TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MOHAMED NASURUDEEN MOHAMED BAHRUDEEN

EXTRINSIC NOISE EFFECTS REGULATION AT THE SINGLE GENE AND SMALL GENE NETWORK LEVELS

Master of Science Thesis

Examiner: Professor Andre Ribeiro

Examiner and topic approved by the Faculty Council of Computing and Electrical Engineering on 09.08.2017

ABSTRACT

MOHAMED NASURUDEEN MOHAMED BAHRUDEEN: Extrinsic noise effects regulation at the single gene and small gene network levels
Tampere University of Technology
Master of Science Thesis, 60 pages
October 2017
Master's Degree Programme in Electrical Engineering
Major: Biomedical Engineering
Examiner: Professor Andre Ribeiro

Keywords: transcription initiation kinetics, gene regulatory networks, stochastic simulation algorithm, extrinsic noise, cell-to-cell variability.

Recent studies of gene expression in *Escherichia coli* using novel *in vivo* measurement techniques revealed that protein and RNA numbers from a gene differ between genetically identical cells. To unravel the causes for this, measurements were conducted and models were developed. These studies revealed that this diversity arises from extrinsic and intrinsic noise. The former is due to cell-to-cell variability in numbers of molecules involved, such as RNA polymerase (RNAp), transcription factors, etc. The latter is due to the stochastic nature of the chemical reactions combined with the fact that the molecules and genes involved exist in small numbers.

One aspect that has not been given much attention so far, is the unique nature of the dynamics of transcription of each promoter of the gene regulatory network (GRN). This process has multiple rate-limiting steps whose duration differs between promoters. How this may diversify the variability in RNA and protein numbers between genes is unknown.

To address this, we use single-cell empirical data and stochastic models with empirically validated parameter values and study how the kinetics of transcription of a gene affects the influence of extrinsic noise on the kinetics. Interestingly, we find that promoters whose open complex formation is longer lasting tend to suppress the propagation of extrinsic noise that affects only the steps prior to initiation of the open complex formation.

In particular, our studies indicate that the cell-to-cell variability in RNA numbers depends on the transcription kinetics. As such, it is sequence-dependent. Further, in a 2-gene toggle switch, we find that its mean switching frequency depends on the transcription kinetics of the promoters but not on the cell-to-cell RNAp variability. On the other hand, the cell-to-cell variability in switching frequency is affected by these two variables. Meanwhile, in a Repressilator network (3 genes where each gene represses the next), we measured the mean and standard deviation of the period of oscillation. From these measurements *in silico*, we found that both parameters are independent of the RNAP cell-to-cell variability, but are strongly controlled by the transcription kinetics of each of its genes.

We conclude that the transcription kinetics of the component genes is a key regulator of small genetic circuits, as it can be used as a tunable filter of extrinsic noise. Overall, the kinetics of the rate-limiting steps in transcription of individual genes act as 'master regulators' of the expression of individual genes and the behavior of genetic circuits', such as switching dynamics, period of oscillation, etc.

PREFACE

This Master's thesis was carried out at the Laboratory of Biosystem Dynamics, a research group of BioMediTech, Tampere University of Technology.

First, I would like to thank my supervisor, Professor Andre Ribeiro, for giving me the opportunity to work on this topic. I am very thankful for guiding and showing me different dimensions of approaching the problems. He taught me various concepts and techniques, which eased me to understand and simulate different models.

My sincere thanks to Samuel Oliveira, a good friend and colleague, who has taught me various image analysis tools and methods, which are part of a fundamental, useful knowledge essential to perform the work required to complete this thesis. Also, I would like to thank Sofia Startceva, for assisting me in developing models and fixing its errors.

My warm thanks to all my colleagues in the Laboratory of Biosystem Dynamics for the help and guidance given to me to complete this work. They are very helpful and friendly, while clarifying my research and technical questions. Also, during the time I have worked with them, they have created and maintained an excellent working atmosphere, which has contributed substantially to the completion of my work.

Finally, I would like to thank my parents for the motivation and encouragement that they have given me, during difficult times living abroad. Without their encouraging words, this work would never have been finished.

Tampere, 9.10.2017

Mohamed Nasurudeen Mohamed Bahrudeen

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CME	Chemical Master Equation
CV^2	Squared Coefficient of Variation
DNA	Deoxyribonucleic Acid
E. coli	*Escherichia coli*
GFP	Green Fluorescence Protein
GRN	Gene Regulatory Network
In vivo	Latin word which means "within the living"
In vitro	Latin word which means "Within the glass"
In silico	Expression used to mean "perfumed via computer simulation"
In situ	Latin word which means "in its original position"
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KDE	Kernel Density Estimation
mRNA	messenger RNA
MS2	Bacteriophage MS2 viral coat protein
ODE	Ordinary Differential Equations
RBS	Ribosome Binding Site
RNA	Ribonucleic Acid
RNAp	RNA polymerase
SGNS2	Stochastic Gene Network Simulator v.2
SSA	Stochastic Simulation Algorithm
TSS	Transcription Start Site
YFP	Yellow Fluorescence Protein

P_{CC}	Promoter in closed complex
P_{OC}	Promoter in open complex
P_{ON}	Promoter in active state
P_{rep}	Promoter in repressed state
Rep	Repressor
Rib	Ribosome
Rp	RNAp numbers per cell

1. INTRODUCTION

Escherichia coli undergoes behavioral changes by tuning the quantities of its regulatory molecules, such as transcription and σ factors, etc. This tuning process requires changes in the kinetics of transcription of its genes and, in some cases, their translation kinetics. This is made possible by changing the numbers of molecules such as RNA polymerase (RNAp) core enzymes, gene-specific activator and repressor molecules, σ factors and ribosomes, among others [1] [2].

For example, in the case of σ factors, since the amount of RNAp core enzymes is limited [3], increasing the numbers of a specific σ factor causes an increase in the number of RNAp molecules carrying that σ factor, while decreasing the number of RNAp molecules carrying other σ factors. Consequently, the activity of the promoters associated with that σ factor will increase (direct positive regulation), whereas the activity of the promoters associated with other σ factors is reduced (indirect negative regulation) [1] [2].

Interestingly, it has been observed that changes in σ factors concentrations do not affect the activity of some genes [3]. Further, those genes that do respond to changes in σ factors numbers, do so in a heterogeneous way, i.e., differ in the degree of change. This heterogeneity in responses is found to occur even between genes associated with the same σ factor.

This diversity in behavioral responses is due to diversity in promoters' selectivity of the σ factors [4], and the influence of transcription factors [3], which were first noticed using *in vitro* measurement techniques (for a review see [5]). Another cause for this diversity of responses, recently acknowledged, are the differences in the dynamics of the rate limiting steps in transcription initiation of the various promoters [6] [7].

Specifically, promoters preferentially transcribed by σ^{70} show lesser responsiveness to changes in σ^{38} as their closed complex formation time-length is increasingly shorter than the open complex formation time-length. This is due to the fact that the concentration of σ^{38} affects the kinetics of the closed complex formation but not the kinetics of the open complex formation.

Based on this hypothesis, experimentally validated by tests in several promoters and when employing different measurement techniques, Kandavalli and colleagues concluded that, in *E. coli*, the responsiveness of promoters to indirect regulation by σ factors' competition is determined by the kinetics of their rate-limiting steps in transcription initiation [7].

Given that σ factors' competition affects mean transcript production rates, it is reasonable to assume that they may affect also the noise levels in transcription. Similarly, if the mean number of RNAp's per cell in a population affects the mean transcript production rates of those cells, then the degree of cell-to-cell variability in RNAp numbers should also affect the cell-to-cell variability in transcription rates.

Based on the above, here we investigate the hypothesis that the effects of extrinsic noise sources on the cell-to-cell variability in RNA and protein numbers of a gene are influenced by the dynamics of the rate-limiting steps in transcription initiation of that gene.

To investigate this hypothesis, we start by creating a stochastic model of transcription with multiple rate limiting steps, based on the modelling strategy first proposed in (Ribeiro et al, 2006). By providing each cell with its own number of RNAp's, this strategy also takes cell-to-cell variability in RNAp numbers into account. Currently, this variability can be measured using state-of-the-art single-cell microscopy, combined with image and data analysis tools to extract the information from the images.

Meanwhile, the stochastic simulations of model cells were done using the software SGNS2 (Stochastic Gene Network Simulator v.2) [8], which operates in accordance with the Stochastic Simulation Algorithm [9]. To generate cell to cell variability in RNAp numbers, for each cell, RNAp numbers are drawn randomly from a normal distribution and then remain constant over the simulation time of the cell gene expression dynamics.

Using this framework, by changing the values of certain parameters of the model of gene expression, within realistic intervals, we studied the extent to which cell-to-cell variability in RNAp affects the cell-to-cell variability in RNA numbers as a function of the transcription initiation kinetics of genes [10].

Furthermore, we extend our studies to small genetics circuits, particularly, genetic switches, whose switching behavior is generated by stochastically-driven changes in the RNA numbers over time [11] [12] [13]. In this regard, we hypothesized that the effects of extrinsic noise sources on a circuit's behavior is affected by the kinetics of the rate-limiting steps in transcription initiation of the genes composing the circuit. In particular, we study the extent to which the responsiveness of a genetic toggle switch and of a repressilator are affected by various degrees of extrinsic noise sources (i.e. degree of cell-to-cell variability in RNAp numbers), as a function of transcription initiation kinetics of the genes.

To assess this, following the same approach described above (for the study of individual genes), we create the stochastic models of a genetic toggle switch and of a repressilator, each having component genes whose transcription dynamics has multiple rate limiting steps. Further, at the cell population level, we account for the cell-to-cell diversity in RNAp numbers. As previously, the cell-to-cell variability in RNAp numbers in a cell

population, and the required model parameters, are measured using state-of-the-art measurement, image and data analysis methods. Then, using the empirically validated parameter values, we performed several stochastic simulations of model cells, each with a number of RNAp's drawn randomly from a normal distribution and kept constant throughout the simulation time. Finally, to assess the influence of cell-to-cell variability in RNAp numbers on the behavior of the switch (switching frequency) and of the repressilator (period of oscillations) as a function of the promoters initiation kinetics of the component genes, we performed simulations for various values of the rate constants of the model controlling the transcription initiation kinetics [14].

This thesis work was carried out at the Laboratory of Biosystem Dynamics (LBD), led by Professor Andre S. Ribeiro, from the BioMediTech Institute (BMT) of Tampere University of Technology (TUT). The results of this work were published in two international conferences, namely, the 9th International Conference on Bioinformatics and Biomedical Technology [10], and the European Conference on Artificial Life [14]. In addition, continuation of these studies, consisting of a study of the multi-scale effects of extrinsic noise (i.e. on the activity of a gene, of small and of large gene networks), as a function of the kinetics of transcription initiation of the component genes, has been accepted for oral presentation and for publication in another international conference, the 12th Workshop on Artificial Life and Evolutionary Computation, with me as co-author [15].

Following introduction (Chapter 1), Chapter 2 provides a summary of the present knowledge on the structure of DNA and RNA, on the dynamics of gene expression and gene regulatory networks, and on the sources of intrinsic and extrinsic noise in gene expression. In addition, several open questions on the observed cell-to-cell phenotypic diversity at the single gene, single cell levels are presented. Next, Chapter 3 presents a description of the most recent live cell microscopy measurement techniques, such as techniques on fluorescent probing of proteins for *in vivo* detection of individual RNA molecules in live cells, signal processing methods for image analysis and data extraction, and a detailed description on stochastic modelling techniques of single genes and gene regulatory network models. Chapter 4 presents the results of *in silico* studies of the dynamics of a single gene and small regulatory networks. Finally, Chapter 5 includes a discussion and main conclusions that can be drawn from the results.

2. BACKGROUND

2.1 Biological background

A brief overview of biological concepts associated with this thesis is provided in this chapter. First, we provide information about the DNA structure and about gene expression dynamics in prokaryotes. Finally, we describe noise sources in gene activity.

2.1.1 Structure of the DNA and RNA

DNA, Deoxyribonucleic Acid, consists of two covalently linked two-polynucleotide chains or strands, each composed of nucleotide subunits. Each of these nucleotides is made up of a sugar phosphate group and a nitrogen base. There are four types of nitrogen bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Adenine binds to thymine and cytosine binds to guanine. The order of arrangement of these nitrogen bases in the DNA strand determines the 'genetic code' (Figure 1). This code has most (if not all) of the information necessary to create the complete organism. Every living organism has a DNA sequence, except for viruses, which instead of DNA, carry their genetic code in an RNA (Ribonucleic acid) molecule.

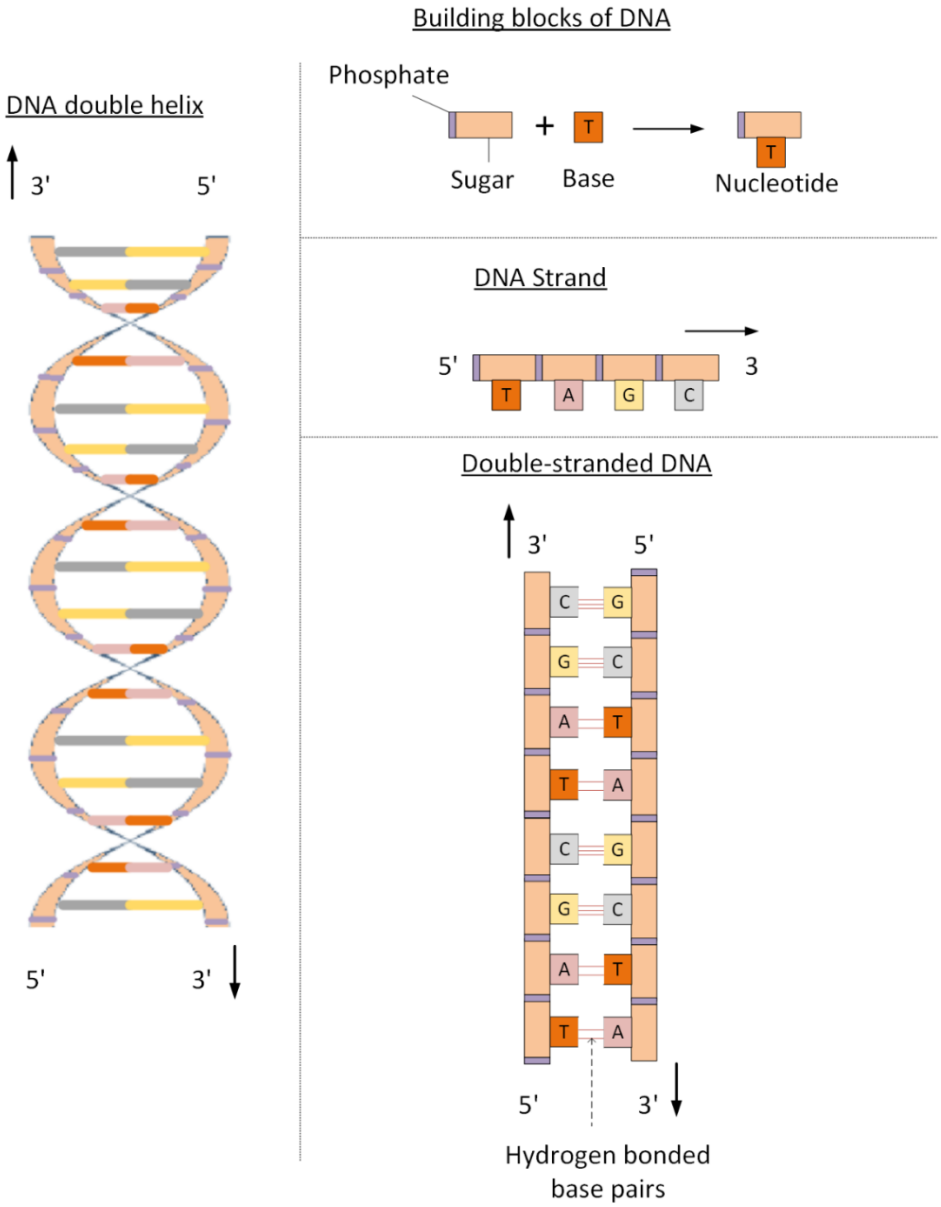

Figure 1. *Double strand DNA and its building blocks. The DNA strand is made up of 4 different nucleotide bases, adenine (A), thymine (T), guanine (G) and cytosine (C), which are covalently linked with sugar-phosphate to form a polynucleotide chain. Each DNA molecule has 2 chemical polarities; that is, its two ends are chemically different. The 3' end carried an unlinked –OH group attached to the 3' position on the sugar ring, while the 5' end carries a free phosphate group attached to the 5' position on the sugar ring.*

RNA is the covalently linked single polynucleotide chain or strand. Like the DNA, the nucleotides composing the RNA are also made up of sugar phosphates and 4 different nitrogen bases. These are Adenine (A), Guanine (G), Cytosine (C), and differently from DNA, Uracil (U) instead of Thymine. The primary function of RNA is to code for protein synthesis, which carry out specific functions in the cell (Figure 2).

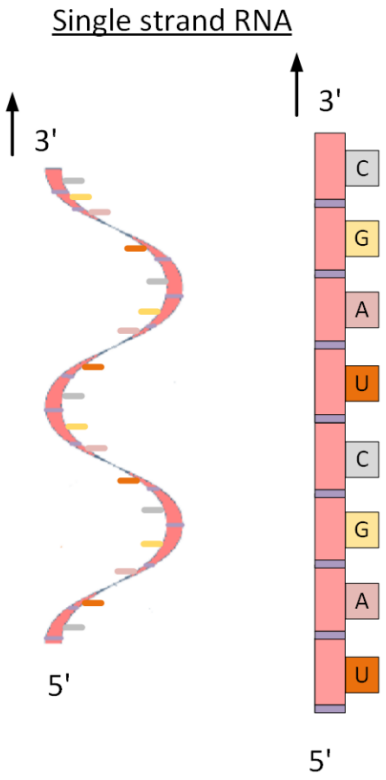

Figure 2. *Single strand RNA and its building blocks. The RNA strand is made up of 4 different nucleotide bases, adenine (A), uracil (U), guanine (G) and cytosine (C), which can covalently link with sugar-phosphate and form a polynucleotide chain.*

2.1.2 Gene expression in prokaryotes

Genes are hereditary units [16]. They consist of segments of DNA, coding for the necessary information to produce proteins, the functional components of cells. The process through which cells propagate the information from genes in DNA strands into functional proteins is named as 'gene expression'. It is carried out in two sequential steps, transcription and translation, which together constitute the central dogma of molecular biology (Figure 3).

