TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Jarno Mäkelä

On the Variability of the Transcription Process in *Escherichia coli*

Julkaisu 1407 • Publication 1407

Tampere 2016

Tampereen teknillinen yliopisto. Julkaisu 1407
Tampere University of Technology. Publication 1407

Jarno Mäkelä

On the Variability of the Transcription Process in *Escherichia coli*

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Sähkötalo Building, Auditorium SA203, at Tampere University of Technology, on the 9th of September 2016, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2016

Supervisor: Assoc. Prof. Andre S. Ribeiro
Tampere University of Technology
Finland

Pre-examiners: Prof. Jeff Gelles
Brandeis University
United States of America

Assist. Prof. Alvaro Sanchez
Yale University
United States of America

Opponent: Prof. Achillefs N. Kapanidis
University of Oxford
United Kingdom

ISBN 978-952-15-3794-3 (printed)
ISBN 978-952-15-3797-4 (PDF)
ISSN 1459-2045

Abstract

In all organisms, cellular functions, such as growth and differentiation, are coordinated by gene networks. These networks control both which genes are transcribed as well as when these events occur, based on intracellular and environmental information. Due to the often small number of specific regulatory molecules in the cell, stochastic fluctuations in molecular numbers tangibly affect the control of transcription. The stochasticity has consequences on the phenotype of the cell and the course of future cellular events. To obtain a detailed understanding of the dynamics of these processes, one must use techniques that allow observing individual events in time. Recent advancements in single-molecule detection techniques in live cells have made this possible and studies using these techniques are beginning to shed light on the functioning of cellular processes at a molecular level.

In this thesis, the dynamics of the multi-step transcription process in *Escherichia coli* was characterized using a combination of *in vivo* single-RNA detection techniques and single-nucleotide level stochastic models. Fluctuations at different stages of the transcription process and their propagation were investigated.

First, intake and transcription dynamics were investigated in different promoters and various induction schemes. Following the beginning of induction, waiting times for the first transcription event and the time intervals between consecutive ones were measured. The measurements were conducted using an MS2-GFP RNA tagging technique to detect single RNA molecules *in vivo*. To accurately measure the time moments when novel transcripts are produced, an automatic method for detecting non-spurious changes in time series data was developed. The stochasticity of the intake dynamics of inducers was found to be responsible for a large transient variability in RNA numbers that gradually vanishes, as the fluctuations from active transcription on the intracellular RNA numbers accumulate.

Next, contributions from the promoter dynamics and steps in transcription and translation elongation to fluctuations in RNA and protein numbers were studied. For this, stochastic single-nucleotide-level models to observe the dynamics of initiation at the promoter region and the dynamics of coupled transcription and translation elongation were constructed. In the closely spaced promoter regions,

interference between RNA polymerases was shown to affect the dynamics and create transient correlations in transcription initiations. During coupled elongation phases, the propagation of fluctuations from transcription to translation was shown to depend on both transcription and translation processes. For example, sequence-dependent transcriptional pauses were shown to affect simultaneously transcription and translation elongation. Together these findings suggest that the dynamics of transcript production is sensitive to the sequence-dependent mechanisms of initiation and elongation.

These results contribute to understanding how different sources of fluctuations contribute to the outcomes of gene expression. While the *in vivo* single-molecule detection techniques allow quantifying the fluctuations in principal components of the process at a molecular level, stochastic modeling contributes to the study by explaining how they fluctuate, as different mechanisms can give rise to similar behaviors. Combination of these methodologies will be crucial in future efforts for better understanding of biological systems.

Preface

This study was carried out at the Department of Signal Processing, Faculty of Computing and Electrical Engineering, Tampere University of Technology under the supervision of Associate Professor Andre S. Ribeiro.

First, I would like to express my sincere gratitude to my supervisor, Associate Professor Andre S. Ribeiro, for his persistent guidance and support during my doctoral studies. It has been an honor to be a member of your group and learn the ways of science along the way.

Next, I would like to thank all co-authors and members of the Laboratory of Biosystem Dynamics, including alumni. The LBD has been a fantastic working environment with great atmosphere. The discussions and knowledge from the interactions have made this thesis possible in the current form, holding fragments of each contribution in it. I would especially like to thank Jason Lloyd-Price and Antti Häkkinen for their contributions on the theoretical aspects, and Samuel Oliveira for the practical aspects.

I would also like to thank the TUT President's graduate programme for supporting my doctoral studies.

I am also grateful to the two pre-examiners, Jeff Gelles and Alvaro Sanchez, for their insightful suggestions which have helped improve this thesis.

Finally, I am deeply grateful to my family, for all their love and support during the years, which have made working on something like this possible.

Tampere, June 2016
Jarno Mäkelä

Contents

List of Abbreviations

CME	chemical master equation
DM	direct method
FISH	fluorescence *in situ* hybridization
FRM	first reaction method
FSP	finite state projection
GFP	green fluorescent protein
HILO	highly inclined and laminated optical sheet
KCpA	kernel change point analysis
KLIEP	Kullback-Leibler importance estimation procedure
LDM	logarithmic direct method
NRM	next reaction method
RNAp	RNA polymerase
RBS	ribosome binding site
SSA	stochastic simulation algorithm
TEC	transcription elongation complex
TF	transcription factor
TIRF	total internal reflection fluorescence
TMG	thiomethyl-β-D-galactoside
tmRNA	transfer-messenger RNA
uLSIF	unconstrained least-squares importance fitting

List of Publications

This thesis is a compilation of the following publications. In the text, these are referred to as **Publication I**, **Publication II** and so on.

I J. Mäkelä, H. Huttunen, M. Kandhavelu, O. Yli-Harja, and A.S. Ribeiro, "Automatic detection of changes in the dynamics of delayed stochastic gene networks and *in vivo* production of RNA molecules in *Escherichia coli*", *Bioinformatics*, 27(19):2714-2720, 2011.

II J. Mäkelä, M. Kandhavelu, S.M.D. Oliveira, J.G. Chandraseelan, J. Lloyd-Price, J. Peltonen, O. Yli-Harja, and A.S. Ribeiro, "*In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter", *Nucleic Acids Research*, 41(13):6544-6552, 2013.

III L. Martins*, J. Mäkelä*, A. Häkkinen, M. Kandhavelu, O. Yli Harja, J.M. Fonseca and A.S. Ribeiro, "Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time", *Journal of Theoretical Biology*, 301:83 – 94, 2012. (*equal contribution)

IV J. Mäkelä, J. Lloyd-Price, O. Yli-Harja, and A.S. Ribeiro, "Stochastic sequence-level model of coupled transcription and translation in prokaryotes", *BMC Bioinformatics* 12:121, 2011.

The author of this thesis contributed to these publications as follows. In **Publication I**, the author performed the simulations, analyzed the results with H. Huttunen and A.S. Ribeiro, and contributed to the writing of the manuscript. In **Publication II**, the author conceived the study with A.S. Ribeiro, designed the experiments, carried out the microscopy experiments and image analysis with S.M.D. Oliveira, designed and implemented the models with J. Lloyd-Price, analyzed the results with A.S. Ribeiro, and contributed to the writing of the manuscript. In **Publication III**, the author conceived the model with A.S. Ribeiro, simulated the model and analyzed the results with L. Martins, and contributed to the writing of the manuscript. In **Publication IV**, the author conceived the model, implemented and simulated the model with J. Lloyd-Price,

analyzed the results with J. Lloyd-Price and A.S. Ribeiro, and contributed to the writing of the manuscript.

Publication II has been used by J.G. Chandraseelan in his PhD dissertation.

1 Introduction

1.1 Background and Motivation

In all organisms, from viruses to mammals, gene regulatory networks coordinate cellular functions, such as growth and differentiation (Arkin et al. 1998; Süel et al. 2007; Acar et al. 2005; Takahashi and Yamanaka 2006). These networks control both which genes are transcribed as well as when these events occur, based on intracellular and environmental information. Such regulation is essential for cells to alternate between different physiological and morphological states, in order to cope with changing environmental conditions (Errington 2003; Süel et al. 2006; Balaban et al. 2004). Notably, not all cellular decisions to change phenotypic state are driven by environmental or internal signals (Arkin et al. 1998; Lewis 2007; Kearns and Losick 2005). Namely, evidence suggests that monoclonal cells in a homogenous environment can exhibit a mixture of different phenotypes in a stochastic manner (Norman et al. 2015). The main source of phenotypic diversity has been identified to be stochasticity in gene expression (Elowitz et al. 2002).

Many regulatory molecules in cells exist in very low copy-numbers (Kaern et al. 2005; Taniguchi et al. 2010). Molecular events involving such low-abundance molecules, such as in gene expression, are poised with randomness in the timing of events. Stochasticity in gene expression causes identical cells in the same environment to exhibit different numbers of RNA and proteins (Paulsson 2004). Aside from this, small differences in cell size, cellular history, etc., have been shown to contribute to the diversity of cell fates (St-Pierre and Endy 2008; Zeng et al. 2010; Robert et al. 2010). Relevantly, not all stochasticity is detrimental, as in some cases, it is the mean by which cellular organisms adjust to challenges posed by the competition and environmental fluctuations (Leibler and Kussell 2010; Süel et al. 2006; Balaban et al. 2004; Raj et al. 2010; Ribeiro et al. 2008).

Gene expression consists of transcription, the reading of DNA and production of a specific RNA molecule, and translation, the reading of the RNA sequence and engineering of a correspondent peptide. Both are complex, multi-step processes with sequence-dependent dynamics (McClure 1985; Saecker et al. 2011; Ramakrishnan 2002). For example, the promoter sequence has been shown to control both the mean and variability in constitutive expression in *Escherichia coli* (Jones

et al. 2014). The regulation of the steps in transcription initiation, especially the closed and open complex formations, allow cells to regulate the rate of RNA production (Lutz et al. 2001; McClure 1980; Sanchez et al. 2011). The coupling between transcription and translation in prokaryotes allows this regulation also to be extended to peptide production dynamics (Yarchuk et al. 1992).

Recent advancements in single-molecule detection techniques have made possible measurements of fluctuations in the RNA and protein numbers in individual cells (Golding et al. 2005; Yu et al. 2006; Fusco et al. 2003; Taniguchi et al. 2010). The fluorescence *in situ* hybridization (FISH) and tagging of RNAs with fluorescent proteins, e.g., MS2-GFP method (Fusco et al. 2003), have rapidly become popular due to their ability to probe variability in endogenous RNA numbers, which is not attainable from averaged measurements of abundance (Raj and van Oudenaarden 2009). By measuring fluctuations in the number of molecules, quantitative information about the underlying processes responsible for the observed fluctuations and even the dynamics, can be determined. Such measurements have been used to probe different stages of transcription, such as RNAp binding, transcription initiation, and elongation, which has proven to be insightful for the understanding of the transcription process (Larson et al. 2011; Friedman et al. 2013; Muthukrishnan et al. 2012).

From the measurements, models of gene expression have been constructed. These seek to explain how the fluctuations in the numbers of involved components arise and how they contribute to the overall dynamics (Sanchez et al. 2011; Garcia et al. 2012). To incorporate stochasticity, gene expression has been modeled using stochastic modeling approaches (Arkin et al. 1998; Ribeiro et al. 2010). The stochastic simulation algorithm (SSA) is a common way to simulate exact sample trajectories from the distribution described by the chemical master equation (CME), which captures the dynamics of molecular scale interactions (Gillespie 1992; Gillespie 2007). Other methods, such as finite state projection (FSP) algorithm (Munsky and Khammash 2006), can provide approximations or exact analytical solutions for biological systems with small number of species. Meanwhile, dynamics of a system with a large number of interacting species can be only simulated with SSA.

1.2 Thesis Objectives

This thesis focuses on studying variability in the transcription process in *E. coli*. First, an automatic methodology to detect changes from time series data is presented. The method aimed to be general and applicable to different biological systems, e.g., single genes, small gene networks, and large-scale networks. Relevantly, it can be used to detect changes in simulated time series and in measurements from time-lapse microscopy. Second, the induction kinetics and subsequent transcription dynamics in live cells for a few promoters and induction

schemes were studied. Specifically, time intervals between transcription events and the waiting time for the first transcription event were measured using an MS2-GFP RNA tagging method. Third, closely spaced promoters, common in *E. coli* and other organisms, were studied assuming various configurations and localizations of promoter start sites using stochastic, single-nucleotide level models. Also, the co-regulation of the promoters with shared transcription factor binding sites was characterized. Finally, a stochastic transcription and translation elongation model at the single nucleotide and codon level was developed. This model was used to study the propagation of fluctuations in transcription kinetics into translation kinetics as a function of the underlying processes.

The thesis has three objectives:

I Propose a novel automatic method to detect changes in the dynamics of gene regulatory networks. This method has to be general so as to be applicable to the study of a broad type of changes in synthetic and empirical time series data.

II Study the timing of promoter activation and consequent transcription dynamics using single-RNA measurement techniques. Characterize the consequences of asynchronous promoter dynamics on the temporal population variability.

III Characterize the dynamics of transcription initiation and coupled transcription and translation elongation using single nucleotide level stochastic models. The models should account for the detailed processes occurring during initiation and elongation of both transcription and translation.

Objective I was completed in **Publication I**. Objective II was completed in **Publication II**. Finally, Objective III was completed in **Publication III** and **Publication IV**.

1.3 Thesis Outline

This thesis is organized as follows. Chapter 2 introduces the biological background by describing the current knowledge on transcription and translation initiation, elongation and regulation in *E. coli*. Chapter 3 introduces the models, modeling strategies and simulation algorithms employed in the publications of this thesis. In particular, the CME, the SSA and the FSP algorithms to model biochemical systems are described in this chapter. Chapter 4 presents fluorescence microscopy techniques to measure *in vivo* single RNA numbers in individual cells. In particular, the measurements and analysis of *in vivo* single RNA detection experiments used in this thesis are discussed. Finally, conclusions and discussion are presented in Chapter 5.

2 Biological Background

This chapter gives an overview on the biological processes studied in this thesis. It includes a description of transcription and translation in *Escherichia coli* along with a more detailed view into transcription initiation, elongation, regulation, closely spaced promoters and noise in gene expression.

2.1 Gene Expression in *Escherichia coli*

Gene expression is the process of reading-out or expressing the genetic information stored in the genome. There are two main steps by which the genetic information is expressed in all cells, namely transcription and translation that together form the central dogma of molecular biology (presented in Figure 2.1) (Crick 1970). In the first step, transcription, a particular region of the DNA nucleotide sequence, a gene, is replicated into a complementary mRNA (messenger RNA) nucleotide sequence (McClure 1985). Following transcription, an mRNA template is used to synthesize the corresponding amino-acid sequence by a process called translation (Ramakrishnan 2002). Each gene in the genome can be transcribed and translated with a given efficiency, allowing the cell to express different, gene-dependent quantities of mRNAs and proteins (Alberts et al. 2002). Moreover, the amount of expression from a gene can be regulated to address changes in the demand of gene products. In prokaryotes this commonly happens through controlling the rate of RNA production. Finally, in the case of many genes, the RNA is the final product of gene expression, and it can have either a structural, catalytic or regulatory role (Alberts et al. 2002).

One of the most studied organisms in the fields of biochemistry and molecular biology, *E. coli*, has been the main source of information on the basic mechanisms involving genes, such as DNA replication, gene expression, and protein synthesis (Blattner et al. 1997; Lee and Lee 2003). The genome of *E. coli*, consisting of a single circular chromosomal double stranded DNA, contains over 4000 genes coding for structural and regulatory proteins (Blattner et al. 1997). Additionally, the cell can contain extra-chromosomal DNA, called plasmids, with additional genes that code for, e.g. antibiotic resistance (Eliasson et al. 1992).

The genes in prokaryotes mainly consist of three components: a promoter, operator

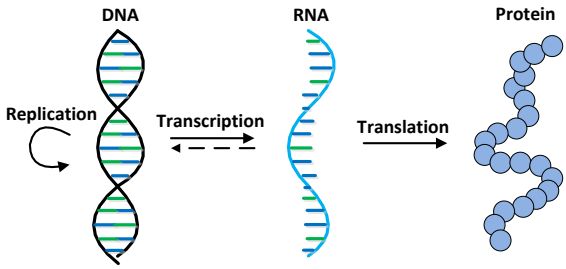


Figure 2.1: The central dogma of molecular biology. The information stored in the DNA can be transferred to mRNA by transcription and from the mRNA to proteins by translation. Information in the DNA can be replicated in the process of DNA replication. Additionally, in special cases, information from the RNA can be transferred to the DNA in a process called reverse transcription.

site(s) and structural gene(s) (Alberts et al. 2002). The promoter is a specific region of DNA recognized by RNA polymerase (RNAp) to initiate transcription. The operator sites are small segments of DNA recognized by regulatory molecules that control the expression from the promoter, e.g. repressor molecules prevent binding of RNAp. Structural genes in prokaryotes are usually organized into operons, in which a set of genes is controlled by a single promoter (Osbourn and Field 2009)(Figure 2.2). Consequently, the operon is transcribed into a single mRNA molecule often containing multiple genes (polycistronic mRNA). The first operon to be described was the *lac* operon by Jacob and colleagues (Jacob et al. 1960). Aside the operon structure, some bacterial genes are organized as closely spaced promoters, which is believed to allow further coordination of the gene products (Beck and Warren 1988).

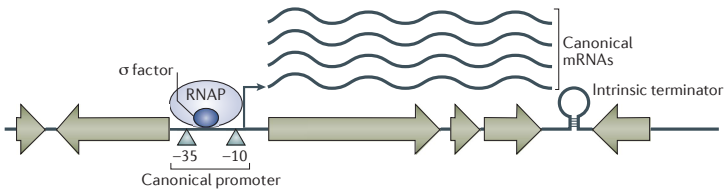


Figure 2.2: Genes in prokaryotes are organized as operons. A promoter in a region between genes initiates mRNA synthesis by recruiting RNAp and facilitate the formation of a transcription elongation complex, which produces mRNAs that terminate at an intrinsic terminator. The genes within operon are transcribed as a single mRNA. Adapted with permission from (Wade and Grainger 2014).

2.2 Mechanisms of Transcription and Translation

The main enzyme involved in transcription is the core RNAp. It consists of several subunits ($\beta\beta$'$\alpha_2\omega$) and contains all necessary enzymatic components required for the synthesis of RNA but it cannot initiate transcription from a promoter (Young et al. 2002). To bind specifically to the promoter and initiate transcription, the core RNAp must be bound by one of the σ-subunits (Murakami et al. 2002). This produces an RNAp in the holoenzyme form (Eσ), which contains exactly one σ-subunit that has affinity for specific promoters in the genome. E.g., σ^{32} is a heat shock sigma factor in that it allows the RNAp to express the genes associated with the response of *E. coli* to heat shock conditions (Alberts et al. 2002).

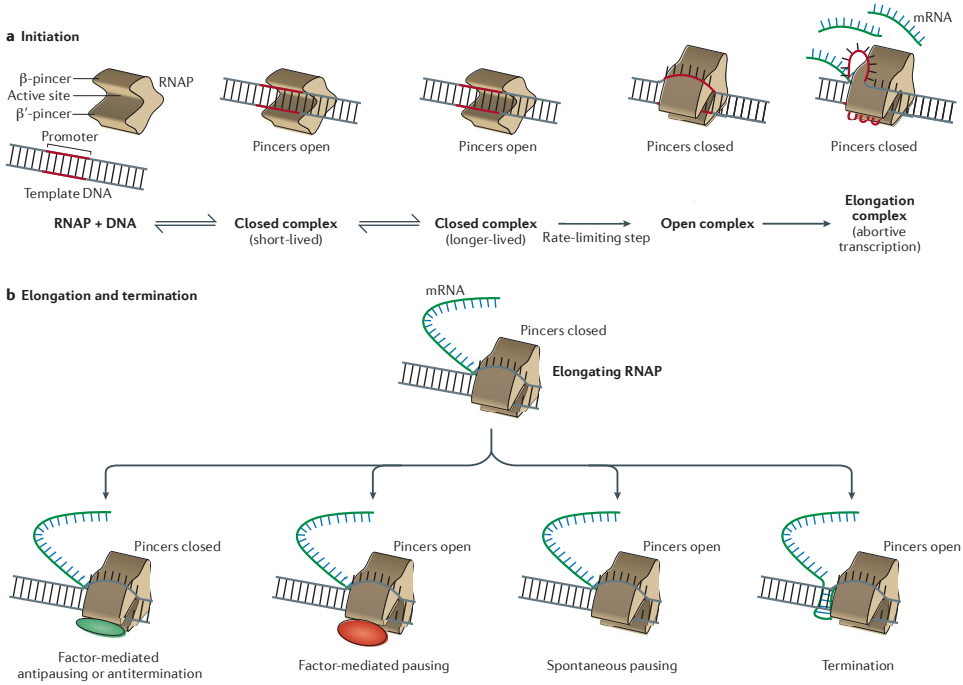


Figure 2.3: Transcription process in *E. coli*. (a) Transcription initiates as an RNAp binds the promoter region and forms a closed complex. Next, in the process of open complex formation, the DNA's double helix is opened and a short stretch of nucleotides are exposed. Finally, the RNAp enters a productive elongation state following an abortive initiation cycle. (b) During elongation the RNAp can go through alternative pathways, such as spontaneous or transcription factor-mediated pausing. Elongation ends in the process of transcription termination. Adapted with permission from (Robinson and Oijen 2013).

The transcription process consists of initiation, elongation and termination as depicted in Figure 2.3 (Alberts et al. 2002). Initiation consists of the RNAp holoenzyme finding a promoter, unwinding the DNA, and initiating the elongation. Next, in elongation, mRNA is synthesized by the transcription elongation complex

(TEC) moving along the DNA template in 3' to 5' direction. Reaching a specific termination signal encoded into the DNA, elongation is terminated and a newly transcribed mRNA is released. The termination signal typically destabilizes the TEC on the DNA by forming an secondary RNA structure. In prokaryotes, translation can initiate as soon as the 5' end of an mRNA including a ribosome binding site (RBS) is synthesized (Miller et al. 1970; Yarchuk et al. 1992).

A translation cycle, similar to transcription, consists of initiation, elongation and termination. Translation in prokaryotes is conducted by ribosomes that are highly complex molecular machines consisting of ribosomal proteins and specialized RNA molecules (rRNAs). *E. coli* ribosome (70S) consists, as in other species, of two main sub-units: a small (30S) and a large (50S) subunit (Ramakrishnan 2002). The small subunit contains a 16S RNA subunit and 21 proteins. The large subunit contains a 5S RNA subunit, a 23S RNA subunit and 31 proteins.

Translation is initiated at the start codon AUG, which is recognized by a special initiator tRNA carrying *N*-formylmethionine (fMet). mRNA contains the RBS consisting of a specific Shine-Dalgarno sequence which is located upstream of the initiation codon (Ramakrishnan 2002). The RBS is recognized by the 16S rRNA of the ribosome. To initiate translation, first, the small ribosomal subunit (30S) of the ribosome binds to the RBS of the mRNA and fMet-tRNA binds to the P-site forming a 30S-RNA complex (Ramakrishnan 2002). Next, the large ribosomal subunit (50S) binds to the complex to complete the ribosome (70S) and initiates the translation elongation. During the elongation, transfer RNAs (tRNAs), carrying specific amino acids, bind to the appropriate codons in mRNA and with the help of the ribosome, an amino acid is added to the growing polypeptide until stop codon is reached (Ramakrishnan 2002). Following this, a release factor binds to the ribosome releasing both the ribosome and the completed polypeptide.

