laser (Melles-Griot) and a 515/30 nm detection filter. Images of cells are taken from each slide using C1 with Nikon software EZ-C1.

3.4.3 Image processing

We detect cells from raw images according to the method in [91] that divides a grayscale image in three classes: background, cell border and cell region. An iterative cell segmentation process identifies and segments clumped cells based on size and edge information. The performance of detection of cells degrades in regions where several cells are clumped together. This can be avoided by applying a threshold based on cell size and discarding the cells whose size goes beyond the threshold.

Detection of MS2d-GFP-RNA spots was made both by inspection and with an automated method. The number of disagreements between the two methods was negligible, and in these few cases, the cells were not used. The automatic spot detection method segments the MS2d-GFP-RNA spots with the kernel density estimation method for spot detection proposed in [92]. This method estimates the probability density function over the image from local information, and processes an image f by filtering it with a kernel:

$$\hat{f}(i,j) = \frac{1}{card(C(i,j))h} \sum_{(k,l)\in C(i,j)} K\left(\frac{f(i,j)-f(k,l)}{h}\right) \tag{3.31}$$

where h is the smoothing parameter or bandwidth, (k, l) represents pixel location in the kernel, card is the cardinality of the set, and K(u) is the kernel. We used a Gaussian kernel [93], and then applied Otsu's threshold [94] to segment spots from the kernel density estimated image, highlighting the spots. After removing the outliers, we subtracted the background autofluorescence from the fluorescence levels of the cells. The background intensity is estimated by measuring the autofluorescence of λRED cells without the GFP insertion and then determining the mean background dependence on cell size.

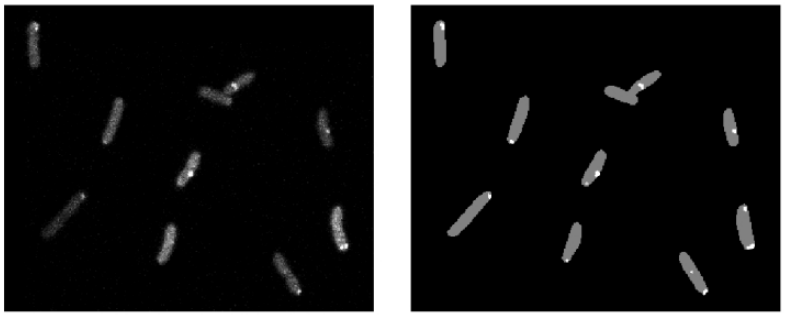

Figure 3.3. *MS2d-GFP-tagged RNA molecules in E. coli cells. Unprocessed gray-scale image of E. coli cells (left) and the corresponding segmented image showing the detected cells (grey) and the spots (white) inside the cells (right).*

To obtain the total fluorescence of tagged RNA spots, one needs to discount the cellular background. Let FGI be the total (sum) foreground (spots) intensity, FGA the total foreground area, BGI the total background (cell) intensity, and BGA the cell area. The total intensity I of a spot is given by:

$$I = FGI - FGA\frac{BGI - FGI}{BGA - FGA} \tag{3.32}$$

Finally, the number of RNA molecules in each spot is quantified using the spot intensity distribution slicing approach [14], that assumes that the first peak of the distribution of intensities of many RNA spots from cells on the same slide correspond to individual RNA molecules. Subsequent peaks in the distribution of intensities correspond to spots of multiple RNA molecules.

4 RESULTS AND DISCUSSION

4.1 Transcription and translation dynamics

4.1.1 Dynamics of mRNA production

Given the number of chemical reactions per nucleotide in the model and that one gene can have thousands of nucleotides, the dynamics are considerably complex. To illustrate this, we show examples of the kinetics of multiple RNAps on a DNA strand within a short time interval, and the dynamics of multiple ribosomes on one of the RNA strands as it is transcribed. Parameter values were obtained from measurements in *E. coli* for *LacZ* (see methods section), since the dynamics of transcription and translation have been extensively studied for this gene. *LacZ* has 3072 nucleotides and its transcription is controlled by the lac operon.

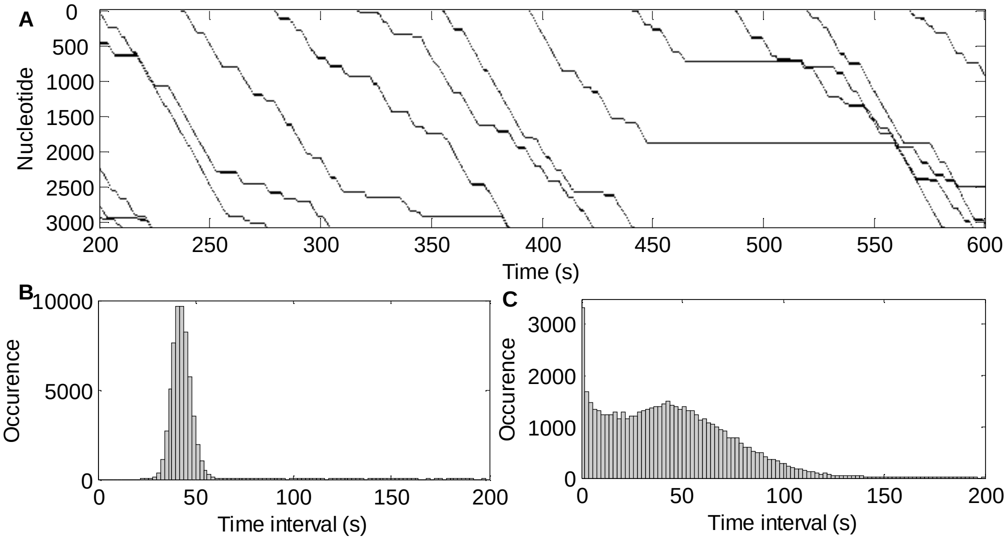

Figure 4.1. *Kinetics of RNA polymerases on the DNA strand (A) Example of the kinetics of multiple RNAp molecules on the DNA template over 400 s. Note that, on several occasions, the RNAp molecules pause and that one RNAp never overtakes another on the DNA template. (B) Distribution of time intervals between consecutive transcription initiation and (C) completion events. Data is from 57 000 initiation events.*

In this simulation, transcription is not repressed. Thus, provided that the promoter is available for transcription, the expected time for a transcription event to start is approximately 2.5 s, given the value of the rate constant of reaction (3.1) in Table 3.1 and that there are 28 RNAp molecules available in the system [5]. The promoter open complex formation step, with a mean duration of 40 s [65] and a standard deviation of 4 s [21] is the major limiting factor of transcription events in these conditions.

Figure 4.1 (A) shows, for a time window of 400 seconds, the positions (y-axis) over time (x-axis) of several RNAp molecules on the DNA template. In real time, this simulation takes ~30 s, on an Intel Core 2 Duo processor. Transcription elongation is visibly stochastic, with events such as arrests (e.g. at t = ~450 s), ubiquitous pauses and pyrophosphorolysis. Several collisions between RNAp molecules are also visible, caused in part by these events. Note that one RNAp never overtakes another on the template.

Figure 4.1 (B) shows the distribution of the time intervals between transcription initiation events, which is Gaussian-like, due to the open complex formation step. The longer tail on the right side of the distribution is mainly due to the contribution of the time it takes for the RNAp to bind to the template, a bimolecular reaction whose expected time to occur follows an exponential distribution with a mean of 2.5 s [25, 62].