The first step in gene expression is transcription. In this, the information of a gene is transcribed by an RNAp enzyme complex into a single stranded RNA molecule, which codes for proteins, which are produced by the translation process (see below). The RNA polymerase holoenzyme is a combination of RNA polymerase core enzyme and a DNA binding protein, named 'σ factor', which can bind to specific nucleotides of the promoter regions named Transcription Start Site (TSS). These promoter regions are specific nucleotide sequences in the DNA strand, which can regulate the expression of a gene, or a group of genes.

In transcription, the RNA polymerase holoenzyme attaches itself to a DNA molecule, slides along the nucleotides (through nonspecifical binding) until it locates itself at the

promoter region of the gene, where it specifically binds to, leading to the unwinding of the two strands of the DNA. After this, the nucleotides of the genes become ‘open’ for transcription. The complex process of transcription initiation is considered to be the most important regulatory step of gene expression in prokaryotes, as it undergoes a series of time-demanding conformational changes that do not occur in subsequent steps [17].

After the transcription of the first 10 nucleotides, the polymerase is out of the promoter region and can move along the DNA towards the end of the DNA coding sequence of the gene, in a process named transcription elongation. When at the elongation mode, the RNAp forms an RNA strand, from free floating nucleotides, which contains the same genetic information (in terms of nucleotides sequence) as the DNA strand. The elongation mode continues until the RNA polymerase reaches the termination site in the DNA, after which it is released. The transcribed RNA then conforms into a three-dimensional structure, by folding.

There are two main reasons why, in prokaryotes, transcription initiation is considered to be the main regulatory step in gene [17]. First, subsequent steps, such as elongation and termination, are much faster and less ‘stochastic’ than transcription initiation. Also, mRNA translation occurs while the mRNA is being transcribed and has no significant rate-limiting steps [17] [18] [19] [20] [21]. The relatively slow nature of transcription initiation and its significance in regulation of RNA and protein production dynamics are due to its multi-stepped nature [17] [22].

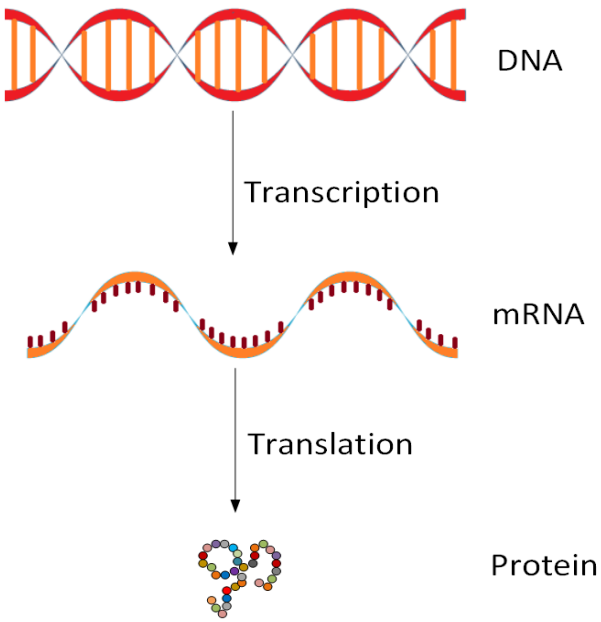

Figure 3. *The Central Dogma of Molecular Biology. The image describes, from top to bottom, the sequence of steps in gene expression, in which transcription is the process through which DNA produces RNA and translation is the process through which the RNA produces polypeptide and protein structures.*

In translation, the information coded in the mRNA is used to produce specific amino-acids by a triplet-wise degenerated universal code (codon) of nucleotides with the help of a complex molecular structure named Ribosome [23]. Thus, gene expression is not spontaneous, rather, it depends on the availability of molecules such as RNAp and Ribosomes, which causes fluctuations in proteins levels over time.

2.1.3 Gene regulatory networks

Gene regulatory networks (GRN) are groups of genes that form a network of interactions (based on proteins) that are capable to perform complex functions. The topology of the network is determined by the regulatory links between the genes of the network. In bacteria, small sets of genes collectively perform a biological function. These are usually clustered into operons [24] [25].

In natural GRNs, such sets of genes, whose activities are directly linked, are called motifs [26] . These motifs perform complex actions, sometimes in response to internal and external stimuli, such as switching between possible states or keeping track of time. Several such natural motifs have been studied recently [27] [28] [29].

Genes can interact in various ways. For example, there are 'positive' interactions, where the expression of gene (A) activates the expression of gene (B), and 'negative' interactions, where the expression of gene (A) reduces the expression of gene (B). In general, these gene networks can be represented in simple forms, to assist the understanding of their behavior. For instance, the schematic representation of a 2-gene toggle switch network and of a repressilator network are shown in Figure 4 and Figure 5 respectively.

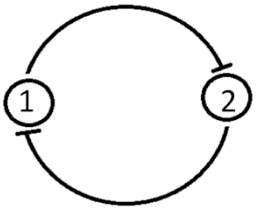

Figure 4. *Schematic representation of a genetic toggle switch, where gene "1" represses gene "2" and vice versa. In this network, gene "1" activity represses gene "2", keeping gene "1" in a 'dominant' position, and vice-versa.*

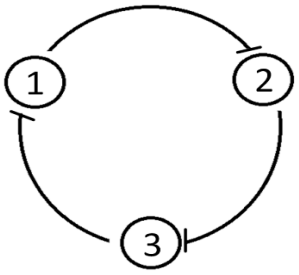

Figure 5. *Schematic representation of a 3-gene Repressilator, where gene "1" represses gene "2", gene "2" represses gene "3" and gene "3" represses gene "1". In a closed system of 3 genes displayed in a loop, where each gene represses the next, it is expected that the activity of each of the genes will oscillate regularly.*

In gene regulatory networks, 'dominant' gene refers to a gene whose activity suppresses the activity of others, as it exhibits higher protein expression levels than a 'recessive' gene, which will have lower protein expression levels.

2.1.4 Intrinsic and extrinsic noise in gene expression

There is a significant variability in cellular phenotype, even among populations of genetically identical cells in the same environment [30] [31] [32] [33].

This diversity is due to, first, the stochastic nature of gene expression and the small number of molecules involved within the same cell (intrinsic noise). Also, cells differ in number of components, which cause differences in the rates of the processes of transcription and translation (extrinsic noise) [34] [31].

Interestingly, some genetic circuits can suppress the effects of fluctuations in molecules species for robust functioning, while other genetic circuits can amplify this noise to increase the cell-to-cell heterogeneity [35] [36].

The level of noise in gene expression also differs between various *E. coli* strains [31], which implies that gene expression is regulatable or the level of extrinsic noise is different.

Noise can be either beneficial or detrimental. Since stochasticity in gene expression causes phenotypic differentiation [33], it might allow at least some cells to be better fit to some environmental fluctuations [37] [38], which is beneficial. Meanwhile, these fluctuations also imply that some cells might not make the proper decision, which is detrimental.

Many questions remain open about noise regulation, both intrinsic and extrinsic. Only some of these questions are being addressed now. Answers to these questions (such as, are there mechanisms of their regulation and, if so, how do they operate) will provide

much better understanding of the phenotypic diversity observed in cell populations, ranging from bacteria to cancer cells.

2.2 Open questions on the observed cell-to-cell phenotypic variability levels at the single gene level

The main questions on the observed levels of phenotypic variability in RNA and protein numbers are: why do these levels differ between genes if the sources of variability are identical for all genes? Also, why do these levels of phenotypic variability in RNA and protein numbers of each gene change by different degrees when changes occur, e.g., in the numbers of master regulator molecules such as RNAp, ribosomes and σ factors.

As mentioned in the introduction, we explore the possibility that the answer to these questions lies in the fact that, in general, the effects of cell-to-cell variability in the numbers of some molecule affecting transcription rates depends on the kinetics of the rate-limiting steps in transcription initiation and on which step that molecule affects.

3. MATERIALS AND METHODS

3.1 Microscopy

The application of state-of-art microscopy techniques has facilitated significantly the understanding of the complex behaviors of various cellular mechanisms. Here, we use confocal and phase contrast microscopy to study the *in vivo* dynamics of transcription. More specifically, we use these to quantify RNAp and RNA molecules inside the cells. A brief explanation about these microscopy techniques and their application in this thesis work is provided in the following chapters.

3.1.1 Confocal time-lapse microscopy

Confocal microscopy is a fluorescence microscopy technique. The term 'confocal' is defined as 'having the same focus' and this microscope creates a final image from a same point of focus. In short, first, the specimen is excited with laser beam at a particular wavelength, which is chosen depending on the fluorophores present in the specimen. The fluorophores emit light, whose wavelength is different from that of the excitation light beam. The thickness of the specimen causes the light to be emitted also from outer regions. To get rid of this out of focus signal there is a pin hole arrangement in front of the image plane, which filters the out of focus signal. After this filtering, the resulting signal is smaller in amplitude, which is then amplified by a photomultiplier tube whose gain is customizable. As these imaging pixels are created point by point, which requires point-to-point excitation, a complete image is formed.

One significant feature of this microscopy technique is its efficient rejection of out of focus fluorescent light, which reduces the degradation of image quality due to out of focus light signals.

Here, we study *E. coli* strains grown over agar gel. This causes emission of out-of-focus fluorescent light. Due to this, we use confocal microscopy to image these *E. coli* cells.

In this project, confocal microscopy is used to capture time-lapse images of *E. coli* cells to study their RNA production dynamics, since it has much better resolution in comparison with other conventional wide field microscopy techniques. In this method, the laser light source is restricted to the volume of observation, so that the out of focus fluorescence signal is ignored from the detected signal. Another main advantage of this method is the enhanced contrast, especially when specimens are thick. Meanwhile, it has the disadvantage of longer imaging time, due to its point-to-point excitation and scanning, and thus cannot be used to image weak signals that degrade rapidly.

The confocal microscopy images that give fluorescence information of the cells are taken every 1 minute so as to provide accurate information on the kinetics of RNA production. However, these images do not provide information on the cells' morphology. To segment the cells, i.e. to define cell boundaries, we make use of phase contrast images.

The RNA molecules, produced by the cells, since they are tagged with MS2-GFP (see section 3.2.2), can be detected through the green fluorescent channel of the microscope (see example images in Figure 6). Meanwhile, another fluorescent protein, mCherry, also used in our studies, are detected through the red fluorescent channel.

The RNA molecules can be seen as fluorescent spots, and move around in the cytoplasm, tending to aggregate in the cell's poles, due to a nucleoid-exclusion phenomenon [39].

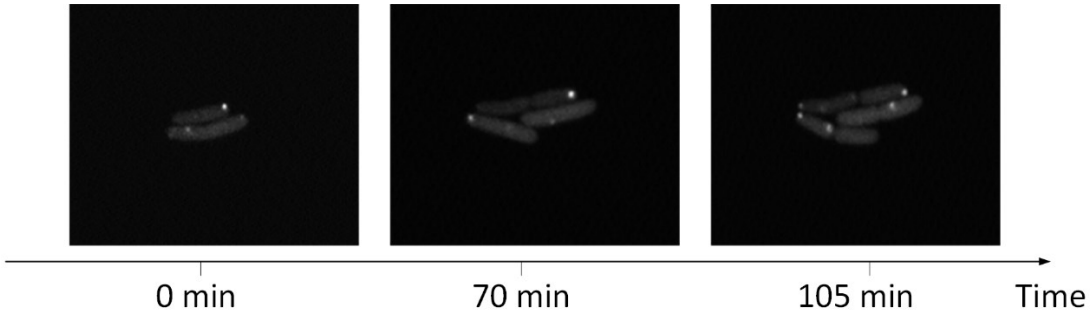

Figure 6. *Time-lapse confocal image examples of E. coli cells expressing MS2-GFP and GFP-tagged target mRNA molecules. Here, fluorescent images were taken once every 1 minute for 180 minutes. Tagged RNAs are visible as bright spots.*

3.1.2 Phase-contrast microscopy

Phase contrast microscopy is a technique used to obtain high contrast microscopy images from transparent samples by converting light phase differences into light amplitude differences. The phase difference is generated by differences in optical path length, which depends upon the refractive index and the thickness of the sample. Different cellular components in the sample have different refractive index, causing the phase of light to change over different regions of the sample, which provides contrast information. Interestingly, even small differences in refractive index between cellular structures, result in large differences in the phase contrast channel.

During the time that light rays are crossing the cells in the sample, they travel relatively slower than those that do not cross the cells. This reduction of the speed of light will cause phase difference of nearly -90˚ with the rays of light crossing only the background. This leads to defocusing and does not give more detailed image. Meanwhile, in phase contrast images, the light incident on the background also phase shift due to crossing a phase-shift ring. Namely, using the positive phase contrasting technique, the phase shift ring shifts the un-diffracted background light by +90˚ causing destructive interference when the

background light and diffracted light rays meet. As a result, cells become darker than the background. In our lab, the positive phase contrasting technique is being used for phase contrast imaging.

In this project, both confocal and phase-contrast microscopy technologies are used simultaneously, in order to capture, respectively, the level of fluorescence in the cells (Figure 6), which is used to measure gene expression, and the cell boundaries, which are obtained from cell segmentation (Figure 7). In general, here, the fluorescence images are taken every minute, while the phase contrast microscopy images are taken every 5 minutes (to reduce the effects of photo toxicity). This is made possible by the fact that the cells in agarose gel move slowly.

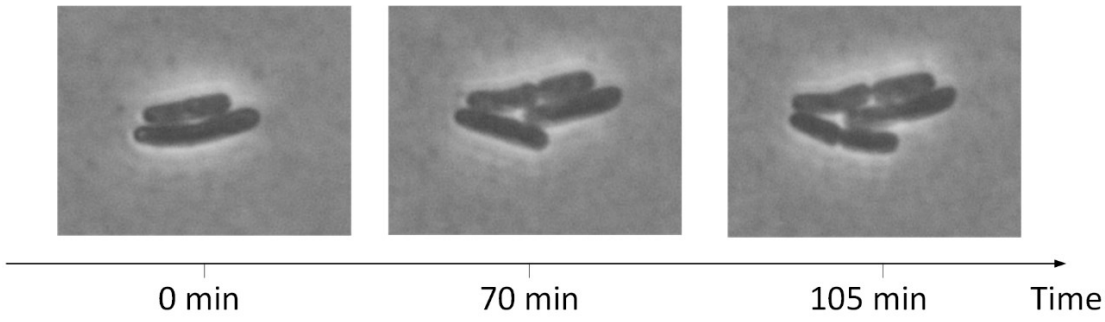

Figure 7. *Example time-lapse of phase contrast images of E. coli cells. In time series measurements, these images are usually taken simultaneously with fluorescent time-lapse images (see example images in Figure 6). Then, the two channels are merged, to allow observing where the fluorescent spots locate (i.e. in which cells).*

3.2 *In vivo* detection of individual RNA molecules in live cells

Researchers have long been using various techniques to study the mechanisms of transcription. These techniques include X-ray crystallography [40], FRET [41], foot printing based on gel electrophoresis [42] and, e.g., fluorescence *in situ* hybridization [43], which could only provide a static picture of a dynamic process. To best understand the dynamics of transcription, *in vivo* single RNA-molecule studies are required, as these studies provide a more complete picture of the kinetics of transcriptional dynamics. For example, by observing only how many RNAs exist in each cell at a given moment in time following induction of a gene, it is not possible to determine when these molecules were produced. E.g. they could have all being produced at the end of the observation time, as well as at the beginning, and still result in the same total number of RNA molecules at the end of the observation time. As such, only observing when each molecule was produced, it is possible to produce models of the kinetics of their production.

3.2.1 Fluorescent proteins

In 1962, while conducting a study in the jellyfish *Aequorea*, Osamu Shimumura and colleagues discovered the presence of a luminescent substance, *aequorin.* This substance was then found to be a fluorescent protein, which has the potential to store high amounts of energy, which can then be released in the presence of calcium. This results in the emission of a bright blue light. Given this unique feature, this protein was first used as a calcium probe. Next, in the process of purifying this fluorescent protein, *aequorin*, another protein with bright green fluorescence was also extracted and named as Green Fluorescence Protein (GFP) [44].

The significance of GFP was realized later on, namely, once it was found that it could be used as a fluorescent marker for gene expression. With the parallel development of protein engineering methods, since then, several fluorescent proteins have been developed, covering almost the entire visible spectrum of light. [45]. As a result, nowadays, fluorescent probing is widely used to detect and quantify proteins by using *in vivo* live cell imaging.

For fluorescent probing to be an effective method, the binding of fluorescent proteins with target molecules should not affect their normal functioning. Despite fluorescent probing being a powerful tool to perceive the dynamics at a spatial and temporal level, there is still scope for improvements regarding, e.g., maturation time, photo bleaching and blinking. For example, shorter maturation time enables the detection of targeting molecules sooner, following their production. Further, the detection of targets is more reliable if the fluctuations in fluorescence intensities and photo bleaching of the molecules can be reduced.

Finally, for precise detection of fluorescent proteins, the light emitted from those fluorescent proteins should be of higher intensity than the background's auto-fluorescence. As such, the fluorescent proteins to be used should be selected based upon prior knowledge of the system (i.e. its autofluorescence levels, etc.), in order to avoid, e.g., having the same excitation wavelength as the elements of the background responsible for its autofluorescence.

In our study, we imaged RNA molecules containing sequences to which the MS2 viral protein can bind to. Namely, each RNA contains 48 tandem repeating binding sites for MS2. In addition, the cells contain a plasmid capable of expressing MS2 fused with a GFP protein. Combining these two systems, 48 MS2-GFP fusion proteins can bind to the binding sites carried in each RNA produced coding for the 48 binding sites. As a result, the RNA-MS2-GFP molecules, emit a fluorescent signal that is much brighter than the background fluorescence, allowing a clear discrimination of these RNA molecules from the image.

3.2.2 MS2-GFP tagging method

To study the dynamic nature of transcription, we need methods to track over time, the process of gene expression in individual cells. Since the finding and acknowledgment of the significance of fluorescent proteins as potential sensors of this process, there has been many developments in the methods to image biological processes *in vivo*.

The first method to detect RNA molecules in real-time *in vivo* was developed by Singer and associates in 1998. They developed a novel approach to visualize mRNA molecules in eukaryotic cells [46]. Later, this method was used to visualize the production of individual mRNA molecules in *E. coli* for several hours [18]. Ever since, the usage of this technique to explore the *in vivo* dynamics of processes at the single cell, single molecule level has been increasing. One of the reasons for this is that it has made possible the quantification of RNA molecules from the fluorescent intensities in live cells over time.