2.2.1 Transcription Initiation

Transcription initiation in prokaryotes is a complex, multi-stepped process that has been observed to include three main steps: binding, isomerization and promoter clearance (McClure 1985; Saecker et al. 2011):

$$\mathrm{R} + \mathrm{P} \overset{K_B}{\longleftrightarrow} \mathrm{RP}_C \overset{k_f}{\longrightarrow} \mathrm{RP}_O \rightarrow\rightarrow\rightarrow \mathrm{RNA} \tag{2.1}$$

This scheme was first suggested by Walter, Zillig and colleagues (McClure 1985; Walter et al. 1967; Chamberlin 1974). It involves binding of a holoenzyme (R) to a promoter (P) with an equilibrium binding constant, K_B, to form a closed complex, RP_c, which subsequently isomerizes with a rate constant k_f to form a stable open complex, RP_o. After an initial RNA synthesis, the RNAp breaks its interactions with the promoter and enters into an elongation phase. Various alterations of this

scheme exist in different promoters and these may include additional steps and equilibrium reactions.

A promoter region in *E. coli* is defined by a consensus sequence at -10 and -35 positions upstream of the transcription start site (Cho et al. 2009; Harley and Reynolds 1987). The highly conserved consensus sequence is required for the RNAp holoenzyme to recognize the transcription initiation site (Hippel et al. 1984). To initiate transcription, the holoenzyme must first find and bind to the promoter. DNA binding proteins have been shown to find the target site faster than the 3D diffusion limit (Riggs et al. 1970). To achieve this, additional search mechanisms such as 1D sliding, 1D hopping and inter-segment transfer are likely necessary (Dangkulwanich et al. 2014; Hammar et al. 2012). The holoenzyme is known to adhere only weakly to non-specific DNA and to slide rapidly along the DNA molecule until it dissociates from it, unless a start site is found (McClure 1985). Recent studies tracking single fluorescent holoenzymes *in vitro* reported that long-range 1D sliding does not significantly affect the search times (Friedman et al. 2013; Wang et al. 2013). Nevertheless, the exact contribution of different mechanisms for the binding of the holoenzyme to the promoter region have not been thoroughly quantified in live cells. Once the holoenzyme finds the promoter, it recognizes the promoter site by making specific contacts with the bases that are exposed on the outside of the helix in the consensus sequence (Alberts et al. 2002).

Following the binding of the holoenzyme to the promoter, the holoenzyme unwinds the DNA's double helix and exposes a short stretch of nucleotides on each strand. This does not require ATP energy in σ^{70} promoters as it is achieved through a reversible structural change of the holoenzyme-DNA complex that is more favorable than the initial state (Alberts et al. 2002). The consequent isomerization into the open complex form is found to contain at least three intermediate steps, namely DNA loading, DNA unwinding, and assembly of the polymerase clamp (Saecker et al. 2011). To enter the elongation phase of transcription, the holoenzyme goes through an abortive initiation cycle before committing to elongation (Goldman et al. 2009; Hsu 2002). Consequently, the holoenzyme synthesizes small transcripts of length up to 10 nucleotides. The abortive initiation is shown to occur via a 'scrunching' mechanism, in which the holoenzyme remains stationary while the downstream DNA is pulled into the active site (Revyakin et al. 2006; Kapanidis et al. 2006). After initial RNA synthesis, the holoenzyme breaks its interaction with the promoter and finally enters into elongation phase. The exact moment of σ^{70} release remains unclear and it has been suggested to remain bound to promoter, be released in beginning of elongation, or be retained in TEC (Bar-Nahum and Nudler 2001; Kapanidis et al. 2005; Raffaelle et al. 2005; Harden et al. 2016). A recent study using *in vitro* techniques showed that a substantial fraction of elongating TEC retained the σ^{70}-factor throughout elongation (Harden et al. 2016).

Regulation of the steps in transcription initiation have been traditionally studied with abortive initiation and *in vitro* transcription assays (Buc and McClure 1985; McClure et al. 1978; McClure 1980; Lutz et al. 2001). The open complex formation rate can be derived from the delay of reaching the steady-state production of the abortive product assays (McClure et al. 1978; McClure 1980). The closed complex formation is dependent on the RNAp concentration, which allows it to be distinguished from the open complex formation (Buc and McClure 1985). This dependence of the lag time for the RNAp concentration allows drawing a τ-plot, which portrays a direct relationship between lag times and the reciprocal RNAp concentration (McClure 1980; Patrick et al. 2015). From this plot, the slope yields the mean time for the closed complex formation and the intercept gives the mean time for the open complex formation. Compared to the time required for elementary steps in enzyme catalyzed reactions, the observed lags are much longer, spanning from a few seconds to several minutes. As such, these processes are rate-limiting for transcription initiation (McClure 1985; Saecker et al. 2011). Studies of transcription reactions *in vitro* have also showed that the reactions times were sequence-dependent as they differed between promoters (Bertrand-Burggraf et al. 1984; McClure 1985; Saecker et al. 2011).

More recent techniques, based kinetic and intermediate trapping experiments, as well as footprinting and crystallographic analysis, have identified multiple intermediate steps during the initiation (Sclavi et al. 2005; Davis et al. 2007; Saecker et al. 2011). Recently, *in vitro* single-molecule fluorescence spectroscopy was used to visualize the rate-limiting steps in transcription initiation including binding, open complex formation, transcript production, and σ^{54} dissociation (Friedman and Gelles 2012). The main steps of the initiation process, including reversible intermediates, were characterized. The isomerization step was found to limit the initiation rate, in agreement with previous findings of DNA supercoiling altering initiation rates in certain σ^{54}-dependent promoters (Amit et al. 2011; Huo et al. 2006).

Compared to *in vitro* environment, *in vivo* measurements of the transcription initiation rates are much more complicated to execute. Most studies quantifying the transcription process have used measurements of the heterogeneity in number of RNAs per cell e.g., using FISH (Jones et al. 2014; So et al. 2011). The measurements have shown that the sequence-dependent transcription initiation process dictates both the mean and variability in mRNA numbers (Jones et al. 2014). Aside from the observed population variability, recent *in vivo* single-RNA level measurements have quantified the time intervals between consecutive production events in single cells for various promoters (Kandhavelu et al. 2011; Kandhavelu et al. 2012a; Kandhavelu et al. 2012b; Muthukrishnan et al. 2012). These studies proposed that the distributions of time intervals could not be explained by a single elementary step. Thus, it was suggested that the dynamics could be explained by multiple rate-limiting steps in transcription in line with the results from *in vitro* studies. Alternatively, the nature of the rate-limiting steps

may correspond to other mechanisms than elementary steps in the transcription cycle.

In **Publication II**, measurements of time intervals between consecutive production events were conducted to study the transcription process. In **Publication III** the closed complex formation, open complex formation and abortive initiation were included in the models to accurately depict the transcription initiation process.

2.2.2 Transcription and Translation Elongation

The transcription elongation phase initiates as the RNAp clears the promoter region. In this phase, the transcription elongation complex (TEC) incorporates nucleotides into the nascent RNA chain while advancing on the DNA template. In active translocation, the TEC has been shown to move up to 50 bp/s (Greive and Von Hippel 2005; Proshkin et al. 2010). The movement of TEC occurs in discontinuous manner as the TEC, also exhibits pausing or even backward diffusion on the template (Greive and Von Hippel 2005). As such, pausing can significantly reduce the overall transcription rate during elongation. The duration of pauses has been shown to vary from less than a second to minutes (Herbert et al. 2006; Herbert et al. 2010; Landick 2009). The pauses can be divided into two categories: short 'ubiquitous' pauses and longer-lived pauses that often are stabilized by backtracking or formation of a hairpin structure in the nascent RNA (Landick 2006). Transcription factors and the DNA sequence have been shown to affect the dynamics of pausing, e.g. the transcription factor NusG can increase the overall transcription rate by both enhancing elongation rate and decreasing the entry rate into long-lifetime, backtracked pause states (Herbert et al. 2010). Also, retained σ-factors have been shown to affect the recognition of a class of transcriptional pause sequences while appearing similar in elongation rates (Harden et al. 2016). Aside pauses, transcription elongation has the alternative pathways of premature termination, pyrophosphorolysis, misincorporation or editing (Arndt and Chamberlin 1988; Greive and Von Hippel 2005; Erie et al. 1993).

Translation elongation occurs in a discontinuous manner as series of translocation – pause events takes place in the movement (Wen et al. 2008)(see Figure 2.4). The ribosome moves three bases (which corresponds to one codon) at a time, followed by a peptide-bond formation between amino-acids. The latter defines the overall rate of translation and is also dependent on the secondary structure of the mRNA. Additionally, longer pauses during elongation were observed and these may lead to translational frameshifting (Farabaugh 1996) and protein misfolding (Kimchi-Sarfaty et al. 2007). Ribosomes stalled on the mRNA can be rescued by transfer-messenger RNA (tmRNA), which releases the ribosome by terminating translation prematurely (Moore and Sauer 2005). Approximately 0.4 per cent of all *in vivo* translations are prematurely terminated.

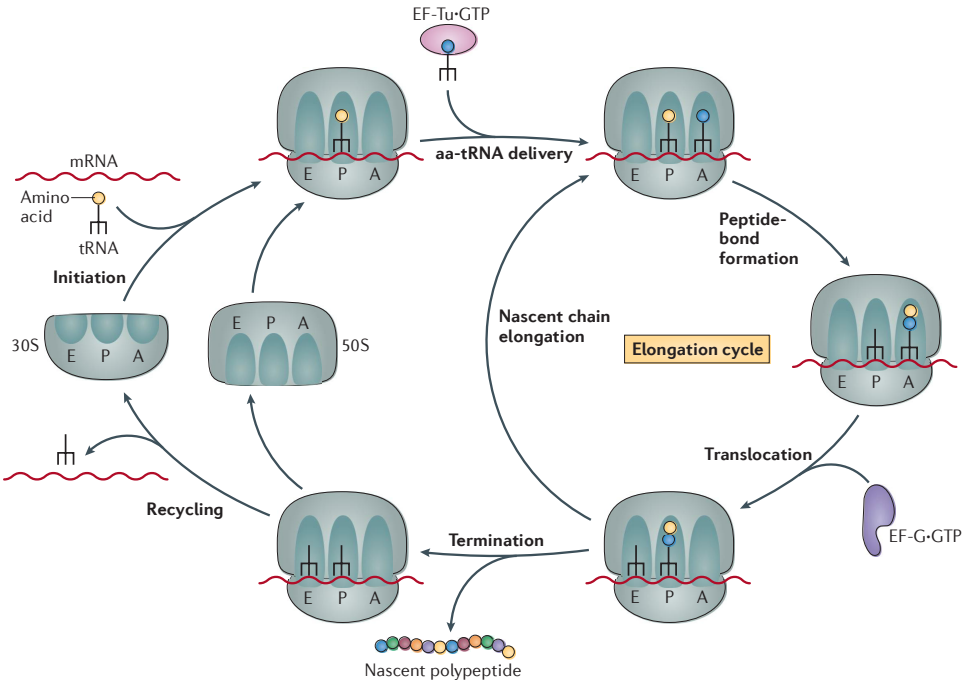


Figure 2.4: Translation cycle. Initiation of protein synthesis involves the formation of a 70S ribosome from a 30S and a 50S subunits in start codon of the mRNA with the initiator tRNA positioned at the P-site. The elongation cycle consists of the delivery of the aminoacylated-tRNA to the A-site of the ribosome by EF-Tu, peptide-bond formation between tRNAs of the A- and P-sites, translocation of the tRNAs, which is catalysed by EF-G, and elongation of the nascent chain. In the termination, the polypeptide chain and subsequent dissociation of the 70S ribosome are released, followed by recycling of the components for the next translation event. Adapted with permission from (Wilson 2014).

Transcription and translation elongation in prokaryotes are dynamically coupled. The majority of genes initiate translation as soon as the RBS emerges from the TEC (Miller et al. 1970). Implications of this coupling have been mostly studied in specific cases of transcription attenuation and polarity (Adhya and Gottesman 1978; Yarchuk et al. 1992). In both mechanisms, slow translating ribosomes increase the distance to the TEC and allow the premature termination of transcription. The premature transcription termination occurs through the formation of a hairpin in the leader RNA sequence that destabilizes the RNAp. This occurs if ribosome binding on the nascent RNA is not fast enough (Yanofsky 1981). In the polarity effect, the growing gap between TEC and the first ribosome allows the termination factor Rho to access the nascent RNA, which results in premature termination of TEC and, in polycistronic mRNAs, reduces expression of the downstream genes (Richardson 1991).

The transcription elongation rate is strongly affected by the rate of translation elongation (Proshkin et al. 2010). Slowing down translation elongation using antibiotics or slow-to-translate codons reduces the transcription elongation rate as well. The first ribosome in the nascent RNA has been proposed to assist TEC during elongation, by preventing backward translocation and pausing (Proshkin et al. 2010). This cooperative mechanism is believed to prevent discrepancy between transcription and translation efficiencies in different genes and environments.

Translation rates have been shown to be codon-specific (Sørensen et al. 1989; Sørensen and Pedersen 1991). The redundancy between codons (64) and amino acids (20) allows an additional level of regulation for translation. E.g. synonymous codons do not change the encoded protein but they can affect translation elongation. Two synonymous codons, read by the same tRNA species, were translated with a threefold difference in rate, which implies that the difference in translation rates are not caused only due to differences in tRNA abundances (Pedersen 1984). The extent of slow translating codons promoting queue formation and causing collisions between ribosomes was studied using stochastic models of translation with different codon translation rates (Mitarai et al. 2008). The simulations suggest that traffic and collisions frequently affect the efficiency of translation.

The average translation efficiency of a sequence, i.e. the protein yield, has been traditionally analyzed with the Codon Adaptation Index (CAI) (Sharp and Li 1987). To estimate the translation efficiency of a specific sequence, a given sequence is compared to a reference set of highly expressed genes. Nevertheless, a study utilizing a synthetic library of 154 synonymous sequences of GFP found no strong correlation with the CAI and gene expression levels (Kudla et al. 2009). This was proposed to be due to the CAI increasing the elongation rate but the translation initiation remaining the rate-limiting step in translation. The expression levels of individual genes were also proposed to be more affected by the mRNA secondary structures. The CAI was proposed to have more influence on global translation efficiency and cellular fitness.

To investigate translation efficiency profiles, native codon sequences and tRNA pools were compared in various organisms (Tuller et al. 2010). Interestingly, most genes were shown to have a ramp of slow-to-translate codons in the beginning of genes, which is proposed to reduce collision between ribosomes and improve the efficiency of translation (Tuller et al. 2010). The length of the ramp was measured to be approximately 30-50 codons. Increasing the translation initiation rate decreases the average spacing between ribosomes and creates collisions between the ribosomes that can stall or even abort translation. The ramp could prevent collisions, by spacing the ribosomes more evenly, especially in abundantly translated genes (Tuller et al. 2010). Also, it would allow some genes to be especially sensitive to the low abundance of amino acid-loaded tRNAs.

The functioning of slow ramps on single genes and small gene circuits were investigated in a recent modeling study (Potapov et al. 2012). This study examined

the effects of codon sequences on the fluctuations of gene expression using stochastic models of coupled transcription and translation at the codon level (first published in **Publication IV**). The model supports the hypothesis of slow ramps reducing ribosomal jams by reducing the rate of translation initiation. Also, the model proposed that the mean and noise in the protein numbers can be separately regulated by the coding sequence.

Noise propagation from transcription initiation to protein expression was studied in **Publication IV**. The stochastic model incorporates the transcription model at the nucleotide level, that includes transcription initiation, pausing, premature termination, and accounts for the RNAp footprint in the DNA template (Ribeiro et al. 2009). The translation model at the codon level includes translation initiation, codon-specific translation rates, stalling, and accounts for the ribosome footprint (Mitarai et al. 2008). The model in **Publication IV** coupled the transcription and translation models to allow events to simultaneously affect both processes.

2.3 Regulation of Transcription

Transcription in *E. coli* is a relatively rare event at the genome level (Taniguchi et al. 2010). Transcription is the main regulator of mRNA abundance, as mRNA degradation rates cannot explain the observed abundance (Bernstein et al. 2002; Chen et al. 2015a). Degradation has been proposed to have an alternative role as a regulator of abundance, e.g. in response to environmental perturbations. Additionally, at the ensemble level, the mean mRNA and respective protein levels were found to be only moderately correlated (Taniguchi et al. 2010).

The regulation of transcription primarily occurs during the main steps of initiation: promoter binding, isomerization and promoter escape (Browning and Busby 2004). The most common mechanism by which regulation occurs is the binding of a transcription factor at the promoter region. Globally, the concentration and activity of RNAp can be used to regulate transcription initiation (Bremer and Dennis 1996; Klumpp et al. 2009). Transcription can also be regulated during elongation in specific leader sequences that can terminate the elongation. E.g., in tryptophan attenuation, if the concentration of charged tRNA^{trp} is high enough, transcription is terminated by a RNA hairpin structure (Simao et al. 2005). An example of transcription regulation is shown in Figure 2.5.

2.3.1 Transcription Factor Dynamics

The regulation of transcription initiation by transcription factors is traditionally described by the operator occupancy model. The transcription factor's state, bound or not bound, determines the state of a gene, which will be not expressing or expressing, depending on the mode of regulation. Namely, the association and dissociation of the transcription factor will turn the gene off and on. The architecture of the promoter, i.e. the location of binding sites and their affinities

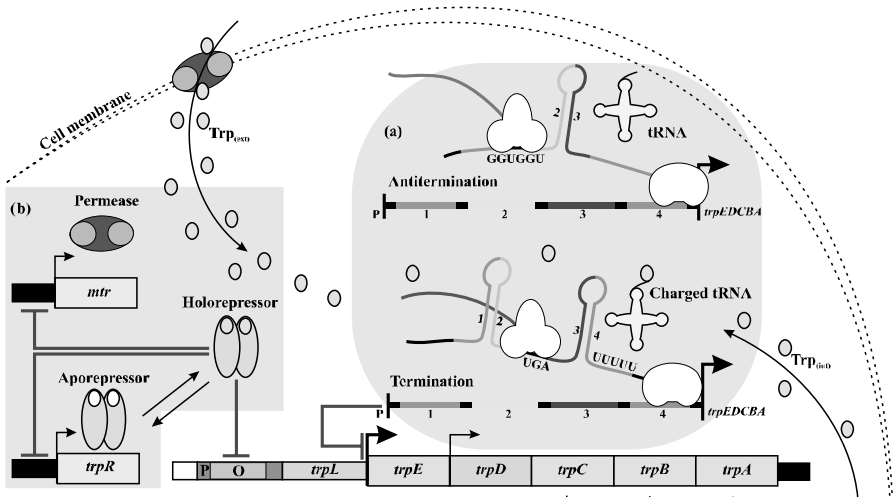


Figure 2.5: Example of transcription regulation. Tryptophan biosynthesis is subject to (a) transcription attenuation, and (b) transcription repression. In transcription attenuation, leader regions of biosynthetic operon serve to synchronize the progress of RNAP with ribosomes. The transcription inhibition of trpEDCBA operon by the dimeric holorepressor results from the combination of the product of the repressor gene trpR with the amino acid Trp. Reproduced with permission from (Simao et al. 2005).

for transcriptional regulators, determines the transcriptional responses of the promoter to changes in transcription regulators numbers and the consequent protein numbers in the cells.

The most common way of regulating the promoter activity in *E. coli* is by repression (Garcia et al. 2010). The exact mechanism by which the repression of transcription initiation occurs varies between promoters. First, the repressor can directly compete with the RNAp in binding to the promoter (Hawley et al. 1985; Schlax et al. 1995). Alternatively, the repressor can prevent the open complex formation (Heltzel et al. 1990; Sanchez et al. 2011). Finally, the repressor can inhibit promoter escape, in which the open complex can be formed but elongation is blocked (Krummel and Chamberlin 1989; Lee and Goldfarb 1991). The promoters in *E. coli* exhibit a wide range of locations for the repressor binding sites in respect to the transcription start site (Garcia et al. 2012; Gama-Castro et al. 2011). Interpreting the repression mechanism can be difficult from population measurements only. Recent single-molecule spectroscopy measurements and statistical analysis approaches have allowed a direct quantification of individual RNAp-DNA interactions in the presence and absence of the repressor molecule (Sanchez et al. 2011; Friedman and Gelles 2012).

The different repression mechanisms lead to qualitatively distinct regulatory behaviors (Sanchez et al. 2011). In the case of inhibition of promoter binding,

the transcription initiation rate is proportional to the RNAp binding rate to the promoter, which can be reduced simply by increasing the repressor numbers. By inhibiting the subsequent steps of transcription initiation, the transcription rate is controlled by the dissociation rate of the repressor from the promoter, which is independent of the repressor numbers in the cell. In this case, the promoter kinetics, including the rate-limiting steps in initiation, have a major contribution on the dynamics of the repression, which makes the equilibrium occupancy model not always valid.

This is also supported by a recent experiment that characterized repression, by having binding sites artificially placed either upstream or downstream from a promoter in *E. coli* (Garcia et al. 2012). The strength of repression could not be explained by the occupancy of binding site alone. In another study, a direct measurement of transcription factor association and dissociation in live *E. coli* cells showed also inconsistencies of the operator occupancy model of gene regulation (Hammar et al. 2014). These findings suggest that these inconsistencies are most likely due to non-equilibrium mechanisms in transcription initiation i.e. its multiple rate-limiting steps. To accurately dissect the regulation of transcription, the effect of promoter dynamics must thus be taken into account, e.g the locations of binding sites, rate-limiting steps in transcription initiation, etc. (McClure 1985; Friedman and Gelles 2012; Garcia et al. 2012).

To investigate this issue, in **Publication III** a single nucleotide level model of the promoter region incorporated a mechanism of repression of transcription to study transcription initiation. In the model, regulatory molecules reserved specific space on the DNA template and thus, depending on the location of the binding site, they either inhibited binding, opening of the DNA template or promoter escape.

2.3.2 Transcription Induction

In fluctuating environments, a single phenotype or behavior of a cell cannot be optimal. To cope with this, cells developed the ability to adapt to different environments by changing phenotypic state. In many cases, these adaptations are triggered directly by signals from environment. In other cases, the switching is stochastic, in that the choices between phenotypes are, for the most part, made randomly (Süel et al. 2007). A common example is persistence in *E. coli*: while antibiotics kill most cells, a small sub-population of genetically identical but persister cells survives (Lewis 2007). The commitment to these phenotypes is usually transient, i.e if a cell is allowed to grow long enough, the mixture of all phenotypes will be restored.

Novick and Weiner studying the *lac* operon in *E. coli*, proposed that the cells switched from a non-producing to a producing state through a single random event (Novick and Weiner 1957). Later, this event was found to be related to the crossing of a critical threshold in permease concentration (Choi et al. 2008). The

phenomena was described as a 'all-or-none phenomenon', as only a fraction of the population, proportional to extracellular thiomethyl-β-D-galactoside (TMG) concentration, produced β-galactosidase. The high variability in the response times was a consequence of a variability in permease molecule numbers prior to the induction event in addition to the inherent stochasticity of chemical reactions at low concentrations (Rao et al. 2002).

The number of permease molecules in uninduced cells have been measured with single molecule sensitivity (Choi et al. 2008). Half of the cells were found to contain at least one permease molecule (Choi et al. 2008). This basal-level expression was proposed to result from a partial dissociation of the tetrameric lactose repressor from one of its operator sites on a looped DNA. The complete dissociation of the repressor from the DNA produces a large production burst of permease molecules fully inducing the *lac* utilization system (Choi et al. 2008). This was verified by disabling the DNA looping mechanism which was shown to be the main regulator of these events.