Figure 4.1 (C) shows the distribution of time intervals between transcription completion events in the same simulation as Figure 4.1 (B). This distribution is strikingly different from that of Figure 4.1 (B) due to the stochastic events in transcription elongation. Pauses, arrests and other stochastic events cause the distribution to be bimodal due to the bursty dynamics (many short intervals and some long intervals). When these probabilistic events occur to some RNAp molecules, they significantly alter the distances in the strand between consecutive RNAps. For example, when one RNAp pauses, its distance to the preceding RNAp increases, while the distance to subsequent RNAps shortens, allowing completion events to be separated by intervals shorter than the promoter delay.

4.1.2 Dynamics of protein production

Figure 4.2 (A) exemplifies the dynamics of ribosomes on one RNA strand. Stochastically, the transcription elongation process of this particular mRNA was halted at t = 50 s for a long period, and was thus selected to illustrate how long pauses in transcription affect the dynamics of translation of the multiple ribosomes on the RNA strand. The solid gray region in the bottom left part of the figure corresponds to the as-of-yet untranscribed sequence of the mRNA. When the RNAp pauses or is arrested (e.g. at t = 50 s), ribosomes accumulate in the region of the mRNA preceding the leading edge of transcription. Stochasticity in the translation elongation process is also visible. However, this process, modeled with realistic parameter values, appears to be less stochastic than transcription elongation, in that the stepwise elongation of ribosomes on the RNA template is more uniform than that of the RNAps on the DNA template. This is especially visible after the effects of the long arrest disappeared (at t > 230 s), at which point the distributions of time intervals between consecutive ribosomes at the start and at the end of translation elongation do not differ significantly.

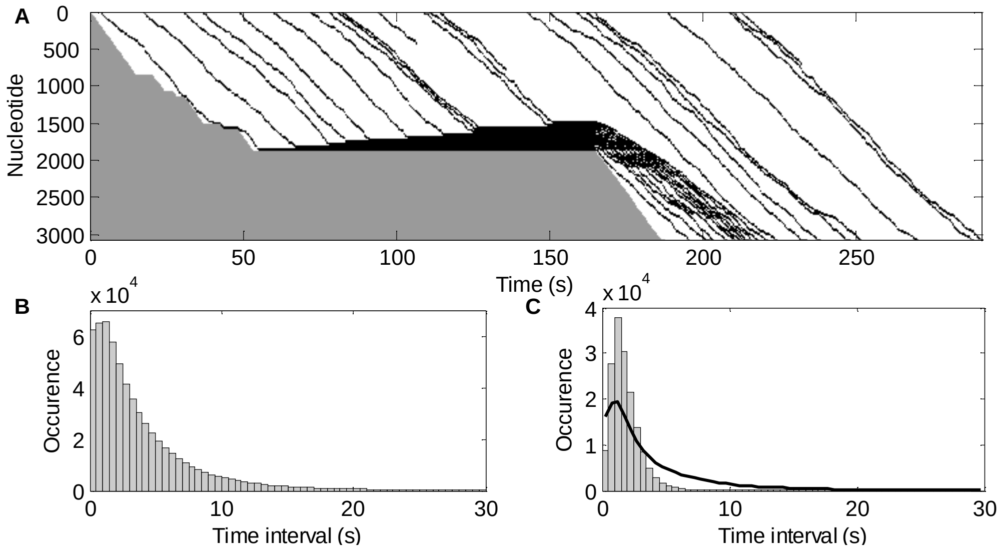

Figure 4.2. *Kinetics of ribosomes on an RNA strand (A) Example of the kinetics of several ribosomes along an mRNA template that suffered an arrest at nucleotide 1850, from the moment the ribosome binding site is formed to the degradation of the mRNA. The continuous gray region in the bottom left corresponds to the untranscribed sequence of the mRNA. (B) Distribution of time intervals between consecutive translation initiation events. (C) Distribution (grey bars) of time intervals between consecutive translation completion events given the presence of a sequence dependent arrest site at nucleotide 1850. The solid black line shows the distribution of time intervals between consecutive translation completion events without the sequence-dependent arrest site, normalized to the same scale. Data is from 600 000 initiation events.*

Figure 4.2 (B) shows the distribution of intervals between translation initiation events. Since there is no significant delay in translation initiation (as the one due to the promoter open complex formation), this distribution is exponential-like. Figure 4.2 (C) shows the corresponding distribution of intervals between translation completion events (grey bars), given the presence of a sequence dependent arrest site at nucleotide 1850. This distribution, while resembling that of Figure 4.2 (B), shows more short time intervals, due to the long arrest in transcription elongation. For comparison, we also show a distribution of intervals between translation completion events drawn from cases without the sequence dependent arrest in transcription (solid black line). The difference between the two distributions illustrates how events in transcription elongation (e.g. a sequence dependent arrest site) can significantly affect the dynamics of translation.

4.1.3 Comparing the model with measured translation dynamics

Recently, the real-time expression of a lac promoter was directly monitored in *E. coli* with single-protein resolution [13]. The proteins were found to be produced in bursts (i.e. several proteins being produced from each RNA), with the distribution of intervals between bursts fitting an exponential distribution, while the number of proteins per burst followed a geometric distribution [13]. These distributions were measured for a gene that was kept strongly repressed and for which the ribosome binding site (RBS) was engineered so that translation was also very weak [13]. Under these conditions, our model reproduces these dynamics (data not shown). Nevertheless, we note that it is possible to match these measurements with a simpler model than the one proposed here, where transcription and translation are modeled as single step events [21, 23].

We next compare the kinetics of translation in our model with measurements of the translation elongation speed in three engineered *E. coli* strains designed to enhance queue formation and traffic in translation [17]. Each strain contains a different mutant of *LacZ*. The pMAS23 strain corresponds to the wild-type *lacZ*. The other two sequences differ in that a region of slow-to-translate codons was inserted (~24 in pMAS-24GAG and ~48 in pMAS-48GAG). The speed of protein chain elongation was measured by subjecting the cells to a pulse of radioactive methionines, and then measuring the level of radioactivity in cells of each population, every 10 s after the pulse. Each strand contained 23 methionines, spread out unevenly on the DNA sequence, causing the incorporation curve to be non-linear.

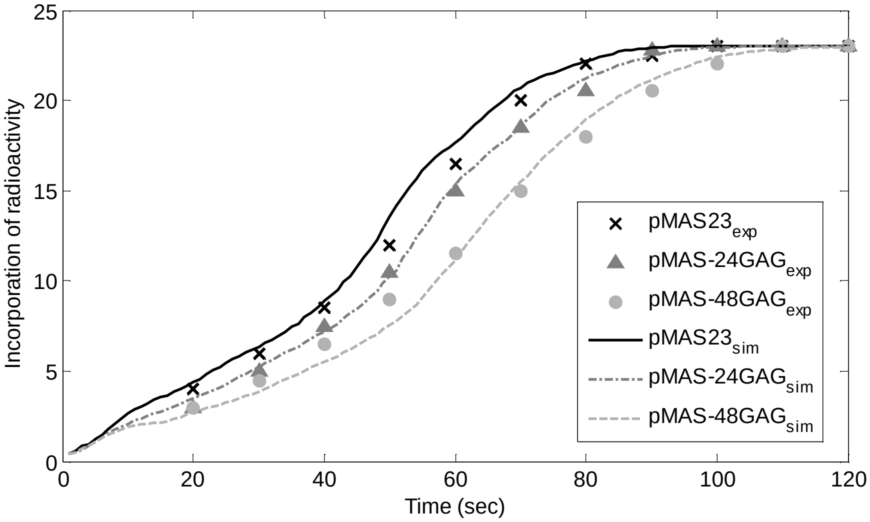

Figure 4.3. *Appearance of radioactivity in β-galactosidase. Appearance of radioactivity incorporated from the three different mRNA strands, at different times after initiation of translation elongation in the models (lines) and in the measurements (crosses, triangles and circles) [17]. Values of radioactivity are normalized such that the maximum corresponds to 23 radioactive methionines.*

Given that they differ in the nucleotide sequence, it was hypothesized that the translation elongation speed of the three strands would differ, as the speed of incorporation of an amino acid depends on which synonymous codon is coding for it [17]. The cells where translation is faster will thus be expected to have higher levels of radioactivity in the translated proteins, as more labeled amino acids have been incorporated in a fixed time interval. If the translation speeds of the three strands were identical, they would exhibit identical levels of radioactivity at the same point in time.