The empirical data used in this thesis, was obtained by using a two-plasmid system, as in [47] [18], which allows to quantify the RNAs produced in the cells over time. The two plasmids are a 'reporter plasmid' and a 'target plasmid'. The reporter plasmid codes for a GFP sequence fused with a tandem dimer of RNA bacteriophage MS2 coat protein. Its production is regulated by the promoter P_{BAD}. Meanwhile, the target plasmid codes for mRNA containing 48 tandem repeats of MS2-binding sites that is under the control of the $P_{lacO3O1}$ promoter. Each binding site consists of a stem loop structure of viral RNA with 19 nucleotides. The schematic description of the two plasmid system of single RNA detection used in this study is shown in Figure 8.

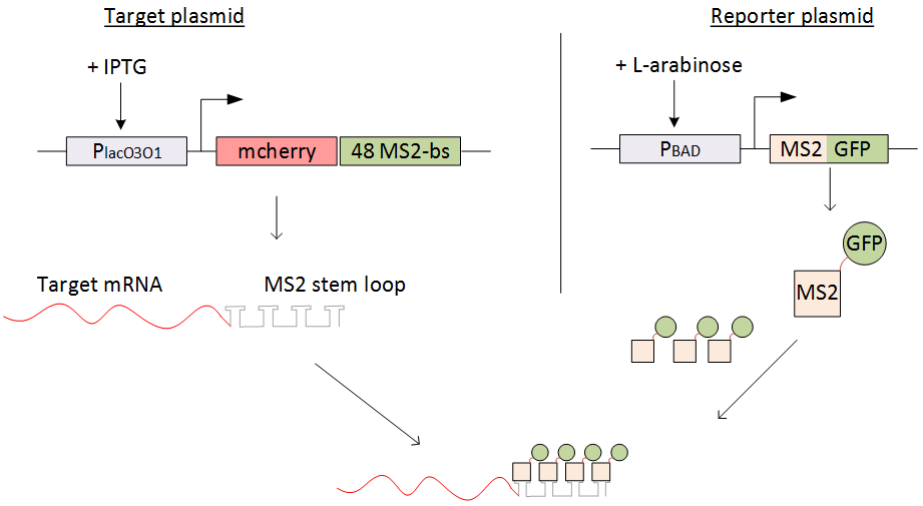

Figure 8. *Schematics of the genetic components of the mRNA detection system. On the left, controlled by the $P_{lacO3O1}$ promoter (whose activity is regulated by the inducer IPTG) is the target RNA, constructed on a single-copy F-plasmid. It consists of a coding region for mCherry, red fluorescent protein, followed by an array of 48 MS2-binding sites. On the right is the reporter system, constructed on a medium-copy vector, which codes for MS2-GFP tagging proteins, whose production is controlled by P_{BAD} (inducible by L-arabinose).*

Induction of the promoter in the reporter plasmid results in the expression of multiple copies of MS2-GFP proteins into the cytoplasm, making the cells greenish. As soon as the target RNA is produced, the reporter proteins, MS2-GFP, binds to its binding sites in RNA, creating a very bright green spot, which is clearly discriminated from the background green fluorescence.

One of the interesting property of the viral MS2 coat protein is that it has a very long lifetime. Further, when RNA molecules are bound with this MS2-GFP, they also become highly robust to degradation (as that is the natural purpose of MS2), due to which the RNA does not degrade over the course of measurement period. This allows not considering the possibility of RNA degradation, which facilitates quantifying more precisely how many RNAs exist in the cell over time from how many appeared [47].

3.3 Image analysis and data extraction

After obtaining time-lapse microscopy images, image analysis methods are employed to extract, e.g., RNAp and RNA numbers over time, mean RNA intervals and other variables of interest to the study of gene expression dynamics.

For this, in our studies, cells are first segmented by a semiautomatic method: first, cells in phase contrast images are segmented by an automated method [48] followed by manual correction. The automated method also measures the dimensions of cells and their orientation. Second, the phase contrast images are aligned on top of the confocal images. Third, the segmentation of RNA spots in the cells is done by automatic kernel density estimation(KDE) [48]. Next, the total fluorescent intensity inside the spots of the cells are calculated, followed by background subtraction.

From the corrected spot intensity (total spot intensity minus background intensity) in each cell, we determine the number of RNA produced by the cell. As the RNA tagged with MS2-GFP is ‘immortal’, the spot intensity should increase monotonically with time. The increase in spot intensities should thus corresponds to the production of more target RNAs by the occurrence of new transcriptional events. From the consecutive transcription events, by measuring the time between two consecutive productions, we get the distribution of RNA production intervals. From the time interval distribution, rate limiting steps in transcription initiation, their number of occurrences and their respective durations can be inferred. The steps involved in this process is explained in the following subchapters.

3.3.1 Cells and spots segmentation

After acquiring the fluorescent and confocal images, we performed image alignment using cross-correlation method. This process is required to remove the movement of cells in image frames over time, as they cause difficulties in cell tracking over time.

Then, we segment the cells using image analysis techniques. For cell segmentation, phase contrast images are preferred over fluorescent images because the morphology of cells are more clearly visible in phase contrast images. To detect cell boundaries and to segment them, we make use of a semi-automated tool that performs cell segmentation and cell tracking [48]. The algorithm works by, first, identifying the cell region, followed by creating a mask over the cells. The automatically generated masks have small errors, which are corrected manually from visual inspection. From the segmented cell masks, cell location, orientation and its morphological features such as shape and dimensions are obtained using principal component analysis (PCA). Those cells which cross border of the image are ignored from masking.

The cell segmentation process is followed by alignment of phase contrast images over fluorescent images. This alignment is done with a semi-automated tool, which aligns the phase contrast images over fluorescent images. The automatic alignment is not perfect and it has some offset, which is corrected manually by visual inspection. An example of this alignment process can be seen in Figure 9.

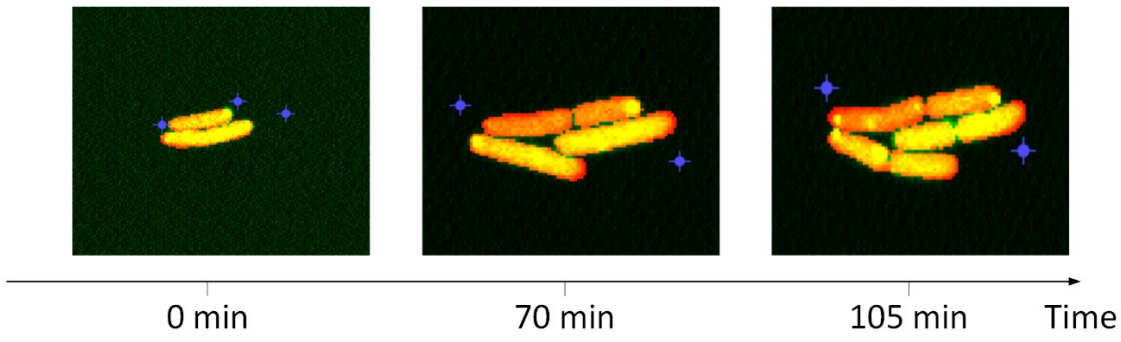

Figure 9. *Segmented phase contrast images aligned over confocal time-lapse images. In this, the blue dots correspond to the regions of the overlapped image that should be manually aligned to extract the fluorescence intensities of each cell detected in the corresponding phase-contrast image.*

3.3.2 RNA quantification

To quantify the RNAs in cells, the spots intensity in the cells need to be calculated. For that, first, the region where the MS2-GFP RNA spots are located, should be segmented. These spots are segmented automatically using a kernel Density Estimation (KDE) method [49]. In short, this method estimates the probability density function from the distribution of pixel intensities of each spot, and then it finds a cut-off point, which corresponds to the first local minimum of the KDE. Then, each pixel is checked and those pixels whose intensities are above the cut-off value are segmented as spots [50].

The total spot intensity of the cell is calculated by adding all the pixel values of the spots in the cell. In addition, the unbound MS2-GFP molecules in the cells constitute background fluorescence, which need to be subtracted from the total spot intensity of the cells.

To perform this background correction, the mean background intensity of the cell is multiplied by the area of the spot and then this value is subtracted from the total spot intensity. The corrected spot intensity is quantified into RNAs by normalizing the spot intensity histogram of cell population by the difference in intensity of the first two peaks of the distribution, which corresponds to the intensity of a single RNA [47], as represented in Figure 10.

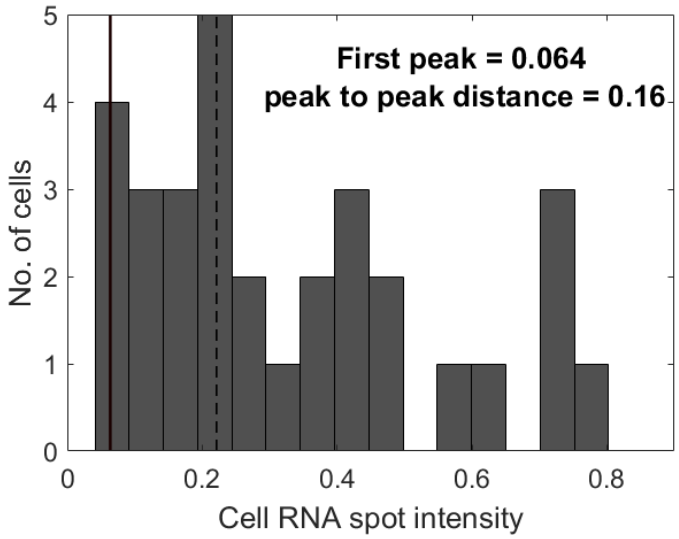

Figure 10. *Manual RNA rounding method **[47]**, here referred to as "peak selection" method, of a distribution of spot intensities. In this, the number of RNAs, per total spot intensity value, is estimated by manually selecting the first peak of intensity that most likely corresponds to 1 RNA molecule.*

3.3.3 RNA polymerases quantification

RNA polymerase numbers inside the cells are known to vary by changing the media richness [6]. Media richness can be changed by, e.g., varying glycerol concentration in the media.

Once having a set of conditions where cells differ in RNAp concentrations, these differences can be determined, e.g. by RNAp fluorescence intensity measurements, which can quantify the changes in RNAp concentrations relative to a control condition. It is expected that, according to standard models of transcription (see e.g. (McClure, 1985)), such changes in RNAp concentrations inside the cell will cause changes in the rate of occurrence of transcription events. To directly correlate the changes in fluorescence density levels of the RNAp molecules to changes in the transcription initiation rates, it is assumed that the RNA polymerase numbers available to bind with the promoter to initiate transcription are proportional to the mean RNAp fluorescence density within a cell.

Based on this, after segmentation and alignment of cells from microscopy images, the fluorescence density of each cell is measured by calculating the mean fluorescence intensity of the cell. Then, the relative RNAp concentration of different conditions is quantified by first calculating the mean fluorescence intensity per pixel of all the cells in the population for each condition. Afterwards, the mean of the mean fluorescence intensity per pixel of the cells for each condition is calculated. Next, the resultant fluorescence intensities are normalized with respect to the control condition, so as to obtain the relative fluorescence between a condition and the control. This relative fluorescence intensity values can then be used in τ plots, a plotting technique that allows dissecting the duration of the rate-limiting steps in transcription initiation subsequent to the initiation of the open complex formation (i.e. that do not depend on the concentration of RNAp in the cell).

Meanwhile, the RNAp numbers variability between cells of a population can be estimated by fitting a normal distributed curve over the relative RNAp fluorescence intensity values of individual cells [1]. From the best fitting curve, the distribution parameters, such as mean and standard deviation, are extracted. We use this distribution parameters to estimate the empirical levels of extrinsic noise in RNAp numbers.

Finally, we introduce those empirical numbers in our stochastic model of gene expression, which is implemented in each cell of a population by randomly drawing an RNAp amount from that distribution for each cell. This is the main innovation of our model, when compared with previous stochastic models [7] [51] [52].

3.4 Measurement of RNA production time intervals

From the RNA production events, the mean RNA production time is calculated. We used two methods to calculate this RNA production intervals. Both methods have their own pros and cons and they are explained briefly in the following sub chapters.

3.4.1 Time intervals from consecutive RNA production events

From the RNA production events estimated from time-lapse microscopy images, we can extract precise information about the RNA production dynamics. As the MS2 viral coat protein and their binding to the target RNA are ‘near-immortal’, the MS2-GFP tagged RNA molecules do not degrade over time. Thus, as more RNAs are created, we expect the total spots intensity in the cell to increase over time. This increase in spots intensity is expected due to the occurrences of new transcriptional events over time.

We extract the time intervals between consecutive RNA production events using an automated method. In this, the total spots intensity of each cell, over time, is fitted with a monotone piecewise-constant function by least squares. The order of the fitted model is selected using an F-test (p-value 0.01) and, for better fitting to the data, higher order to

be chosen. From the results of model fitting, the distribution of intervals between consecutive RNA production events is obtained for each condition. An example of RNA production events obtained from fluorescent spots intensity of a cell over time is shown in Figure 11.

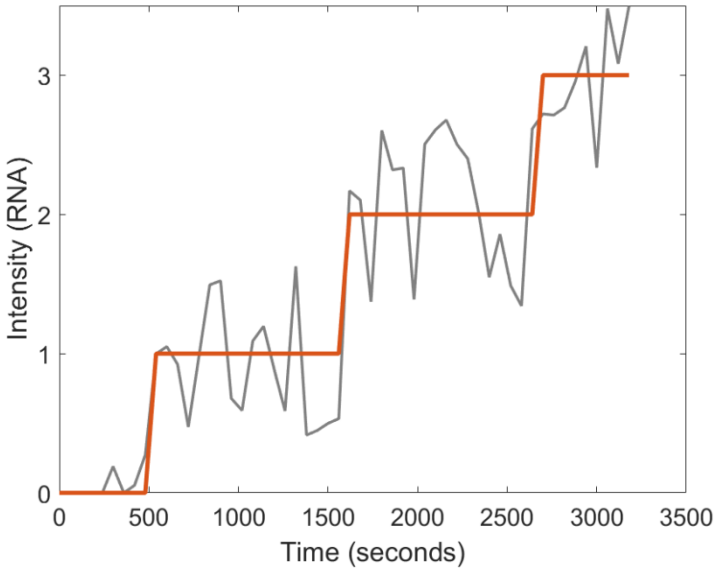

Figure 11. *Example plot of the time course of the total corrected intensity levels of spots in a cell (grey line), from time-lapse confocal microscopy images, and the monotone piecewise-constant fit (orange line) that assigns RNA numbers to the intensity levels in this cell time-series.*

3.4.2 The first and last frame method

In this method, an approximate value of the mean of the time intervals between successive RNA production is obtained from the RNA fluorescence intensities of the cells in the first and last frames of the time lapse microscopy images. This method considers two assumptions: i) all the cells in the population have the same cell division rate and, ii) all cells have the same RNA production rate (as represented in Figure 12). In comparison to the method described in section 3.4.1, this method is advantageous as much less time is consumed in obtaining the final results. Namely, it significantly reduces the time taken for automatic segmentation followed by manual correction of the cells. The disadvantage of this method is that it is less informative of the RNA production kinetics when compared to the previous method.

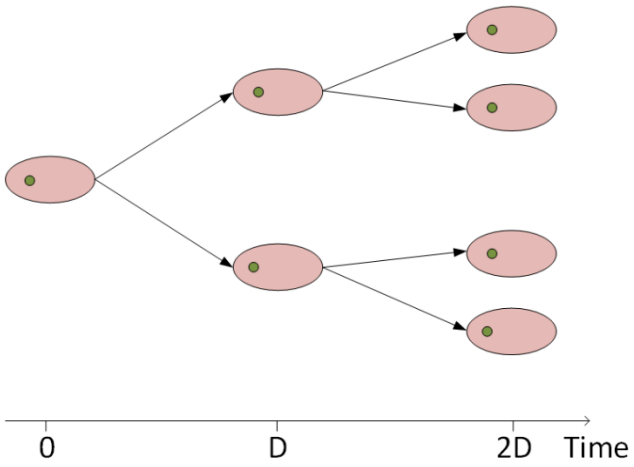

Figure 12. *Model of formation of a cell lineage by cell division. In this, a new cell generation occurs at each doubling interval, and all cellular components of the mother cell, such as RNAs, are equally divided by the two daughter cells.*

Consider N_0 to be the number of cells present at the beginning of the time series, and N to be the number of cells at any given subsequent time moment t. Thus, if D is the doubling time, we have that:

$$N(t) = N_0 2^{\frac{t}{D}} \tag{3.1}$$

Assuming $R(t)$ to be the number of RNAs in the entire cell population at the moment t, the rate at which RNAs are produced by the population is:

$$\frac{dR}{dt} = kN \tag{3.2}$$

Where, k is the RNA production rate constant.

From (3.1) and (3.2),

$$\frac{dR}{dt} = kN_0 2^{\frac{t}{D}} \tag{3.3}$$

Solving this linear differential equation, one obtains:

$$R = kN_0 2^{\frac{t}{D}} \frac{D}{\ln 2} + C \tag{3.4}$$

Applying the following conditions (i) and (ii) to equation (3.4), the values of the constants C and k can be found:

i) When $t = 0$, $R = R_0$, where R_0 is the initial number of RNAs in the population

ii) When $t = D$, $R\text{-}R_0 = nN_0$, where D is the doubling time, and n is the number of RNAs produced per cell in 1 doubling time.