The process of inducing gene expression differs between genetic motifs (Choi et al. 2008; Schleif 2010; Schnappinger and Hillen 1996). In general, the process by which a cell becomes induced has been described as a single rate-limiting event or a chain of many molecular steps (Choi et al. 2010). At the molecular level, the activation of gene expression consists of multiple molecular steps, such as the uptake of the inducer molecules, dissociation of repressor(s), association of activator(s) etc. The details vary from gene to gene and can include transitions between multiple different phenotypes (Ozbudak et al. 2004). Understanding the process in detail requires a model of the process built using an experimental approach to measure the fluctuations in the components.

In **Publication II** the variability in the response times between individual cells in the arabinose utilization system were observed using a single RNA detection technique. Previously, such dynamics have been observed using population level techniques (Johnson and Schleif 1995; Siegele and Hu 1997) or following single cell trajectories of fluorescent protein products (Megerle et al. 2008; Fritz et al. 2014). The RNA detection techniques used in **Publication II** allowed also the measurement of time intervals between transcription events following the induction. Further, to compare with the arabinose promoter, additional measurements were conducted on a synthetic promoter (Lutz and Bujard 1997) and under various induction schemes.

2.3.3 Arabinose Operon

The arabinose utilization system is used by *E. coli* for catabolizing L-arabinose as a source of carbon and energy (Helling and Weinberg 1963; Englesberg et al. 1965). This system imports pentose L-arabinose from the environment into the cell by AraFGH, a high-affinity ABC transporter, and by a low-affinity transporter, AraE, which binds to the inner cell membrane and makes use of an electrochemical

potential to intake the arabinose (see Figure 2.6)(Hogg and Englesberg 1969; Schleif 2000; Lee et al. 1981). The AraJ protein of the utilization system is poorly characterized but it is thought to act as a transporter or an exporter of arabinose containing polymers (Schleif 2010). The dimeric AraC protein is the regulatory protein for all genes in the arabinose system with a copy-number of approximately 20 molecules per cell (Schleif 2010).

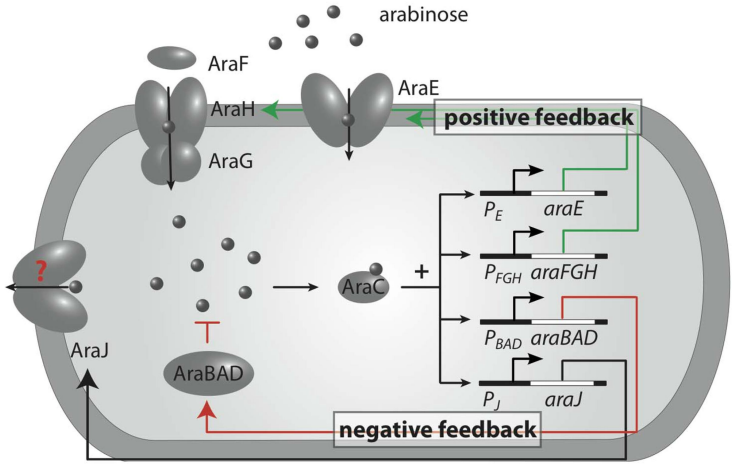


Figure 2.6: Scheme of the arabinose utilization in *E. coli*. Arabinose is imported via the arabinose transporters AraE and AraFGH. AraC, once bound by arabinose, activates the promoters P_E, P_{FGH}, P_{BAD} and P_J, expressing proteins araE, araFGH, araBAD and araJ, respectively. AraBAD encodes arabinose catabolism proteins, while AraJ is assumed to efflux arabinose. Arrows indicate arabinose transport, negative (red) and positive (green) regulation. T-shaped arrow represents arabinose metabolization. Reproduced with permission from (Fritz et al. 2014).

The AraC protein both activates and represses the genes responsible for the intake and catabolism of arabinose (Englesberg et al. 1965; Sheppard and Englesberg 1967; Johnson and Schleif 1995; Schleif 2010). In the presence of high intracellular arabinose, AraC binds to the I_1 and I_2 half-sites close to the promoter which activates the transcription initiation at P_{BAD} (Schleif 2010). Otherwise, AraC promotes the DNA loop formation between two AraC binding sites on the DNA (I_1 and O_2), which prevents access of the RNAp to the promoters region (P_{BAD} and P_C) (Schleif 2010).

The response of the arabinose pathway has been traditionally described as a 'all-or-nothing' response to induction (Schleif 2010; Siegele and Hu 1997). This is a simplification of the overall dynamics and a recent study on the bacterial sugar utilization described the response to be all-or-nothing at low concentrations and graded at high concentrations (Afroz et al. 2014). At low concentrations of arabinose, the fraction of cells expressing the gene products defined the overall

expression. When exceeding the concentration when most cells are induced, further increases in the concentration lead to an increase in enzyme expression in a graded manner.

Recent studies have observed the activation dynamics of the arabinose utilization system by following the gene expression trajectories in single cells (Megerle et al. 2008; Fritz et al. 2014). The timing of activation and in the rates of accumulation of gene products have been shown to exhibit a wide cell-to-cell variability and this timing variability has been shown to be dependent on the arabinose concentration. The variability in the importer molecules have been proposed to have a contribution to the diverse activation dynamics (Siegele and Hu 1997). Replacement of the promoter responsible for the expression of AraE caused the population to produce more uniformly (Khlebnikov et al. 2001; Morgan-Kiss et al. 2002). Interestingly, the variability in timings of de-activation of gene expression upon removal of arabinose was shown to be more homogeneous than the activation (Fritz et al. 2014).

2.4 Closely Spaced Promoters

The genome of *E. coli* contains various configurations of promoters with closely spaced transcription start sites (TSSs) (Gama-Castro et al. 2011). Approximately 15 per cent of the promoters in *E. coli* are closely spaced (Gama-Castro et al. 2011). Such arrangements have been commonly observed in bacterial genomes and in other organisms (Beck and Warren 1988; Häkkinen et al. 2011; Wang et al. 2011).

The geometry of the promoters with closely spaced TSSs can be tandem ($\rightarrow\rightarrow$), divergent ($\leftarrow\rightarrow$), or convergent ($\rightarrow\leftarrow$) (McClure 1985; Beck and Warren 1988; Korbel et al. 2004). The closely spaced promoters can also be classified according to the function of the gene products (Beck and Warren 1988). In the first type, both transcripts code for structural proteins, e.g. bioA-bioBFCD (Nath and Guha 1982). In the second type, one transcript codes for a regulatory molecule while the other codes for a structural protein, e.g. araC-araBAD (Schleif 2010). In the third type, both transcripts code for regulatory molecules, e.g. cI-cro (Arkin et al. 1998). In addition to differing in geometry, closely spaced promoters also differ in the number of nucleotides between promoters and the location of the transcription factor binding sites in respect to the TSSs (Gama-Castro et al. 2011).

A hypothesis for the existence of closely spaced promoters is that the proximity of the genes facilitates their transfer between species, especially for genes that are non-essential (Lawrence and Roth 1996; Lawrence 2003). Also, the proximity of essential genes could make them less likely to be disrupted by deletion or insertion of DNA (Fang et al. 2008). Nevertheless, the small distance between the promoters provides unique opportunities for the regulation of the gene expression. RNAps can interact between each other directly or indirectly by affecting the binding of

transcription factors (Shearwin et al. 2005). Such interactions are likely to affect the transcription initiation kinetics in one or both promoters. Finally, it may allow the same transcription factor to regulate the transcription of both promoters, especially in divergent promoters, where the transcription factor binding sites are often centrally located (Beck and Warren 1988).

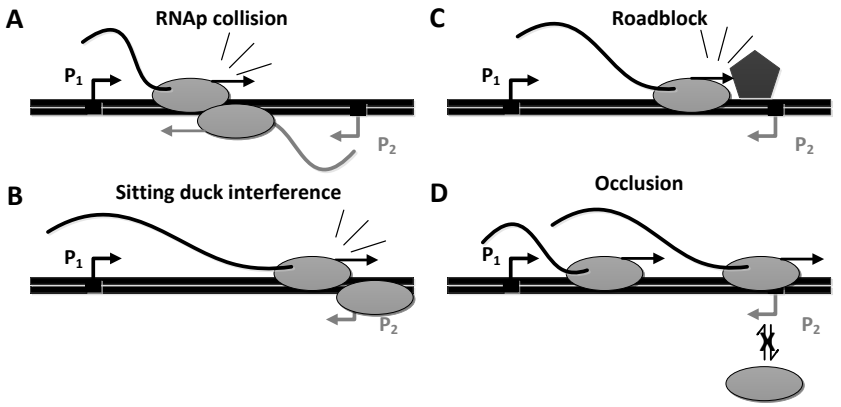


Figure 2.7: Mechanisms of transcriptional interference. Schematic of a general system of closely spaced promoters. (A) RNAp collision during elongation. (B) Sitting duck mechanism, where an elongating RNAp collides with a promoter bound RNAp. (C) Roadblock, where a DNA bound protein prevents elongation of RNAp. (D) Occlusion mechanism, where an elongating RNAp prevents binding of another RNAp to a promoter. Adapted with permission from (Courtney and Chatterjee 2014).

In overlapping promoters, transcription factors are not always needed for accurate regulation. The prosigma factor Crl in *E. coli* stimulates the interaction between RpoS (σ^{38}) and core RNA polymerase (RNAp). This makes it an important factor for global gene regulation (Lelong et al. 2007). The control of the expression of Crl is based on two overlapping promoters transcribing two mRNAs of which, one of them is lacking the RBS and cannot be translated (Pratt and Silhavy 1998; Zafar et al. 2014). The RNAp transcribing this RNA blocks the expression of Crl without the need to produce any trans-acting regulatory molecules. The regulatory response of this mechanism was found to be near-instantaneous making it even faster than an sRNA. The mechanism might also be economical as the protein synthesis is found to require far more energy than the transcription (Neidhardt et al. 1990).

Transcriptional interference in closely spaced promoters has been studied in different configurations (Sneppen et al. 2005; Bendtsen et al. 2011). These models of traffic between the RNAps have been shown to match with measurements from convergent promoters (Sneppen et al. 2005; Bendtsen et al. 2011; Callen et al. 2004). RNAps in closely spaced promoters interact by several mechanisms, depicted in Figure 2.7. The occlusion mechanism in which the RNAp momentarily prevents

binding of an another RNAp to the promoter, was originally proposed to explain an upstream promoter inhibiting the activity of a downstream promoter (Adhya and Gottesman 1982). This mechanism can cause a high level of interference between convergent promoters and overlapping divergent promoters (Sneppen et al. 2005). The sitting duck mechanism describes the removal of promoter-bound complexes by the elongating RNAp from the opposing promoter (Sneppen et al. 2005; Callen et al. 2004). Finally, collisions between the RNAps elongating in opposite directions causes termination of one or both RNAps (Ward and Murray 1979; Prescott and Proudfoot 2002; Sneppen et al. 2005). The amount of interference is also defined by the promoter-dependent kinetics of transcription initiation.

The dynamics of gene expression from closely spaced promoters depends on many factors such as the transcription initiation kinetics, promoter orientation and distances. Also, empirical data suggests that, in principle, any DNA binding protein can be used for both activation and repression of transcription, depending on the promoter architecture (Bendtsen et al. 2011). Also, small changes in the location of the promoter sites and transcription factors can cause drastic changes in the behavior, suggesting that not only the sequence determining the location of the binding sites but also the sequence between adjacent promoters may be subject to strong selective pressure (Garcia et al. 2012; Bendtsen et al. 2011).

In **Publication III**, stochastic single nucleotide models of closely spaced promoters were used to study the activity of the promoters as a function of the distance between TSSs, geometry and locations of repressor binding sites. Also, coordination between the promoter sites and the favorable orientations were investigated.

2.5 Noise in Gene Expression

Genetically identical cells in the same environment can exhibit significant amount of variation in molecular species and in the phenotype of the cell (Neubauer and Calef 1970; McAdams and Arkin 1997; Elowitz et al. 2002; Kaern et al. 2005). This variability is often linked to stochasticity in gene expression caused by low copy number fluctuations. The regulation of transcription is mediated by molecular events, such as binding of a molecule to a promoter, resulting from random encounters between molecules that are inherently stochastic. Further, molecular fluctuations in one molecular species will act as a source of fluctuations for all other species it interacts with (Paulsson 2005; Elowitz et al. 2002). The fluctuations in molecular species can be suppressed by some genetic motifs for more robust functioning or amplified to enhance cell-to-cell heterogeneity (Paulsson and Ehrenberg 2001; Paulsson 2004).

To better understand the sources of variability that contribute to the overall cell-to-cell variability in gene expression, Elowitz and colleagues constructed strains

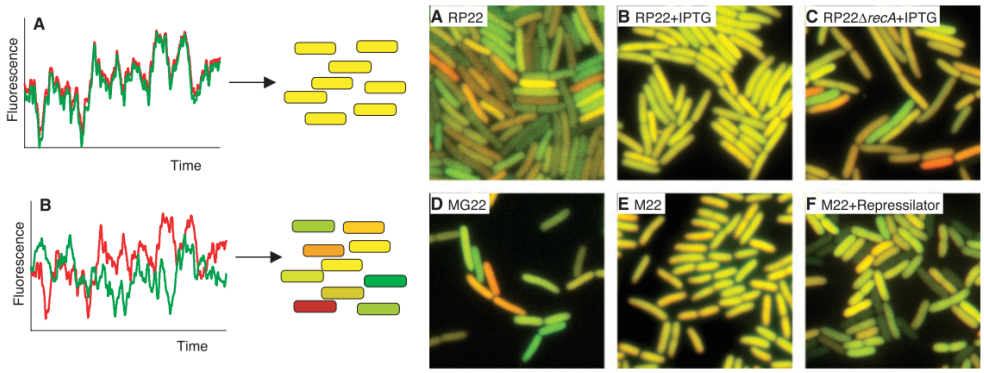


Figure 2.8: Variability in gene expression between genetically identical cells using a double reporter system in (Elowitz et al. 2002). Left: Intrinsic and extrinsic noise sources can be distinguished with two different fluorescent proteins controlled by identical regulatory sequences. Cells with the equal amount of proteins appear yellow, while cells with different expression of the two proteins appear as red or green. (A) Without intrinsic noise, the two fluorescent proteins fluctuate in a correlated fashion over time. (B) Expression of the two genes become uncorrelated due to the intrinsic noise. Right: Microscopy images of cells expressing CFP and YFP were combined in the green and red channels, respectively. Different strains exhibit widely different levels of noise. Reproduced with permission from (Elowitz et al. 2002).

of *E. coli* with a dual reported system (shown in Figure 2.8) (Elowitz et al. 2002). In this setup, two identical promoters coding for two different color fluorescent proteins are integrated at equal distance from the origin of replication but on opposite sides of the chromosome. The relative difference in fluorescence intensity of the two reporters indicates the inherent stochasticity in the process of gene expression, referred as intrinsic noise. The correlated component between the two reporters indicates the contribution of other cellular components to overall variation, referred to as extrinsic noise. Interestingly, different strains of *E. coli* varied in the levels of noise (Elowitz et al. 2002) implicating that noise in gene networks is subject to regulation and/or have different levels of extrinsic noise.

Stochasticity in gene networks can act as a mechanism for phenotypic differentiation (Kaern et al. 2005), e.g. cells can adapt to fluctuating environments in stochastic manner in opposition to responsive switching (Leibler and Kussell 2010; Norman et al. 2015). A classical example of stochastic switching between phenotypes is the lysis-lysogeny regulation circuit in *E. coli* (Neubauer and Calef 1970; McAdams and Arkin 1997). The same cell lineage can transiently switch back and forth between two distinct states: immune (im^+), where the negative control over virus growth is present, and the non-immune (im^-), in which a superinfecting λ is specifically channeled towards the lytic cycle (Neubauer and Calef 1970). A lineage can persist for many generations in one of the states. The

choice between the lysogenic or lytic pathways in individual cells have been shown to result from fluctuations in the protein numbers due to stochasticity in gene expression (Arkin et al. 1998). This cell-to-cell variability in protein numbers is present even in cells that have not gone through the differentiation pathway.

Cell-to-cell variability in gene expression products, namely in mRNAs and proteins, have been intensively studied. The methodologies have improved to allow measurement of RNA and protein numbers with single molecule sensitivity in single cells (Taniguchi et al. 2010; Golding et al. 2005; So et al. 2011; Jones et al. 2014; Yu et al. 2006; Hensel et al. 2012). These measurements show that the stochasticity in the transcription and translation processes can only partially explain the observed variability and part of the variability arises from extrinsic sources (Elowitz et al. 2002; Taniguchi et al. 2010). The exact contribution of different sources of fluctuations on the RNA and protein numbers is still unclear. Independent fluctuations from molecular species can contribute to the overall fluctuations by interacting with the transcription machinery. Also, fluctuations can be propagated through molecular species, e.g. fluctuations in RNA numbers causing protein numbers to fluctuate (Paulsson 2005).

Transcription and translation are often assumed to follow Poisson processes where the production probabilities per time unit depend on the promoter occupancy and mRNA numbers, respectively (Paulsson 2005). However, transcription and translation are also known to be complex multi-step processes that exhibit wide sequence-dependent dynamics (Saecker et al. 2011; Lutz et al. 2001; Jones et al. 2014). Also, regulation of transcription has been shown to contribute to the observed dynamics independently of TF occupancy (Garcia et al. 2012). Finally, steps in the transcription and translation elongation can fluctuate greatly (Herbert et al. 2006; Tuller et al. 2010). Unless a single elemental step in the overall process is rate-limiting, gene expression dynamics would exhibit non-exponential time intervals between production events. Recent measurements of time intervals between transcription events in live *E. coli* cells have reported non-Poissonian dynamics in various promoters (Kandhavelu et al. 2011; Kandhavelu et al. 2012b; Muthukrishnan et al. 2012). The shape of time interval distribution was shown to be less dispersed than Poisson process and depend on the promoter sequence, environmental conditions and induction conditions.

Additional diversity in RNA and proteins numbers have been proposed to arise from fluctuations in molecule species involved in gene expression such as σ-factors, transcription factors, ribosomes, and RNAps (Taniguchi et al. 2010; Bakshi et al. 2012; Yang et al. 2014; Jones et al. 2014; Hensel et al. 2012). Other mechanisms not directly related to gene expression such as DNA replication, negative DNA supercoiling, DNA condensation by nucleoid proteins and asymmetries in protein and mRNA partitioning during cell division have been also shown to contribute to the observed variability (Huh and Paulsson 2011; Sanchez and Golding 2013; So et al. 2011; Chong et al. 2014).

Finally, cellular physiology strongly affects gene expression dynamics (Bremer and Dennis 1996; Klumpp et al. 2009). A recent study in *E. coli* reported that fluctuations in gene expression of metabolic enzymes can perturb cell growth, which in turn can propagate back to gene expression, influence even genes unrelated to metabolism (Kiviet et al. 2014). The interdependence between growth and gene expression fluctuations was proposed to be important in coordinating metabolic activities and growth homeostasis. It could also act as a generic source of cellular heterogeneity for the cell population (Balazsi et al. 2011).

3 Theoretical Background

This chapter is an overview of the theoretical concepts of simulation and modeling approaches used in this thesis. It includes the basics about modeling biological systems, a description of stochastic simulation methods and concepts of incorporating complex biological processes involved in transcription and translation.

3.1 Chemical Master Equation

Many biochemical processes involved in gene expression result from the interaction between chemical species that are present in very low copy numbers. E.g., DNA, RNA and regulatory proteins generally have only a few copies per cell (Taniguchi et al. 2010). Regarding the dynamics of interactions between such species, a description of concentration alone is meaningless, and deterministic approaches are not valid, which entails that discrete models are needed (Munsky and Khammash 2008).

To accurately model the time evolution of a system of chemically reacting species, one would have to track each individual molecule through space, detect collisions between the molecules and once a chemical reaction occurs change the populations of the species. Chemical reactions are considered instantaneous and can be divided into two categories: unimolecular reactions, which are internal processes of individual molecules, and bimolecular reactions, which result from the collision and interaction between two molecules. In both cases, the exact timing of the reaction cannot be deduced (Gillespie 2007).

The dynamics of such systems cannot be described by a single trajectory of the system through the state space. Given the discrete nature and the stochastic time evolution of the population, to accurately describe the dynamics of a such system, one must consider the probability distribution of states the system occupies at a certain time moment. For a discrete population of chemically reacting species, the time evolution of this probability distribution is described by the stochastic chemical kinetics (Gillespie 2007).

In the stochastic formulation, a system of molecules of N chemical species homogeneously spread at a time t is represented by an N-dimensional vector $\mathbf{x}$. These chemical species interact through M chemical reactions that can occur between

the species and result in a change in the populations of the species. The system is assumed to have a constant volume and to be well-stirred, which allows the exact trajectories of the particles and non-reactive collisions between them to be ignored (Gillespie 1977). As such, only molecular events that change the populations of the species need to be considered.

The change in the population of the species is a consequence of chemical reactions which are characterized by two quantities. One is the state-change vector v_μ that defines the change in the population species $\mathbf{x}$. The other is the propensity function a_μ of reaction R_μ, which is defined as the following (Gillespie 2007):

$$a_\mu(\mathbf{x})dt = \text{the probability that a particular combination of the molecules that are presently in the system will react via reaction } R_\mu \text{ in the next infinitesimal time interval } [t, t+dt). \tag{3.1}$$

The rationale behind the propensity function depends on which of the two categories the reaction belongs to. For unimolecular reactions, the underlying physics, which often can only be described in quantum mechanical terms, defines the existence of a constant c_μ that gives a probability that this particular molecule will go through the reaction R_μ in the next infinitesimal time moment dt (Gillespie 2007). Overall, the propensity function for X molecules of this species is:

$$a_\mu(\mathbf{x}) = c_\mu X \tag{3.2}$$

For bimolecular reactions, the assumption of a well-stirred system and the kinetic theory define the existence of a constant c_μ that is the probability that single random pair of X_1 and X_2 molecules will react according to the reaction R_μ in the next infinitesimal time window dt (Gillespie 2007). The propensity function of this event is:

$$a_\mu(\mathbf{x}) = c_\mu X_1 X_2 \tag{3.3}$$

In the case of two molecules of the same species reacting together, the propensity function is (Gillespie 2007):

$$a_\mu(\mathbf{x}) = \frac{c_\mu X(X-1)}{2} \tag{3.4}$$

From 3.1 and the probability $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$ of having a given state vector $\mathbf{x}$ at time t after the initial conditions $\mathbf{x} = \mathbf{x}_0$ at $t = t_0$, the time-evolution equation for stochastic chemical kinetics can be derived according to the laws of probability

(Gillespie 2007). The result is a partial differential equation for P called the chemical master equation (CME):

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \sum_{\mu=1}^{M} [a_\mu(\mathbf{x} - v_\mu)P(\mathbf{x} - v_\mu, t|\mathbf{x}_0, t_0) - a_\mu(\mathbf{x})P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \quad (3.5)$$

The CME determines the probability that each species will have a specified molecular population at a given time in the future. The CME simultaneously describes the probability of all possible trajectories as a set of coupled ODEs with one equation for every possible combination of the reactant species. Consequently, the CME can only be analytically solved for the probability density function of $\mathbf{X}(t)$ for a few, very simple systems. To circumvent this problem, the Monte Carlo approach can be used. Namely, multiple numerical realizations of $\mathbf{X}(t)$ trajectories over t can be constructed, in order to sample the distribution of $\mathbf{X}(t)$. This approach was proposed by Gillespie to simulate chemical or biochemical systems of reactions (Gillespie 1976; Gillespie 1977).

3.2 Stochastic Simulation Algorithm

The approach of simulating individual trajectories from $\mathbf{X}(t)$ is not based on $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$ but on a probability function $p(\tau, \mu|\mathbf{x}, t)$ (Gillespie 2007). This function defines the probability that given $\mathbf{X}(t) = \mathbf{x}$, the next reaction to occur in the system will be R_μ and it will occur in the next infinitesimal time interval $[t, t + dt)$.