To model this, we simulate the transcription and translation processes of the three sequences [17]. We model the incorporation of radioactive methionines at the same locations as in these sequences. The three model strands differ only in sequence, as in the measurements. During the simulations, we measure the number of incorporated radioactive methionines at the same points in time as in the experiment. Results of our simulations and of the measurements [17] are shown in Figure 4.3, showing good agreement between model and measurements.

4.1.4 Propagation of fluctuations in RNA levels to protein levels

We simulate the model for varying effective rates of transcription initiation (denoted k_{eff}). This rate is determined by the basal rate of transcription initiation (k_{init}), which sets the binding affinity of the RNAp to the transcription start site, and by the strength of repression of transcription. Thus, to vary k_{eff}, we vary the number of repressor molecules present in the system. Three sets of simulations are performed, differing in rate of translation initiation (k_{tr}). This rate is one of the kinetic parameters of the model, thus can be changed directly, and not by indirect means as k_{eff}. In *E. coli* genes, this rate is believed to be determined by the RBS sequence [68]. mRNA and protein degradation rates are set so that the mRNA and protein mean levels are identical for all cases, allowing us to study how the level of noise in mRNA and protein levels changes.

For each set of values of k_{eff} and k_{tr} we perform 100 independent simulations. Depending on these rates, the mean time to reach steady state differs. Each case is simulated for long enough to reach steady state and for an additional 100 000 s after that. The time series of the 100 simulations for each set of parameter values is concatenated into one time series, from which the noise is quantified by the square of the coefficient of variation, CV^2 (variance over the mean squared) [59]. This number of long simulations is necessary to properly sample the system due to the stochasticity of the underlying processes.

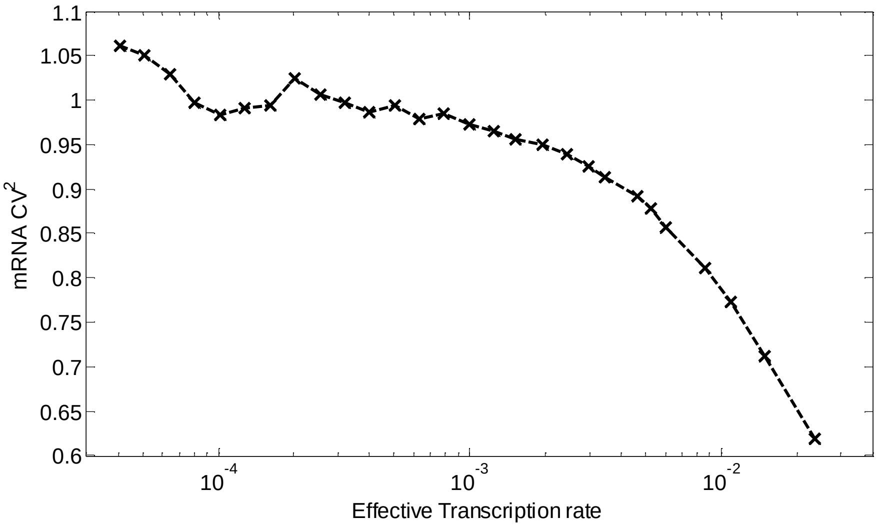

Figure 4.4. *Noise in mRNA as a function of the transcription initiation rate. Noise (CV^2) in mRNA levels for varying effective transcription initiation rates. The mRNA degradation rate is set so that the mean mRNA levels at steady state are identical in all cases.*

In Figure 4.4, we first show the CV^2 of mRNA time series for varying k_{eff}. Noise decreases as k_{eff} increases due to the promoter open complex formation step [6]. Without this event, the distribution of time intervals between transcription initiation events would be exponential, and the CV^2 would not vary. However, with this step, if the expected time for an RNAp to bind to the free promoter is faster than the duration of the promoter open complex formation, then the distribution of time intervals becomes Gaussian-like [6].

No measurements have yet been made to study experimentally the relation between the noise in mRNA levels and the corresponding protein levels. Nevertheless, it is possible to create a robust estimate, provided reasonable assumptions on the nature of the underlying processes [8]. Our model allows for a direct assessment, and it additionally includes realistic events such as RNAp and ribosome traffic, in transcription and translation elongation, which are not included in the aforementioned estimations [8]. Figure 4.5 shows the noise (CV^2) in protein levels, for varying k_{eff} and three values of k_{tr}. The data was obtained from the same simulations used to generate the results in Figure 4.4.

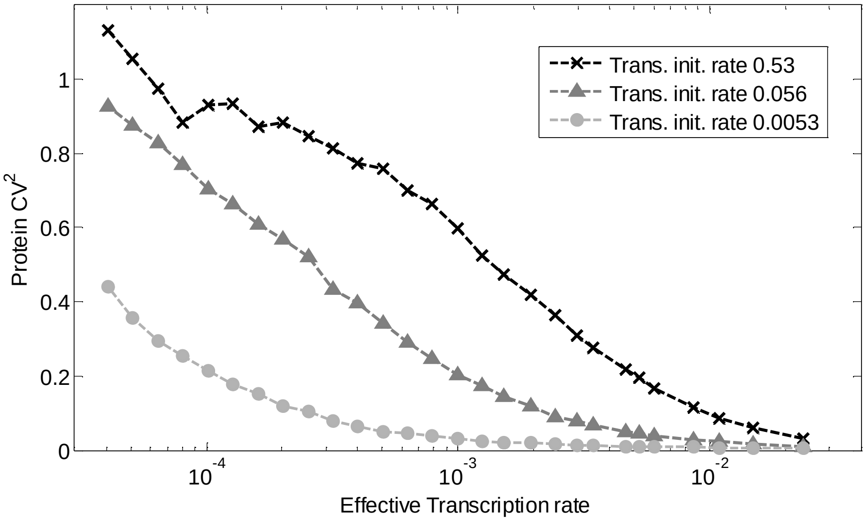

Figure 4.5. *Noise in protein levels for varying transcription and translation initiation rates. Noise (CV^2) in protein levels for varying effective transcription initiation rates and three different rates of translation initiation. mRNA and protein degradation rates are set so that the mean mRNA and mean protein levels at steady state are identical in all cases.*

In general, we find that increasing k_{eff} decreases the noise in protein levels due to the decrease of noise in mRNA levels. Increasing k_{tr} increases the noise in protein levels, due to the increased size of the bursts in the protein level [8, 59]. This finding has not yet been experimentally validated by direct means.

An interesting observation from Figures 4.4 and 4.5 is that, for $k_{eff} < 5x10^{-4}$ s^{-1}, as k_{eff} is increased, the noise in protein levels decreases significantly, while the noise in RNA levels does not noticeably change. This is due to the decrease in mean protein burst size, i.e., the mean number of proteins produced from each RNA molecule, as both k_{eff} and the degradation rate of RNA molecules are varied.