Applying condition (i) into equation (3.4), we obtain:

$$R_0 = kN_0(1)\frac{D}{\ln 2} + C \tag{3.5}$$

Thus, the constant C is found to be:

$$C = R_0 - kN_0\frac{D}{\ln 2} \tag{3.6}$$

Replacing the expression of constant C into equation (3.4), we get:

$$R = kN_0 2^{(\frac{t}{D})}\frac{D}{\ln 2} + R_0 - kN_0\frac{D}{\ln 2} \tag{3.7}$$

The above equation can be rewritten as:

$$R - R_0 = kN_0\frac{D}{\ln 2}2^{(\frac{t}{D})}\left[1 - 2^{(-\frac{t}{D})}\right] \tag{3.8}$$

Applying condition (ii) in equation (3.8):

$$nN_0 = kN_0 \cdot \frac{D}{\ln 2} \cdot 2^{(1)}\left[1 - 2^{(-1)}\right] \tag{3.9}$$

From the above expression, constant 'k' is found as:

$$k = \frac{n \cdot \ln 2}{D} \tag{3.10}$$

Introducing the expression of the constant 'k' in equation (3.8), we get:

$$R - R_0 = nN_0 \cdot 2^{(\frac{t}{D})}\left[1 - 2^{(-\frac{t}{D})}\right] \tag{3.11}$$

The above expression can be rewritten as:

$$\frac{R}{N_0 2^{(\frac{t}{D})}} - \frac{R_0}{N_0 2^{(\frac{t}{D})}} = n\left[1 - 2^{(-\frac{t}{D})}\right] \tag{3.12}$$

Next, let $M = R/N_0 2^{(\frac{t}{D})}$ be the mean number of RNA per cell after time t, and $M_0 = M_0/N_0$ be the mean number of RNAs per cell at the initial time moment ($t = 0$). Replacing these simplified terms in equation (3.12), the number of RNAs produced per cell (n) in 1 doubling time (D) is found to be:

$$n = \frac{M - M_0 2^{(-\frac{t}{D})}}{1 - 2^{(-\frac{t}{D})}} \tag{3.13}$$

From equation (3.13), the RNA production rate (P_{rna}), i.e. the number of RNAs produced per cell per unit time, is given by:

$$P_{rna} = \frac{n}{D} = \frac{M - M_0 2^{(\frac{-t}{D})}}{D\left[1 - 2^{(\frac{-t}{D})}\right]} \tag{3.14}$$

3.5 Dissection of RNA production time intervals

The dissection of RNA production time interval is done to quantify the duration of open and closed complex formation. This is done using a τ plot (Lloyd-Price et al, 2016), which includes a line fitting procedure, which allows extracting the time-length of the open complex formation (McClure, 1985).

3.5.1 τ Plots

The duration of the rate limiting steps in transcription initiation have been calculated by using an 'abortive initiation method' as demonstrated by McClure in 1985 [22]. As expected, there is a mean time for an RNAp to successfully bind to a promoter and the initiation of transcription. This process is named as closed complex formation and we represent here the time it takes as t_{cc}.

To dissect this binding and the subsequent isomerization steps required for the production of a RNA and to quantify their respective time duration, McClure considered a model of a two-step reversible transcription initiation process (3.15):

$$R_p + P \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} Pcc \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} Poc \tag{3.15}$$

where, R_p is free RNAp available for transcription, P is free promoter, and P_{cc} and P_{oc} are promoter states in closed and open complex formation, respectively. Using this method,

it is possible to quantify the duration of these rate-limiting steps in transcription initiation [53].

Applying the steady state condition to P_{cc} and considering $k_2 \gg k_{-2}$:

$$k_{obs} = \frac{k_1[R_p]k_2}{k_1[R_p] + k_{-1} + k_2} \tag{3.16}$$

Here, k_{obs} is the rate at which P_{oc} is formed, which can be obtained from empirical measurements. Meanwhile, the average time (τ_{obs}) the promoter takes for completion of one promoter initiation process is:

$$\tau_{obs} = \frac{1}{k_2} + \frac{k_{-1}+k_2}{k_1[R_p]k_2} \tag{3.17}$$

Using equation (3.17), it is possible to separate the open and closed complex formation durations from τ_{obs}, using a 'τ plot', where the inverse of RNAp concentration on the x-axis and the respective τ_{obs} in the y-axis are plotted (Figure 13).

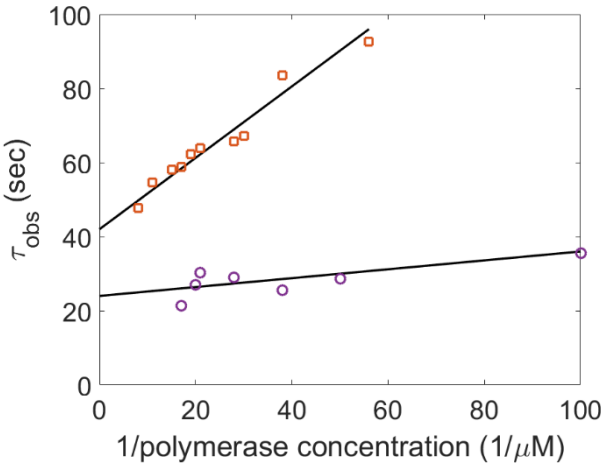

Figure 13. *τ plot of lag times (τ_{obs}) for D and A2 promoters of T7 bacteriophage. The lag times observed (τ_{obs}) for pGpUpu synthesis from the D promoter (in squares) and for pGpC synthesis from A2 promoter (in circles), are plotted versus the inverse of RNAp concentrations.*

As the closed complex formation time is inversely proportional to the concentration of RNAp, a linear relationship is expected between the closed complex formation time and the inverse of the RNAp concentration. After plotting the data points, a line that best fits is drawn through those data points as shown in Figure 13.

Relevantly, the y-intercept (c) of the best fitting line is approximately equal to the mean time the promoter spends for the formation of open complex (because at this point, the

number of RNAp is 'infinite'), which is equal to $1/k_2$. Meanwhile, for a given finite concentration of RNAp, the closed complex formation time is approximately equal to the product of inverse of RNAp concentration and the slope (m) of the best fitting line. Using these ideas, McClure and colleagues performed the dissection of the *in vitro* kinetics of transcription initiation.

Recent studies [6] have shown that the rate limiting steps of transcription initiation in *in vivo* could also be similarly dissected, from the mean RNA production intervals in *in vivo* single cell measurements. It was observed that the mean RNA production intervals decrease with increasing media richness, which increase the RNAp numbers in the cells.

Then, a plot is made between the inverse of RNAp concentration in x-axis and mean RNA production interval in y-axis (τ plot). Then, slope (m) and y-intercept (c) of the line that best fits to those data points is obtained using the maximum likelihood algorithm explained in 3.5.2. It is assumed that at infinite RNAp numbers, the rate at which the formation of closed complex is assumed to be infinitely fast and the time required for this formation is therefore ~0. Thus, the y-intercept is approximately equal to the open complex formation time. Also, the closed complex formation time is equal to the product of inverse of RNAp concentration and the slope (m) of the best fitting line. In this thesis, this methodology is applied to dissect the duration of rate limiting steps in transcription initiation, from which the corresponding rate constants are inferred.

3.5.2 Fitting line procedure

For dissecting the promoter initiation kinetics using a τ plot, a straight line is to be fitted to the data points (x_i, y_i), where x_i is the inverse of RNAp concentration and its corresponding mean RNA production interval is y_i. To find the line that fit best to the data points (x_i, y_i), the maximum likelihood estimation method is used. The equation of the straight line that is to be fitted with the data is:

$$y_{fit} = mx_i + c \tag{3.18}$$

Where, m is the slope and c is the y-intercept of the best-fitted line.

Given this, the slope and y-intercept of the best-fitting straight line can be calculated through the minimization of the sum of the squares of the residuals. Considering that each data point has some uncertainty σ_i, the maximum likelihood estimate of the parameters can be obtained by minimizing the chi-squared function (χ^2) in equation (3.19).

$$\chi^2 = \sum_{i=1}^{i=N}\left[\frac{(y_i - y_{fit})}{\sigma_i}\right]^2 = \sum_{i=1}^{i=N}\left[\frac{(y_i - mx_i - c)}{\sigma_i}\right]^2 \tag{3.19}$$

After minimization, the parameters of the best fitting line, the slope (m) and y-intercept(c) are calculated from equations (3.20) and (3.21) respectively.

$$m = \frac{\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2}\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2} - \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\sum_{i=1}^{N}\frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)^2} \tag{3.20}$$

$$c = \frac{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\sum_{i=1}^{N}\frac{x_i y_i}{\sigma_i^2} - \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2}\right)^2} \tag{3.21}$$

3.6 Multi-step models of gene expression

3.6.1 Modelling transcription

We consider the model of transcription initiation in which the RNA production kinetics can vary from sub-Poissonian to super-Poissonian dynamics, depending on the rate constant values of the rate limiting steps of the model. This model was derived after reviewing various studies, encompassing genome-wide variability in RNA numbers [54] [55] and the dynamics of transcription in individual genes [56].

The model includes important rate-limiting steps in transcription initiation in *E. coli* such as, open complex formation [5], closed complex formation, ON/OFF process [57]. From the empirical RNA production interval measurements and the single cell RNAp measurements, we find the best fitting model and their parameters as reported in [6].

The reaction equations of the best-fitted model are in (3.22) and (3.23), which can also be applied to other general promoters in *E. coli*. However, the rate constant values vary between promoters and conditions, due to different transcription initiation kinetics.

$$P_{\mathrm{ON}} + R_p \underset{k_{rev}}{\overset{k_{cc}}{\rightleftarrows}} P_{CC} \xrightarrow{k_{oc}} P_{OC} \xrightarrow{k_e} R_p + P_{\mathrm{ON}} + \mathrm{RNA} \tag{3.22}$$

$$P_{\mathrm{ON}} \underset{k_{ON}}{\overset{k_{OFF}}{\rightleftarrows}} P_{\mathrm{OFF}} \tag{3.23}$$

Reactions in equation (3.22) represent the multi-step transcription model of an active promoter [58], where P_{ON} is the state of promoter in unrepressed active state. First, the RNAp (R_p) binds with free active promoter (P_{ON}) and forms closed complex (P_{CC}) with rate con-

stant k_{cc}. Then, the promoter undergoes several intermediate steps (association and dissociation of RNAp with the promoter) with rate constant of reversibility k_{rev}, before it forms open complex (P_{OC}) with rate constant k_{oc} [58] [59]. And after that, the promoter clearance happens, followed by transcription elongation process with a rate constant k_e, which produces an RNA.

As the number of RNAp (R_p) in the cells is high, it is assumed that its concentration remains approximately constant over time and the equation (3.22) is rewritten as:

$$P_{\mathrm{ON}} \underset{k_{rev}}{\overset{k^*_{cc}}{\rightleftarrows}} P_{CC} \xrightarrow{k_{oc}} P_{OC} \xrightarrow{k_e} R_p + P_{\mathrm{ON}} + \mathrm{RNA} \tag{3.24}$$

Where, $k^*_{cc} = k_{cc} R_p$

Reactions in equation (3.23) represent the back and forth transitioning of the free promoter (P_{ON}) to an inactive state (P_{OFF}) due to various reasons such as binding/unbinding of repressors/activators [60], accumulation of positive DNA supercoiling [61], etc. In our model, we assume that the back and forth transition between the ON and OFF states are due to binding/unbinding of repressors and so, the equation (3.23) is rewritten as:

$$P_{\mathrm{ON}} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P_{\mathrm{rep}} \tag{3.25}$$

Where, P_{rep} is the promoter at repressed state, k_{rep} is the rate constant at which the Promoter goes from P_{ON} to P_{rep} state and k_{urep} is the rate constant at which the Promoter goes from P_{rep} to P_{ON} state.

To calculate the time that the promoter spends in different states in the transcriptional process, we make use of the concept of reaction competition. To understand this, a simple example model is assumed, where a reaction specie A produces species B and C at rate constants k_b and k_c respectively, as shown in equation (3.26):

$$B \xleftarrow{k_b} A \xrightarrow{k_c} C \tag{3.26}$$

From equation (3.26), it is inferred that, on average, by the time reaction $A \xrightarrow{k_c} C$ occurs, reaction $A \xrightarrow{k_b} B$ should have occurred k_b/k_c times. On the other hand, by the time reaction $A \xrightarrow{k_b} B$ occurs, reaction $A \xrightarrow{k_c} C$ should have occurred k_c/k_b times. Using this idea, the times that the promoter spends before and after committing to the formation of open complex are derived as follows.

From equation (3.24), the average time the promoter spends in elongation process (t_e) and open complex formation (t_{oc}) in complete transcription are:

$$t_e = \frac{1}{k_e} \tag{3.27}$$

$$t_{oc} = \frac{1}{k_{oc}} \tag{3.28}$$

Also, from equation (3.24), the open complex formation step occurs only once in every transcription. Having said that, by the time that a reaction $P_{CC} \xrightarrow{k_{oc}} P_{OC}$ occurs, reaction $P_{CC} \xrightarrow{k_{rev}} P_{ON}$, should have occurred k_{rev}/k_{oc} instants. Meanwhile, reaction $P_{ON} \xrightarrow{k^*_{cc}} P_{CC}$ should have occurred $k_{rev}/k_{oc} + 1$ instants. So, the average time that the promoter spends in the closed complex formation (t_{cc}) to complete transcription is:

$$t_{cc} = \left(\frac{k_{rev}}{k_{oc}} + 1 \right) \cdot \frac{1}{k^*_{cc}} \tag{3.29}$$

From equation (3.24) and (3.25), if the reaction $P_{ON} \xrightarrow{k^*_{cc}} P_{CC}$ occurs once, the reaction $P_{ON} \xrightarrow{k_{rep}} \mathrm{P}_{rep}$ should have occurred k_{rep}/k^*_{cc} instants. Since, the reaction $P_{ON} \xrightarrow{k^*_{cc}} P_{CC}$ occurs $k_{rev}/k_{oc} + 1$ instants, the reaction $P_{ON} \xrightarrow{k_{rep}} \mathrm{P}_{rep}$ should have occurred $(k_{rev}/k_{oc} + 1) \cdot k_{rep}/k^*_{cc}$ instants and the reaction $P_{rep} \xrightarrow{k_{urep}} P_{ON}$ also should have occurred $(k_{rev}/k_{oc} + 1) \cdot k_{rep}/k^*_{cc}$ instants. Hence, the average time the promoter stays in repressed state (t_{rep}) for a transcription even to occur is, on average:

$$t_{rep} = \left(\frac{k_{rev}}{k_{oc}} + 1 \right) \cdot \frac{k_{rep}}{k^*_{cc}} \cdot \frac{1}{k_{urep}} \tag{3.30}$$

The above reaction is rewritten as:

$$t_{rep} = \frac{(k_{rev} + k_{oc}) \cdot k_{rep}}{k_{oc} k^*_{cc} k_{urep}} \tag{3.31}$$

As the mean RNA production interval (Δt) is equal to the time taken for 1 complete transcription, Δt is equal to the sum of average repression time (t_{rep}), average closed complex formation time (t_{cc}), average open complex formation time (t_{oc}) and average elongation time (t_e), and is expressed as in equation (3.32).

$$\Delta t = t_{rep} + t_{cc} + t_{oc} + t_e \tag{3.32}$$

Replacing equations (3.27), (3.28), (3.29) and (3.31) in equation (3.32):

$$\Delta t = \frac{(k_{rev} + k_{oc}) \cdot k_{rep}}{k_{oc} k^*_{cc} k_{urep}} + \left(\frac{k_{rev}}{k_{oc}} + 1 \right) \cdot \frac{1}{k^*_{cc}} + \frac{1}{k_{oc}} + \frac{1}{k_e} \tag{3.33}$$

$$\Delta t = \frac{k_{rev} + k_{oc}}{k_{oc} k^*_{cc}} \left[\frac{k_{rep}}{k_{urep}} + 1 \right] + \frac{1}{k_{oc}} + \frac{1}{k_e} \tag{3.34}$$

$$\Delta t = \frac{\left(k_{rev} + k_{oc}\right)\left(k_{rep} + k_{urep}\right)}{k_{oc} k^*_{cc} k_{urep}} + \frac{1}{k_{oc}} + \frac{1}{k_e} \tag{3.35}$$

As the elongation time is relative very small, $1/k_e$ tends to zero. And, the mean RNA interval approximately equals:

$$\Delta t \approx \frac{\left(k_{rev} + k_{oc}\right)\left(k_{rep} + k_{urep}\right)}{k_{oc} k^*_{cc} k_{urep}} + \frac{1}{k_{oc}} \tag{3.36}$$

The equation (3.36) is rewritten by replacing $k^*_{cc} = k_{cc} R_p$

$$\Delta t \approx \frac{\left(k_{rev} + k_{oc}\right)\left(k_{rep} + k_{urep}\right)}{k_{oc} k_{cc} R_p k_{urep}} + \frac{1}{k_{oc}} \tag{3.37}$$

Thus, the relationship between the mean RNA production interval (Δt) and rate constants of rate limiting steps in transcription is expressed in equation (3.37). From the above expression, the time the promoter spends before committing to the formation of open complex (t_{prior}) and the time the promoter spends after committing to the formation of open complex (t_{after}) are expressed as:

$$t_{prior} = \frac{\left(k_{rev} + k_{oc}\right)\left(k_{rep} + k_{urep}\right)}{k_{oc} k_{cc} R k_{unrep}} \tag{3.38}$$

$$t_{after} = \frac{1}{k_{oc}} \tag{3.39}$$

According to our model, increasing k_{cc}, decreases the duration of closed complex formation and increasing k_{oc}, reduces the duration of open complex formation. The duration of open and complex formation of the model is altered by varying k_{oc} and k_{cc}, while keeping the mean RNA production interval (Δt) and all the other rate constant values as constant.

Finally, the mean RNA production interval that is kept constant is calculated using equation (3.37) and the rate constants are obtained empirically [6]. These parameter values are

listed in Table 1. The relative time that the promoter spends before its commits to the formation of the open complex is varied by changing k_{cc} and k_{oc} using equation (3.40).

$$\frac{\tau_{prior}}{\Delta t} = \frac{\left(k_{rev} + k_{oc}\right)\left(k_{rep} + k_{unrep}\right)}{k_{oc} k_{cc} R k_{urep}} \times \Delta t^{-1} \tag{3.40}$$

Table 1. *Parameter values of the model of transcription under control conditions. k_{cc} value is set, assuming that the number of available RNAp equals 1 (and are never depleted).*

Parameter	Value	Reference
k_{rep}	281 s^{-1}	[6]
k_{urep}	0.01 s^{-1}	[6]
k_{cc}	6469 s^{-1}	[6]
k_{oc}	0.005 s^{-1}	[6]
k_e	∞ s^{-1}	[6]
k_{rev}	1 s^{-1}	[6]
R_p (Relative mean RNAp per cell)	1*	[6]

3.7 Modelling a 2-gene toggle switch

We consider a dynamic model of a 2-gene toggle switch, in which in addition to the model of transcription explained in chapter 3.8.4, a translation step is also added for each gene in the network. This model allows RNA and protein production kinetics to differ widely in noise levels, depending on the rate constants of the rate limiting steps. The translation step of gene expression is modelled as a result of detailed studies, including translational kinetics at the single protein level [62] [63] [64], protein folding and activation kinetics [65] and the structure of natural genetic switches [32] [11].