This joint probability density function in a state $\mathbf{x}$ is a function of two random variables: the time to the next reaction (τ) and the index of the next reaction (μ). The exact formula for $p(\tau, \mu|\mathbf{x}, t)$ can be derived as before by applying the laws of probability to the aforementioned premise 3.1 (Gillespie 1977):

$$p(\tau, \mu|\mathbf{x}, t) = a_\mu(\mathbf{x})e^{-a_0(\mathbf{x})\tau} \quad (3.6)$$

where,

$$a_0(\mathbf{x}) = \sum_{\mu=1}^{M} a_\mu(\mathbf{x}) \quad (3.7)$$

These equations (3.6 and 3.7) are the mathematical basis for the SSA. The time to the next reaction τ is an exponential random variable with a mean of $1/a_0(\mathbf{x})$ while the index of the next reaction μ is a statistically independent integer random variable with point probabilities $a_\mu(\mathbf{x})/a_0(\mathbf{x})$.

This formula is based on the fact that the distribution of the earliest next reaction time is the distribution of the minimum of all next reaction times (see derivation of

this method (Gillespie 1976)). The minimum of a set of independent exponential distributions with different rates is an exponential distribution with a rate equal to the sum of the individual exponentials' rates (Gillespie 1976).

Several Monte Carlo procedures exist for generating samples of τ and μ according to these distributions. In the original formulation of the SSA two methods, the direct method (DM) and the first reaction method (FRM) were proposed (Gillespie 1976; Gillespie 1977). Since then, other sampling procedures such as the next reaction method (NRM) (Gibson and Bruck 2000) and the logarithmic direct method (LDM) (Li and Petzold 2006) have been proposed.

The original DM follows the standard inversion generating method of the Monte Carlo theory (Gillespie 1992): two random numbers r_1 and r_2 are generated from the uniform distribution. These two random numbers are used to generate τ and μ as follows:

$$\tau = \frac{1}{a_0(\mathbf{x})} ln(\frac{1}{r_1}) \tag{3.8}$$

$$\mu = \text{the smallest integer satisfying } \sum_{\mu'=1}^{\mu} a_{\mu'}(\mathbf{x}) > r_2 a_0(\mathbf{x}) \tag{3.9}$$

With help of these formulas (or any other mentioned methods for generating samples of τ and μ), the exact distribution described by the CME can be sampled. The SSA can be used for constructing the exact numerical realization of the process $\mathbf{X}(t)$. Given a start time t_0, a stop time t_{stop}, and an initial vector of species populations $\mathbf{x}_0$, the procedure of the SSA is given in Algorithm 1 (Gillespie 1977).

Algorithm 1 : Stochastic Simulation Algorithm

1: Set $t \leftarrow 0$, and $\mathbf{x} \leftarrow \mathbf{x}_0$.
2: With the system in state $\mathbf{x}$ at time t, evaluate all the $a_\mu(\mathbf{x})$ and their sum $a_0(\mathbf{x})$.
3: Using a suitable sampling procedure, generate a random pair (τ, μ) according to the joint probability distribution defined above by $p(\tau, \mu|\mathbf{x}, t)$.
4: Output the system state for each of the sampling points in the time interval $[t, t + \tau)$.
5: If $t + \tau \geq t_{stop}$, terminate.
6: Set $t \leftarrow t + \tau$, and $\mathbf{x} \leftarrow \mathbf{x} + v_\mu$.
7: Go to step 2.

3.2.1 Delayed SSA

The SSA does not allow explicit delays to be simulated, as the time evolution would no longer be a pure Markov process. However, complex biological processes

such as protein maturation take a non-negligible time to be completed (Cormack et al. 1996; Megerle et al. 2008). Also, transcription elongation, during which the RNAp transcribes thousands of nucleotides, can last up to several minutes (Greive and Von Hippel 2005). Importantly, delays in many cellular processes have been shown to affect the dynamics of gene regulatory networks and, as such, they need to be considered in the simulations (Ribeiro et al. 2010).

The construction of explicit models for complex biological events might not even be possible, as all steps in the process might not be known even though the overall duration of such events can be measured. Adding delays based on such measurements or following arbitrary distributions allow to model these events without knowing all details of the system. A version of the SSA was proposed that allow introducing such events as 'delayed reactions' (Gibson and Bruck 2000). In delayed reactions, the substrates are consumed instantaneously (i.e. removed from the system to avoid further reactions) but the products are released only after a specific time lag. Using delays potentially allows removing many reactions from the system without affecting its dynamics, which speeds up simulations (Gibson and Bruck 2000) without necessarily affecting the realism of the results. A later implementation of the delayed SSA allows multiple delays for each reaction (Roussel and Zhu 2006).

To implement such simulations, reactions with delays can be stored in a wait list L, which is sorted by the time of occurrence. The products of a reaction are placed on the wait list L as a tuple (t_r, i, n), where t_r is the time at which the n molecules of the species S_i are to be released. A heap-based priority queue with the same runtime boundaries can be implemented to run alongside the DM implementation. The steps in this variant of the SSA execution (Roussel and Zhu 2006) are shown in Algorithm 2. The NRM is also well befitted to run wait lists, as both reactions and a wait list can share the same priority queue (Gibson and Bruck 2000).

Algorithm 2 : Delayed SSA

1: Set $t \leftarrow 0$, and $\mathbf{x} \leftarrow \mathbf{x}_0$. Create an empty wait list L for delayed reactions.
2: Perform the normal SSA procedure to generate (τ, μ).
3: If $t_{min} < t + \tau$, where t_{min} is the earliest entry in L then
    Add the earliest molecule in L and add it to $\mathbf{x}$.
    Set $t \leftarrow t + t_{min}$.
4: Else
    Set $t \leftarrow t + \tau$, and $\mathbf{x} \leftarrow \mathbf{x} + v_\mu$.
    If reaction μ has delayed products then
        Add them to L.
5: Go to step 2.

3.3 Modeling Gene Expression

The models built in the stochastic formulation, for simplicity, are represented here as a set of chemical reactions. In general, reactions are presented as follows:

$$\mathrm{A} + \mathrm{B} \xrightarrow{k} \mathrm{C} \tag{3.10}$$

Here, one molecule of the species A and and one molecule of species B react to form a molecule of species C, with a stochastic rate constant $c_\mu = k$.

When modeling gene expression, only the most important steps affecting the overall behavior are included (Ribeiro et al. 2006). For example, the binding of an RNAp to a promoter, the transcription of an RNA followed by the translation of proteins can be modeled as a single, compact reaction, as follows:

$$\mathrm{Pro} + \mathrm{RNAp} \xrightarrow{k_t} \mathrm{Pro} + \mathrm{RNAp} + n\mathrm{P} \tag{3.11}$$

Here, Pro represents the promoter of the gene, RNAp is an RNA polymerase holoenzyme, P is the protein produced and n is the number of proteins produced from a single RNA molecule. Note that the promoter is always available as no regulation have been imposed on the gene making the gene produce in a constitutive manner. Also note that this compaction results in an oversimplification of the model, in that the number of proteins translated from a single RNA is a constant, when this is not true in real biosystems. Another oversimplification is that this model does not account for, e.g. noise in translation.

In reality, the transcription and translation processes are separate and conducted by different macromolecules. To account for this in the model, these processes need to described as separate reactions. This model is already capable of recreating variability in number of proteins produced per mRNA (Yu et al. 2006; Zhu et al. 2007).

$$\mathrm{Pro} + \mathrm{RNAp} \xrightarrow{k_t} \mathrm{Pro} + \mathrm{RNAp} + \mathrm{RBS} \tag{3.12}$$

$$\mathrm{RBS} + \mathrm{Rib} \xrightarrow{k_{tr}} \mathrm{RBS} + \mathrm{Rib} + \mathrm{P} \tag{3.13}$$

In this model, RBS is the binding site for ribosomes in the mRNA, and Rib is the ribosome. This model, while far more realistic than the former one, it still lacks some features of gene expression, such as, e.g., effects of traffic in elongation ribosomes and RNA polymerases.

To model the temporal RNA and protein number in cells, the degradation of these species must accounted for. Both production and degradation processes affect the mean and noise in the RNA and protein numbers (Paulsson 2005). Degradation of RNAs and proteins is also an important factor in the dynamics of genetic circuits

as it dictates the time to reach a steady-state. In prokaryotes, both mRNAs and proteins have been shown to exhibit exponential-like degradation (Taniguchi et al. 2010; Bernstein et al. 2002; Chen et al. 2015a). In other species, this might not be true (Pedraza and Paulsson 2008). Degradation of mRNAs and proteins are thus modeled as follows:

$$\mathrm{RBS} \xrightarrow{k_{rbsd}} \oslash \tag{3.14}$$

$$\mathrm{P} \xrightarrow{k_d} \oslash \tag{3.15}$$

The stochastic mRNA production and decay, in which mRNA production events are uncorrelated and memoryless, is described by a Poisson distribution, for which the variance is equal to the mean. For proteins, the variance is usually higher than the mean as they are produced in bursts from mRNA, i.e. each mRNA is translated several times and these events are separated by much shorter intervals than the intervals between transcription events. Further, the propagation of fluctuations from RNA to proteins is dependent on the lifetimes of both RNA and proteins (Paulsson 2005).

Previous models presented here assume transcription and translation to be instantaneous. In reality, transcription initiation, elongation in both transcription and translation, and protein maturation involves series of chemical reactions which takes a considerable time and thus affect the numbers and fluctuations in mRNA and protein species (McClure 1985; Pedraza and Paulsson 2008; Cormack et al. 1996). The dynamics of these events are gene specific and are known to affect the dynamics of genetic motifs (e.g. with feedback loops) (Bratsun et al. 2005; Ribeiro et al. 2006; Gaffney and Monk 2006). A complex process of transformation from species A to species B can be described by n-step reaction:

$$\mathrm{A} \xrightarrow{k_1} \mathrm{S}_1 \xrightarrow{k_2} \mathrm{S}_2 \ldots \mathrm{S}_{n-2} \xrightarrow{k_{n-1}} \mathrm{S}_{n-1} \xrightarrow{k_n} \mathrm{B} \tag{3.16}$$

$$\mathrm{A} \xrightarrow{k} \mathrm{B}(\tau) \tag{3.17}$$

Multi-step reaction 3.16 can be shortened as reaction 3.17. Here, while the reacting molecule is immediately removed from the system, the produced molecule B is not available to react until τ time has passed. The delays with specific distributions allow to model complex dynamics without explicitly knowing the reactions underlying the dynamics, provided that the overall dynamics is known. Note that, during the delay, the product cannot interact with other species or go through unimolecular reactions such as degradation. In general, this approximation has little effect on the dynamics of genetic motifs and can thus be made.

Gene expression with delayed products can be written as the follows:

$$\text{Pro} + \text{RNAp} \xrightarrow{k_t} \text{Pro}(\tau_{pro}) + \text{RNAp}(\tau_{transc}) + \text{RBS}(\tau_{pro}) \tag{3.18}$$

$$\text{RBS} + \text{Rib} \xrightarrow{k_{tr}} \text{RBS}(\tau_{rbs}) + \text{Rib}(\tau_{transl}) + \text{P}(\tau_{folding}) \tag{3.19}$$

Here, τ_{pro} represents the delay in transcription initiation consisting of e.g. the open complex formation and promoter escape, τ_{transc} represents the transcription elongation time, τ_{rbs} is the time to initiate the translation after binding the RBS, τ_{transl} is the time to complete the translation elongation and finally, $\tau_{folding}$ is time for the folded protein to appear. The gene expression as a complex process with delays no longer is a simple Poisson process. The delays that are gene specific differentiate the dynamics between genes and allow a wide spectrum of behaviors to be expressed (Lutz et al. 2001; Jones et al. 2014).

Most genes in live cells are not constitutively expressing and thus the regulation of gene expression must be taken into account. Most genes in *E. coli* are regulated by means of transcription factors binding the promoter region (Gama-Castro et al. 2011). E.g., the lac promoter in *E. coli* can be bound and regulated by lacI and Crp (Schlax et al. 1995).

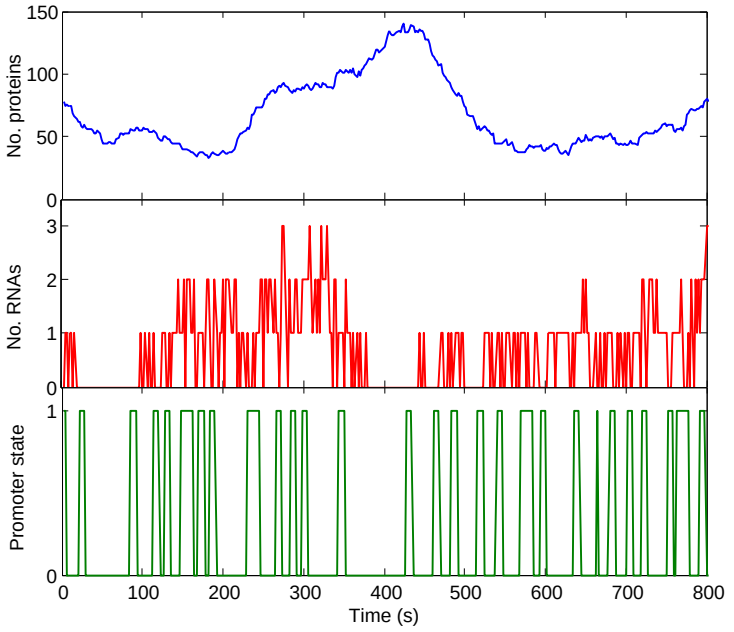


Figure 3.1: Example state trajectory from a model containing transcription and translation with delays, repression and degradation. Top: protein numbers. Middle: RNA numbers. Bottom: promoter state.

Regulation of transcription initiation by transcription factor (TF) can be modeled as follows (Roussel and Zhu 2006; Ribeiro and Kauffman 2007):

$$\text{Pro} + \text{TF} \xrightarrow{k_b} \text{Pro} \cdot \text{TF} \tag{3.20}$$

$$\text{Pro} \cdot \text{TF} \xrightarrow{k_{ub}} \text{Pro} + \text{TF} \tag{3.21}$$

Here, TF is the transcription factor that binds and unbinds promoter region. While bound, e.g. repressor prevents transcription events from initiating. The repression here occurs by a occlusion mechanism (Garcia et al. 2010). Note that during promoter delay, τ_{pro} (in reaction 3.18), repressors cannot bind the promoter. Alternatively, if the TF activates transcription, this can be modeled by an additional reaction:

$$\text{Pro} \cdot \text{TF} \xrightarrow{k_{act}} \text{Pro} \cdot \text{TF} + \text{RNAp}(\tau_{transc}) + \text{RBS}(\tau_{proact}) \tag{3.22}$$

Here, $k_{act} > k_t$. Two types of activators can be considered: activators that recruit RNAp to the promoter, and activators that stimulate the transition rate of bound RNAp from a closed to an open complex the transition from a closed to an open complex form (Lutz et al. 2001). In the model, the first type is based on k_{act} rate and the second type on the promoter delay τ_{proact}.

The models of gene expression shown here provide a simple description of gene expression that can be useful for higher-level studies of, e.g., gene regulatory networks. The model captures the dynamics of transcripts and protein production without the need to explicitly model details of every component. See an example time series from the model in Figure 3.1. Aside the notion of fitting experimental data, the model can be used to explain how different mechanisms affect the fluctuations and further it can be used to generate hypothesis that, in turn, can be tested with experiments.

The delayed model of transcription and translation is used in **Publication I** to model gene expression and interactions between the genes.

3.4 Detailed Model of Transcription and Translation

To model detailed aspects of the dynamics of transcription and translation, more underlying steps than those shown above must be represented explicitly. Compared to the usage of delays, this representation allows regulation of the underlying events (e.g. at the nucleotide level). This also allows fine-grain tuning of the dynamics of the processes (e.g. of transcription and translation elongation). The current knowledge of detailed events that occur during transcription (Greive and Von Hippel 2005; Davenport et al. 2000; Herbert et al. 2006; McClure 1985) and translation (Sørensen and Pedersen 1991; Moore and Sauer 2005; Keiler 2008; Wen et al. 2008) is extensive, and its known that they can have tangible effects on the fluctuations in RNA and protein numbers. Additionally, there is an

interdependence between transcription, translation and mRNA degradation that should be considered (Yarchuk et al. 1992; Proshkin et al. 2010; Yanofsky 2004).

Transcription initiation	Reaction
RNAp binding to the DNA	$RNAp + U_{[n-\Delta_D,\, n+\Delta_D]} \xrightarrow{k_b} O_n$
RNAp unbinding from the DNA	$O_n \xrightarrow{k_u} RNAp + U_{[n-\Delta_D,\, n+\Delta_D]}$
RNAp diffusion on the DNA	$O_n + U_{n+\Delta_D+1} \xrightarrow{k_{1D}} O_{n+1} + U_{n-\Delta_D}$
Closed complex formation	$O_{TSS+\Delta_D} \xrightarrow{k_c} RP_c$
Isomerization	$RP_c + U_{[TSS+1,\, TSS+19]} \xrightarrow{k_i} RP_i$
Open complex formation	$RP_i \xrightarrow{k_o} RP_o$
Elongation complex formation	$RP_o \xrightarrow{k_{ef}} E_{TSS}$
Initial elongation (Scrunching)	$E_{TSS+n} \xrightarrow{k_{el}} E_{TSS+n+1}$
Abortive initiation	$E_{TSS+n} \xrightarrow{k_a} RP_o$
TSS clearance	$E_{TSS+12} + U_{TSS+\Delta_E+12} \xrightarrow{k_{cl}} E_{TSS+13} + U_{[TSS+12,\, TSS+2\Delta_D+12]}$

Figure 3.2: Example reactions from the transcription initiation model. O_n stands for occupied nucleotides (by RNAp or a repressor), where n denotes its location in the DNA sequence, and U_n stands for the nth unoccupied nucleotide. Ranges of nucleotides are denoted as $U_{[start,end]}$. The range occupied by the RNAp is referred to as [n-Δ_D, n+Δ_D]. TSS refers to transcription start site location. RP_c, RP_i and RP_o refers to closed, isomerized and open complex form of RNAp, respectively. E refers to elongation complex. The stochastic rate constants are shown above each reaction arrow. For more detailed description, see **Publication III**.

The transcription initiation process at the promoter region involves several sequential steps: finding of the promoter site through different forms of diffusion, opening of the DNA, and initiation of the RNA synthesis. These reactions can be modeled explicitly by adding reactions for each nucleotide on the promoter region. For example reactions, see Figure 3.2. The first step is the non-specific binding of the RNAp to random location of the DNA template. The footprint of the RNAp covers a specific range of nucleotides on the DNA, which prevents binding of other molecules in that location. 1D diffusion on the template is modeled to occur one nucleotide at a time in a random direction chosen initially. The

contribution of 1D diffusion to the finding of the start site might not be relevant when compared to 3D diffusion, depending on the reaction rates (Friedman et al. 2013; Dangkulwanich et al. 2014). The diffusing RNAp can unbind randomly from the template at any position. Multiple RNAps moving on the DNA template can cause collisions, e.g. in convergent promoters (Callen et al. 2004), and in the collision events, the model assumes that one or both RNAps are removed from the template (Sneppen et al. 2005).

Binding of the RNAp to the transcription start site forms the closed complex between RNAp and the promoter. The dynamics of the subsequent, consecutive initiation reactions, including isomerization and open complex formation, follow elemental steps with means extracted from *in vitro* measurements (Buc and McClure 1985; Lutz et al. 2001; Saecker et al. 2011). The promoter escape occurs through a step-wise abortive initiation process that has been shown to release from small to very high amounts of abortive transcripts, depending on the promoter. The latter case reflects the instability of the initial elongation complex (Hsu 2002; Hsu et al. 2003). Following this, RNAp commences the production of nascent RNA.

The model also allows binding of repressor molecules on the specific sites in the promoter region. Depending on the specific DNA region occupied by the repressor, different reactions can be prevented by steric occlusion. The repressor can prevent RNAp binding, isomerization or promoter escape and each mode of repression exhibits different dynamics (Sanchez et al. 2011). In **Publication III** stochastic nucleotide level model of the promoter region was used to study the dynamics of transcription initiation in closely spaced promoters.

The elongation phase of transcription and translation proceeds in steps, each of which involving the addition of a specific nucleotide or amino-acid. RNAp molecules have been shown to move at varying rates along the template strand (Herbert et al. 2006; Landick 2009). This is due to alternative reaction pathways, e.g. pausing, pyrophosphorolysis or editing, which compete with normal elongation and can reduce the overall elongation rate (Greive and Von Hippel 2005). Additionally, reactions such as transcriptional pausing can generate traffic on the template, as they force multiple trailing RNAps to stop at very close distances and then proceeding jointly forward in elongation. Traffic slows down the overall transcription rate to an extent, but its main effect is the induction of bursty production of RNAs. Stochastic transcription elongation models have been shown to be reproduce these results (Dobrzynski and Bruggeman 2009; Klumpp and Hwa 2008; Rajala et al. 2010). The probability of RNAp entering alternative pathways not only depends on the nucleotide sequence but also on transcription elongation factors (Davenport et al. 2000; Herbert et al. 2006; Herbert et al. 2010).

To accurately model elongation, reactions for each pathway with specific rates in each template position must be stated. In Figure 3.3 example reactions for coupled transcription and translation elongation processes are shown. RNAps

Transcription	Reaction	Translation	Reaction
Elongation	$A_n + U_{n+D_{RNAp}+1} \xrightarrow{k_e} O_{n+1} + U_{n-D_{RNAp}} + U^R_{n-D_{RNAp}}$	Initiation	$Rib + U^R_{[1,D_{Rib}+1]} \xrightarrow{k_{ini}} O^R_1 + Rib^R$
Activation	$O_n \xrightarrow{k_a} A_n$	Stepwise translocation	$A^R_{n-3} + U^R_{[n+D_{Rib}-3,n+D_{Rib}-1]} \xrightarrow{k_e} O^R_{n-2}$ $O^R_{n-2} \xrightarrow{k_e} O^R_{n-1}$ $O^R_{n-1} \xrightarrow{k_e} O^R_n + U^R_{[n-D_{Rib}-2,n-D_{Rib}]}$
Pausing	$O_n \xrightarrow{k_a/f_p} O_{n_p}$	Activation	$O^R_n \xrightarrow{k_{ini}\cdot k_a} A^R_n$
Premature termination	$O_n \xrightarrow{k_{prem}} RNAp + U_{[n-\Delta_{RNAp},n+\Delta_{RNAp}]}$	Back-translocation	$O^R_n + U^R_{[n-D_{Rib}-2,n-D_{Rib}]} \xrightarrow{k_b} A^R_{n-3} + U^R_{[n+D_{Rib}-3,n+D_{Rib}-1]}$
Pyrophosphorolysis	$O_n + U_{n-D_{RNAp}-1} + U^R_{n-D_{RNAp}-1} \xrightarrow{k_{pyro}} O_{n-1} + U_{n+D_{RNAp}-1}$	Drop-off	$O^R_n \xrightarrow{k_{do}} Rib + U^R_{[n-D_{Rib},n+D_{Rib}]}$
Completion	$A_{n_{last}} \xrightarrow{k_e} RNAp + U_{[n_{last},n_{last}-D_{RNAp}]}$	Completion	$A^R_{n_{last}} \xrightarrow{k_e \cdot k_{ini}} Rib + U^R_{[n_{last},n_{last}-D_{Rib}]} + P_{prem}$

Figure 3.3: Example reactions from the transcription and translation elongation model. In the transcription model, A_n, O_n and U_n depict for the nth nucleotide when activated, occupied, and unoccupied, respectively. Ranges of nucleotides are denoted such as $U_{[start,end]}$. Each RNAp occupies $(2\Delta_{\mathrm{RNAp}}+1)$ nucleotides. U^R_n denotes transcribed ribonucleotides (denoted by the R superscript) which are free, i.e. not under the RNAp's footprint. In the translation model, each ribosome occupies $(2\Delta_{\mathrm{Rib}}+1)$ ribonucleotides. A^R_n denotes that a ribosome has created peptide bond for the peptide coded by the codon at position $[n-2, n]$, where n is a multiple of 3 ($n = 3, 6, 9, \ldots$). The activation reaction has a codon-specific rate (Sørensen and Pedersen 1991). The stochastic rate constants are shown above each reaction arrow. For more detailed description, see **Publication IV**.