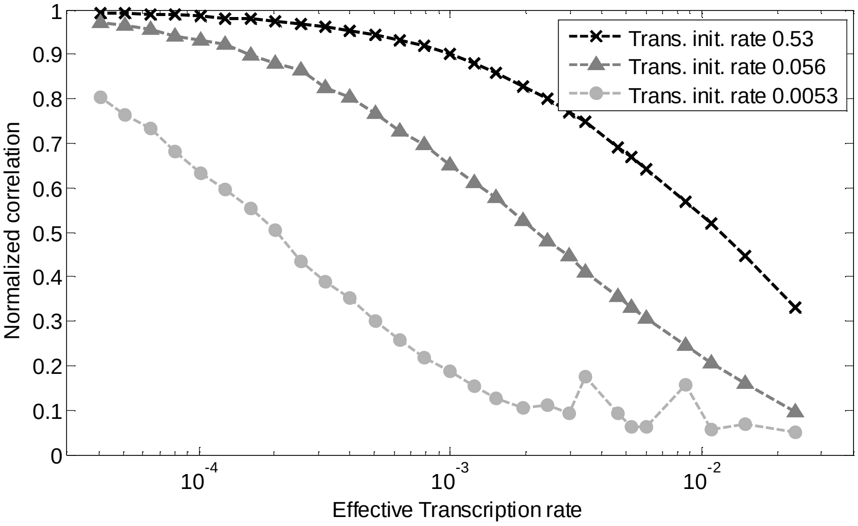

Figure 4.6. *Normalized maximum correlation between RNA and protein time series. The higher the rate of translation initiation (and thus higher protein degradation to keep the mean the same), the more correlated the fluctuations in protein and RNA levels become, as measured by the normalized maximum correlation. This is because the protein levels follow any fluctuations in the RNA levels faster. Similarly, increasing the rate of transcription initiation, while maintaining the rate of translation initiation constant, decreases the correlation between fluctuations in protein and RNA levels.*

From these results, we conclude that the degree of coupling between transcription and translation is likely to be a key determining factor of the noise in protein levels. This can be verified by computing the normalized maximum correlation between time-series of protein and mRNA levels for each set of parameter values (Figure 4.6**).** Comparing Figures 4.5 and 4.6, we see that higher correlation values are obtained for the regime of higher noise in the protein levels. This implies that the principal source of this noise is the fluctuations in RNA levels.

The correlation value is largely determined by the rates of mRNA and protein degradation and production. For example, both increasing the mRNA degradation rate and/or decreasing the protein degradation rate increases the time averaging constant of the mRNA fluctuations, and thus decreases the correlation between mRNA and protein levels. In general, if the mean mRNA and protein levels and kept unchanged by tuning their degradation rates accordingly, the correlation between RNA and protein time series can be increased by lowering the mRNA production rate and/or increasing the protein production rate.

4.1.5 Transcriptional pauses and the fluctuations in protein levels

Recent work [1] reported that long transcriptional pauses enhance the noise in mRNA levels. We next investigate to what extent the fluctuations in RNA levels caused by long transcriptional pauses propagate to protein levels. Long sequence-dependent pauses [16, 47, 95] in transcription elongation may cause the ribosome to stall in the mRNA chain. This will likely cause subsequent ribosomes to accumulate in the preceding sequence. When the RNAp is spontaneously released from the pause [47], translation of the stalled ribosomes likely resumes but the distribution of intervals between them will differ significantly from what it would have been without the pause event. Consequently, the protein production is likely to become more bursty, especially if the long pause site is located near the end of the sequence. An increase in burstiness ought to increase the noise in protein levels.

To verify this, we perform two simulations. We introduce a long-pause sequence with mean pause durations of 500 s in one case, and 100 s in the other (both values are within realistic intervals [95]). In both cases, we set the probability that an RNAp will pause at that site to 70% (identical to the value for *his* pause sites [16]).

Measuring the protein noise levels, we find that the CV^2 is ~5% higher for the 100 s pause site and ~10% higher for the 500 s pause site, in comparison to the same sequence without any sequence specific long-pause site. These relative differences can be biologically relevant in that such a change may, in some cases, cause the degree of phenotypic diversity of a monoclonal cell population to change.

The effects of several pause sites on the same strain are cumulative, namely, the higher the number of pause sites, the higher the noise in RNA levels [96]. Combined with the present results, this leads us to the conclusion that the sequence-dependent transcriptional pausing mechanism likely exists to allow a wide variation of both RNA and protein noise levels.

4.2 Measurements of RNA and Protein levels

We use recently developed methods to measure gene expression *in vivo* in individual cells, at the single RNA and protein molecule levels. Such measurements, attained in various conditions, as well as the proposed modeling strategy, are used to study the dynamics of transcription and translation at the single event level and to estimate noise sources in these processes.

The Fano factor, defined as variance divided by the mean, is a common measure of diversity of RNA and protein numbers across a cell population [97]. The reason is that, for a Poisson process, the variance equals the mean, i.e. Fano factor is one. The comparison with the Poissonian process only works well for univariate discrete random processes, where the variance is proportional to the average, in which scenario the proportionality constant describes the overall behavior of the process [59].

We observed the distribution of GFP expression levels in the cells for each concentration of aTc and studied how the diversity in gene expression changes. As one increases the strength of induction, the Fano factor remains approximately constant for weak induction strengths, but then it increases for the two highest levels of induction measured.

Relevantly, this quantity would not vary if transcription remained a Poissonian process for all levels of induction [98].The distributions from the stochastic model are shown in Figure 4.7, superimposed with the distributions attained from the measurements. The only difference between this model and the previously described one is that it includes an extrinsic noise source in the gene activation process. We model this noise source by having a variable number of RNAps transcribing the mRNA, rather than a constant population size. The number of RNAps available to express a fully induced gene in *E. coli* has been estimated to vary between 5 and 20 [58]. This variability contributes to the variance in gene expression, as it affects from gene activation to mRNA numbers and further to protein levels. The fluctuations in RNAp numbers was set to be a very slow process i.e. in the order of 45 minutes, which corresponds to the timescale of cell division, which ought to be the one of the main sources for these fluctuations [99]. The mean number of RNAps over time is the same as in the previous models.