The illustrative image of two-gene toggle switch network is shown in Figure 4. To better understand, the modelling of this genetic circuit is split into 3 steps, namely active transcription, repression and translation. First, in active transcription, the RNAp (R_p) binds with free active promoter (P_{ON}^i) [59] and forms closed complex (P_{cc}^i) with rate constant k_{cc}. Then, the promoter undergoes several intermediate steps (Association and dissociation of RNAp with the promoter) with rate constant of reversibility k_{rev}, before it forms open complex (P_{oc}^i) with rate constant k_{oc} [58] [59]. After that, promoter clearance happens, during which RNAp is released from the promoter, followed by transcription elongation [66] with a rate constant k_e, which produces RNA^i. For simplicity of the model, we assume a symmetric toggle switch network, with the genes having the same rate constant values. This active transcription step of gene expression of individual genes in the network is modelled as reactions in equation (3.41), where i=1, 2, index of the individual promoter in the network.

$$P^i_{\mathrm{ON}} + R_p \underset{k_{rev}}{\overset{k_{cc}}{\rightleftarrows}} P^i_{cc} \xrightarrow{k_{oc}} P^i_{oc} \xrightarrow{k_e} R_p + P^i_{\mathrm{ON}} + \mathrm{RNA}^{\mathrm{i}} \tag{3.41}$$

In translation part, the Ribosome (*Rib*) binds with $\mathrm{RNA}^{\mathrm{i}}$, and produces inactive protein molecules($Prot^i_{unfolded}$) with a rate constant k_{rbs}, where *i* represents the index of the individual promoter in the network, which takes part in this translational process. Then it is followed by subsequent post-translational process, in which the inactive protein($Prot^i_{unfolded}$) converts to active protein($Prot^i_{folded}$) with a rate constant k_{fold}. These processes are modelled as reactions in equations (3.42) and (3.43).

$$RNA^i + Rib \xrightarrow{k_{rbs}} RNA^i + Rib + \mathrm{Prot}^i_{unfolded} \tag{3.42}$$

$$\mathrm{Prot}^i_{unfolded} \xrightarrow{k_{fold}} \mathrm{Prot}^i_{folded} \tag{3.43}$$

In repression, the free active promoter (P^i_{ON}) goes to inactive repressed state (P^i_{rep}) state after it binds with promoter specific repressor protein ($Prot^j_{folded}$), where *i* represents the index of the promoter being repressed, where *j* represents the index of the repressor protein. This is modelled as in reactions in equations (3.50) and (3.51).

$$P^1_{\mathrm{ON}} + \mathrm{Prot}^2_{folded} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P^1_{\mathrm{rep}} \tag{3.44}$$

$$P^2_{\mathrm{ON}} + \mathrm{Prot}^1_{folded} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P^2_{\mathrm{rep}} \tag{3.45}$$

Then, RNA and protein numbers in a cell gets decayed in two ways. First, the RNA and protein degrades over time with a degradation rate constant (*Deg*). Second, the dilution of RNA and protein molecules due to cell division at a dilution rate constant (*Dil*). Here, the mean lifetime of the cell (*div*) is assumed to be 1 hour [6] and dilution rate constant (*Dil*) is calculated using equation (3.46). Then the rate at which RNA and protein decay is modelled as reactions in equations (3.48) and (3.49), whose decay rate constant (k_d) is the sum of degradation constant (*Deg*) and dilution constant (*Dil*) shown in equation (3.47).

$$Dil = div^{-1} \times \log(2) \tag{3.46}$$

$$k_d = Dil + Deg \tag{3.47}$$

$$rna^i \xrightarrow{k_d} \varnothing \tag{3.48}$$

$$\mathrm{Prot}^i_{folded} \xrightarrow{k_d} \varnothing \tag{3.49}$$

3.7.1 Modelling a 3-gene Repressilator

The dynamic model of 3 gene Repressilator genetic circuit is considered, which includes both transcription and translation steps for individual genes in the network. The network works in a way that gene 1 represses gene 2, gene 2 represses gene 3 and gene 3 represses gene 1. The schematics of this network is shown in Figure 5. Depending upon the rate constant values of the rate limiting steps in transcription initiation of genes in the network, there is diversity in RNA and protein production over the cell population. The translational step of the genes in the network is modelled as a result of detailed studies, including translational kinetics at a single protein level [62] [63] [64], protein folding and activation kinetics [65] and the structure of natural genetic switches [32] [11].

For better understanding, the modelling of this gene circuit is split into three steps, namely active transcription, repression and translation. First, in active transcription, the RNAp (R_p) binds with free active promoter (P^i_{ON}) [59], it goes to a state of closed complex (P^i_{cc}) with rate constant k_{cc}. Then, the promoter undergoes several intermediate steps (association and dissociation of RNAp with the promoter) with rate constant of reversibility k_{rev}, before it forms open complex (P^i_{oc}) with rate constant k_{oc} [58] [59]. After that, promoter clearance happens, during which RNAp is released from the promoter, followed by transcription elongation [66] with a rate constant k_e, which produces RNA^i in the end. For simplicity of the model, we assume a 3 gene Repressilator network, with all the genes having the same rate constant values. This active transcription part of gene expression of individual gene in the network is modelled as reactions in equation (3.41), where i=1,2,3, is the index of the individual promoter in the network.

In repression, the free active promoter (P^i_{ON}) goes to inactive repressed state (P^i_{rep}) state after it binds with promoter specific repressor protein ($Prot^j_{folded}$), where i represents the index of the promoter being repressed, where j represents the index of the repressor protein. This is modelled in reactions in equations (3.50), (3.51) and (3.52).

$$P^1_{\mathrm{ON}} + \mathrm{Prot}^3_{folded} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P^1_{\mathrm{rep}} \tag{3.50}$$

$$P^2_{\mathrm{ON}} + \mathrm{Prot}^1_{folded} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P^2_{\mathrm{rep}} \tag{3.51}$$

$$P^3_{\mathrm{ON}} + \mathrm{Prot}^2_{folded} \underset{k_{urep}}{\overset{k_{rep}}{\rightleftarrows}} P^3_{\mathrm{rep}} \tag{3.52}$$

In translation, the Ribosome (Rib) binds with RNA^i, and produces inactive protein molecules($Prot^i_{unfolded}$) with a rate constant k_{rbs}, where i represents the index of the individual promoter in the network, which takes part in this translational process. Then it is fol-

lowed by subsequent post-translational process, in which the inactive protein($Prot^{i}_{unfolded}$) converts to active protein($Prot^{i}_{folded}$) with a rate constant k_{fold}. These processes are modelled as reactions in equations (3.42) and (3.43).

Finally, RNA and protein numbers in a cell decay in two ways. First, the RNA and protein degrades over time with a degradation rate constant (*Deg*). Second, the dilution of RNA and protein molecules due to cell division at a dilution rate constant (*Dil*). Here, the mean lifetime of the cell (*div*) is assumed as 1 hour [6] and dilution rate constant ($\underline{Dil}$) is calculated using equation (3.46). RNA and protein decays are modelled as reactions in equations (3.48) and (3.49), whose decay rate constant (k_d) is the sum of degradation constant (*Deg*) and dilution constant (*Dil*) shown in equation (3.47).

3.8 Stochastic simulation of models

The models that are constructed in section 3.6 are simulated using the stochastic simulation algorithm. Before that, the concept of stochastic simulation of chemical kinetics is explained, followed by a simplified stochastic simulation algorithm. The simulation of individual models and the ways to measure their parameters are described.

3.8.1 Stochastic simulation of chemical kinetics

This chapter briefly explains the theoretical concepts of modelling and simulating stochastic chemical kinetics and its application to gene expression, specifically transcription and translation.

In biochemical processes, such as gene expression, the number of reacting species is small. E.g. DNA, RNA, regulatory proteins generally have few copies per cell [54]. Having a small number of reactive species, deterministic methods are not the proper approach to simulate the dynamics of gene expression. Due to this, a discrete model is needed [67].

To accurately model the time evolution of reacting species in the system, it is required to track the movement of each individual molecule in the system space, to detect collisions between molecules and to update the concentration of the molecular species in the system, after each reactive collision, i.e. collisions which lead to the formation of the new molecule. The timing of the reactive collisions cannot be deduced exactly [68].

Having said that, the dynamics of such system is not deterministic, as it cannot be described by a single trajectory in the state space. Considering the discrete nature and stochastic time evolution of the reactive species in the population, it is required to consider the probability distribution of the possible states of the species at a certain time moment. The time evolution of the reactive species, whose consequent states in discrete population are determined by probabilistic distribution, is well explained by the stochastic chemical kinetics [68].

In stochastic time evolution, a system with N different molecular species homogeneously spread at time t is represented by an N-dimensional vector x. These species interact through M number of possible chemical reaction that can occur between them, which results in changes in the population of species. It is assumed that the system volume is constant and to be well-stirred, such that there are no non-reactive collisions [9].

The change in the population of species is determined by two quantities, namely state-change vector (v_μ) and propensity function (a_μ). State-change vector (v_μ) defines the change in the population of species x. Propensity function(a_μ), which is the probability at which reaction R_μ occurs, is defined as follows [68]:

$$a_\mu(x)dt = \text{the probability at which molecules of the system react via reaction } R_\mu \text{ in the next infinitesimal time interval } [t,\ t+dt). \tag{3.53}$$

The propensity function depends upon the nature of the reacting species in the reaction. For unimolecular reactions, the constant c_μ is the probability of a molecule of X will react through reaction R_μ in the next infinitesimal time window dt [68] and the propensity function of a molecular specie of concentration of X, is defined as follows:

$$a_\mu(x) = c_\mu X \tag{3.54}$$

For bimolecular reactions between two species X_1 and X_2, the constant c_μ is the probability that a single random pair of molecules from X_1 and X_2, react in accordance with reaction R_μ in the next infinitesimal time window dt [68]. The propensity function of this kind of reactions is:

$$a_\mu(x) = c_\mu X_1 X_2 \tag{3.55}$$

For bimolecular reactions between the same species X, the constant c_μ is the probability that a single random pair of molecules from X, react in accordance with reaction R_μ in the next infinitesimal time window dt [68] and propensity function of this kind of reaction is:

$$a_\mu(x) = \frac{c_\mu X(X-1)}{2} \tag{3.56}$$

From (3.53) and the probability $P(x, t|x_0, t_0)$ of reaching a state vector x at time t, after having the initial conditions $x=x_0$ at $t=t_0$, the equation of time evolution can be derived for stochastic chemical kinetics using the laws of probability [68]. The chemical master equation (CME), which is the partial differential equation of P:

$$\frac{\partial P(x,t|\mathrm{x}_0,\mathrm{t}_0)}{\partial t} = \sum_{\mu=1}^{M}[a_\mu(x-v_\mu)P(x-v_\mu,t|\mathrm{x}_0,\mathrm{t}_0) - a_\mu(x)P(x,t|\mathrm{x}_0,\mathrm{t}_0)] \tag{3.57}$$

The CME determines the population of the reactive species after time t in the future. This equation can only be solved analytically for probability density function of *X(t)* for simpler systems. To solve this problem, a Monte Carlo approach is applied, in which multiple numerical realizations of *X(t)* trajectories over time t can be constructed, in order to sample the distribution of *X(t)*. This technique was proposed by Gillespie to simulate the reactions in chemical and biochemical systems [9] [69].

3.8.2 Stochastic simulation algorithm

The evolution of *X(t)* is simulated not based on deterministic approach but on probabilistic function $p(\tau, \mu|x,t)$ [68]. Given the state vector *X(t)* at time instant *t*, this probabilistic function defines the probability of reaction R_μ to occur in the next infinitesimal time interval [*t, t+dt*). This joint density function at state *X* is a function of two random variables. They are the time taken for the next reaction to occur (τ) and the index of the next reaction (μ). By applying the probabilistic laws to equation (3.53), the formula for $p(\tau, \mu|x,t)$ can be derived [9] and the equations (3.60) and (3.61) are the mathematical basis of SSA.

$$p(\tau, \mu \mid x,t) = a_\mu(x) e^{-a_0(x)\tau} \tag{3.58}$$

where,

$$a_0(x) = \sum_{\mu=1}^{M} a_\mu(x) \tag{3.59}$$

$$\tau = \frac{1}{a_0(x)} \ln\left(\frac{1}{r_1}\right) \tag{3.60}$$

$$\mu = \text{the smallest integer satisfying } \sum_{\mu'=1}^{\mu} a_{\mu'}(x) > r_2 a_0(x) \tag{3.61}$$

The next reaction (μ) and the time taken for it to happen (τ) are determined by generating two random numbers r_1 and r_2 from uniform distribution using the equations (3.59), (3.60) and (3.61). The steps of SSA is given in Algorithm 1 [9]:

Algorithm 1: Stochastic Simulation Algorithm

1: Set $t = 0$, and $x = x_0$, where the x is the state vector, which consists of the numbers of all the molecular species present in the system at moment *'t'*, x_0 is the state vector consisting initial concentration of molecular species.

2: Estimate the propensity value of all the reaction $a_\mu(x)$ at time moment *'t'* and calculate its sum $a_0(x)$.

3: Using an appropriate sampling procedure, the next reaction (μ) and the time taken for the next reaction (τ) to occur is calculated.

4: If $t + \tau \geq t_{stop}$ (stop time), terminate the simulation.

5: Set $t = t + \tau$ and update the state vector (x) considering the reaction (μ) just occurred.

6: Go to step 2.

3.8.3 Simulation tools

We have carried out the simulations using SGNS2 (Stochastic Gene Network Simulator v.2) [8], a simulator that runs chemical systems based on the delayed Stochastic Simulation Algorithm, which allows multi-time delayed reactions [51]. SGNS2 also allows creating, destroying and dividing hierarchical, interlinked compartments at runtime. This unique feature is utilized here to generate independent model cells.

3.8.4 Simulation of transcription

We hypothesized that the rate limiting steps in promoter initiation act as an important regulator of the extrinsic noise in transcription. Using the SGNS simulator, we study the cell-to-cell variability in the number of RNA produced in a population as a function of transcription initiation kinetics and cell-to-cell variability in RNAp numbers.

For the simulation of models, we use empirically obtained model parameters, shown in Table 1. Empirical observations have showed that the relative time of open and closed complex formation varies between different promoters kept under different conditions (Table 2). To mimic this behavior of variation in transcription initiation kinetics in our models, we varied the relative time the promoter spends before and after committing to the formation of open complex, which can be done by varying the rate constants k_{cc} and k_{oc}, while keeping the mean RNA production interval constant, using equation (3.40).

Table 2. *Empirical values of $t_{prior}/\Delta t$ of different promoters kept under different conditions.*

Promoter and induction	**$t_{prior}/\Delta t$**
P_{BAD} (0.1% arabinose)	0.71
P_{BAD} (0.01% arabinose)	0.55
P_{BAD} (0.001% arabinose)	0.17
$P_{lac\text{-}O1O3}$ (1 mM IPTG)	0.55
$P_{lac\text{-}O1O3}$ (0.05 mM IPTG)	0.46

$P_{lac\text{-}O1O3}$ (0.005 mM IPTG)	0.12
P_{tetA} (no inducers)	0.07
$P_{lac\text{-}O1}$ (1 mM IPTG)	0.05
$P_{lac\text{-}ara1}$ (1 mM IPTG and 0.1% arabinose)	0.49

Here, we model each cell with one promoter and RNAp molecules, which interact via reactions in equations (3.24) and (3.25). The parameters of the model for actual control conditions are shown in Table 1. Then, we ran the simulation of model cells under varying conditions using SGNS2 [8]. To observe RNA production events over time, we simulated 5 individual model cells, with a lifetime of 2000 s shown in Figure 14. From this figure, it is apparent that most of the cells produced 2 RNAs during their lifetime, as expected [6].

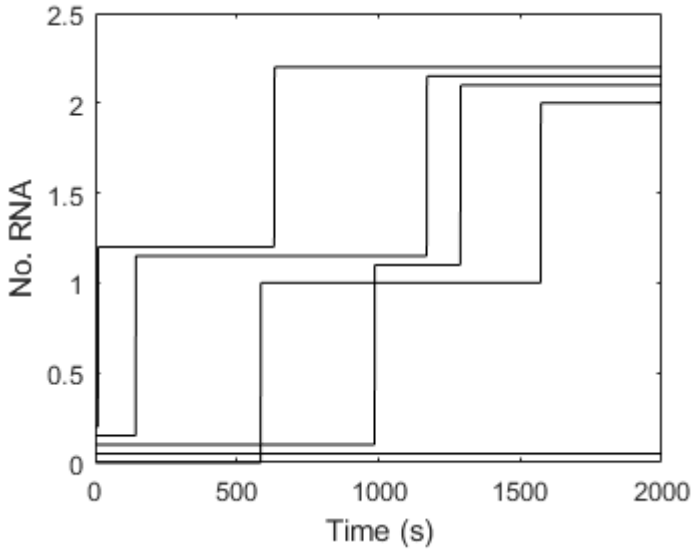

Figure 14. *Time series of 5 individual model cells, with lifetime of 2000s, showing the production of new RNA molecules overtime. The representation of these numbers in the plot are offset, on the y-axis, for good visualization of the lines of different cells (note that only integer RNA numbers are possible).*

From the empirical range of $t_{prior}/\Delta t$, shown in Table 2, we selected 10 values between 0.05 and 0.95, with an increment of 0.1. Also, from the empirical range of the squared coefficient of variation (CV^2) in RNAp shown in Table 4, we select 7 values between 0 and 0.09, with an increment of 0.015. So, we have 70 different model conditions. Each model condition is simulated for 1000 cells and cell lifetime is assumed as 20 000 s. The simulations are carried out in SGNS2 [8], a stochastic simulator which works based on the Stochastic Simulation Algorithm [9]. From the simulation results, the number of RNA produced by the cells are extracted and from which the mean and CV^2 is calculated for the total population of cells.

3.8.5 Simulation of 2 gene toggle switch

We hypothesized that the rate limiting steps in transcription initiation acts as an important factor to regulate the extrinsic noise in 2 gene toggle switch network. Using SGNS simulator, we study the cell-to-cell mean and variability in switching frequency of toggle switch in a population as a function of transcription initiation kinetics and cell-to-cell variability in RNAp numbers.

The transcriptional parameters of the genes in this network model are obtained from empirically validated data, shown in Table 1, except for k_{cc} and k_{oc}. The parameters k_{cc} and k_{oc} varied with transcription initiation conditions and they are calculated using equation (3.40) as explained in the section 3.8.4. Also, the translational parameters of genes in the network model are listed in Table 3. For simplicity of the model, we assume a symmetric switch, with both genes having the same rate constant values.