(and ribosomes) occupy a specific range of nucleotides on the template. This is modeled with occupied nucleotides that are occupied and released as the elongation ensues. During elongation, ribonucleotides are released from each RNAp loci, allowing ribosomes to translate the sequence. Each reaction in elongation states the requirement of different molecular species for each pathway. The stochastic rates for each pathway can be defined for each template position separately, e.g. codons are translated with variable rates independent of the tRNA species (Sørensen and Pedersen 1991). The impact of slow-translating codons on ribosome collisions and translation efficiency has been studied in different codon sequences (Mitarai et al. 2008). This study showed that neither deterministic nor stochastic models with uniform translation rates are able to reproduce the experimental results of translation elongation rates. For simulated movement of RNAps and ribosomes on the DNA and mRNA, respectively, from the model in **Publication IV**, see Figure 3.4.

In **Publication IV** the transcription and translation elongation models were

combined to study the coupling between the two elongation processes. In combination with the model of transcription initiation at the promoter region, the final model can used to study complex regulatory patterns of transcription and translation dynamics and the effects of stochastic events during these processes.

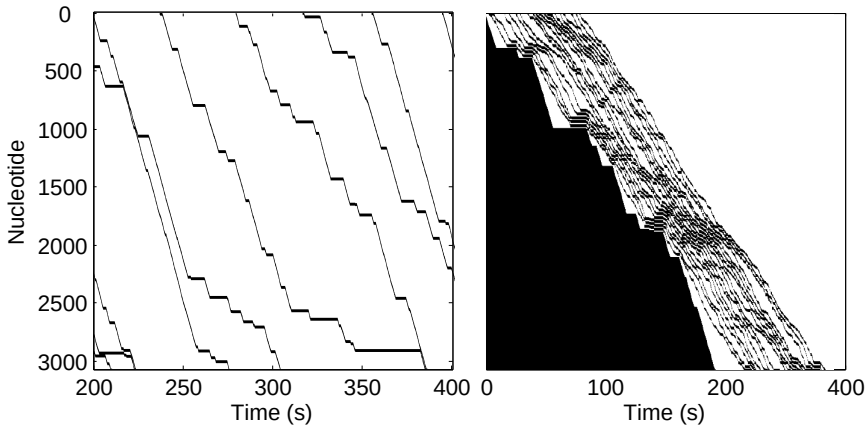


Figure 3.4: Movement of RNAps and ribosomes during the elongation processes observed from the elongation model. (Left) RNAps transcribing a DNA template over time. Horizontal lines are momentary pauses during the elongation processes. (Right) Ribosomes translating nascent RNA. The continuous black area depicts the transcription progression of the RNAp, which is followed by a number of ribosomes. In both processes, pauses cause traffic between elongating macromolecules. For more detailed description of the model, see **Publication IV**.

3.4.1 SGNS2

To accurately model the coupled transcription and translation elongation, each RNA molecule state must be explicitly modeled to correctly account for gradual degradation of RNA, traffic, and interaction between RNAp and the ribosomes simultaneously in multiple RNAs. This can be achieved by utilizing transient compartments that contain the necessary reactions. These transient compartments allow molecules to react with other molecules of the same compartment or between compartments and their containing compartments, for a certain amount of time. The same set of reactions can be utilized in multiple compartments (e.g. two different compartments containing identical RNA molecules will have the same set of possible reactions). See illustration in Figure 3.5.

A simulator of chemical reaction systems (SGNS2) have been proposed, which can simulate reactions according to the Stochastic Simulation Algorithm with multi-delayed reactions (Lloyd-Price et al. 2012). Importantly, this simulator utilizes the concept of transient compartments as in (Spicher et al. 2008). SGNS2 is based on NRM which allows to incorporate other simulation algorithms as

sub-simulations (Gibson and Bruck 2000). The NRM priority queue of each compartment defines the 'next reaction to happen' for the overall NRM priority queue which can add and remove entire sub-simulations at runtime (Lloyd-Price et al. 2012).

Two main features of SGNS2 are: transient, interlinked, hierarchical compartments that can be created, destroyed and divided at run time; and support for multiple molecule and compartment partitioning schemes, applicable separately for each molecular species (Lloyd-Price et al. 2012). Aside the coupled transcription and translation elongation, these novel features allow to simulate, e.g. biased partitioning of protein aggregates in cell division that affect cell fitness or aggregation and the functioning of small genetic networks, among other (Lindner et al. 2008; Lloyd-Price et al. 2012; Gupta et al. 2015).

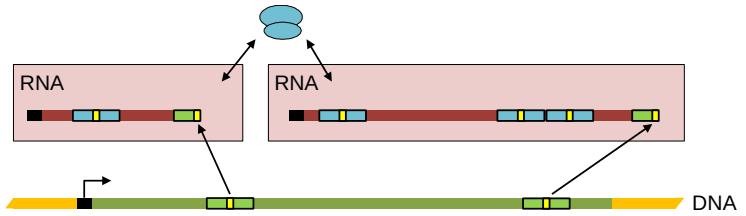


Figure 3.5: A transient compartment consists of a virtual independent space where a set of reactions take place. The single-nucleotide transcription model places the transcribed ribonucleotides into RNA compartments. Meanwhile, multiple single-nucleotide translation models can operate inside each RNA compartment. The dynamics of the RNAps guarantees the temporal ordering of the RNA production.

An early version of SGNS2 was used in **Publication IV** to simulate the dynamics of coupled transcription and translation model.

3.5 Finite State Projection Algorithm

The finite state projection (FSP) algorithm differs from the SSA in that it provides a direct solution or approximation of the CME without requiring the computation of large numbers of sample time traces (Munsky and Khammash 2006). In the case of any Markov process containing only a finite number number of states, the FSP method provides an exact analytical solution. When the number of possible states is infinite, the approximate solution provided by the FSP method guarantees, unlike simulation based methods, its own accuracy. The size of the finite state projection can be systematically increased until a specific accuracy is achieved (Munsky and Khammash 2006; Munsky et al. 2015).

From the CME (see equation 3.5), the probability mass functions for all possible

states can be collected into vector form (Munsky and Khammash 2006):

$$\mathbf{P}(t) = \left[\mathbf{P_0^T P_1^T} \dots\right] \tag{3.23}$$

which allows to write the CME in a simplified matrix:

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{A}\mathbf{P}(t) \tag{3.24}$$

In this expression, $\mathbf{A}$, the infinitesimal generator, has the following elements (Munsky and Khammash 2006):

$$A_{j,i} = \begin{cases} -\sum_{\mu=1}^{M} \alpha_\mu(\mathbf{x}_i) & \text{if } i = j \\ \alpha_\mu(\mathbf{x}_i) & \text{for every } j \text{ such that } \mathbf{x}_j = \mathbf{x}_i + \nu_\mu \\ 0 & \text{elsewhere} \end{cases} \tag{3.25}$$

The FSP algorithm provides an approximation to the CME solution by, instead of analyzing infinite set of all possible states, selecting a finite subset of states that still captures most of the probability for the specified finite time interval (Munsky and Khammash 2006). E.g. in transcription only include states where the RNA number is less than some integer N_m, while the rest of the states are transformed into a single absorbing state. The reduced master equation has the form:

$$\frac{d}{dt}\begin{bmatrix} \mathbf{P}^{FSP}_{\leq N_m}(t) \\ \mathbf{g}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{\leq N_m} & \mathbf{0} \\ \mathbf{-1}^T\mathbf{A}_{\leq N_m} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{P}^{FSP}_{\leq N_m}(t) \\ \mathbf{g}(t) \end{bmatrix} \tag{3.26}$$

The FSP solution, $P^{FSP}_{\leq N_m}(t)$, is an approximation of the CME and $g(t)$ is the computable error in the approximation (Munsky et al. 2015). Theorems guarantee that the FSP is a lower bound of the true solution $\mathbf{P}^{FSP}_{\leq N_m} \leq \mathbf{P}_{\leq N_m}$, and the total error in the approximation is $\left|\mathbf{P}_{\leq N_m} - \mathbf{P}^{FSP}_{\leq N_m}\right| \leq g(t)$ (Munsky et al. 2015). In practice, the matrix can be extended until the solution is within some error tolerance ($g(t) < \epsilon$) and in some scenarios such as the maximum number of RNAs, this truncation can be applied directly. An illustration of modeling transcription using the FSP approach is shown in Figure 3.6.

While the order of the CME is significantly reduced by the FSP algorithm, the reduction might not be sufficient for more complicated systems. For this, the method can be further improved or modified to extend it's capabilities (Munsky 2008). E.g. the dynamics can be projected onto a lower dimensional slow reactions manifold, multiple periods of time can be solved using different projections, or part of the states can be interpolated at the cost of loss of accuracy. Nevertheless, even with the aforementioned improvements, the FSP cannot be applied to systems with many interacting chemical species. Instead, the advantages of the FSP method are the speed and the precision of the solution for systems with small number of possible configurations.

The FSP has been used along with single-cell experiments to study the dynamics of various systems, such as in a recent study measured the osmotic stress response pathway in Saccharomyces cerevisiae using the FISH method for RNA detection (Neuert et al. 2013). Models with varying complexity were generated with FSP and parameter estimation and cross-validation were used to select the most predictive model. Additionally, the method has been used to quantify e.g. the transcriptional activity of the proto-oncogene c-Fos at individual endogenous alleles (Senecal et al. 2014).

In **Publication II**, the FSP approach was used to model the intake kinetics of inducer and the production dynamics of RNAs.

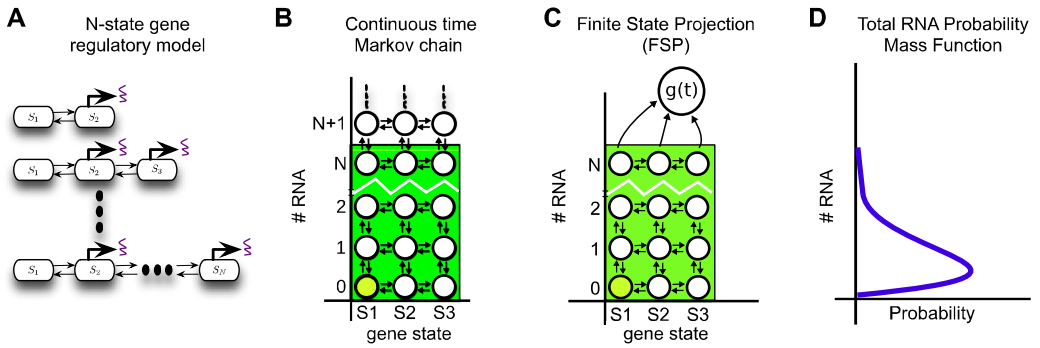


Figure 3.6: Illustration of modeling transcription using the FSP approach. (A) Multiple states of a gene with different rates of transcription. Adding more gene states in the model will increase the complexity of behaviors. (B) A lattice for all possible combinations of gene (x-axis) and number of RNAs (y-states) states. The lattice is infinite as the number of RNAs can, in theory, exceed any finite bound. (C) The FSP algorithm truncates the infinite state space at N RNAs. Reactions exceeding the truncated value are absorbed into a sink state with a probability $g(t)$. (D) As a result, a finite state space is used so as to estimate the probability of the system to be in each state within that space. Consequently, the RNA distribution at each time moment can be produced. Reproduced with permission from (Munsky et al. 2015).

4 Measurements and Analysis

This chapter is an overview of the fluorescence microscopy techniques and analysis methods employed in this thesis. These methods include single-RNA detection methods, cell and spot segmentation, RNA quantification and change-point detection algorithms.

4.1 Fluorescent Proteins and Microscopy

Fluorescent proteins were discovered in the early 1960s (Shimomura et al. 1962) and were successfully cloned in the 1990s (Prasher et al. 1992). Since then, fluorescent proteins have become one of the most used tools in biological sciences. This is mostly due to the simplicity of fusing fluorescent proteins with cellular target proteins (Tsien 1998). The availability of fluorescent proteins has evolved to cover most of the visible spectrum of light (Shaner et al. 2004; Day and Davidson 2009). Also, development in the field have led to the appearance of novel properties of fluorescent probes, such as photoactivation and photoconversion (Day and Davidson 2009; Wu et al. 2011). These fluorescent proteins can be switched on and off or be converted to a different emission wavelength by using specific wavelengths of excitation light. These properties are the foundation for many advanced imaging techniques in microscopy, e.g. super-resolution microscopy (Huang et al. 2009).

An optimal fluorescent protein for single cell microscopy has high quantum yield and brightness, favorable photo-physical properties, and sufficient inertness so as not to interfere with the functioning of the target molecule (Shaner et al. 2004). Drawbacks of many fluorescent proteins include blinking (intensity fluctuations) and limited photo-stability (Ha and Tinnefeld 2012). Organic fluorophores, compared to fluorescent proteins, have a smaller size, and superior stability and brightness (Pitchiaya et al. 2014). Nevertheless, the possibility of expressing fluorescent proteins fused to desired cellular target proteins inside a cell makes fluorescent proteins convenient for many studies.

To accurately detect fluorescent proteins, the emitted fluorescence signal have to be significantly above the cellular background fluorescence, referred as autofluorescence. To achieve this, first, the selection of a bright fluorophore that absorbs

and emits light outside the spectrum of the autofluorescence is recommended (Ha and Tinnefeld 2012). Second, media exhibiting a low autofluorescence should be selected. Third, the choice of light source, optics, illumination scheme, and detector contribute directly to the signal-to-noise ratio in microscopy experiments.

The most commonly used illumination scheme in fluorescence microscopy is a wide-field epi-illumination. The epi-illumination excites the entire depth of the sample causing the out-of-focus fluorescent molecules also to contribute towards the background fluorescence. To avoid out-of-focus illumination, several methods have been developed to restrict the illumination volume of the sample: confocal microscopy (Pawley 2006), total internal reflection fluorescence (TIRF) microscopy (Axelrod 1981), and highly inclined and laminated optical sheet (HILO) microscopy (Konopka and Bednarek 2008). Confocal microscopy is based on reducing the focal volume and consequently the out-of-focus light with a pinhole (Pawley 2006). A drawback is that as the sample is illuminated only one volume at a time and has to be scanned, making the imaging slower than in wide-field techniques. The scan-speed can be improved with setups, such as the spinning-disc confocal microscopy, which simultaneously illuminate multiple regions of the sample (Nakano 2002).

In TIRF microscopy, only a thin section at the sample surface is illuminated (Axelrod 1981). Light in total internal reflection creates a thin lamina of evanescent wave that penetrates the coverglass-sample surface and excites molecules within approximately 150 nm from the surface. The low penetration depth of the TIRF microscopy allows only to probe molecules close to the coverglass surface, e.g. in the cell membrane. TIRF has been more commonly used to study processes *in vitro*. To increase the penetration depth without significantly reducing the signal-to-noise ratio, the HILO microscopy was developed (Konopka and Bednarek 2008; Tokunaga et al. 2008). In the HILO microscopy, light is refracted into the sample at high inclination angle (less than the critical angle in TIRF) only illuminating an angled layer within the sample, resulting in lower out-of-focus fluorescence.

4.2 Single-Molecule Approaches for RNA Detection

One of the earliest implementations of a single molecule technique to study biological processes was the observation of single β-galactosidase molecules trapped into microscopic droplets by using fluorogenic substrate to measure the quantity (Rotman 1961). Another example is the measurement of the unidirectional movement of kinesin driving plastic beads along microtubules *in vitro* (Gelles et al. 1988). Also, enzymatic reactions of single cholesterol oxidase molecules have been observed using a real-time single-molecule approach (Lu et al. 1998). Among other findings, the analysis of single-molecule measurements have shown fluctuations in the rate of reactions.

In light microscopy, Abbe's law defines the limit of ability to distinguish two features located closer than half of the wavelength of the illuminating or the

emitted light. This is directly related to the ability of detecting single fluorescent molecules in biological samples. In order to accurately detect fluorescent molecules in the observation area, the number of the target molecules must be low enough to distinguish them. This is problematic as many molecules are present in high quantities, e.g. RNAps and ribosomes in *E. coli*.

Various strategies have been utilized to limit the number of molecules in the observation area (Pitchiaya et al. 2014). For example, the expression of the target can be limited (Yu et al. 2006), the delivery of the probes can be controlled (Santangelo et al. 2009), or the visualization of the probes at any given time can be limited to a subset of the probes (Huang et al. 2009). The latter strategy is used in super-resolution microscopy. These methods are based on turning only a few fluorescent proteins into a bright state simultaneously and cycling through the total population of fluorescent proteins in a stochastic manner. This is achieved by using photoactivatable or photoconvertible fluorephores (Day and Davidson 2009). Stochastic optical reconstruction microscopy (STORM) or photoactivation localization microscopy (PALM) can achieve up to 10 nm spatial resolution, even with a high copy numbers of target molecules (Huang et al. 2009; Walter et al. 2008).

Methods have been developed to probe RNA numbers *in vivo* with fluorescent probes. Techniques of labeling RNAs generally use two different schemes, direct and indirect labeling. Indirect labeling involves sequence-complementary oligonucleotides or fluorophore labeled RNA binding probes, which bind to a specific RNA motif (Pitchiaya et al. 2014). Direct labeling uses chemically reactive functional groups or structural motifs in the RNA, which can be naturally present or introduced by chemical synthesis and RNA modifying proteins, for fluorophore attachment (Pitchiaya et al. 2014). Currently, indirect labeling methods are more popular due to the possibility of probing endogenous RNAs, in addition to exogenous constructs (Raj and van Oudenaarden 2009).

One of the first methods to achieve single RNA sensitivity was the fluorescence *in situ* hybridization method (Raj and van Oudenaarden 2009). The method is based on the fluorescently labeled oligonucleotide probes that specifically hybridize to its complementary sequence on the RNA. It can be used to probe the cell-to-cell variability in endogenous RNAs, which cannot normally be quantified with population level measurements such as qPCR, microarrays, or deep-sequencing (Raj and van Oudenaarden 2009). Additionally, FISH can be used to measure the spatial localization of RNAs inside the cell (Montero Llopis et al. 2010; Lecuyer et al. 2007). The protocol for FISH generally entails fixation of the cells, permeabilization of the cell membrane, hybridization of probes to their target RNA sequences, extensive washing of the cells to remove non-bound probes, and image acquisition (Gasnier et al. 2013).

The difficulties in the FISH methodology that affect the quantification of RNA numbers are the following: the variability in fluorophores labeled to the oligonu-

cleotides, long probes with a poor cell-membrane permeability, and low signal-to-noise ratio caused by the unbound and non-specifically bound probes (Pitchiaya et al. 2014). To avoid these problems, various approaches have been proposed. E.g. multiple short probes that bind to the adjacent sequences within the target RNA for improved cell-membrane permeability, probes that minimize the proximity mediated fluorescence self-quenching (Femino 1998; Raj et al. 2008), and single fluorophore labeled oligonucleotides that improve the labeling homogeneity of the RNA (Raj et al. 2008; Taniguchi et al. 2010)

4.2.1 MS2-GFP Tagging Method

Since RNA binding proteins can be fused with fluorescent proteins, using them instead of oligonucleotides to label RNAs has been proposed to detect single RNAs. This method allows detection and tracking of single RNAs in living cells (Fusco et al. 2003; Golding et al. 2005; Coulon et al. 2013). The method is based on fusing an endogenous RNA with multiple copies of an RNA motif. This motif is then bound by a specific protein fused to a fluorescent protein. The RNA binding fusion protein should be extensively expressed in the cell prior to the measurement (see Figure 4.1) so as to be able to detect any target RNA. Both constructs (the target RNA and the RNA binding fusion protein) can be genetically engineered into a plasmid transfected into the cell or integrated into the genome.

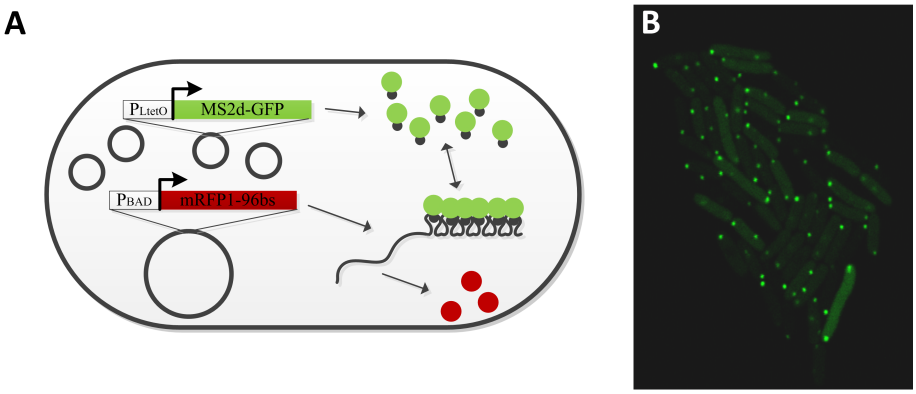


Figure 4.1: *In vivo* RNA detection with MS2-GFP method. (A) Illustration of the MS2-tagging system for RNA detection. Target RNA carrying 96 MS2 biding sites is produced under the control of a promoter in a single copy F-plasmid (large circle). MS2-GFP molecules (green balls) are produced by a high copy-number plasmid (small circles). Once the target RNA is transcribed, MS2-GFP molecules bind to it. The target RNA also has a coding region for a red fluorescent protein (red balls). (B) Example fluorescence microscope image of *E. coli* cells expressing both target RNAs and MS2-GFP proteins. Individual RNA molecules are visible as fluorescent spots. The uniform background of the cells is due to the unbound MS2-GFP diffusing inside the cells.

The most common high-affinity RNA binding protein used in RNA detection is the MS2 coat protein derived from the MS2 bacteriophage (Fusco et al. 2003). The protein binds to a 21 nt long RNA fragment that spontaneously forms a stem-loop secondary structure (Keryer-Bibens et al. 2008). Other proteins used for the RNA binding are the PP7, derived from the PP7 bacteriophage (Larson et al. 2011; Chao et al. 2008), and the λ_N peptide, derived from the lambda bacteriophage (Lange et al. 2008; Daigle and Ellenberg 2007). The binding sites of the tagging proteins are orthogonal between MS2, PP7 and λ_N, i.e. the MS2 does not bind the PP7 binding site and vice-versa (Lange et al. 2008; Chao et al. 2008). This allows to simultaneously imaging up to three independent RNA targets, or probe three different regions of a single RNA, using a combination of MS2, PP7 and λ_N systems (Hocine et al. 2013; Lange et al. 2008).