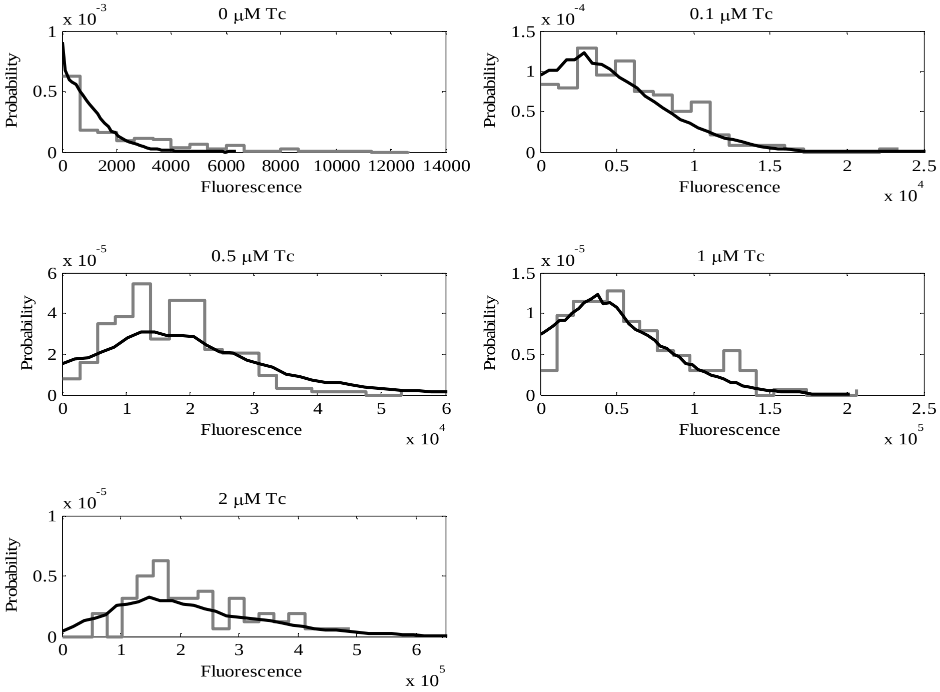

Figure 4.7. *Measured distributions of cells with a given protein level compared with model estimations. Binned distribution from measurements (grey line) of the cells with given GFP expression levels for aTc of 0, 0.1, 0.5, 1, and 2 (ng/ml). The distribution of expression levels as predicted by the model is shown for each case (black line). In each model, mean expression level was imposed to be the same as in the measurements.*

Figure 4.8 shows the Fano factor for the protein numbers measured in the cell populations and measured from simulations of single cells dynamics with the stochastic model. The measurements show a constant Fano factor for the low induction regime but an increase as it enters the high induction regime. The stochastic model behaves similarly as the measurements showing a minimum noise level, that in are due to noise in the gene activation process, which is always present. In Poisson processes the variance is inversely proportional to the mean but the population level noise in gene activation is not, causing the Fano factor of proteins to increase for higher protein numbers.

If the noise in the gene activation step causes the increase in Fano factor of GFP intensities in individual cells as induction strength is increased, then its effects ought to be visible also in the variance of RNA numbers of the cell population. To study this, we measured the transcriptional activity of a tetracycline inducible promoter, P_{tet}, at the single RNA molecule level, as described in the methods section. The measurements were conducted under the same levels of induction as in study of the expression levels of GFP. The Fano factor of RNA numbers in individual cells is shown in Figure 4.9. For induction strengths of 0, 0.1, 0.5, 1 and 2 ng/ml, the number of cells analyzed was 128, 185, 83, 124 and 248, respectively.

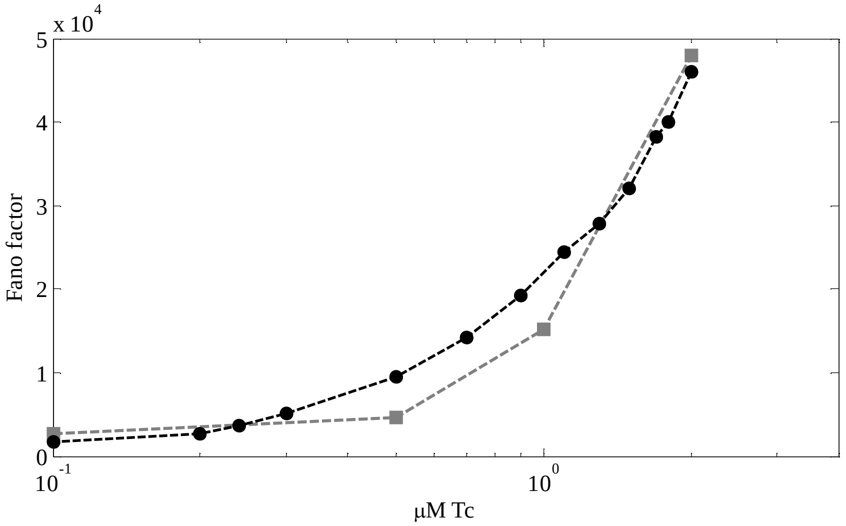

Figure 4.8. *Fano factors in experiments and models. Fano factors for increasing induction strength in model (o) and measurements (□).*

Figure 4.9 shows the Fano factor for the RNA numbers measured from the cell populations and for the same stochastic model that was used to estimate the expected Fano factor of protein numbers. The results from the direct measurements of RNA numbers show a slowly rising Fano factor within the low induction regime but then a fast increase in the regime of high induction. In the stochastic model, the increase in Fano factor starts approximately for the same induction levels, because the noise in RNA numbers is the main source of diversity in protein levels. If the Fano factor of RNA numbers increased in a different regime, it would indicate an independent noise source, located at the translation process (in the model, there are no subsequent noise sources such as during protein folding and activation).

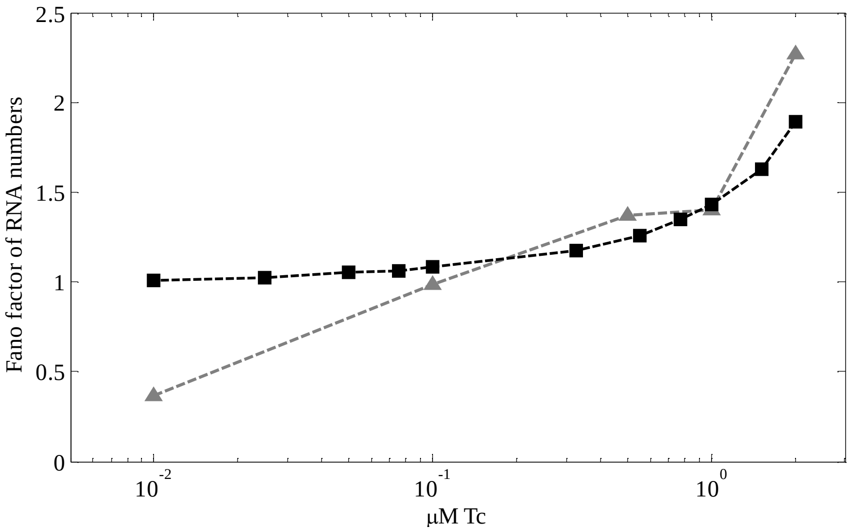

Figure 4.9. *Fano factors in measurements of RNA numbers in individual cells. Fano factors in RNA numbers for increasing induction in model (□) measurements (Δ).*

Comparing Figures 4.8 and 4.9, the Fano factors of RNA and protein levels have a visible resemblance, providing strong evidence that the increase observed in the Fano factor of protein levels arises at the transcription stage, more precisely, due to the variability in intervals between consecutive transcription events. From all of the above, the source of this variability is likely the gene activation dynamics, as the variability in protein numbers follows that of the RNA numbers.

5 CONCLUSIONS

We proposed a new delayed stochastic model of prokaryotic transcription and translation at the nucleotide and codon level, where the processes of transcription and translation are dynamically coupled in that translation can initiate immediately upon the formation of the ribosome binding site region of the nascent mRNA. Simulations of the dynamics show that, within realistic parameter values, the protein noise levels are determined, to a great extent, by the fluctuations in the RNA levels, rather than from sources in translation, in agreement with indirect measurements [14], as translation elongation was found to be less stochastic than transcription elongation. Specifically, the distributions of intervals between translation initiation and translation completion events only differ significantly if the sequence possesses long sequence-dependent pauses or clusters of slow-to-translate codons. The sequence dependence of several mechanisms that can act as generators of strong fluctuations in RNA levels [15], the propagation of these fluctuations to protein levels, and the ability of fluctuations in protein levels to affect cellular phenotype [100], suggest that these mechanisms may be subject to selection and, thus, are evolvable.