Table 3. *Translational parameter values of the model switch under control conditions. k_{rbs} is set assuming the ribosome equals 1 and it never depletes.*

Parameter	Value	Reference
kd_{rna}	0.002 s^{-1}	[55]
k_{rbs}	0.637 s^{-1}	[62] [63] [64]
k_{fold}	0.0024 s^{-1}	[65]
kd_p	0.0019 s^{-1}	[65]

From the empirical range of $t_{prior}/\Delta t$, shown in Table 2, we selected 10 values between 0.05 and 0.95, with an increment of 0.1. Also, from the empirical range of CV^2 in RNAp shown in Table 4, we selected 7 values between 0 and 0.09, with an increment of 0.015. So, we have 70 different model conditions. Each model condition is simulated for 100 cells, with a simulation time of 5×10^7s and the protein numbers are sampled at every 10^4 s. The simulations are carried out in SGNS2 [8], a stochastic simulator based on the Stochastic Simulation Algorithm [9].

From the simulation results, the macro and microscale dynamics of the switch is studied. For macro scale dynamics, the number of proteins produced by each gene over time are extracted, from which the cell-to-cell mean, and CV^2 of switching frequency are calculated for all 70 conditions. The switching frequency of the network (F) is calculated from equation (3.62), where n is the sum of number of instants either one of the protein switches from ON to OFF and OFF to ON states. The ON and OFF states of the proteins are determined so that if *protein 1* level is higher than *protein 2* at a given time instant, then *protein 1* is in ON state ($Prot^1_{ON}$) and *protein 2* is in OFF state ($Prot^2_{OFF}$) at that time instant. On the other hand, if the level of *protein 1* is lower than *protein 2* at a given

time instant, then the *protein 2* is in ON state ($Prot_{ON}^{2}$) and *protein 1* is in OFF state ($Prot_{OFF}^{1}$) at that time instant.

From these states of the proteins, the number of switches (n) over the time series is calculated by adding the number of instants the protein goes from ON to OFF state and the number of instants the promoter goes from OFF to ON states. This is done by first subtracting either *protein 1* from *protein 2* or *protein 2* from *protein 1*. Then the number of times the difference in proteins levels changes from positive to negative and negative to positive, is counted, which is the number of switches (n). This n also includes short transient switches due to its stochastic nature and these short transient noisy switches need to be filtered before the calculating n. For this, we assign a filter, which assigns the difference in protein levels as 0 if the absolute difference in protein levels is less than a threshold value of 100. After this filtering step, the n is calculated and from which we calculate the switching frequency (F) as:

$$F = \frac{n+1}{\Delta t} \tag{3.62}$$

Where, n is number of switches and Δt is observation time.

Next, to study the microscale dynamics of a toggle switch, we measured the cell-to-cell variability in protein levels of the 'dominant' (i.e. ON) and 'recessive' (i.e. OFF) genes. For this, we selected time windows where no transition between the states occurs. Then, for each of these genes, we obtained the mean and cell-to-cell variability in protein numbers of dominant and recessive genes. The measured mean and variability in protein levels are plotted as surface plots in Figure 18 of results section 4.3.1.

3.8.6 Simulation of 3-genes Repressilator

We hypothesize that the rate limiting steps in transcription initiation acts as a main regulating parameter of extrinsic noise in 3-gene Repressilator network, due to cell-to-cell variability in RNAp molecules. To prove our hypothesis, we created several models with varying $t_{prior}/\Delta t$, which are within the empirical range. We varied $t_{prior}/\Delta t$ by changing the rate constant values of only k_{cc} and k_{oc}, while maintaining the mean RNA production time (Δt) and other rate constant values as constant using equation (3.40) as explained in section 3.8.4. To each of these models, we introduced different levels of cell-to-cell variability in RNAp, which are within the empirical range.

From the empirical range of $t_{prior}/\Delta t$, shown in Table 2, we selected 10 values between 0.05 and 0.95, with an increment of 0.1. Also, from the empirical range of CV^2 in RNAp shown in Table 4, we select 7 values between 0 and 0.09, with an increment of 0.015. So, we have 70 different model conditions. Each model condition is simulated for 10 cells, with a simulation time of 5×10^5s and the protein numbers are sampled at every 10 s.

The transcriptional parameters of individual genes in all the network models are obtained from empirically validated data, shown in Table 1, except for k_{cc} and k_{oc}, where these 2 parameters are varied in accordance with transcription initiation conditions. In addition, the translational parameters of genes in the network model are listed in Table 3. For simplicity of the model, we modelled it as symmetric oscillator circuit, in which all the genes have the rate constant values.

The simulations are carried out in SGNS2 [8]. After simulations, the protein time series of all the 3 proteins are extracted and with either one of these protein time series, the cell-to-cell mean, and CV^2 of period of oscillation is calculated for all 70 model conditions. Since all genes in the network have the same transcription and translation parameters, it is expected that all its expressed proteins should have almost same period of oscillation.

Having said that, we measured the period of oscillation from the autocorrelation measurements of one of the proteins time series. In general, the time interval between two consecutive peaks is considered to be the period of oscillation. In this thesis, we calculated it in different way. First, one of the protein time series of a model cell with a predefined sampling interval (*ts*) are extracted from the stochastic simulation. Second, from that protein time series, we calculate the autocorrelation measure (A) with time lag (τ). Third, A is subtracted with its median value. Fourth, the time lag moments (τ_i), where A intersects zero axis is identified using interpolation of the nearest positive and negative values to the zero axis. From this we find the series of time lag moments (τ_i), $1 \leq i \leq NI$, where i be the index of the time moments, *NI* be total number of instants where the A intersects zero axis. Fifth, the distribution of oscillation periods (*P*) is calculated from the relation $P = \text{ind}(i+2) - \text{ind}(i)$, $1 \leq i \leq NI-2$. Sixth, the distribution of oscillation periods obtained from all the model cells of a particular condition are concatenated and we get its combined distribution. Next, the mean and CV^2 of the distribution is calculated. Similarly, the mean and CV^2 of distribution of period of oscillations for all the model conditions are calculated.

4. RESULTS AND CONCLUSIONS

4.1 Cell-to-cell variability in RNAp

The diversity in RNAp numbers over the cell population is obtained from cell-to-cell fluorescence intensity measurements in *E. coli* cells with fluorescently tagged β'subunits, reported in [6]. From the distribution of RNAp fluorescence intensity of cells, fluorescent intensities which far away from the mean are considered it to be outliers and are discarded. The trimmed distribution of fluorescence intensities is then normalized with its mean, such that relative mean RNAp numbers is obtained as 1. Next, to measure the variability of RNAp, a normal distribution curve is fitted on the relative RNAp fluorescence intensity distribution as shown in Figure 15. The CV^2 of the best fitting curve on the distribution is measured.

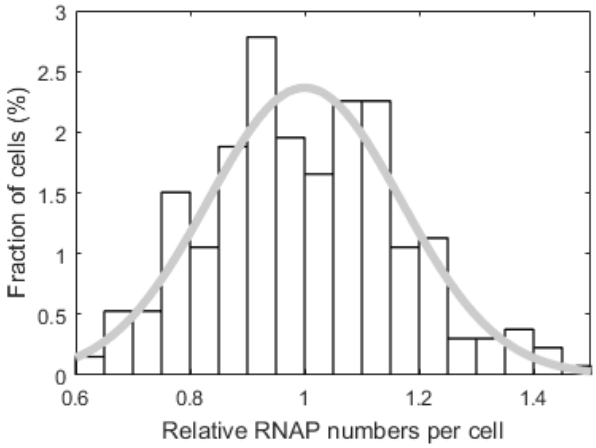

Figure 15. *Relative RNAp fluorescence intensity distribution of E. coli cells with fluorescently tagged β' subunits measured by microscopy [1]. The mean of the distribution is set as 1. Also shown is the best-fitted normal distribution curve (grey).*

By measuring it for different promoters under different conditions, we observed that the RNAp variability differed between conditions. These results are listed in Table 4. To measure the closeness of the fitting with the empirical distribution, we did a Kolmogorov-Smirnov (KS) test. If the p-values of the fitting are above 0.01, it is assumed that the two distributions cannot be distinguished in a statistical sense. The CV^2 and the p-value of the fit of the promoters under different conditions are listed in Table 4.

Table 4. *Squared coefficient of variance of RNAp (CV^2 of RNAp) measured for different promoters at different conditions.*

Condition	CV^2(RNAp)	KS test (p value)

3X	0.074	0.130
0.5 X	0.069	0.0145
2X	0.050	0.075
1X	0.047	0.0116
1.5 X	0.039	0.0107
Acidic stress + 1X	0.024	0.8484
Acidic stress + 0.25 X	0.021	0.0373
Oxidative stress + 0.75 X	0.017	0.2702

4.2 Cell-to-cell variability in RNA

To study the cell-to-cell variability in RNA production as a function of $t_{prior}/\Delta t$, we carried out the simulation as explained in the section 3.8.4. From the simulation results, we generated a surface plot showing the CV^2 of RNA for each model condition, which is shown in Figure 16.

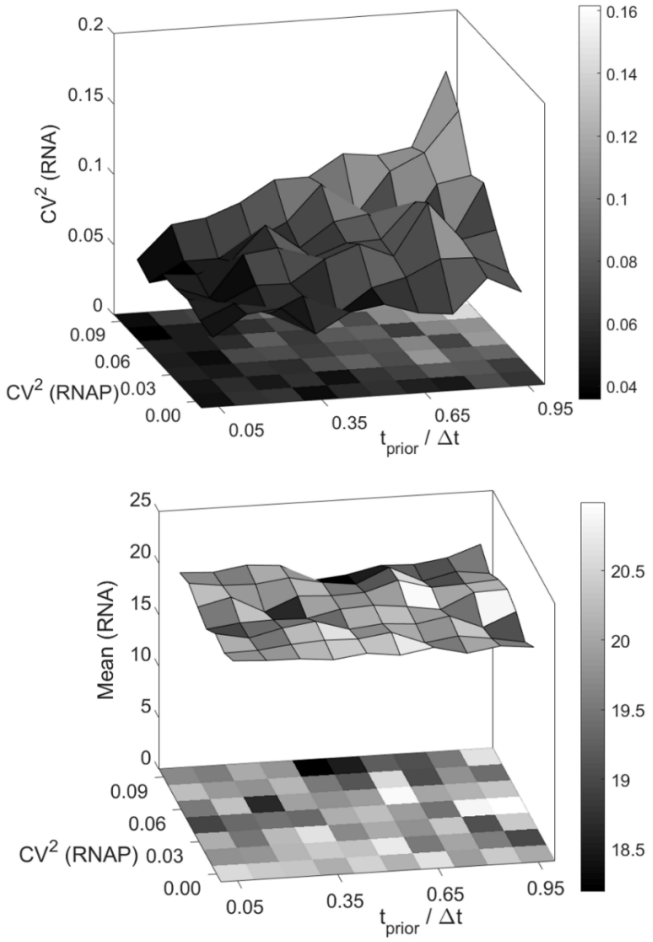

Figure 16. *Mean and Squared coefficient of variance (CV^2) of number of produced RNAs in model cells during their lifetime as a function of relative duration of the time spent in the steps prior to initiation of the open complex formation and of the cell-to-cell variability in RNAp numbers.*

From Figure 16, the cell-to-cell variability in RNA increases with increase in $t_{prior}/\Delta t$, by maintaining same RNAp variability condition. Similarly, the increase in cell-to-cell variability in RNAp increases the cell-to-cell variability in RNA, considering same $t_{prior}/\Delta t$ condition. By increasing both RNAp variability and $t_{prior}/\Delta t$, the CV^2 of RNA increased even to a greater extent in comparison with varying only one parameter and keeping the other parameter constant. But the mean RNA production does not vary with $t_{prior}/\Delta t$ and CV^2 of RNAp.

From the above results, we conclude that the cell-to-cell variability in RNAp numbers propagates to the cell-to-cell diversity in RNA and the extent to which this variability affects the variability in RNA production depends upon the transcription initiation kinetics.

4.3 Toggle switch

From the simulations of a 2-genes toggle switch, the time-series of proteins of individual genes in the network in the control conditions is shown in Figure 17.

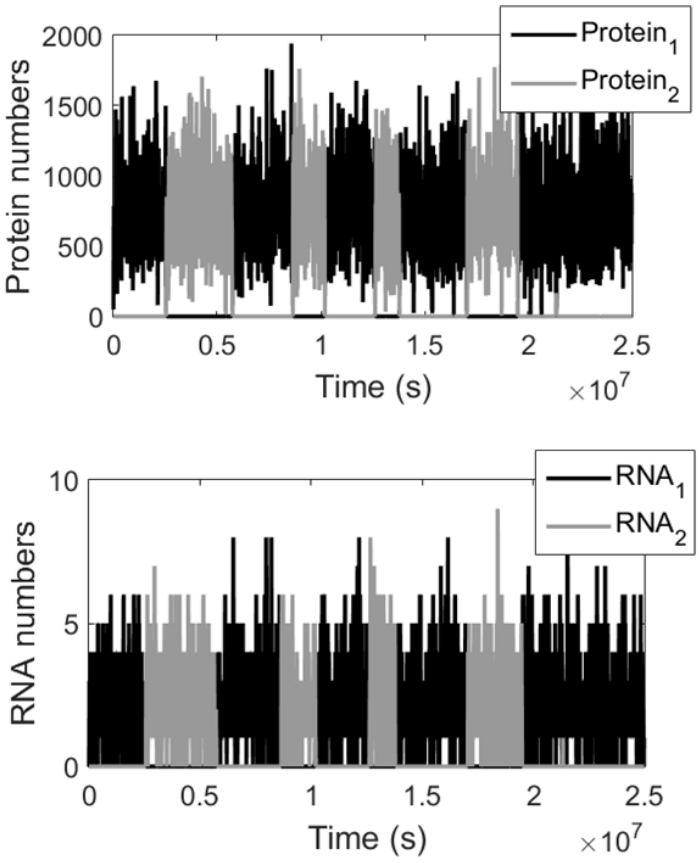

Figure 17. *Time series of protein (top) and RNA (bottom) number of a 2-gene toggle switch from a single stochastic simulation.*

4.3.1 $t_{after}/\Delta t$ acts as a tunable filter of cell-to-cell variability in RNAp numbers affects the toggle switch dynamics

We performed simulations for various values of $t_{prior}/\Delta t$, by the changing the parameters, k_{cc} and k_{oc} and CV^2 of RNAp as explained in the section 3.7. From the protein levels over time, the cell-to-cell mean and variability in switching frequency (F) for each condition is measured, whose surface plot is shown in Figure 18.

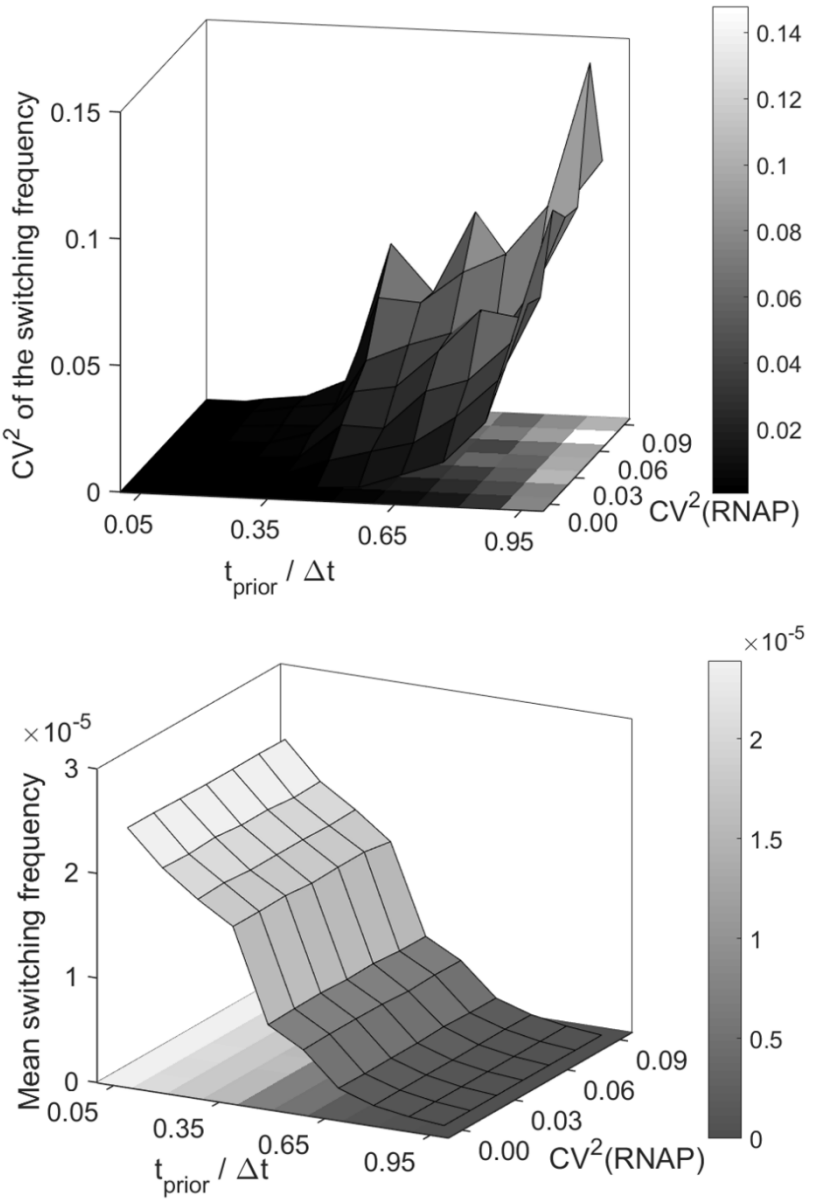

Figure 18. *Cell-to-cell mean (bottom) and variability (CV^2) (top) of switching frequency as a function of $t_{prior}/\Delta t$ and CV^2 (RNAp). 100 independent cells per condition.*

From Figure 18, one can see that the mean switching frequency (F) is greatly affected by the $t_{prior}/\Delta t$. Increase in $t_{prior}/\Delta t$, decrease the mean switching frequency of the switch. Also, the mean F is not affected by the various degrees of variability in RNAp numbers.

As expected, the CV^2 (RNAp) should not affect the mean behavior of the population and it should affect only the variability in the behavior of the population. In addition, one can also observe that CV^2 of switching frequency is low when the $t_{prior}/\Delta t$ is less and the cell-to-cell variability in RNAp numbers does not have much effect in those cases. And, when the $t_{prior}/\Delta t$ is larger than ~0.35, the CV^2 of switching frequency increases with increase in RNAp variability.

4.3.2 Micro-scale dynamics of the switch is controlled by $t_{prior}/\Delta t$

We measure the cell-to-cell variability in protein levels at 1 millionth time instant when the gene is in 'ON' and 'OFF' state separately as explained in chapter 3.8.5 and their respective results are represented as surface plots in Figure 19 and Figure 20.