The binding of multiple tagging proteins (an RNA binding protein fused with a fluorescent protein) to the same target RNA renders it much brighter than the fluorescence from freely diffusing unbound tagging proteins and the cellular autofluorescence (Fusco et al. 2003; Golding et al. 2005; Xie et al. 2008). Further, the unbound tagging proteins normally diffuse much faster than the time resolution of the image acquisition, blurring the fluorescence over a large area inside the cell. On the other hand, the target RNA bound by multiple tagging proteins moves slowly and can be detected as a distinct diffraction limited spot. However, the highly expressed, unbound tagging proteins contribute significantly to the background, creating need for a high number of binding sites in the target RNA. In the first implementation of the method, 24 binding sites of the tagging protein were used in the target RNA, resulting in the binding of 48 fluorescent proteins, as the MS2 protein binds as a dimer (Valegard et al. 1994). Versions with a lower (Haim et al. 2007; Fusco et al. 2003) or a higher number of binding sites have been used (Golding and Cox 2004). As an alternative technique to improve the signal-to-noise ratio, recent studies have utilized split green fluorescent protein (GFP) fragments to reduce the background fluorescence (Kerppola 2006; Wu et al. 2014).

Certain drawbacks of the method hinder its usage as a probe for RNA dynamics. These are to be considered in the analysis of the microscopy experiment. First, an incomplete and heterogeneous binding of tagging proteins to target RNA affect the quantification as the amount of fluorescence fluctuates between tagged RNAs (Fusco et al. 2003; Wu et al. 2012). Next, the binding of a large number of the tagging proteins to the target RNA can affect the mobility, functioning or localization of target RNAs (Wu et al. 2012; Oliveira et al. 2016). Also, the target RNA becomes protected against natural degradation by the bound tagging proteins (Tran et al. 2015; Muthukrishnan et al. 2012). Lastly, this method can only be used to probe genetically engineered RNAs due to the requirement of the RNA binding sites.

In **Publication I** and **Publication II** MS2-GFP tagging method was used to measure the production of target RNAs from a single promoter in live cells.

4.3 Image Analysis and Data Extraction

To accurately estimate the number of RNAs in each cell from fluorescence microscopy images, image analysis and signal processing methods must be used. This section describes the methods used in **Publication I** and **Publication II**.

4.3.1 Image Analysis and RNA Quantification

The first step in the analysis is to segment the cells from a background. The image is manually divided into separate regions occupied by each cell. From these regions, the location, the orientation and the dimensions of the cell are extracted by using principal component analysis (PCA) (Kandhavelu et al. 2012a). This segmentation method requires the cells not to be too clustered. For cell clusters, a more sophisticated method of cell segmentation has been proposed, which employs a multi-scale morphological edge detection with an image denoising algorithm (Chowdhury et al. 2013). For time series data, the images are temporally aligned using a cross-correlation to remove any drift during the image acquisition process, e.g. due to temperature changes or movement of the stage, since small drifts altering the position of the cells can complicate significantly the tracking of cells over time. The tracking of cells during the time series is conducted by associating a cell with the most overlapping cell in the previous frame. If there exist two cells with significant overlap in the previous frame beyond a threshold, the cell is expected to have divided.

To measure the production of RNA molecules in each cell, the diffraction limited spots of the MS2-GFP tagged RNAs must be segmented. The intensity distribution of a diffraction-limited spot can be mathematically described by a point spread function (PSF) and can be approximated by a two-dimensional Gaussian function (Ruusuvuori et al. 2010). The spots inside each cell area are automatically segmented using a Kernel Density Estimation (KDE) method for spot detection (Ruusuvuori et al. 2010). A probability density of intensity values is used to compare the likelihood of pixel intensities with a threshold to obtain a binary image of the spots. A circular window and a Gaussian kernel were used for the detection. The threshold was obtained with Otsu's method (Otsu 1979).

The spot intensities are corrected for the background fluorescence as follows. The average intensity of the cell outside the spots, consisting of unbound MS2-GFP molecules, is multiplied by the area of the spot and then subtracted from the total intensity of each spot. This background corrected spot intensity can then be used to quantify the number of RNAs in each cell. The number of RNA molecules in the cell can be extracted from the intensity histogram from all spot intensities by normalizing it with the intensity of a single tagged RNA molecule (equivalent to the first peak in the intensity histogram) (Golding et al. 2005). The consecutive peaks correspond to the integer valued number of RNAs.

Recently, an automatic method for quantifying the fluorescent spot intensities

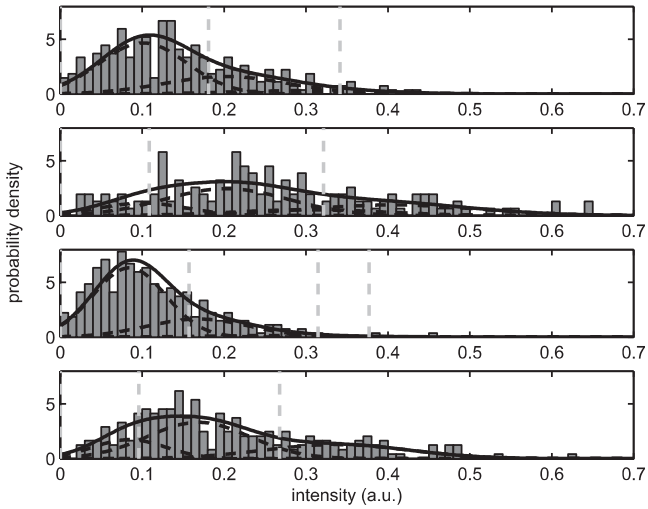


Figure 4.2: Distribution of measured intensities from MS2-GFP-tagged RNAs with different number of binding sites (96 or 48 bs). Panels from top to bottom: spot intensities (96 bs), cell intensities (96 bs), spot intensities (48 bs) and cell intensities (48 bs). The solid black lines shows the overall estimated distributions, the dashed black lines their components and the dashed gray lines the decision boundaries. Reproduced with permission from (Häkkinen et al. 2014).

was proposed (Häkkinen et al. 2014). This other method consists of a numerical maximum likelihood parameter estimation followed by a maximum a posterior classification. This method is applicable to any fluorophore-tagged molecule quantification if the molecules in a cell are present in low-copy numbers. The advantage of an automatic method is that it does not rely on a human intervention, which in many cases complicates comparison between the experiments (Häkkinen et al. 2014). Also, the distribution of the number of RNA molecules from a single molecule experiment is often noisy and a simple rounding can cause errors. This quantification error is expected to increase with the number of tagged RNAs as the variance increases. An example of RNA number detection is shown in Figure 4.2.

The cell and spot segmentation methods were used **Publication I** and **Publication II**. The RNA quantification methods were used in **Publication II** to extract the RNA numbers over time in the cells.

4.3.2 Measurement of Intervals

Information from time series data can be used to quantify the RNA production dynamics in single cells. The time series data contains more information than a stationary RNA distribution from a cell population. Since the MS2-GFP tagged

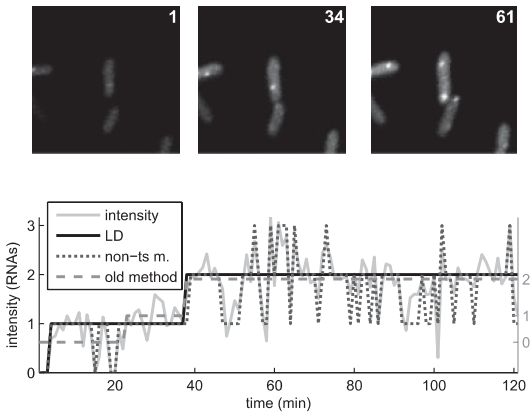


Figure 4.3: Detection of RNA production events from temporal data. Upper panel: florescence microscope images at 1, 34 and 61 min after the start of time series. Lower panel: example intensity series and fit curves using the least-deviations (LD) method and two other methods. RNA numbers estimated by the LD method are shown by the curves. Reproduced with permission from (Häkkinen and Ribeiro 2015).

RNAs do not degrade in the time span of several hours (Golding et al. 2005; Muthukrishnan et al. 2012; Tran et al. 2015), the total spot intensity increases over time as the production of RNAs ensues (Kandhavelu et al. 2012a). Thus, the moment of appearance of a novel RNA molecule in a cell results in a discrete jump in the background corrected total spots intensity of the cell, given an accurate sampling.

These discrete jumps in the total spots intensity of the cell are used to measure the time intervals between consecutive production events of novel RNAs. The method does not provide the absolute number of RNAs if the production does not initiate from zero RNAs. From a time series of a cell population, this method can be used to extract a distribution of time intervals in RNA production (Muthukrishnan et al. 2012; Kandhavelu et al. 2012b). The automatic method in (Kandhavelu et al. 2012a) fits the total spots intensity over time to a monotone piecewise-constant function by least squares (LSQ) fitting. The model order is selected using the F-test. Each jump has been shown to correspond to the production of a single RNA molecule (Kandhavelu et al. 2012a). Recently, an improved method for countering outliers observed in the intensity time series, e.g. spots moving out-of-focus, proposes least-deviations (LD) cost for the detection (Häkkinen and Ribeiro 2015). An example of the jump detection in RNA production is shown in Figure 4.3.

This method can be also used to quantify the time for the first RNA production event to occur following induction. This requires usage of microfluidics, as the cells have to be induced under the microscope observation. The time series acquisition

is initiated simultaneously to the induction of target promoter and the duration for the first RNA production events to occur is counted from the start of induction. Examples of distributions of time intervals between transcription events and of the waiting times for the first transcription event following induction are shown in Figure 4.4

The methods of measurement of time intervals between RNA production events and the waiting time for the first production event were used in **Publication II** to quantify these processes in different promoters and induction schemes.

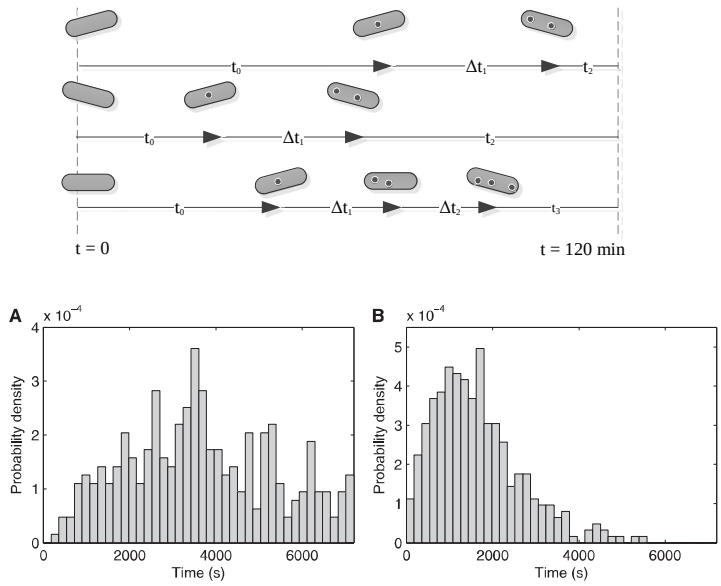


Figure 4.4: Distributions of time intervals. Top: Description of the waiting time for the first RNA production (t_0) and intervals between subsequent transcription events (Δt). Bottom: Example kinetics of the intake and production. (A) Probability density distribution of waiting times for the first RNA to be produced in cells (B) Probability density distribution of intervals between transcription events.

4.4 Change Point Detection Methods

Automatic change point detection methods are often general and can be used to detect changes in the dynamics of the system without explicitly knowing the exact nature of the change. These methodologies recognize candidates for the moment of changes in the dynamics. The earliest approaches were based on the Behrens–Fisher problem, a statistical hypothesis test of equal means (Belloni and Didier 2008; Fisher 1939). These approaches, e.g. the Welch's t-test, assume normal distributions which makes them sensitive for heavy-tailed distributions that many biological processes exhibit (Taniguchi et al. 2010). The detection

in non-parametric cases is still, in general, an open problem. The different approaches are based on statistics, density estimation, theory of kernel machines and classification (Kawahara and Sugiyama 2009; Harchaoui et al. 2009).

The problem of change point detection can be formulated as the following. Given a multidimensional time series $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathrm{R}^n$, where the time moment K represents a change in the dynamics. Given the data samples in the M-point backward window $\mathbf{X}_B = (\mathbf{x}_{K-M}, \ldots, \mathbf{x}_{K-1})$ and the M-point forward window $\mathbf{X}_F = (\mathbf{x}_{K+1}, \ldots, \mathbf{x}_{K+M})$, the dissimilarity of the two windows can be posed as a hypothesis testing problem:

$$\begin{cases} H_0 : p_{\mathbf{X_F}}(\mathbf{x}) = p_{\mathbf{X_B}}(\mathbf{x}) \\ H_1 : p_{\mathbf{X_F}}(\mathbf{x}) \neq p_{\mathbf{X_B}}(\mathbf{x}) \end{cases} \tag{4.1}$$

where $p_{\mathbf{X_F}}(\mathbf{x})$ and $p_{\mathbf{X_B}}(\mathbf{x})$ denote the probability density functions of the forward and backward windows, respectively.

Two recent change point detection methods, namely, the unconstrained least-squares importance fitting (uLSIF) (Kawahara and Sugiyama 2009) and kernel change point analysis (KCpA) (Harchaoui et al. 2009) have been reported to exhibit good performance, when compared with alternative methods. The first detection method attempts to model the densities based on the data (Kawahara and Sugiyama 2009). Instead of estimating both windows separately and measuring their similarity with e.g. the Kolmogorov-Smirnov test, this method considers the likelihood of a change using the density ratios directly. For this, there are estimation methods, such as Kullback-Leibler importance estimation procedure (KLIEP) (Sugiyama et al. 2008) and uLSIF (Kanamori et al. 2009). The density ratio $w(\mathbf{x}) : \mathbf{R}^n \to \mathbf{R}$ is the following:

$$w(\mathbf{x}) = \frac{p_{\mathbf{X_F}}(\mathbf{x})}{p_{\mathbf{X_B}}(\mathbf{x})}, \tag{4.2}$$

where $p_{\mathbf{X_F}}(\mathbf{x})$ and $p_{\mathbf{X_B}}(\mathbf{x})$ are the probability densities of the forward and backward windows, respectively.

The second method, the KCpA method, as all kernel methods, is based on mapping the data into a higher dimensional feature space, which allows the data in the backward and forward windows to be modeled with linear models. The kernel methods were originally developed for pattern recognition but they have since been applied to the problem of change point detection (Harchaoui et al. 2009; Desobry et al. 2005; Schölkopf and Smola 2002). Harchaoui and colleagues used a kernel-based binary classifier to separate the forward and backward windows based on the dissimilarity of the two sets (Harchaoui et al. 2009). Their separability can be directly measured by using the Kernel Fisher Discriminant, which defines the ratio of the between-class-variance and the within-class-variance.

Most parameters of change point detection algorithms can be inferred from the data by cross-validation. However, the window length cannot be determined from the training data due to multiple time scales of the change and the algorithm cannot know which time scale is relevant. Thus, the time scale of the processes to be studied must be known beforehand. Provided a prior knowledge on the time scale of the changes, automatic methods to detect changes in time series data accurately and robustly are important for future analysis of gene regulatory networks.

In **Publication I** these methods were used to detect changes in simulations of gene expression in small and large gene networks and in measurements of transcription in live cells using the MS2-GFP tagging method.

5 Conclusions and Discussion

This thesis has studied the transcription process in *Escherichia coli* using stochastic modeling approaches and RNA measurements with single molecule sensitivity. The four publications contribute to this by, first, presenting a new tool for automatically detecting changes in time series data from simulations or measurements (**Publication I**). This was followed by a study that quantifies the dynamics of induction and subsequent transcription process (**Publication II**), a study of transcription initiation in closely spaced promoters (**Publication III**), and a study of coupled transcription and translation elongation (**Publication IV**).

In **Publication I** a new method for detecting non-spurious changes in simulated data and time-lapse microscopy of gene expression was proposed. The method finds candidate moments when the the dynamics of the system changed from time series data. Two recent methods, the density ratio method (Harchaoui et al. 2009) and the kernel change point analysis (Kawahara and Sugiyama 2009), are based on a non-parametric approach for the detection problems. These methods show good performance compared to alternative methods, as expected, given that they presently represent state-of-the-art approaches to the problem.

To assess the accuracy of these methods, the ground truth of the changes must be known. Thus, the performance of the two methods was tested in detecting changes in the dynamics of delayed stochastic models of small genetic networks, including a toggle-switch (Gardner et al. 2000), and a large genetic network (Chowdhury et al. 2010). The changes that occur during the time series were implemented so as to mimic natural changes in the dynamics, e.g. different mean expression levels and noise levels, or the switching between producing and non-producing states. These changes can occur for various reasons, such as gene copy numbers change as a results of DNA replication (Peterson et al. 2015), a change in the transcription factor concentration (Ozbudak et al. 2004; Kandhavelu et al. 2012b), or a change in the environment (Young et al. 2013). In most tests, the kernel method outperforms the density method. This might be due to the nature of the changes aimed to detect. Overall, the methods showed good accuracy in detecting the moments when changes occurred.

Finally, the methods were used to detect novel transcription events from time-lapse microscopy where the RNAs in live *E. coli* cells were detected by the MS2-GFP

RNA tagging method (Golding et al. 2005). Both methods accurately detected the production moments of novel RNAs.

Most detection parameters can be inferred from the data by cross-validation, but the window length for the detection was found to have a significant effect on the accuracy of detection. Changes in the dynamics in gene networks usually have multiple time scales in which they occur, e.g. gene expression levels are affected by events that can range from a simple repressor unbinding to a complex cell division process. Consequently, some information on the time scales of the process and/or changes must be known beforehand, in order to detect change points accurately. This is especially true for gene regulatory networks capable of exhibiting a wide range of dynamics.

In **Publication II**, the contributions of induction and the subsequent transcription process to the cell-to-cell diversity in RNA numbers were quantified for a few promoters and induction schemes. First, the waiting times for the production of the first RNA following introduction of inducers to the environment and the subsequent intervals between transcription events were measured using the MS2-GFP RNA tagging method (Golding et al. 2005) in live *E. coli* cells. Both processes were found to exhibit broad distributions of intervals. The dynamics of intake is determined by the induction mechanism and the extra-cellular concentration of inducers. Meanwhile, the transcription dynamics of active promoters is mostly dictated by the promoter sequence and the presence of transcription factors.

The induction kinetics showed a surprisingly broad distribution of waiting times in all measured cases. This was shown to have a tangible effect on the diversity in RNA numbers of a cell population, increasing it much above than expected from noise in the transcription process alone. This effect is, however, transient in that as the transcription is initiated, the RNA numbers in cells become defined more and more by the kinetics of transcription and RNA degradation. To estimate the duration of the transient caused by the non-negligible times of the intake process, both the intake and transcription processes were studied using the FSP algorithm (Munsky and Khammash 2006). Both processes were modeled as d-step processes, each step with an exponentially distributed duration corresponding to an elementary step. The models were fitted with experimental data in different experimental conditions and compared with a model where transcription is fully active since the first time moment of the observation window.

The cell-to-cell variability in RNA numbers increased transiently due to the intake process. Relevantly, this effect was found to last for a long period after the start of induction, in that a large fraction of the population was slowly induced. Also, different intake mechanisms exhibited different dynamics which had non-negligible effects on the RNA population statistics. The regulation of gene activation might be an important source of variability in the population, as it allows additional coordination of transcription dynamics that is gene independent.

In **Publication III**, the dynamics of transcription initiation in closely spaced

promoters were investigated using stochastic nucleotide level models. Compared to an isolated promoter, dynamics from closely spaced promoters is affected by RNAps interfering with each other during binding, transcription initiation, and elongation (Sneppen et al. 2005). The mechanisms of interference include occlusion, RNAp collisions, road blocks and sitting duck interference (see section 2.4), which are easily modeled with the nucleotide level model, where bound RNAps and transcription factors reserve a range of nucleotides on the DNA.

First, the binding rate of RNAps to the promoter region was found to be non-uniform due to rate-limiting steps in transcription initiation. Importantly, the localization of the promoters' start sites and the promoter's geometry were found to affect the distributions of intervals between transcription initiations due to the interferences between the promoters. Also, the interference causes transient correlations between consecutive transcription initiations of the dual promoter system, in that, the spatial closeness causes a promoter to be more likely to express more times in a row than if spatially more separated. This correlation is dependent on the existence of a multiple rate-limiting steps transcription initiation process and affected by the orientation and distance between promoter sites.

Next, repression mechanisms of transcription initiation were studied. A change in the location of the repressor binding site leads to qualitatively distinct regulatory behaviors, as it causes the repressor to interact differently with the transcriptional machinery. E.g. inhibiting the promoter binding, the transcription initiation rate is proportional to RNAP binding rate and can be reduced by increasing the repressor concentration. Meanwhile, inhibition of the open complex formation or promoter escape is partially independent of the repressor numbers, as the fraction of time that gene is expressed, is also controlled by the dissociation constant of the repressor and the rate-limiting steps in transcription initiation.

Finally, a single repressor binding site can simultaneously repress both closely spaced promoter sites. The mechanism by which the repression occurs differs between promoters, causing differences in repression strength. Relevantly, using shared repression sites allow coordinating the activity between the promoter sites.

In **Publication IV**, the coupled transcription and translation elongation was studied using stochastic models of transcription and translation at the nucleotide and codon level. During transcription and translation elongation, alternative reaction pathways, including pausing, premature termination, and backtracking, can occur in various DNA template positions, in accordance with experimental data (see section 2.2.2). Consequently, transcription of a gene consisting of a few thousand nucleotides can give rise to complex dynamics. For example, the stochastic events during elongation affect the distribution of distances between elongating RNAps (Dobrzynski and Bruggeman 2009) and between elongating ribosomes (Mitarai et al. 2008), which gives rise to bursts in RNA and peptides production. Further, collisions between elongating macromolecules can attenuate bursts introduced by transcription initiation (Dobrzynski and Bruggeman 2009).

To study the propagation of fluctuations, transcription and translation initiation rates were varied, as these directly affect the average distance between elongating macromolecules. To maintain the mean mRNA and protein levels unaltered with changing initiation rates, the degradation of mRNA and protein species were tuned accordingly. Fluctuations in the RNA numbers were found to decrease with increasing transcription initiation. This was due to the existence of a promoter open complex formation step, that was shown to reduce collisions and bursting between consecutive RNA productions. In the absence of this event or significant reduction of its duration, the distribution of time intervals would become exponential-like. In this case, increasing initiation rates would lead to more collisions with RNAps.

The fluctuations in the transcription process were found to propagate to translation, due to these processes being dynamically coupled. Reduction of the translation initiation rate led to a decrease in ribosome traffic on the RNA, which consequently reduced the fluctuations in the protein numbers. Reduction in the fluctuation was also partly caused by the de-coupling of transcription and translation. This was shown by calculating the normalized maximum correlation between time-series of protein and mRNA numbers for each set of parameters. Finally, transcriptional stalling e.g. due to transcriptional arrests (Davenport et al. 2000) was found to simultaneously affect both transcription and translation elongation processes.

Recent measurements with single molecule sensitivity have shown the extent of heterogeneity in mRNA and protein numbers in *E. coli* cells in homogenous environments (So et al. 2011; Taniguchi et al. 2010; Jones et al. 2014). While the observation of full distributions of molecule numbers is interesting *per se*, dynamics in single cells is more informative about the underlying processes. In particular, the possibility of observing single cell trajectories following single cell events, when complemented with modeling approaches, can provide mechanistic understanding of the processes underlying the kinetics (Norman et al. 2013; Nachman et al. 2007). The methodology of measuring time intervals between events or waiting times from a perturbation, utilized in this thesis, can be used to understand molecular events involved in, e.g., gene activation, repairing of DNA, switching between cell states etc. (Norman et al. 2013; Uphoff et al. 2013). The properties of fluctuations on the time intervals inform about the mechanisms involved in the process and hypothesized mechanisms can be tested with modeling approaches, e.g. how the model responds to perturbations (Uphoff et al. 2016).