As a previous study has suggested [8], the translation initiation rate was found to be key in determining the degree of coupling between the fluctuations in RNA and protein levels, if one assumes that the degradation rate of the proteins is changed accordingly to maintain their mean level unchanged. Varying this sequence-dependent, and thus, evolvable parameter [68] within realistic ranges gave a widely varying degree of coupling between the fluctuations in RNA and protein levels. It is therefore not necessarily true that noisy production of RNA molecules results in noisy protein levels. Interestingly, while decreasing the coupling between transcription and translation by decreasing the rate of translation initiation causes the protein levels to become less noisy, it also takes longer for a change in RNA levels to be followed by the protein levels. This suggests that to be able to change rapidly in response to, e.g., environmental changes, the levels of a protein will be necessarily noisier.

Confirming previous studies [1, 5, 8, 19], we found that the distributions of time intervals between transcription initiation and completion events differ significantly and that the faster the rate of transcription initiation events, the more they differ. This implies that in the regime of fast transcription, both the transcription and translation elongation processes need to be modeled explicitly and coupled, if one is to match the mean and fluctuations in the protein levels at the molecular level. This is of relevance, since bursts in protein levels may trigger many processes, such as phenotypic differentiation [58, 100]. A final justification for using the model proposed here is the complexity of the process of gene expression in *E. coli*, and the fact that many events therein may or may not affect the temporal RNA and protein levels significantly, depending on their specific sequence-dependent features. Such effects, due to the complexity of the system, are not easily predictable without performing explicit numerical simulations and measurements at the single event level.

The model proposed here includes several features not included in previous models such as a gradual degradation event that can be triggered while the RNA is still being transcribed. As its parameter values were extracted from measurements, it should be useful in the study of several aspects of the dynamics of gene expression in prokaryotes that cannot yet be measured directly and to explore the state space of gene expression dynamics by varying any of the physical variables within realistic ranges.

However, the present model does not yet account for known effects of ribosomes on the dynamics of transcription elongation. These might need to be included in future developments of the proposed model as recent results [69, 70] suggest that the rate of translation elongation can affect the rate of transcription elongation, due to possible interactions between the ribosome that first binds to the mRNA and the RNAp transcribing it. Possible effects may include facilitating the release of paused RNAps, which could affect the degree of the contribution of pauses to the noise in RNA and thus protein levels. We do not exclude the possibility that the contrary may occur in specific cases, that is, that the paused state of the RNAp may cause pauses in the ribosome translational dynamics, which would amplify the effect of transcriptional pauses on the fluctuations of protein levels. Whether the pause is ubiquitous or due to loop formations in the nascent RNA may affect the results of the interaction as well. Provided experimental evidence on the nature and consequences of these interactions, once included in the model, we may be able to test, among other things, whether long transcriptional pauses located in an attenuator system provide an additional layer of control over premature transcription terminations, and thus over RNA and protein noise levels.

In the measurements of gene expression, as we increased transcription induction, we observed an increase in cell-to-cell diversity in protein numbers in a gene integrated into *E. coli* genome as the higher levels of induction were reached. This increase would not have been observed if the process of RNA production obeys Poissonian statistics [6, 13, 101]. The observed distribution of protein expression in individual cells indicates that the production of RNAs is not a Poisson process in the regime of strong induction. To verify the source of diversity in protein numbers for strong induction regimes, we measured directly the transcriptional activity by detecting individual RNA molecules as these are produced. For that, we placed the promoter controlling the expression of an RNA sequence target for 96 MS2-GFP proteins. The Fano factor of these RNA numbers in individual cells varied with induction strength in a very similar manner to the Fano factor of protein levels. We can rule out the overall cell-to-cell phenotypic diversity as a cause, as this would likely act at all induction strengths. Further, we can rule out measurement noise and autofluorescence, as this would mainly affect the results in the regime of weak induction.

We compared the dynamics of our stochastic model of gene expression with the measurements. The comparison suggests that the variability in the gene activation dynamics is the most likely source of noise in the dynamics of RNA production in the regime of strong induction, where the effects of low-copy number noise are minimal. This variability in the gene activation dynamics enhances significantly the observed cell-to-cell diversity in protein numbers.

Relevantly, the fano factors of RNA and protein behaved similarly in the model and in the measurements. The increases as one enters the regime of strong induction reflect the existence of a noisy process that is independent of induction strength. In the regime of low induction, low-copy number noise cannot be neglected, however in this case was overshadowed, as the cell to cell diversity in RNA and protein numbers increased with induction rather than decreasing.

REFERENCES

[1] Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O, Ribeiro, A.S., “Effects of transcriptional pausing on gene expression dynamics,” *PLOS Comput Biol* 2010, 6(3): e1000704.

[2] Greive, S.J., von Hippel, P.H., “Thinking quantitatively about transcriptional regulation,” *Nat Rev Mol Cell Biol* 2005, 6: 221-232.

[3] Wen, J.D., Lancaster, L., Hodges, C., Zeri, A.C., Yoshimura, S.H, Noller, H.F., Bustamante, C., Tinoco, Jr. I., “Following translation by single ribosomes one codon at a time,” *Nature* 2008, 452: 598–603.

[4] Landick, R., “The regulatory roles and mechanism of transcriptional pausing,” *Biochem Soc Trans* 2006, 34(6): 1062-1066.

[5] Ribeiro, A.S., Rajala, T., Smolander, O.P., Häkkinen, A., Yli-Harja, O., “Delayed Stochastic Model of Transcription at the Single Nucleotide Level,” *J Comput Biol* 2009, 16: 539-553.

[6] Ribeiro, A.S., Häkkinen, A., Mannerström, H., Lloyd-Price, J., Yli-Harja, O., “Effects of the promoter open complex formation on gene expression dynamics," *Phys Rev E* 2010, 81(1): 011912.

[7] Kaern, M., Elston, T.C., Blake, W.J., Collins, J.J., “Stochasticity in gene expression: from theories to phenotypes,” *Nat Rev Genet* 2005, 6: 451-464.

[8] Pedraza, J., Paulsson, J., “Effects of Molecular Memory and Bursting on Fluctuations in Gene Expression,” *Science* 2008, 319: 339-334.

[9] Murphy, K.F., Balazsi, G., Collins, J.J., “Combinatorial promoter design for engineering noisy gene expression,” *Proc Natl Acad Sci USA* 2007, 104: 12726-12731.

[10] Mayr, E., *What evolution is,* Basic Books, NY, USA, 2001.

[11] Lee, H.H., Molla, M.N., Cantor, C.R., Collins, J.J., “Bacterial charity work leads to population-wide resistance,” *Nature* 2010 467: 82-86.

[12] Acar, M., Mettetal, J., van Oudenaarden, A., “Stochastic switching as a survival strategy in fluctuating environments,” *Nature Genetics* 2008, 40: 471-475.

[13] Yu, J., Xiao, J., Ren, X., Lao, K., Xie, X.S., "Probing gene expression in live cells, one protein molecule at a time," *Science* 2006, 311: 1600-1603.

[14] Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C., "Real-time kinetics of gene activity in individual bacteria," *Cell* 2005, 123: 1025-1036.

[15] Ribeiro, A.S., "Stochastic and delayed stochastic models of gene expression and regulation," *Mathematical Biosciences* 2010, 223(1): 1-11.

[16] Herbert, K.M., La Porta, A., Wong, B.J., Mooney, R.A., Neuman, K.C., Landick, R., Block, S.M.; "Sequence-resolved detection of pausing by single RNA polymerase molecules," *Cell* 2006, 125: 1083-1094.