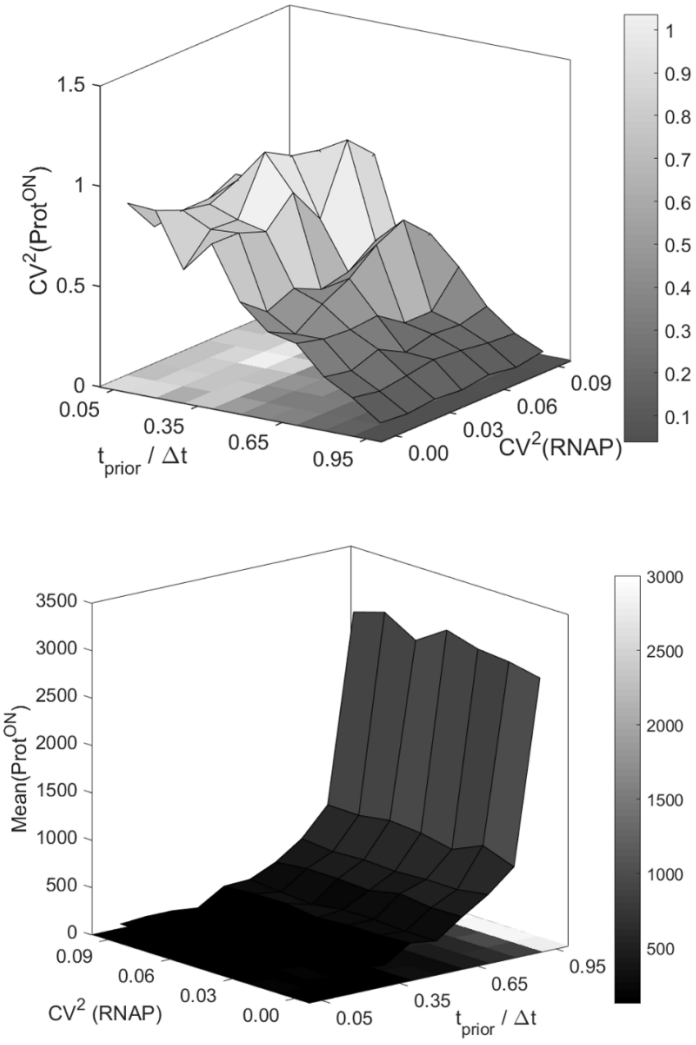

Figure 19. *Cell-to-cell mean (bottom) and diversity (top) in protein numbers in ON state at a given point in time (CV^2 ($Prot^{ON}$)), as a function of $t_{prior}/\Delta t$ and CV^2 of RNAp. 100 independent cells per condition.*

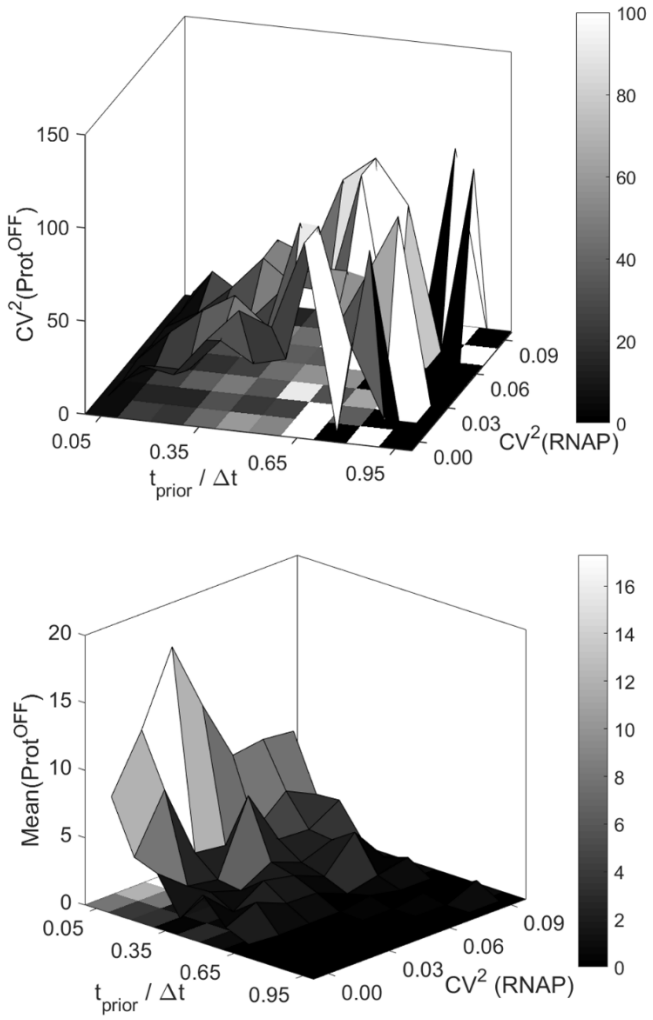

Figure 20. *Cell-to-cell mean (bottom) and diversity (top) in protein numbers in OFF state at a given point in time (CV^2 ($Prot^{OFF}$)), as a function of $t_{prior}/\Delta t$ and CV^2 of RNAp. 100 independent cells per condition.*

From Figure 19, we find that the mean protein level in ON state increases with increase in $t_{prior}/\Delta t$ and the CV^2 of protein level in 'ON' state decreases with increase in $t_{prior}/\Delta t$. From Figure 20, we found that mean protein level in OFF state decreases with increasing $t_{prior}/\Delta t$ and the CV^2 of protein level in 'OFF' state increases with increasing $t_{prior}/\Delta t$ ratios. Having said that, it is apparent that the CV^2 of protein level in 'ON' and 'OFF' states do not depend upon the CV^2 of RNAp. To confirm this further, the Pearson correlation coefficient is calculated between CV^2 of RNAp and the CV^2 of protein level when the gene is in ON and OFF states (listed in Table 5 and Table 6).

Table 5. *Pearson correlation coefficient between CV^2 of RNAp and the CV^2 of proteins in the ON state as a function of $t_{prior}/\Delta t$.*

$t_{prior}/\Delta t$	**Pearson correlation coefficient between $CV^2(Prot^{ON})$ and CV^2(RNAp)**	**p value**
0.05	-0.553	0.198

0.15	-0.650	0.114
0.25	0.404	0.369
0.35	0.495	0.258
0.45	-0.066	0.888
0.55	0.830	0.021
0.65	-0.202	0.664
0.75	-0.087	0.853
0.85	-0.228	0.622
0.95	0.285	0.536

Table 6. *Pearson correlation coefficient between the CV^2 of RNAp and the CV^2 of proteins in the OFF state as a function of $t_{prior}/\Delta t$.*

$t_{prior}/\Delta t$	**Pearson correlation coefficient between $CV^2(Prot^{ON})$ and $CV^2(RNAp)$**	**p value**
0.05	-0.553	0.198
0.15	-0.650	0.114
0.25	0.404	0.369
0.35	0.495	0.258
0.45	-0.066	0.888
0.55	0.830	0.021
0.65	-0.202	0.664
0.75	-0.087	0.853
0.85	-0.228	0.622
0.95	0.285	0.536

From Table 5 and Table 6, it is clear that there is no correlation between the degree of RNAp variability and variability in protein levels whether the gene is active or repressed. Hence, it is concluded that the microscale dynamics, the cell-to-cell variability in proteins levels in 'ON' and 'OFF' states only depends on the transcription initiation kinetics and not on the variability in RNAp numbers.

4.4 Repressilator

From the simulations of a Repressilator network, the time-series of proteins and RNA numbers of its individual genes at control conditions is shown in Figure 21, and the results of the model from 100 simulations are presented in Figure 22.

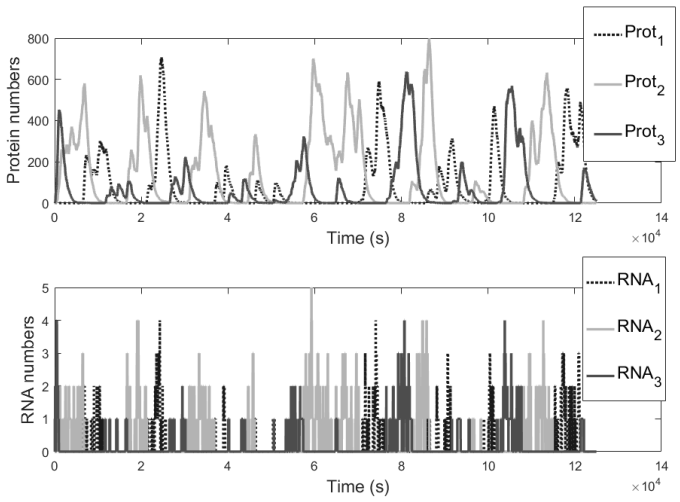

Figure 21. *Time series of protein (top) and RNA (bottom) number of a 3-gene Repressilator from a single stochastic simulation.*

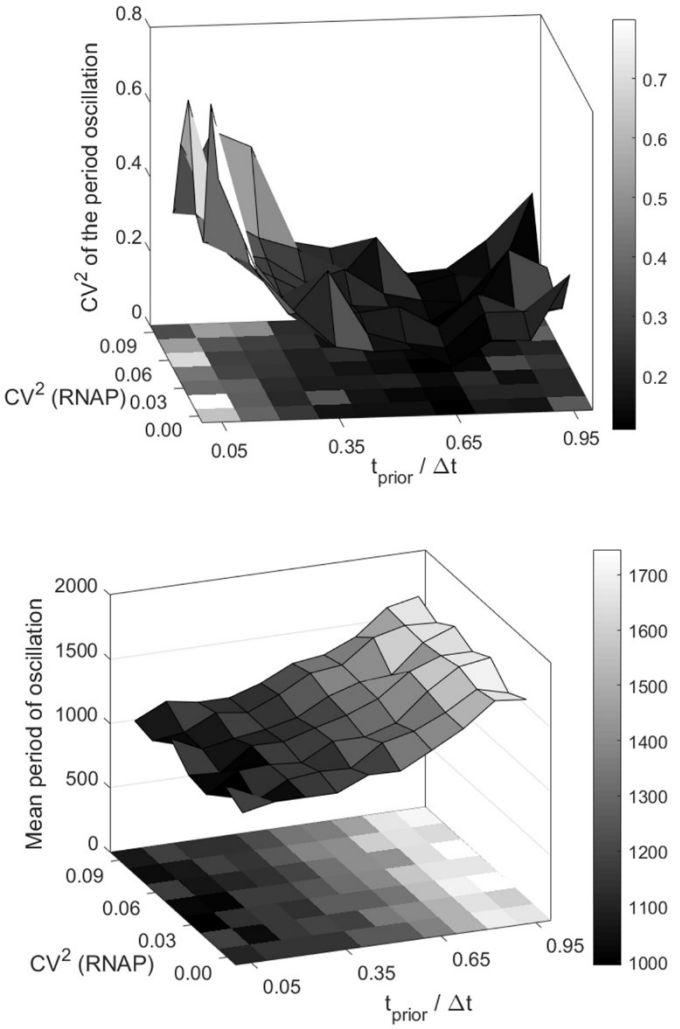

Figure 22. *Cell-to-cell mean (bottom) and diversity (CV^2) (top) of the period of oscillation of Repressilator as a function of $t_{prior}/\Delta t$ and CV^2 of RNAp.*

First, from Figure 22, it is seen that the mean period of oscillation increases with $t_{prior}/\Delta t$ as expected.

Second, The CV^2 of period of oscillation is significantly changed, while varying the $t_{prior}/\Delta t$ of individual genes in the network. The trend is quite surprising that it is high when $t_{prior}/\Delta t$ is low, then it decreases gradually until it reaches ~0.50 and beyond 0.50, it starts to increase again. Also, it is clearly visible, that there is no trend or correlation between CV^2 of period of oscillation and CV^2 of RNAp number.

Hence, it is concluded that cell-to-cell mean and variability in period of oscillation depend only upon the relative duration of rate limiting steps in transcription initiation. And they do not depend upon the cell-to-cell variability in RNAp numbers.

5. DISCUSSION AND FUTURE WORKS

In section 4.2, we explored how the cell-to-cell variability in RNAp numbers (CV^2 of RNAp) affects the cell-to-cell variability in RNA production rates, as a function of transcription initiation kinetics of the promoter ($t_{prior}/\Delta t$). For that, we made use of an empirically-validated model that accounts for a realistic range of values for kinetics rates (Table 1, Table 2 and Table 3) and RNAp variability (Table 4), the latter, in particular, extracted from various environment conditions in this work (Figure 15).

The simulations of the single-gene model predict that the mean RNA production does not vary with relative duration of $t_{prior}/\Delta t$ Figure 16 (bottom). This is because all transcription models are constructed by varying $t_{prior}/\Delta t$ ratios, while keeping the mean RNA production time as constant. In addition, the results of Figure 16 (bottom) show that the mean RNA production is unaffected by CV^2 of RNAp. This is expected because the mean should not vary solely due to the variability in RNAp numbers. Also from Figure 16 (top), the degree at which CV^2 of RNAp affects cell-to-cell variability in RNA numbers (CV^2 of RNA) depends upon the relative duration of rate liming steps in transcription initiation ($t_{prior}/\Delta t$ ratios). Finally, from these results (section 4.2), we showed that CV^2 of RNA increases with both increasing $t_{prior}/\Delta t$ ratios and CV^2 of RNAp. This occurs due to the influence that RNAp numbers have in the closed complex formation step in transcription initiation process. i.e., the higher $t_{prior}/\Delta t$ ratio, the more time the promoter has to spend to form closed complex, which amplifies the CV^2 of RNA due to the noise in RNAp numbers. In conclusion, based on the above arguments, it is inferred that the relative duration of rate limiting steps in transcription initiation acts as a key regulatory parameter of the effects of extrinsic noise (i.e. (RNAp variability) on the RNA variability across cell population.

In the simulation results of the 2-genes toggle switch model (presented in section 4.3), we showed that the mean switching frequency of the network (F) depends solely upon the transcription initiation of individual genes ($t_{prior}/\Delta t$) (Figure 18, bottom), as it is barely affected by CV^2 of RNAp. In that, the mean switching frequency (F) decreases with increasing $t_{prior}/\Delta t$ ratio due to the increasing stability of the system. This occurs because the switching of states of the genes, in this network, depends upon the degree of repression exerted by one gene over the other. Namely, if the relative repression of one gene is high, the switch stability will be also high, and the mean switching frequency (F) low.

Interestingly, this degree of repression is mainly dependent upon two parameters, the number of repressor molecules in the system, and the likelihood of repressor proteins to bind with the promoters. These two parameters can be tuned by varying the relative duration of rate liming steps in transcription initiation ($t_{prior}/\Delta t$) of the component genes of the network.

In the case of the first parameter, the number of repressor molecules in the system increases with increasing $t_{prior}/\Delta t$. It is shown in section 4.3.2 (Figure 18, bottom) that, the mean protein levels of the 'dominant' gene (i.e. the one expressing higher protein levels at a given 'time window') increases with increasing $t_{prior}/\Delta t$ ratio. This is possible because, first, a promoter with high $t_{prior}/\Delta t$ ratio is able to produce higher protein levels in a short period of time, in comparison to low ratio. Thereby, the dominant gene produces an increasing number of repressor molecules, thus repressing the recessive gene more effectively, which increases the stability of the switch (i.e. less switching of states from many cell simulations). Interestingly, as a consequence of this, in a toggle switch with high $t_{prior}/\Delta t$ promoters, the protein level expressed by a 'dominant' gene is proportional to the repression strength realized by the 'recessive' gene.

In the case of the second parameter, the likeliness of repressor molecules to bind to free promoters greatly depends on the relative duration of rate liming steps in transcription initiation ($t_{prior}/\Delta t$). Namely, if the promoter has a small $t_{prior}/\Delta t$ ratio, the promoter spends a relatively small time in the closed complex formation step, i.e. either bound by a repressor molecule or in forming the RNAp closed complex after the first RNAp molecule finds a free promoter. With a relatively large time spent in the open complex formation, the probability of repressor proteins or new RNAp molecules to bind to a free promoter is reduced. On the other hand, e.g., if the promoter has a large $t_{prior}/\Delta t$ ratio, a relatively longer closed complex formation, i.e. the time that it takes for a repressor to free the promoter region and the first RNAp to bind to it, the probability of which a repressor protein can bind to a free promoter is increased.

To conclude, in the first part of section 4.3.2, it is found that the mean protein level of dominant gene is getting increased with increase in $t_{prior}/\Delta t$, on the other hand that the mean protein level of recessive gene is getting decreased with increases in $t_{prior}/\Delta t$. This behavior is due to increase in repression effect of the dominant gene on the recessive gene. The increase in repression effect is due to increase in relative duration of repression time. If the dominant protein (i.e. expressed by the dominant gene) binds more times with the recessive promoter, the protein expression of recessive gene goes low. If the protein level of recessive gene is less in the system, then its repression effect on the dominant gene becomes negligible. This allows the dominant gene to express its proteins freely, due to which its protein levels goes higher with increase in $t_{prior}/\Delta t$. Therefore, the mean protein level of dominant and recessive genes in the network depend only upon the transcription initiation kinetics. And they do not depend upon the RNAp variability over the cell population.

In addition, from the last part of section 4.3.1, in Figure 18 (top), it is shown that the cell-to-cell variability in switching frequency (CV^2 of F) of the toggle switch becomes greatly dependent on transcription initiation kinetics ($t_{prior}/\Delta t$) of the component genes, and only slightly dependent on the CV^2 of RNAp when $t_{prior}/\Delta t$ ratio of the genes is high. When $t_{prior}/\Delta t$ ratios are below 0.5, the network is barely affected by CV^2 of RNAp, presenting

very low CV^2 of F. On the other hand, at higher values of $t_{prior}/\Delta t$, the network presents high CV^2 of F values. There are two possible explanations for that.

First, similarly to the previous discussion, the explanation for that follows from the relationship between the switching stability and its dependence on the $t_{prior}/\Delta t$ ratios of the genes. The increase in CV^2 of RNAp at high $t_{prior}/\Delta t$ ratios regime is proportional to the fold changes observed in the $t_{prior}/\Delta t$ ratios, because when $t_{prior}/\Delta t$ ratio becomes, if not only, the most important rate-limiting step in transcription, the quantity of RNAp numbers becomes directly proportional to the production rate of mRNAs, increasing the effect of proteins/repressors in the system. From that, the variability in RNAp numbers will also greatly affect the noise and duration of the closed complex formation. Namely, if the closed duration of complex formation is slow, the effect of RNAp variability becomes more significant, and vice-versa. Thus, considering this relationship, and the results observed presented in Figure 18 (top), we expected that, for this network, the effect of CV^2 of RNAp on the CV^2 of F increases with increasing $t_{prior}/\Delta t$ ratio of the component genes.