Current methods of tagging RNAs with fluorescent proteins in live cells allow simultaneous measurement of up to three different RNAs or regions of RNA, in addition to a wide spectrum of fluorescent proteins that can be fused with target proteins (Hocine et al. 2013; Lange et al. 2008; Bakshi et al. 2013). Even with limited number of probes, careful selection of molecular species to probe would allow detection of sequential events in transcription, such as TF binding, open complex formation, elongation of the nascent RNA, and degradation of RNA

(Larson et al. 2011). Conceivably, all sequential steps resulting in gene expression could be monitored in individual cells. Advanced microscopy techniques, such as combinatorial probe labeling (Lubeck and Cai 2012; Chen et al. 2015b), could provide methods to visualize all intermediates of the process in a single cell.

Naturally, genes do not act as independent units inside the cell, neither are they the sole controllers of RNA and protein numbers. Several studies have proposed that a large portion of the observed variability in mRNA and protein numbers arises from other sources than the transcription and translation processes themselves (So et al. 2011; Taniguchi et al. 2010; Sanchez and Golding 2013). Non-gene-specific factors involved in transcription and translation, such as σ-factors, transcription regulators, ribosomes and RNA polymerases, have been shown to fluctuate significantly (Taniguchi et al. 2010; Bakshi et al. 2012; Yang et al. 2014). Additional mechanisms, including DNA replication, negative DNA supercoiling, asymmetric protein and mRNA partitioning during cell division, and cellular aging, contribute also to the observed RNA and protein numbers (Peterson et al. 2015; Chong et al. 2014; Huh and Paulsson 2011; Lindner et al. 2008). It will be interesting to quantify the contributions of these sources of fluctuations on the dynamics of gene expression in the future. This would provide a more embedded view of gene expression and its association with the functioning of the organism as a whole.

Overall, gene expression has been shown to exhibit a wide range of dynamics, likely due to the existence of multiple regulatory mechanisms. Further, the possibility of attaining similar kinetics through different mechanisms complicates the characterizing of a particular process. These complications ought to be tackled with a combination of advanced measurement techniques and novel modeling approaches. As most cellular processes depend on gene expression, the more details we understand of it, the more understanding we will have of cellular behaviors in general.

Bibliography

Acar, M., Becskei, A., and van Oudenaarden, A. "Enhancement of cellular memory by reducing stochastic transitions". *Nature* 435.7039 (2005), pp. 228–232.

Adhya, S. and Gottesman, M. "Control of Transcription Termination". *Annual Review of Biochemistry* 47 (1978), pp. 967–996.

Adhya, S. and Gottesman, M. "Promoter occlusion: transcription through a promoter may inhibit its activity". *Cell* 29.3 (1982), pp. 939–944.

Afroz, T., Biliouris, K., Kaznessis, Y., and Beisel, C. L. "Bacterial sugar utilization gives rise to distinct single-cell behaviours". *Molecular Microbiology* 93.6 (2014), pp. 1093–1103.

Alberts, B, Johnson, A, Lewis, J, Raff, M, Roberts, K, and Walter, P. *Molecular Biology of the Cell*. Garland Science, USA, 2002.

Amit, R., Garcia, H. G., Phillips, R., and Fraser, S. E. "Building enhancers from the ground up: A synthetic biology approach". *Cell* 146.1 (2011), pp. 105–118.

Arkin, A., Ross, J., and Mcadams, H. H. "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in". *Genetics* 149 (1998), pp. 1633–48.

Arndt, K. M. and Chamberlin, M. J. "Transcription termination in *Escherichia coli*: Measurement of the rate of enzyme release from rho-independent terminators". *Journal of Molecular Biology* 202.2 (1988), pp. 271–285.

Axelrod, D. "Cell-substrate contacts illuminated by total internal reflection fluorescence". *Journal of Cell Biology* 89.1 (1981), pp. 141–145.

Bakshi, S., Dalrymple, R. M., Li, W., Choi, H., and Weisshaar, J. C. "Partitioning of RNA polymerase activity in live *Escherichia coli* from analysis of single-molecule diffusive trajectories". *Biophysical Journal* 105.12 (2013), pp. 2676–2686.

Bakshi, S., Siryaporn, A., Goulian, M., and Weisshaar, J. C. "Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells". *Molecular Microbiology* 85.1 (2012), pp. 21–38.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., and Leibler, S. "Bacterial Persistence as a Phenotypic Switch". *Science* 305 (2004), pp. 1622–1625.

Balazsi, G., van Oudenaarden, A., and Collins, J. J. "Cellular decision making and biological noise: From microbes to mammals". *Cell* 144.6 (2011), pp. 910–925.

Bar-Nahum, G. and Nudler, E. "Isolation and characterization of σ70-retaining transcription elongation complexes from *Escherichia coli*". *Cell* 106.4 (2001), pp. 443–51.

Beck, C. F. and Warren, R. A. "Divergent promoters, a common form of gene organization." *Microbiological Reviews* 52.3 (1988), pp. 318–326.

Belloni, A. and Didier, G. "On the Behrens-Fisher Problem: A globally convergent algorithm and a finite-sample study of the wald, LR and LM tests". *Annals of Statistics* 36.5 (2008), pp. 2377–2408.

Bendtsen, K. M., Erdssy, J., Csiszovszki, Z., Svenningsen, S. L., Sneppen, K., Krishna, S., and Semsey, S. "Direct and indirect effects in the regulation of overlapping promoters". *Nucleic Acids Research* 39.16 (2011), pp. 6879–6885.

Bernstein, J. A., Khodursky, A. B., Lin-Chao, S, and Cohen, S. N. "Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays". *Proceedings of the National Academy of Sciences of the United States of America* 99.15 (2002), pp. 9697–9702.

Bertrand-Burggraf, E., Lefèvre, J. F., and Daune, M. "A new experimental approach for studying the association between RNA polymerase and the tet promoter of pBR322". *Nucleic Acids Research* 12.3 (1984), pp. 1697–1706.

Blattner, F. R., Plunkett, G. I., Bloch, A. C., Perna, T. N., Burland, V., Riley, M., Collado-Vides, J., Glasner, D. J., Rode, K. C., Mayhew, F. G., Gregor, J., Davis, W. N., Kirkpatrick, A. H., Goeden, A. M., Rose, J. D., Mau, B., and Shao, Y. "The Complete Genome Sequence of *Escherichia coli* K-12". *Science* 277.5331 (1997), pp. 1453–1462.

Bratsun, D., Volfson, D., Tsimring, L. S., and Hasty, J. "Delay-induced stochastic oscillations in gene regulation." *Proceedings of the National Academy of Sciences of the United States of America* 102.41 (2005), pp. 14593–14598.

Bremer, H and Dennis, P. "Modulation of chemical composition and other parameters of the cell by growth rate". *Neidhardt, F. (ed.). Washington, DC: American Society for Microbiology Press* 122 (1996), p. 1553.

Browning, D. F. and Busby, S. J. W. "The regulation of bacterial transcription initiation". *Nature Reviews Microbiology* 2.1 (2004), pp. 57–65.

Buc, H and McClure, W. R. "Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lacUV5 promoter. Evidence for a sequential mechanism involving three steps." *Biochemistry* 24.11 (1985), pp. 2712–2723.

Callen, B. P., Shearwin, K. E., and Egan, J. B. "Transcriptional interference between convergent promoters caused by elongation over the promoter". *Molecular Cell* 14.5 (2004), pp. 647–656.

Chamberlin, M. "The selectivity of transcription". *Ann. Rev. Biochem.* 43.3 (1974), pp. 721–775.

Chao, J. A., Patskovsky, Y., Almo, S. C., and Singer, R. H. "Structural basis for the coevolution of a viral RNA-protein complex". *Nature Structural and Molecular Biology* 15.1 (2008), pp. 103–105.

Chen, H., Shiroguchi, K., Ge, H., and Xie, X. S. "Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*". *Molecular Systems Biology* 11.781 (2015), pp. 1–11.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. "Spatially resolved, highly multiplexed RNA profiling in single cells." *Science* 348.6233 (2015), aaa6090.

Cho, B.-K., Zengler, K, Qiu, Y, Park, Y. S., Knight, E. M., Barrett, C. L., and B, P. "The transcription unit architecture of the *Escherichia coli* genome". *Nature Biotechnology* 27 (2009), pp. 1043–1049.

Choi, P. J., Xie, X. S., and Shakhnovich, E. I. "Stochastic Switching in Gene Networks Can Occur by a Single-Molecule Event or Many Molecular Steps". *Journal of Molecular Biology* 396.1 (2010), pp. 230–244.

Choi, P., Cai, L., Frieda, K, and Xie, X. "A stochastic single-molecule event triggers phenotype switching of a bacterial cell". *Science* 322 (2008), pp. 442–446.

Chong, S., Chen, C., Ge, H., and Xie, X. S. "Mechanism of Transcriptional Bursting in Bacteria". *Cell* 158.2 (2014), pp. 314–326.

Chowdhury, S., Kandhavelu, M., Yli-Harja, O., and Ribeiro, A. S. "Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE)." *BMC Bioinformatics* 14 Suppl 1.10 (2013), S8.

Chowdhury, S., Lloyd-Price, J., Smolander, O.-P., Baici, W. C. V., Hughes, T. R., Yli-Harja, O., Chua, G., and Ribeiro, A. S. "Information propagation within the Genetic Network of Saccharomyces cerevisiae." *BMC systems biology* 4.1 (2010), p. 143.

Cormack, B. P., Valdivia, R. H., and Falkow, S. "FACS-optimized mutants of the green fluorescent protein (GFP)." *Gene* 173.1 (1996), pp. 33–38.

Coulon, A., Chow, C. C., Singer, R. H., and Larson, D. R. "Eukaryotic transcriptional dynamics: from single molecules to cell populations." *Nature reviews. Genetics* 14.8 (2013), pp. 572–584.

Courtney, C. M. and Chatterjee, A. "cis-Antisense RNA and Transcriptional Interference: Coupled Layers of Gene Regulation". *Journal of Gene Therapy* 2.1 (2014), p. 9.

Crick, F. "Central Dogma of Molecular Biology". *Nature* 227 (1970), pp. 561–563.

Daigle, N. and Ellenberg, J. "λN-GFP: an RNA reporter system for live-cell imaging". *Nature Methods* 4 (2007), pp. 633–636.

Dangkulwanich, M., Ishibashi, T., Bintu, L., and Bustamante, C. "Molecular mechanisms of transcription through single-molecule experiments". *Chemical Reviews* 114.6 (2014), pp. 3203–3223.

Davenport, R. J., Wuite, Gijs, J. L., Landick, R., and Bustamante, C. "Single-Molecule Study of Transcriptional Pausing and Arrest by *E. coli* RNA Polymerase". *Science* 287 (2000), p. 2497.

Davis, C. A., Bingman, C. A., Landick, R., Record, M. T., and Saecker, R. M. "Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by *Escherichia coli* RNA polymerase." *Proceedings*

of the National Academy of Sciences of the United States of America 104.19 (2007), pp. 7833–7838.

Day, R. N. and Davidson, M. W. "The fluorescent protein palette: tools for cellular imaging". *Chem Soc Rev* 38.10 (2009), pp. 2887–2921.

Desobry, F, Doncarli, M. D. C., Davy, M., and Doncarli, C. "An online Kernel Change Detection Algorithm". *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 2961–2974.

Dobrzynski, M. and Bruggeman, F. J. "Elongation dynamics shape bursty transcription and translation." *Proceedings of the National Academy of Sciences of the United States of America* 106.8 (2009), pp. 2583–2588.

Eliasson, A, Bernander, R, Dasgupta, S, and Nordström, K. "Direct visualization of plasmid DNA in bacterial cells." *Molecular Microbiology* 6.2 (1992), pp. 165–170.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. "Stochastic gene expression in a single cell." *Science* 297.5584 (2002), pp. 1183–1186.

Englesberg, E., Irr, J., Power, J., and Lee, N. "Positive Control of Enzyme Synthesis by Gene C in the Positive Control of Enzyme Synthesis by Gene C in the L-Arabinose System". *Journal of Bacteriology* 90.4 (1965), pp. 946–957.

Erie, D., Hajiseyedjavadi, O, Young, M., and Hippel, P. von. "Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription". *Science* 262.5135 (1993), pp. 867–873.

Errington, J. "Regulation of endospore formation in Bacillus subtilis". *Nature Reviews Microbiology* 1.2 (2003), pp. 117–126.

Fang, G., Rocha, E. P. C., and Danchin, A. "Persistence drives gene clustering in bacterial genomes". *BMC Genomics* 9 (2008), p. 4.

Farabaugh, P. J. "Programmed translational frameshifting." *Annual Review of Genetics* 30.1 (1996), pp. 507–528.

Femino, A. M. "Visualization of Single RNA Transcripts in Situ". *Science* 280.5363 (1998), pp. 585–590.

Fisher, R. "The comparison of samples with possibly unequal variances". *Ann. Eugenics* 9 (1939), pp. 174–180.

Friedman, L. J. and Gelles, J. "Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation". *Cell* 148.4 (2012), pp. 679–689.

Friedman, L. J., Mumm, J. P., and Gelles, J. "RNA polymerase approaches its promoter without long-range sliding along DNA." *Proceedings of the National Academy of Sciences of the United States of America* 110.24 (2013), pp. 9740–5.

Fritz, G., Megerle, J. A., Westermayer, S. A., Brick, D., Heermann, R., Jung, K., Rädler, J. O., and Gerland, U. "Single cell kinetics of phenotypic switching in the arabinose utilization system of *E. coli*". *PLoS ONE* 9.2 (2014).

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S. M., Blanchard, J. M., Singer, R. H., and Bertrand, E. "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells". *Current Biology* 13.2 (2003), pp. 161–167.

Gaffney, E. A. and Monk, N. A. M. "Gene Expression Time Delays and Turing Pattern Formation Systems Gene Expression Time Delays and Turing Pattern Formation Systems". *Bulletin of Mathematical Biology* 68 (2006), pp. 99–130.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J. S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., and Collado-Vides, J. "RegulonDB version 7.0: Transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units)". *Nucleic Acids Research* 39.SUPPL. 1 (2011), pp. 98–105.

Garcia, H. G., Sanchez, A., Boedicker, J. Q., Osborne, M., Gelles, J., Kondev, J., and Phillips, R. "Operator sequence alters gene expression independently of transcription factor occupancy in bacteria". *Cell Reports* 2.1 (2012), pp. 150–161.

Garcia, H. G., Sanchez, A., Kuhlman, T., Kondev, J., and Phillips, R. "Transcription by the numbers redux: Experiments and calculations that surprise". *Trends in Cell Biology* 20.12 (2010), pp. 723–733.

Gardner, T. S., Cantor, C. R., and Collins, J. J. "Construction of a genetic toggle switch in *Escherichia coli*". *Nature* 403.6767 (2000), pp. 339–342.

Gasnier, M., Dennis, C., Vaurs-Barrière, C., and Chazaud, C. "Fluorescent mRNA labeling through cytoplasmic FISH." *Nature protocols* 8.12 (2013), pp. 2538–47.

Gelles, J., Schnapp, B. J., and Sheetz, M. P. "Tracking kinesin-driven movements with nanometre-scale precision". *Nature* 331 (1988), pp. 450–453.

Gibson, M. A. and Bruck, J. "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels". *The Journal of Physical Chemistry A* 104.9 (2000), pp. 1876–1889.

Gillespie, D. T. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". *Journal of Computational Physics* 22.4 (1976), pp. 403–434.

Gillespie, D. T. "A rigorous derivation of the chemical master equation". *Physica A: Statistical Mechanics and its Applications* 188.1-3 (1992), pp. 404–425.

Gillespie, D. T. "Exact Stochastic Simulation of Coupled Chemical Reactions". *The Journal of Physical Chemistry* 81.25 (1977), pp. 2340–2361.

Gillespie, D. T. "Stochastic simulation of chemical kinetics." *Annual Review of Physical Chemistry* 58 (2007), pp. 35–55.

Golding, I. and Cox, E. C. "RNA dynamics in live *Escherichia coli* cells." *Proceedings of the National Academy of Sciences of the United States of America* 101.31 (2004), pp. 11310–11315.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. "Real-time kinetics of gene activity in individual bacteria." *Cell* 123.6 (2005), pp. 1025–1036.

Goldman, S. R., Ebright, R. H., and Nickels, B. E. "Direct detection of abortive RNA transcripts *in vivo*." *Science* 324.5929 (2009), pp. 927–8.

Greive, S. J. and Von Hippel, P. H. "Thinking quantitatively about transcriptional regulation." *Nature Reviews Molecular Cell Biology* 6.3 (2005), pp. 221–232.

Gupta, A., Lloyd-Price, J., and Ribeiro, A. S. "In silico analysis of division times of *Escherichia coli* populations as a function of the partitioning scheme of non-functional proteins". *In Silico Biology* 12.1-2 (2015), pp. 9–21.

Ha, T. and Tinnefeld, P. "Photophysics of Fluorescence Probes for Single Molecule Biophysics and Super-Resolution Imaging". *Annu Rev Phys Chem* 63.2 (2012), pp. 595–617.

Haim, L., Zipor, G., Aronov, S., and Gerst, J. E. "A genomic integration method to visualize localization of endogenous mRNAs in living yeast". *Nature Methods* 4 (2007), pp. 409–412.

Häkkinen, A. and Ribeiro, A. S. "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data". *Bioinformatics* 31.1 (2015), pp. 69–75.

Häkkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S. "Estimation of fluorescence-tagged RNA numbers from spot intensities". *Bioinformatics* 30.8 (2014), pp. 1146–1153.

Häkkinen, A., Healy, S., Jacobs, H. T., and Ribeiro, A. S. "Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse". *Journal of Theoretical Biology* 281.1 (2011), pp. 74–83.

Hammar, P., Leroy, P., Mahmutovic, a., Marklund, E. G., Berg, O. G., and Elf, J. "The lac Repressor Displays Facilitated Diffusion in Living Cells". *Science* 336.6088 (2012), pp. 1595–1598.

Hammar, P., Walldén, M., Fange, D., Persson, F., Baltekin, O., Ullman, G., Leroy, P., and Elf, J. "Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation." *Nature Genetics* 46.4 (2014), pp. 405–8.

Harchaoui, Z, Bach, F, and Moulines, E. "Kernel Change-point Analysis". *Advances in Neural Information Processing Systems (NIPS)* 21 (2009), pp. 609–616.

Harden, T. T., Wells, C. D., Friedman, L. J., Landick, R., Hochschild, A., Kondev, J., and Gelles, J. "Bacterial RNA polymerase can retain σ70 throughout transcription". *Proceedings of the National Academy of Sciences of the United States of America* 113.3 (2016), pp. 602–607.

Harley, C. B. and Reynolds, R. P. "Analysis of *E. coli* promoter sequences." *Nucleic Acids Research* 15.5 (1987), pp. 2343–2361.

Hawley, D. K., Johnson, A. D., and McClure, W. R. "Functional and physical characterisation of transcription initiation complexes in the bacteriophage lambda O R region". *J.Biol.Chem.* 260.14 (1985), pp. 8618–8626.

Helling, R and Weinberg, R. "Complementation studies of arabinose genes in *Escherichia coli*". *Genetics* 48 (1963), pp. 1397–1410.

Heltzel, A, Lee, I., Totis, P., and Summers, A. "Activator-dependent preinduction binding of sigma-70 RNA polymerase at the metal-regulated mer promoter". *Biochemistry* 29 (1990), pp. 9572–9584.

Hensel, Z., Feng, H., Han, B., Hatem, C., Wang, J., and Xiao, J. "Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis". *Nature Structural and Molecular Biology* 19.8 (2012), pp. 797–802.

Herbert, K. M., Zhou, J., Mooney, R. A., Porta, A. L., Landick, R., and Block, S. M. "*E. coli* NusG inhibits backtracking and accelerates pause-free transcription by promoting forward translocation of RNA polymerase". *Journal of Molecular Biology* 399.1 (2010), pp. 17–30.

Herbert, K. M., La Porta, A., Wong, B. J., Mooney, R. A., Neuman, K. C., Landick, R., and Block, S. M. "Sequence-Resolved Detection of Pausing by Single RNA Polymerase Molecules". *Cell* 125.6 (2006), pp. 1083–1094.

Hippel, P. H. von, Bear, D. G., Morgan, W. D., and McSwiggen, J. A. "Protein-Nucleic Acid Interactions in Transcription: A Molecular Analysis". *Annual Review of Biochemistry* 53 (1984), pp. 389–446.

Hocine, S., Raymond, P., Zenklusen, D., Chao, J. a., and Singer, R. H. "Single-molecule analysis of gene expression using two-color RNA labeling in live yeast." *Nature Methods* 10.2 (2013), pp. 119–21.

Hogg, R. W. and Englesberg, E. "L-arabinose binding protein from *Escherichia coli* B/r." *Journal Of Bacteriology* 100.1 (1969), pp. 423–432.

Hsu, L. M. "Promoter clearance and escape in prokaryotes." *Biochimica et Biophysica Acta* 1577.2 (2002), pp. 191–207.

Hsu, L. M., Vo, N. V., Kane, C. M., and Chamberlin, M. J. "In vitro studies of transcript initiation by *Escherichia coli* RNA polymerase. 1. RNA chain initiation, abortive initiation, and promoter escape at three bacteriophage promoters". *Biochemistry* 42.13 (2003), pp. 3777–3786.

Huang, B., Bates, M., and Zhuang, X. "Super resolution fluorescence microscopy". *Annual Review of Biochemistry* 78 (2009), pp. 993–1016.

Huh, D. and Paulsson, J. "Non-genetic heterogeneity from stochastic partitioning at cell division." *Nature Genetics* 43.2 (2011), pp. 95–100.

Huo, Y. X., Tian, Z. X., Rappas, M., Wen, J., Chen, Y. C., You, C. H., Zhang, X., Buck, M., Wang, Y. P., and Kolb, A. "Protein-induced DNA bending clarifies the architectural organization of the sigma-54-dependent glnAp2 promoter". *Molecular Microbiology* 59.1 (2006), pp. 168–180.

Jacob, F., Perrin, D., Sánchez, C., and Monod, J. "The Operon: A Group of Genes Whose Expression is Coordinated by an Operator". *C R Hebd Seances Acad Sci.* 250 (1960), pp. 1727–1729.

Johnson, C. M. and Schleif, R. F. "*In vivo* induction kinetics of the arabinose promoters in *Escherichia coli*." *Journal Of Bacteriology* 177.12 (1995), pp. 3438–3442.

Jones, D. L., Brewster, R. C., and Phillips, R. "Promoter architecture dictates cell-to-cell variability in gene expression". *Science* 346.6216 (2014), pp. 1533–1537.

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. "Stochasticity in gene expression: from theories to phenotypes." *Nature Reviews Genetics* 6.6 (2005), pp. 451–64.

Kanamori, T., Hido, S., and Sugiyama, M. "A Least-squares Approach to Direct Importance Estimation". *Journal of Machine Learning Research* 10 (2009), pp. 1391 – 1445.

Kandhavelu, M, Häkkinen, A, Yli-Harja, O., and Ribeiro, A. S. "Single-molecule dynamics of transcription of the lar promoter". *Physical Biology* 9 (2012), p. 026004.

Kandhavelu, M., Mannerström, H., Gupta, A., Häkkinen, A., Lloyd-Price, J., Yli-Harja, O., and Ribeiro, A. S. "*In vivo* kinetics of transcription initiation of the lar promoter in *Escherichia coli*. Evidence for a sequential mechanism with two rate-limiting steps." *BMC Systems Biology* 5.1 (2011), p. 149.

Kandhavelu, M., Lloyd-Price, J., Gupta, A., Muthukrishnan, A.-B., Yli-Harja, O., and Ribeiro, A. S. "Regulation of mean and noise of the *in vivo* kinetics of transcription under the control of the lac/ara-1 promoter." *FEBS letters* 586.21 (2012), pp. 3870 – 3875.