[17] Sorensen, M.A., Pedersen, S., "Absolute in vivo translation rates of individual codons in Escherichia coli," *J Mol Biol* 1991, 222: 265-280.

[18] Bernstein, J., Khodursky, A., Lin, P., Lin-Chao, S., Cohen, S., "Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays," *Proc Natl Acad Sci USA* 2002, 99: 9697–9702.

[19] Roussel, M.R., Zhu, R., "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression," *Phys Biol* 2006, 3: 274-284.

[20] Mitarai, N., Sneppen, K., Pedersen, S., "Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization," *J Mol Biol* 2008, 382(1): 236–245.

[21] Zhu, R., Ribeiro, A.S., Salahub, D., Kauffman, S.A., "Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models," *J Theor Biol* 2007, 246: 725-745.

[22] Voliotis, M., Cohen, N., Molina-Paris, C., Liverpool, T.B., "Fluctuations, pauses and backtracking in DNA transcription," *Biophys J* 2008, 94: 334-348.

[23] Ribeiro, A.S., Zhu, R., Kauffman, S.A., "A general modeling strategy for gene regulatory networks with stochastic dynamics," *J Comput Biol* 2006, 13: 1630-1639.

[24] Ribeiro, A.S., Lloyd-Price, J., "SGN Sim, a Stochastic Genetic Networks Simulator," *Bioinformatics* 2007 23(6): 777-779.

[25] Gillespie, D.T., “Exact stochastic simulation of coupled chemical reactions,” *J Phys Chem* 1977, 81: 2340-2361.

[26] Gillespie, D.T., “Stochastic simulation of chemical kinetics,” *Annual Review of Physical Chemistry* 2007, 58: 35-55.

[27] Gillespie, D.T., “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions,” *J Comput Phys* 1976, 22: 403–34.

[28] Gillespie, D.T., “A rigorous derivation of the chemical master equation,” *Physica A* 1992, 188: 404–25.

[29] McQuarrie, D., “Stochastic approach to chemical kinetics,” *J Appl Probab* 1967, 4: 413–78.

[30] Gibson, M.A., Bruck, J., “Exact stochastic simulation of chemical systems with many species and many channels,” *J Phys Chem* 2000, 105: 1876–89.

[31] Li, H., Petzold, L., “Logarithmic Direct Method for Discrete Stochastic Simulation of Chemically Reacting Systems,” Technical Report, 2006.

[32] Davenport, R., White, G., Landick, R., Bustamante, C., “Single-molecule study of transcriptional pausing and arrest by E. coli RNA polymerase,” *Science* 2000, 287: 2497-2500.

[33] Hooshangi, S., Thiberge, S., Weiss, R., “Ultrasensitivity and noise propagation in a synthetic transcriptional cascade” *Proc Natl Acad Sci USA* 2005, 102: 3581–6.

[34] Raser, J.M., O’Shea, E.K., “Control of stochasticity in eukaryotic gene expression,” *Science* 2004, 304: 1811–4.

[35] Smolen, P., “Baxter, D.A., Byrne, J.H., “Modeling circadian oscillations with interlocking positive and negative feedback loops,” *J Neurosci* 2001, 21: 6644–56.

[36] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., “Stochastic gene expression in a single cell,” *Science* 2002, 297(5584): 1183-1186.

[37] McAdams, H.H., Arkin, A., “It’s a noisy business! Genetic regulation at the nanomolar scale,” *Trends Genet* 1999, 15: 65–9.

[38] Bratsun, D., Volfson, D., Tsimring, L., Hasty, J., “Delay-induced stochastic oscillations in gene regulation,” *Proc Natl Acad Sci USA* 2005, 102: 14593-14598.

[39] Barrio, M., Burrage, K., Leier, A., Tian, T., “Oscillatory regulation of Hes1: discrete stochastic delay modelling and simulation,” *PLoS Comput Biol* 2006, 2: e117.

[40] Cai, X., “Exact stochastic simulation of coupled chemical reactions with delays,” *J Chem Phys* 2007, 126: 124108.

[41] Alberts B., Johnson A., Lewis J., Raff M., Roberts K., Walter P., *Molecular biology of the cell,* Garland Science, USA, 2002.

[42] McClure, W.R., “Mechanism and control of transcription initiation in prokaryote,” *Ann Rev Biochem* 1985, 54: 171-204.

[43] Lutz, R., Bujard, H., “Independent and tight regulation of transcriptional units in escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements,” *Nucleic acids research,* 1997, 25(6): 1203-1210.

[44] Walter, G., Zillig, W., Palm, P., Fuchs, E., “Initiation of DNA-Dependent RNA Synthesis and the Effect of Heparin on RNA Polymerase,” *Eur J Biochem* 1967. 3:194-201.

[45] Chamberlin, M.J., “The Selectivity of Transcription,” *Ann Rev Biochem* 1974, 43: 721-75.

[46] Hsu, L.M., “Promoter clearance and escape in prokaryotes,” *Biochimica et Biophysica Acta - Gene Structure and Expression* 2002, 1577(2): 191-207.

[47] Landick, R., “Transcriptional pausing without backtracking,” *Proc Natl Acad Sci USA* 2009, 106(22): 8797-8798.

[48] Sorensen, M.A., Kurland, C.G., Pedersen, S., “Codon usage determines translation rate in Escherichia coli,” *J Mol Biol* 1989, 207: 365-377.

[49] Shoji, S., Walker, S.E., Fredrick, K., “Ribosomal translocation: One step closer to the molecular mechanism,” *ACS Chem Biol* 2009, 4: 93–107.

[50] Qin, Y., Polacek, N., Vesper, O., Staub, E., Einfeldt, E., Wilson, D.N., Nierhaus, K.H., “The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome,” *Cell* 2006, 127: 721–733.

[51] Menninger, J.R., “Peptidyl transfer RNA dissociates during protein synthesis from ribosomes of Escherichia coli,” *J Biol Chem* 1976, 251: 3392–3398.

[52] Keiler, K.C., “Biology of trans-translation,” *Annu Rev Microbiol* 2008, 62: 133–151.

[53] Ozbudak, E., Thattai, M., Kurtser, I., Grossman, A., van Oudenaarden, A., “Regulation of noise in the expression of a single gene,” *Nat Genet* 2002, 31: 69-73.

[54] Blake, W., Kaern, M., Cantor, C., Collins, J., “Noise in eukaryotic gene expression,” *Nature* 2003, 422: 633-637.

[55] Swain, P., Elowitz, M., Siggia, E., “Intrinsic and extrinsic contributions to stochasticity in gene expression,” *Proc Natl Acad Sci USA* 2002, 99: 12795-12800.

[56] Thattai, M., van Oudenaarden, A., ”Intrinsic noise in gene regulatory networks,” *Proc Natl Acad Sci USA* 2001, 98: 8614-8619.

[57] Kierzek, A., Zaim, J., Zielenkiewicz, P., “The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression,” *J Biol Chem* 2001, 276: 8165-8172.

[58] Xie, X.S., Choi, P.J., Li, G.W., Lee, N.K., Lia, G., “Single-molecule approach to molecular biology in living bacterial cells,” *Annu Rev Biophys* 2008, 37: 417-444.

[59] Paulsson, J., “Models of stochastic gene expression,” *Phys Life Rev* 2005, 2(2): 157-175.