Secondly, the increase of CV^2 of F with increasing $t_{prior}/\Delta t$ ratio of the component genes can also be explained from the interpretation of the equation used to calculate CV^2 of F and the mean and standard deviation of switching frequency measured from the simulations of the switch. In that, if CV^2 is the squared ratio of the standard deviation of F over the mean of F, from a cell population, then any increase in the CV^2 can only be created by: (i) a decrease in the mean, while the standard deviation is constant, or (ii) an increase in the standard deviation, while the mean remains constant, or even (iii) an increase in the standard deviation along with a decrease in the mean. Therefore, given that the mean switching frequency (F) decreases (Figure 18, bottom), the increase in the CV^2 of F seen in Figure 18 (top) can only be explained by (iii), in that the standard deviation of F increases when $t_{prior}/\Delta t$ ratio and cell-to-cell RNAp variability increase.

Overall, from section 4.3 results, it is possible to conclude that the CV^2 of protein levels of dominant and recessive gene is only dependent upon the relative duration of transcription initiation kinetics of genes ($t_{prior}/\Delta t$) in the network and they do not depend upon the cell-to-cell RNA variability (CV^2 of RNAp). The CV^2 of dominant protein decreases with increase in $t_{prior}/\Delta t$ and it might be mainly due to increase in its mean value with increase in $t_{prior}/\Delta t$. On the contrary, the CV^2 of recessive protein increases with increase in $t_{prior}/\Delta t$ and it might be mainly due to decrease in its mean value with increase in $t_{prior}/\Delta t$. Finally, from Table 5 and Table 6, we found no correlation between the degree of CV^2 of RNAp and the microscale dynamics of the switch (CV^2 of F), hence, it is concluded that the cell to cell variability in proteins levels, when ON or OFF states, only depends on $t_{prior}/\Delta t$ and does not vary as a function of CV^2 of RNAp.

In the section 4.4, we presented our study on how much the dynamics of a 3-genes Repressilator model is, again, affected by both the transcription initiation kinetics of its genes ($t_{prior}/\Delta t$) and cell-to-cell RNAp variability (CV^2 of RNAp). The simulation results

show that the mean period of oscillation increases with increasing $t_{prior}/\Delta t$, and this is due to an increase in the network stability, similar to the reasons presented above to explain the results observed for the switch dynamics. Namely, due to an increasing repression strength employed by one gene over the other, generated by the increasing $t_{prior}/\Delta t$ ratio of the promoters of some of the models tested. Again, similar to the results obtained from the toggle switch dynamics (section 4.3), first, the results of the 3-genes Repressilator clearly shows that the mean period of oscillation is unaffected by CV^2 of RNAp. Second, the CV^2 of the period of oscillation is higher at low values of $t_{prior}/\Delta t$, gradually decreases when $t_{prior}/\Delta t$ ratio changes from low to ~0.50, but increases again when ratio goes beyond 0.50. Thus, as one can observe from the shape of the curve in Figure 22 (top), the cell-to-cell variability in oscillation period is minimized at around $t_{prior}/\Delta t \sim 0.50$, creating an optimal minimal point for this type of network, when the $t_{prior}/\Delta t$ ratio of the genes are symmetrically close to 0.5.

Again, similar to the toggle switch dynamics, we believe that the reason for high CV^2 of period of oscillation of promoters having low $t_{prior}/\Delta t$ ratios is due to the fact that lesser deviation in standard deviation in period of oscillations, when the system produces period of short durations (i.e. small values). Further, the CV^2 of period of oscillation is barely affected by the CV^2 of RNAp (Figure 22, top), which demonstrated that the noise from CV^2 of RNAp can be regulated by the network in a hyperbolical manner, whereas the toggle switch could instead amplify it, in which its magnitude depends upon the $t_{prior}/\Delta t$ ratios. Therefore, it is quite interesting how the rate limiting steps in transcription initiation of the component genes of a network can act as different regulator of its dynamics, depending on the topology of the network.

Finally, in this thesis work, we considered that all the individual genes in the network have the same transcription initiation kinetics. In the future, first, we plan to study the behavior of more realistic gene network models of component genes having varying $t_{prior}/\Delta t$ ratios. Second, we plan to extend our study on how it may be possible to attain desired levels of noise in the macro dynamics of other genetic circuits, such as clocks and filters, by tuning the kinetics of transcription initiation of their component genes. Third, while in the thesis we considered CV^2 of RNAp and $t_{prior}/\Delta t$ as independent variables. In the future, we plan to check, empirically and with the support of models, whether there is any correlation between CV^2 of RNAp and $t_{prior}/\Delta t$ at various environmental conditions. Finally, we will create more realistic (i.e. empirically validated) models to study the effects of other cellular components such as activators, σ factors and other transcription factors, on the dynamics of small and large gene regulatory networks.

REFERENCES

[1] M. Jishage, A. Iwata, S. Ueda and A. Ishihama, "Regulation of RNA polymerase sigma sub-unit synthesis in Escherichia coli: intracellular levels of four species of sigma subunit under various growth conditions," *J. Bacteriol,* vol. 178, p. 5447–51, 1996.

[2] M. Rahman, M. R. Hasan, T. Oba and K. Shimizu, "Effect of rpoS gene knockout on the metabolism of Escherichia coli during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations," *Biotechnol. Bioeng,* vol. 94, p. 585–95, 2006.

[3] A. Farewell, K. Kvint and T. Nyström, "Negative regulation by RpoS: A case of sigma factor competition," *Mol. Microbiol,* vol. 29, p. 1039–51, 1998.

[4] R. Hengge-Aronis, "Recent insights into the general stress response regulatory network in Escherichia coli," *J. Mol. Microbiol. Biotechnol,* vol. 4, p. 341–346, 2002.

[5] W. R. McClure, "Rate-limiting steps in RNA chain initiation," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 77, p. 5634–5638, 1980.

[6] J. Lloyd-Price, S. Startceva, V. Kandavalli, J. Chandraseelan, N. Goncalves, S. M. Oliveira, A. Häkkinen and A. S. Ribeiro, "Dissecting the stochastic transcription initiation process in live Escherichia coli," *DNA Research,* vol. 23, no. 3, pp. 203-214, 2016.

[7] V. K. Kandavalli, H. Tran and A. S. Ribeiro, "Effects of σ factor competition on the in vivo kinetics of transcription initiation in Escherichia coli," *BBA Gene Regulatory Mechanisms,* vol. 1859, p. 1281–1288, 2016.

[8] J. Lloyd-Price, A. Gupta and A. S. Ribeiro, "SGNS2: A Compartmentalized Stochastic Chemical Kinetics Simulator for Dynamic Cell Populations," *Bioinformatics,* vol. 28, pp. 3004-3005, 2012.

[9] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem,* vol. 81, no. 25, p. 2340–2361, 1977.

[10] M. Bahrudeen, S. Startceva and A. Ribeiro, "Effects of extrinsic noise are promoter kinetics dependent," 2017.

[11] A. Arkin, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage Lambda-Infected Escherichia coli Cells," *Genetics,* vol. 149, no. 4, pp. 1633-1648, 1998.

[12] A. S. Ribeiro, "Effects of coupling strength and space on the dynamics of coupled toggle switches in stochastic gene networks with multiple-delayed reactions," *Phys. Rev. E,* vol. 75, no. 6, p. 061903, 2007a.

[13] A. S. Ribeiro, "Dynamics of a two-dimensional model of cell tissues with coupled stochastic gene networks," *Phys. Rev. E,* vol. 76, no. 5, p. 051915, 2007b.

[14] M. Bahrudeen and A. Ribeiro, "Tuning extrinsic noise effects on a small genetic circuit," 2017.

[15] S. Oliveira, M. Bahrudeen, S. Startceva and A. Ribeiro, "Estimating the multi-scale effects of extrinsic noise on genes and circuits activity from an empirically validated model of transcription kinetics," 2017.

[16] B. Alberts, A. Johnson and J. Lewis, "The Structure and Function of DNA," in *Molecular Biology of the Cell. 4th edition*, New York, Garland Science, 2002.

[17] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, Molecular biology of the cell, 5 ed., USA: Garland Science, 2002.

[18] I. Golding and E. C. Cox, "RNA dynamics in live Escherichia coli cells," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 101, no. 31, p. 11310–11315, 2004.

[19] K. M. Arndt and M. J. Chamberlin, "Transcription termination in Escherichia coli. Measurement of the rate of enzyme release from Rho-independent terminators," *J Mol Biol,* vol. 202, p. 271–285, 1988.

[20] R. Young and H. Bremer, "Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate," *Biochem J,* vol. 160, p. 185–194, 1976.

[21] P. P. Dennis and H. Bremer, "Differential rate of ribosomal protein synthesis in Escherichia coli B/r," *J Mol Biol,* vol. 84, p. 407–422, 1974.

[22] W. R. McClure, "Mechanism and control of transcription initiation in prokaryotes," *Ann Rev Biochem,* vol. 54, pp. 171-204, 1985.

[23] A. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, Molecular Biology of the Cell, 5 ed., New York: Garland Science, 2007.

[24] J. Müller, S. Oehler and B. Müller-Hill, "Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator," *J. Mol.Biol.,* vol. 257, no. 1, pp. 21-29, 1996.

[25] R. Schleif, "Regulation of the l-arabinose operon of Escherichia coli," *Trends in Genetics,* vol. 16, no. 12, pp. 559-565, 2000.

[26] D. M. Wolf and A. P. Arkin, "Motifs, modules and games in bacteria," *Curr.Opin. Microbiol.,* vol. 6, p. 125–134, 2003.

[27] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators.," *Nature,* vol. 403, no. 6767, p. 335–338, 2000.

[28] C. Cheng, K. K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, M. Snyder and M. Gerstein, "Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.," *PLoS Comput. Biol.,* vol. 7, no. 11, p. e1002190, 2011.

[29] A. de la Fuente, P. Brazhnik and P. Mendes, "Linking the genes: inferring quantitative gene networks from microarray data.," *Trends in Genetics,* vol. 18, no. 8, pp. 395-398, 2002.

[30] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," *Natl. Acad. Sci. U. S. A.,* vol. 94, no. 3, p. 814–819, 1997.

[31] M. B. Elowitz, A. J. Levine, E. D. Siggia and P. S. Swain, "Stochastic gene expression in a single cell," *Science,* vol. 297, no. 5584, p. 1183 – 1186, 2002.

[32] Z. Neubauer and E. Calef, "Immunity phase-shift in defective lysogens: Non-mutational hereditary change of early regulation of prophage," *Journal of Molecular Biology,* vol. 51, no. 1, p. 1 – 13, 1970.

[33] M. Kaern, T. C. Elston, W. J. Blake and J. J. Collins, "Stochasticity in gene expression: from theories to phenotypes.," *Nature Reviews Genetics,* vol. 6, no. 6, p. 451 –464, 2005.

[34] J. Paulsson, "Models of stochastic gene expression," *Phys Life Rev,* vol. 2, no. 2, pp. 157-175, 2005.

[35] J. Paulsson, "Summing up the noise in gene networks," *Nature,* vol. 29, p. 415–418, 2004.

[36] J. Paulsson and M. Ehrenberg, "Noise in a minimal regulatory network: plasmid copy number control," *Quarterly reviews of biophysics,* vol. 34, no. 1, p. 1 – 59, 2001.

[37] S. Leibler and E. Kussell, "Individual histories and selection in heterogeneous populations," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 29, p. 13183 – 13188, 2010.

[38] T. M. Norman, N. D. Lord, J. Paulsson and R. Losick, "Stochastic Switching of Cell Fate in Microbes," *Annual Review of Microbiology,* vol. 69, p. 381 – 403, 2015.

[39] A. Gupta, R. Lloyd-Price and R. Neeli-Venkata, "In vivo kinetics of segregation and polar retention of MS2-GFP-RNA complexes in Escherichia coli," *Biophys. J,* vol. 106, no. 9, p. 1928–1937, 2014.

[40] K. S. Murakami, S. Masuda and E. A. Campbell, "Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex," *Science,* vol. 296, no. 5571, p. 1285–1290, 2002.

[41] V. Mekler, E. Kortkhonjia and J. Mukhopadhyay, "Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex," *Cell,* vol. 108, no. 5, p. 599–614, 2002.

[42] M. L. Craig, W. C. Suh and M. T. Record, "HO. and DNase I probing of E sigma 70 RNA polymerase--lambda PR promoter open complexes: Mg2+ binding and its structural consequences at the transcription start site," *Biochemistry,* vol. 34, no. 48, p. 15624–15632, 1995.

[43] S. O. Skinner, L. A. Sepúlveda and H. Xu, "Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization," *Nat. Protoc,* vol. 8, no. 6, p. 1100–1113, 2013.

[44] O. Shimomura, "Structure of the chromophore of Aequorea green fluorescent protein," *FEBS Lett,* vol. 104, no. 2, p. 220–222, 1979.

[45] M. W. Davidson and R. E. Campbell, "Engineered fluorescent proteins: innovations and applications," *Nat. Methods,* vol. 6, no. 10, p. 713–717, 2009.

[46] E. Bertrand, . P. Chartrand and M. Schaefer, "Localization of ASH1 mRNA Particles in Living Yeast," *Mol. Cell,* vol. 2, no. 4, p. 437–445, 1998.

[47] I. Golding, J. Paulsson and S. M. Zawilski, "Real-time kinetics of gene activity in individual bacteria," *Cell,* vol. 123, no. 6, p. 1025–1036, 2005.

[48] A. Hakkinen, A. B. Muthukrishnan and A. Mora, "CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli," *Bioinformatics,* vol. 29, no. 13, p. 1708–1709, 2013.

[49] P. Ruusuvuori, T. Aijo and S. Chowdhury, "Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images," *BMC Bioinformatics,* vol. 11, no. 1, p. 248, 2010.

[50] T. B. Chen, H. H. S. Lu and Y. S. Lee, "Segmentation of cDNA microarray images by kernel density estimation," *J. Biomed. Inform,* vol. 41, no. 6, p. 1021–1027, 2008.

[51] M. R. Roussel and R. Zhu, "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression," *Phys Biol,* vol. 3, no. 4, pp. 274-284, 2006.

[52] A. S. Ribeiro, R. Zhu and S. A. Kauffman, "A General Modeling Strategy for Gene Regulatory Networks with Stochastic Dynamics," *J. of Comput. Biol,* vol. 13, pp. 1630-1639, 2006.

[53] S. Strickland, G. Palmer and V. Massey, "Determination of dissociation constants and specific rate constants of enzyme-substrate (or protein-ligand) interactions from rapid reaction kinetic data," *J. Biol. Chem,* vol. 250, no. 11, p. 4048–4052, 1975.

[54] Y. Taniguchi, P. J. Choi and G. W. Li, "Qunatifyin E. coli proteome and transcriptome with single-molecule sensitivity in single cells," *Science,* vol. 329, pp. 533-538, 2010.

[55] J. A. Bernstein, A. B. Khodursky, L. Pei-Hsun, S. Lin-Chao and S. N. Cohen, "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,* vol. 99, p. 9697–9702, 2002.

[56] A. B. Muthukrishnan, A. Martikainen, R. Neeli-Venkata and A. S. Ribeiro, "In Vivo Transcription Kinetics of a Synthetic Gene Uninvolved in Stress-Response

Pathways in Stressed Escherichia coli Cells," *PLoS ONE,* vol. 9, no. 9, p. e109005, 2014.

[57] L. H. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev and I. Golding, "General properties of transcriptional time series in Escherichia coli," *Nat. Genet,* vol. 43, p. 554–560, 2011.

[58] R. M. Saecker, M. T. Record and P. L. Dehaseth, "Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis," *J. Mol. Biol,* vol. 412, p. 754–771, 2011.

[59] M. J. Chamberlin, "The selectivity of transcription," *Annu. Rev. Biochem,* vol. 43, p. 721–775, 1974.

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

[61] S. Chong, C. Chen, H. Ge and X. S. Xie, "Mechanism of transcriptional bursting in bacteria," *Cell,* vol. 158, p. 314–326, 2014.

[62] N. Mitarai, K. Sneppen and S. Pedersen, " Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization," *J. Mol. Biol,* vol. 382, no. 1, pp. 236-245, 2008.

[63] H. Bremer and P. P. Dennis, "Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate," in *Escherichia Coli and Salmonella: cellular and molecular biology*, vol. 2, 2, Ed., Washington DC, ASM Press, 1996, p. 1553–1569.

[64] D. Kennel and H. Riezman, "Transcription and translation initiation frequencies of the Escherichia coli lac operon," *J. Mol. Biol,* vol. 114, no. 1, pp. 1-21, 1977.

[65] B. P. Cormack, R. H. Valdivia and S. Falkow, "FACS-optimized mutants of the green fluorescent protein (GFP)," *Gene,* vol. 173, no. 1, pp. 33-38, 1996.

[66] P. L. deHaseth, M. L. Zupancic and M. T. Record, "RNA polymerasepromoter interactions: The comings and goings of RNA polymerase," *J. Bacteriol,* vol. 180, p. 3019–3025, 1998.

[67] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *J. Chem. Phys.,* vol. 124, p. 044104, 2006.

[68] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annual Review of Physical Chemistry,* vol. 58, pp. 35-55, 2007.

[69] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics,* vol. 22, no. 4, p. 403 – 434, 1976.

[70] X. S. Xie, P. J. Choi and G. W. Li, "Single-Molecule Approach to Molecular Biology in Living Bacterial Cells," *Annu. Rev. Biophys,* vol. 37, no. 1, p. 417–444, 2008.

[71] M. Kandhavelu, A. Häkkinen and O. Yli-Harja, "Single-molecule dynamics of transcription of the lar promoter," *Phys. Biol,* vol. 9, no. 2, p. 26004, 2012.

[72] A. B. Muthukrishnan, M. Kandhavelu and J. Lloyd-Price, "Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells," *Nucleic Acids Res,* vol. 40, p. 8472–8483, 2012.

[73] M. Kandhavelu, J. Lloyd-Price, A. Gupta, A. Muthukrishnan, O. Yli-Harja and A. S. Ribeiro, "Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter," *FEBS Lett,* vol. 586, p. 3870–3875, 2012.

[74] A. S. Ribeiro and S. A. Kauffman, "Noisy Attractors and Ergodic Sets in Models of Gene Regulatory Networks," *J. of Theor. Biol,* vol. 247, no. 4, pp. 743-755, 2007.