[60] Adhya, S., Gotterman, M., “Control of Transcription Termination,” *Annu Rev Biochem* 1978, 47: 967-96.

[61] Jorgensen, F., Kurland, C.G., “Processivity errors of gene expression in Escherichia coli,” *J Mol Biol* 1990, 215: 511-521.

[62] Arkin, A., Ross, J., McAdams, H., “Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected E. coli cells,” *Genetics* 1998, 149: 1633-1648.

[63] Gaffney, E., Monk, N., “Gene expression time delays and turing pattern formation systems,” *Bull Math Biol* 2006, 68: 99-130.

[64] Ota, K., Yamada, T., Yamanishi, Y., Goto, S., Kanehisa, M., “Comprehensive Analysis of Delay in Transcriptional Regulation Using Expression Profiles,” *Genome Informatics* 2003, 14: 302–303.

[65] McClure, W.R., “Rate-limiting steps in RNA chain initiation,” *Proc Natl Acad Sci USA* 1980, 77: 5634-5638.

[66] Uptain, S., Kane, C., Chamberlin, M., “Basic mechanisms of transcript elongation and its regulation,” *Annu Rev Biochem* 1997, 66: 117-172.

[67] Kennell, D., Riezman, H., “Transcription and translation initiation frequencies of the Escherichia coli lac operon,” *J Mol Biol* 1977, 114: 1–21.

[68] Yarchuk, O., Jacques, N., Guillerez, J., Dreyfus, M., “Interdependence of translation, transcription and mRNA degradation in the *lacZ* gene,” *J Mol Biol* 1992, 226: 581–596.

[69] Phroskin, S., Rachid Rahmouni, A., Mironov, A., Nudler, E., “Cooperation between translating ribosomes and RNA polymerase in transcription elongation,” *Science* 2010, 328(5977): 504-508.

[70] Burmann, B.M., Schweimer, K., Luo, X., Wahl, M.C., Stitt, B.L., Gottesman, M.E., Rösch, P., “A NusE:NusG Complex Links Transcription and Translation,” *Science* 2010, 328(5977): 501-504.

[71] Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., Prasher, D.C., “Green fluorescent protein as a marker for gene expression,” *Science* 1994, 263: 802–805.

[72] Ludin, B., Doll, T., Meili, R., Kaech, S., Matus, A., “Application of novel vectors for GFP-tagging of proteins to study microtubule-associated proteins,” *Gene* 1996, 173: 107–111.

[73] Moss, J. B., Price, A.L., Raz, E., Driever, W., Rosenthal, N., "Green fluorescent protein marks skeletal muscle in murine cell lines and zebrafish," *Gene* 1996, 173: 89–98.

[74] Niedenthal, R.K., Riles, L., Johnston, M., Hegemann, J.H., "Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast," *Yeast* 1996, 12: 773–786.

[75] Andersen, J. B., Sternberg, C., Poulsen, L. K., Bjørn, S. P., Givskov, M., Molin, S., "New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria," *Applied and Environmental Microbiology* 1998, 64(6): 2240-2246.

[76] Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., Long, R.M., "Localization of ASH1 mRNA Particles in Living Yeast," *Mol Cell* 1998, 2: 437–445.

[77] Fusco, D., Accornero, N., Lavoie, B., Shenoy, S.M., Blanchard, J.M., Singer, R.H., Bertrand, E., "Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells," *Curr Biol* 2003, 13: 161–167.

[78] Valencia-Burton, M., McCullough, R.M., Cantor, C.R., Broude, N.E., "RNA visualization in live bacterial cells using fluorescent protein complementation," *Nature Methods* 2007, 4(5): 421-427.

[79] Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A., Tsien, R.Y., "A monomeric red fluorescent protein," *Proc Natl Acad Sci USA* 2002, 99: 7877–7882.

[80] Lutz, R., Lozinski, T., Ellinger, T., Bujard, H., "Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator," *Nuc Ac Res* 2001, 29: 3873–3881.

[81] Taniguchi, Y., Choi, P. J., Li, G., Chen, H., Babu, M., Hearn, J., "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells," *Science* 2010, 329(5991): 533-538.

[82] Balesco, J.G., "All things must pass: Contrasts and commonalities in eukaryotic and bacterial mRNA decay," *Nat Rev Mol Cell Biol* 2010, 11(7): 467-478

[83] Epshtein, V., Nudler, E., "Cooperation between RNA polymerase molecules in transcription elongation," *Science* 2003, 300(5620): 801-805.

[84] Lewin, B., *Genes IX*, 256-299, Jones and Bartlett Publishers, USA, 2008.

[85] Erie, D.A., Hajiseyedjavadi, O., Young, M.C., von Hippel, P.H., “Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription,” *Science* 1993, 262: 867-873.

[86] Greive, S.J., Weitzel, S.E., Goodarzi, J.P., Main, L.J., Pasman, Z., von Hippel, P.H., “Monitoring RNA transcription in real time by using surface plasmon resonance,” *Proc Natl Acad Sci USA* 2008, 105: 3315-3320.

[87] Moore, S.D., Sauer, R.T., “Ribosome rescue: tmRNA tagging activity and capacity in Escherichia coli,” *Mol Microbiol* 2005, 58: 456-466.

[88] Cormack, B.P., Valdivia, R.H., Falkow, S., “FACS-optimized mutants of the green fluorescent protein (GFP),” *Gene* 1996, 173(1): 33-38.

[89] Bracewell, R., *Pentagram Notation for Cross Correlation. The Fourier Transform and Its Applications.* New York: McGraw-Hill, 46-243, 1965.

[90] Peabody, D.S., Lim, F., “Complementation of rna binding site mutations in ms2 coat protein heterodimers,” *Nucleic acids research* 1996, 24(12): 2352–2359.

[91] Wang, Q., Niemi, J., Tan, C., You, L., West, M., “Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy,” *Cytometry Part A* 2010, 77(1): 101–110.

[92] Chen, J. and Gupta, A. *Parametric Statistical Change-point Analysis*. Birkhäuser, Boston, 2000.

[93] Devroye, L. and Györfi, L. *A Probabilistic Theory of Pattern Recognition*, 1st edition, Springer-Verlag, New York, 1996.

[94] Otsu, N., “Thershold Selection Method from Gray-level Histograms,” *IEEE Trans Syst Man Cybern* 1979, SMC-9(1):62-66.

[95] Shaevitz, J.W., Abbondanzieri, E.A., Landick, R., Block, S.M., “Backtracking by single RNA polymerase molecules observed at near-base-pair resolution,” *Nature* 2003, 426: 684–687.

[96] Ribeiro, A.S., Häkkinen, A., Healy, S., Yli-Harja, O., “Dynamical effects of transcriptional pause-prone sites,” *Comput Biol Chem* 2010, 34(3): 143-148.

[97] Zhu, R., Salahub, D., “Delay stochastic simulation of single-gene expression reveals a detailed relationship between protein noise and mean abundance,” *FEBS Letters* 2008, 582:2905–2910.

[98] Paulsson, J., “Summing up the noise in gene networks,” *Nature* 2004, 29: 415–418.

[99] Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., Elowitz, M.B., “Gene regulation at the single-cell level,” *Science* 2005, 307: 1962–1965.

[100] Choi, P.J., Cai, L., Frieda, K., Xie, X.S., “A Stochastic Single-Molecule Event Triggers Phenotype Switching of a Bacterial Cell,” *Science* 2008, 322(5900): 442-446.

[101] Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., Barkai, N., “Noise in protein expression scales with natural protein abundance,” *Nature Genetics* 2006, 38: 636-643.