Kapanidis, A. N., Margeat, E., Ho, S. O., Kortkhonjia, E., Weiss, S., and Ebright, R. H. "Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism." *Science* 314.5802 (2006), pp. 1144 – 1147.

Kapanidis, A. N., Margeat, E., Laurence, T. A., Doose, S., Ho, S. O., Mukhopadhyay, J., Kortkhonjia, E., Mekler, V., Ebright, R. H., and Weiss, S. "Retention of transcription initiation factor σ70 in transcription elongation: Single-molecule analysis". *Molecular Cell* 20.3 (2005), pp. 347 – 356.

Kawahara, Y and Sugiyama, M. "Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation". *Proc of 9th SIAM Int Conf on Data Mining* 1 (2009), pp. 385 – 396.

Kearns, D. B. and Losick, R. "Cell population heterogeneity during growth of Bacillus subtilis". *Genes Dev* 19.24 (2005), pp. 3083 – 3094.

Keiler, K. C. "Biology of trans-Translation". *Annual Review of Microbiology* 62.1 (2008), pp. 133 – 151.

Kerppola, T. K. "Visualization of molecular interactions by fluorescence complementation". *Nature Reviews Molecular Cell Biology* 7.6 (2006), pp. 449 – 456.

Keryer-Bibens, C., Barreau, C., and Osborne, H. B. "Tethering of proteins to RNAs by bacteriophage proteins." *Biology of the Cell* 100.2 (2008), pp. 125 – 138.

Khlebnikov, A, Datsenko, K. A., Skaug, T, Wanner, B. L., and Keasling, J. D. "Homogeneous expression of the P-BAD promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter." *Microbiology* 147.Pt 12 (2001), pp. 3241 – 3247.

Kimchi-Sarfaty, C, Oh, J., Kim, I., Sauna, Z., Calcagno, A., Ambudkar, S., and Gottesman, M. "A "silent" polymorphism in the MDR1 gene changes substrate specificity". *Science* 315.5811 (2007), pp. 525 – 528.

Kiviet, D. J., Nghe, P., Walker, N., Boulineau, S., Sunderlikova, V., and Tans, S. J. "Stochasticity of metabolism and growth at the single-cell level". *Nature* 514.7522 (2014), pp. 376 – 379.

Klumpp, S. and Hwa, T. "Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination." *Proceedings of the National Academy of Sciences of the United States of America* 105.47 (2008), pp. 18159–18164. arXiv: 0811.3163.

Klumpp, S., Zhang, Z., and Hwa, T. "Growth Rate-Dependent Global Effects on Gene Expression in Bacteria". *Cell* 139.7 (2009), pp. 1366–1375.

Konopka, C. A. and Bednarek, S. Y. "Variable-angle epifluorescence microscopy: A new way to look at protein dynamics in the plant cell cortex". *Plant Journal* 53.1 (2008), pp. 186–196.

Korbel, J. O., Jensen, L. J., Von Mering, C., and Bork, P. "Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs". *Nature Biotechnology* 22.7 (2004), pp. 911–917.

Krummel, B. and Chamberlin, M. J. "RNA Chain Initiation by *Escherichia coli* RNA polymerase. Structural Transitions of the Enzyme in Early Ternary Complexes". *Biochemistry* 28 (1989), pp. 7829–7842.

Kudla, G, Murray, A., Tollervey, D, and Plotkin, J. "Coding-Sequence Determinants of Gene Expression in *Escherichia coli*". *Science* 324 (2009), pp. 255–258.

Landick, R. "The regulatory roles and mechanism of transcriptional pausing." *Biochemical Society transactions* 34.Pt 6 (2006), pp. 1062–1066.

Landick, R. "Transcriptional pausing without backtracking." *Proceedings of the National Academy of Sciences of the United States of America* 106.22 (2009), pp. 8797–8798.

Lange, S., Katayama, Y., Schmid, M., Burkacky, O., Brauchle, C., Lamb D.C., D. C., and Jansen, R. P. "Simultaneous transport of different localized mRNA species revealed by live-cell imaging". *Traffic* 9.8 (2008), pp. 1256–1267.

Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A., and Singer, R. H. "Real-time observation of transcription initiation and elongation on an endogenous yeast gene." *Science* 332.6028 (2011), pp. 475–478.

Lawrence, J. G. and Roth, J. R. "Selfish operons: Horizontal transfer may drive the evolution of gene clusters". *Genetics* 143.4 (1996), pp. 1843–1860.

Lawrence, J. "Gene organization: Selection, selfishness, and serendipity". *Annual Review of Microbiology* 57 (2003), pp. 419–440.

Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. "Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function". *Cell* 131.1 (2007), pp. 174–187.

Lee, J. H., Al-Zarban, S., and Wilcox, G. "Genetic characterization of the araE gene in Salmonella typhimurium LT2". *Journal of Bacteriology* 146.1 (1981), pp. 298–304.

Lee, J. and Goldfarb, A. "Lac repressor acts by modifying the initial transcribing complex so that it cannot leave the promoter". *Cell* 66.4 (1991), pp. 793–798.

Lee, P. S. and Lee, K. H. "*Escherichia coli* - A Model System That Benefits from and Contributes to the Evolution of Proteomics". *Biotechnology and Bioengineering* 84.7 (2003), pp. 801–814.

Leibler, S. and Kussell, E. "Individual histories and selection in heterogeneous populations." *Proceedings of the National Academy of Sciences of the United States of America* 107.29 (2010), pp. 13183–13188.

Lelong, C., Aguiluz, K., Luche, S., Kuhn, L., Garin, J., Rabilloud, T., and Geiselmann, J. "The Crl-RpoS regulon of *Escherichia coli*." *Molecular and Cellular Proteomics* 6.4 (2007), pp. 648–659.

Lewis, K. "Persister cells, dormancy and infectious disease". *Nature Reviews Microbiology* 5.1 (2007), pp. 48–56.

Li, H. and Petzold, L. R. "Logarithmic Direct Method for Discrete Stochastic Simulation of Chemically Reacting Systems". *Technical Report* (2006), pp. 1–11.

Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J., and Taddei, F. "Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation." *Proceedings of the National Academy of Sciences of the United States of America* 105.8 (2008), pp. 3076–81.

Lloyd-Price, J., Gupta, A., and Ribeiro, A. S. "SGNS2: A Compartmentalized Stochastic Chemical Kinetics Simulator for Dynamic Cell Populations". *Bioinformatics* 28.22 (2012), pp. 3004–3005.

Lu, H. P., Xun, L, and Xie, X. S. "Single-molecule enzymatic dynamics." *Science* 282 (1998), pp. 1877–1882.

Lubeck, E. and Cai, L. "Single-cell systems biology by super-resolution imaging and combinatorial labeling". *Nature Methods* 9 (2012), pp. 743–748.

Lutz, R. and Bujard, H. "Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I 2 regulatory elements". *Nucleic Acids Research* 25.6 (1997), pp. 1203–1210.

Lutz, R., Lozinski, T., Ellinger, T., and Bujard, H. "Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator". *Nucleic Acids Research* 29.18 (2001), pp. 3873–3881.

McAdams, H. H. and Arkin, A. "Stochastic mechanisms in gene expression." *Proceedings of the National Academy of Sciences of the United States of America* 94.3 (1997), pp. 814–9.

McClure, W. R. "Mechanism and control of transcription initiation in prokaryotes." *Annual Review of Biochemistry* 54 (1985), pp. 171–204.

McClure, W. R. "Rate-limiting steps in RNA chain initiation." *Proceedings of the National Academy of Sciences of the United States of America* 77.10 (1980), pp. 5634–5638.

McClure, W. R., Cech, C. L., and Johnston, D. E. "A steady state assay for the RNA polymerase initiation reaction." *The Journal of Biological Chemistry* 253.24 (1978), pp. 8941–8.

Megerle, J. A., Fritz, G., Gerland, U., Jung, K., and Rädler, J. O. "Timing and dynamics of single cell gene expression in the arabinose utilization system". *Biophysical Journal* 95.4 (2008), pp. 2103–2115.

Miller, O. L., Hamkalo, B. A., and Thomas, C. A. "Visualization of bacterial genes in action." *Science* 169.943 (1970), pp. 392–395.

Mitarai, N., Sneppen, K., and Pedersen, S. "Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization". *Journal of Molecular Biology* 382.1 (2008), pp. 236–245.

Montero Llopis, P., Jackson, A. F., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T., and Jacobs-Wagner, C. "Spatial organization of the flow of genetic information in bacteria." *Nature* 466.7302 (2010), pp. 77–81.

Moore, S. D. and Sauer, R. T. "Ribosome rescue: tmRNA tagging activity and capacity in *Escherichia coli*". *Molecular Microbiology* 58.2 (2005), pp. 456–466.

Morgan-Kiss, R. M., Wadler, C., and Cronan, J. E. "Long-term and homogeneous regulation of the *Escherichia coli* araBAD promoter by use of a lactose transporter of relaxed specificity". *Proceedings of the National Academy of Sciences of the United States of America* 99.11 (2002), pp. 7373–7377.

Munsky, B. E. "The Finite State Projection Approach for the Solution of the Master Equation and its Applications to Stochastic Gene Regulatory Networks". PhD thesis. University of California, 2008.

Munsky, B. and Khammash, M. "The finite state projection algorithm for the solution of the chemical master equation." *Journal of Chemical Physics* 124.4 (2006), p. 044104.

Munsky, B. and Khammash, M. "The finite state projection approach for the analysis of stochastic noise in gene networks". *IEEE Transactions on Automatic Control* 53 (2008), pp. 201–214.

Munsky, B., Fox, Z., and Neuert, G. "Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics". *Methods* 85 (2015), pp. 12–21.

Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., and Darst, S. A. "Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex." *Science* 296.5571 (2002), pp. 1285–1290.

Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., Kudasov, F., Chowdhury, S., Yli-Harja, O., and Ribeiro, A. S. "Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells." *Nucleic Acids Research* 40.17 (2012), pp. 8472–8483.

Nachman, I., Regev, A., and Ramanathan, S. "Dissecting Timing Variability in Yeast Meiosis". *Cell* 131.3 (2007), pp. 544–556.

Nakano, A. "Spinning-disk confocal microscopy – a cutting-edge tool for imaging of membrane traffic." *Cell Structure and Function* 27.5 (2002), pp. 349–355.

Nath, S. and Guha, A. "Abortive termination of bioBFCD RNA synthesized in vitro from the bioABFCD operon of *Escherichia coli* K-12". *Proceedings of the National Academy of Sciences of the United States of America* 79.6 (1982), pp. 1786–1790.

Neidhardt, F, Ingraham, J, and Schaechter, M. *Physiology of the Bacterial Cell: A Molecular Approach*. Sunauer Associates, Sunderland, MA., 1990.

Neubauer, Z. and Calef, E. "Immunity phase-shift in defective lysogens: Non-mutational hereditary change of early regulation of prophage". *Journal of Molecular Biology* 51.1 (1970), pp. 1–13.

Neuert, G., Munsky, B., Tan, R. Z., Teytelman, L., Khammash, M., and van Oudenaarden, A. "Systematic Identification of Signal-Activated Stochastic Gene Regulation". *Science* 339.6119 (2013), pp. 584–587.

Norman, T. M., Lord, N. D., Paulsson, J., and Losick, R. "Memory and modularity in cell-fate decision making." *Nature* 503.7477 (2013), pp. 481–6.

Norman, T. M., Lord, N. D., Paulsson, J., and Losick, R. "Stochastic Switching of Cell Fate in Microbes." *Annual Review of Microbiology* 69 (2015), pp. 381–403.

Novick, A. and Weiner, M. "Enzyme induction as and all-or-nothing phenomenon". *Proc Natl Acad Sci U S A* 43.7 (1957), pp. 553–566.

Oliveira, S. M. D., Neeli-Venkata, R., Goncalves, N. S. M., Santinha, J. A., Martins, L., Tran, H., Mäkelä, J., Gupta, A., Barandas, M., Häkkinen, A., Lloyd-Price, J., Fonseca, J. M., and Ribeiro, A. S. "Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli*". *Molecular Microbiology* 99.4 (2016), pp. 686–699.

Osbourn, A. E. and Field, B. "Operons". *Cellular and Molecular Life Sciences* 66.23 (2009), pp. 3755–3775.

Otsu, N. "A threshold selection method from gray-level histograms". *IEEE Trans. Sys. Man. Cyber.* 9.1 (1979), pp. 62–66.

Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and van Oudenaarden, A. "Multistability in the lactose utilization network of *Escherichia coli*." *Nature* 427.6976 (2004), pp. 737–740.

Patrick, M., Dennis, P. P., Ehrenberg, M., and Bremer, H. "Free RNA polymerase in *E. coli*". *Biochimie* 119 (2015), pp. 80–91.

Paulsson, J and Ehrenberg, M. "Noise in a minimal regulatory network: plasmid copy number control." *Quarterly reviews of biophysics* 34.1 (2001), pp. 1–59.

Paulsson, J. "Models of stochastic gene expression". *Physics of Life Reviews* 2.2 (2005), pp. 157–175.

Paulsson, J. "Summing up the noise in gene networks." *Nature* 427.6973 (2004), pp. 415–8.

Pawley, J. E. *Handbook of Biological Confocal Microscopy (3rd ed.)* Berlin: Springer, 2006.

Pedersen, S. "*Escherichia coli* ribosomes translate *in vivo* with variable rate." *The EMBO journal* 3.12 (1984), pp. 2895–2898.

Pedraza, J. M. and Paulsson, J. "Effects of molecular memory and bursting on fluctuations in gene expression." *Science* 319.5861 (2008), pp. 339–343.

Peterson, J. R., Cole, J. A., Fei, J., Ha, T., and Luthey-Schulten, Z. A. "Effects of DNA replication on mRNA noise". *Proceedings of the National Academy of Sciences of the United States of America* 112.52 (2015), pp. 15886–15891.

Pitchiaya, S., Heinicke, L. a., Custer, T. C., and Walter, N. G. "Single molecule fluorescence approaches shed light on intracellular RNAs". *Chemical Reviews* 114.6 (2014), pp. 3224–3265.

Potapov, I., Mäkelä, J., Yli-Harja, O., and Ribeiro, A. S. "Effects of codon sequence on the dynamics of genetic networks." *Journal of Theoretical Biology* 315 (2012), pp. 17–25.

Prasher, D. C., Eckenrode, V. K., Ward, W. W., Prendergast, F. G., and Cormier, M. J. "Primary structure of the Aequorea victoria green-fluorescent protein". *Gene* 111.2 (1992), pp. 229–233.

Pratt, L. and Silhavy, T. "Crl stimulates RpoS activity during stationary phase". *Molecular Microbiology* 29.5 (1998), pp. 1225–1236.

Prescott, E. M. and Proudfoot, N. J. "Transcriptional collision between convergent genes in budding yeast." *Proceedings of the National Academy of Sciences of the United States of America* 99.13 (2002), pp. 8796–801.

Proshkin, S., Rahmouni, A. R., Mironov, A., and Nudler, E. "Cooperation between translating ribosomes and RNA polymerase in transcription elongation." *Science* 328.5977 (2010), pp. 504–8.

Raffaelle, M., Kanin, E. I., Vogt, J., Burgess, R. R., and Ansari, A. Z. "Holoenzyme switching and stochastic release of sigma factors from RNA polymerase *in vivo*". *Molecular Cell* 20.3 (2005), pp. 357–366.

Raj, A., Bogaard, P. van den, Rifkin, S. A., van Oudenaarden, A., and Tyagi, S. "Imaging individual mRNA molecules using multiple singly labeled probes". *Nature Methods* 5.1 (2008), pp. 877–879.

Raj, A. and van Oudenaarden, A. "Single-molecule approaches to stochastic gene expression." *Annual Review of Biophysics* 38 (2009), pp. 255–270.

Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. "Variability in gene expression underlies incomplete penetrance." *Nature* 463.7283 (2010), pp. 913–918.

Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O., and Ribeiro, A. S. "Effects of transcriptional pausing on gene expression dynamics". *PLoS Computational Biology* 6.3 (2010), pp. 29–30.

Ramakrishnan, V. "Ribosome structure and the mechanism of translation". *Cell* 108.4 (2002), pp. 557–572.

Rao, C. V., Wolf, D. M., and Arkin, A. P. "Control, exploitation and tolerance of intracellular noise." *Nature* 420.6912 (2002), pp. 231–237.

Revyakin, A, Liu, C, Ebright, R. H., and Strick, T. R. "Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching". *Science* 314.5802 (2006), pp. 1139–1143.

Ribeiro, A. S. and Kauffman, S. A. "Noisy attractors and ergodic sets in models of gene regulatory networks". *Journal of Theoretical Biology* 247 (2007), pp. 743–755.

Ribeiro, A. S., Zhu, R., and Kauffman, S. A. "A General Modeling Strategy for Gene Regulatory Networks with Stochastic Dynamics". *Journal of Computational Biology* 13.9 (2006), pp. 1630–1639.

Ribeiro, A. S., Smolander, O.-P., Rajala, T., Häkkinen, A., and Yli-Harja, O. "Delayed stochastic model of transcription at the single nucleotide level". *Journal of Computational Biology* 16.4 (2009), pp. 539–553.

Ribeiro, A. S., Häkkinen, A., Mannerström, H., Lloyd-Price, J., and Yli-Harja, O. "Effects of the promoter open complex formation on gene expression dynamics". *Physical Review E* 81.1 (2010), p. 011912.

Ribeiro, A. S., Lloyd-Price, J., Kesseli, J., Häkkinen, A., and Yli-harja, O. "Quantifying local structure effects in network dynamics". *Physical Review E* 78 (2008), p. 056108.

Richardson, J. P. "Preventing the synthesis of unused transcripts by rho factor". *Cell* 64.6 (1991), pp. 1047–1049.

Riggs, A., Bourgeois, S, and Cohn, M. "The lac repressor-operator interaction. 3. Kinetic studies". *Journal of Molecular Biology* 53.3 (1970), pp. 401–417.

Robert, L., Paul, G., Chen, Y., Taddei, F., Baigl, D., and Lindner, A. B. "Predispositions and epigenetic inheritance in the *Escherichia coli* lactose operon bistable switch." *Molecular Systems Biology* 6.357 (2010), p. 357.

Robinson, A. and Oijen, A. M. van. "Bacterial replication, transcription and translation: mechanistic insights from single-molecule biochemical studies". *Nature Reviews Microbiology* 11 (2013), pp. 303–315.

Rotman, B. "Measurement of activity of single molecules of beta-D-galactosidase." *Proceedings of the National Academy of Sciences of the United States of America* 47 (1961), pp. 1981–1991.

Roussel, M. R. and Zhu, R. "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression." *Physical Biology* 3.4 (2006), pp. 274–284.

Ruusuvuori, P., Äijö, T., Chowdhury, S., Garmendia-Torres, C., Selinummi, J., Birbaumer, M., Dudley, A. M., Pelkmans, L., and Yli-Harja, O. "Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images." *BMC Bioinformatics* 11 (2010), p. 248.

Saecker, R. M., Record, M. T., and DeHaseth, P. L. "Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis". *Journal of Molecular Biology* 412.5 (2011), pp. 754–771.

Sanchez, A. and Golding, I. "Genetic determinants and cellular constraints in noisy gene expression." *Science* 342.6163 (2013), pp. 1188–93.

Sanchez, A., Osborne, M. L., Friedman, L. J., Kondev, J., and Gelles, J. "Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules". *The EMBO Journal* 30.19 (2011), pp. 3940–3946.

Santangelo, P., Lifland, A., Curt, P, Sasaki, Y, Bassell, G., Lindquist, M., and Crowe, J. J. "Single molecule-sensitive probes for imaging RNA in live cells". *Nature Methods* 6.5 (2009), pp. 347–349.

Schlax, P. J., Capp, M. W., and Record, M. T. "Inhibition of transcription initiation by lac repressor." *Journal of Molecular Biology* 245.4 (1995), pp. 331–350.

Schleif, R. "AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action." *FEMS Microbiology Reviews* 34.5 (2010), pp. 779–796.

Schleif, R. "Regulation of the L-arabinose operon of *Escherichia coli*". *Trends in Genetics* 16.12 (2000), pp. 559–565.

Schnappinger, D and Hillen, W. "Tetracyclines: antibiotic action, uptake, and resistance mechanisms." *Archives of microbiology* 165.6 (1996), pp. 359–369.

Schölkopf, B. and Smola, A. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Sclavi, B., Zaychikov, E., Rogozina, A., Walther, F., Buckle, M., and Heumann, H. "Real-time characterization of intermediates in the pathway to open complex formation by *Escherichia coli* RNA polymerase at the T7A1 promoter." *Proceedings of the National Academy of Sciences of the United States of America* 102.13 (2005), pp. 4706–4711.

Senecal, A., Munsky, B., Proux, F., Ly, N., Braye, F. E., Zimmer, C., Mueller, F., and Darzacq, X. "Transcription factors modulate c-Fos transcriptional bursts". *Cell Reports* 8.1 (2014), pp. 75–83.

Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and Tsien, R. Y. "Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein." *Nature Biotechnology* 22.12 (2004), pp. 1567–1572.

Sharp, P. M. and Li, W.-h. "The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications". *Nucleic Acids Research* 15.3 (1987), pp. 1281–1295.

Shearwin, K. E., Callen, B. P., and Egan, J. B. "Transcriptional interference - A crash course". *Trends in Genetics* 21.6 (2005), pp. 339–345.

Sheppard, D and Englesberg, E. "Further evidence for positive control of the L-arabinose system by gene araC". *Journal of Molecular Biology* 25 (1967), pp. 443–454.

Shimomura, O, Johnson, F. H., and Saiga, Y. J. "Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea." *J Cell Comp Physiol* 59 (1962), pp. 223–239.

Siegele, D. A. and Hu, J. C. "Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations". *Proceedings of the National Academy of Sciences of the United States of America* 94.15 (1997), pp. 8168–8172.

Simao, E., Remy, E., Thieffry, D., and Chaouiya, C. "Qualitative modelling of regulated metabolic pathways: Application to the tryptophan biosynthesis in E. Coli". *Bioinformatics* 21.SUPPL. 2 (2005), pp. 190–196.

Sneppen, K., Dodd, I. B., Shearwin, K. E., Palmer, A. C., Schubert, R. A., Callen, B. P., and Egan, J. B. "A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*". *Journal of Molecular Biology* 346.2 (2005), pp. 399–409.

So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R., and Golding, I. "General properties of transcriptional time series in *Escherichia coli*." *Nature genetics* 43.6 (2011), pp. 554–560.

Sørensen, M. A. and Pedersen, S. "Absolute *in vivo* translation rates of individual codons in *Escherichia coli*: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate". *Journal of Molecular Biology* 222.2 (1991), pp. 265–280.

Sørensen, M. A., Kurland, C. G., and Pedersen, S. "Codon usage determines translation rate in *Escherichia coli*". *Journal of Molecular Biology* 207.2 (1989), pp. 365–377.

Spicher, A., Michel, O., Cieslak, M., Giavitto, J. L., and Prusinkiewicz, P. "Stochastic P systems and the simulation of biochemical processes with dynamic compartments". *BioSystems* 91.3 (2008), pp. 458–472.

St-Pierre, F. and Endy, D. "Determination of cell fate selection during phage lambda infection." *Proceedings of the National Academy of Sciences of the United States of America* 105.52 (2008), pp. 20705–20710.

Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M., and Elowitz, M. B. "An excitable gene regulatory circuit induces transient cellular differentiation." *Nature* 440.7083 (2006), pp. 545–550.

Süel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M. B. "Tunability and noise dependence in differentiation dynamics." *Science* 315.5819 (2007), pp. 1716–1719.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bunau, P., and Kawanabe, M. "Direct importance estimation for covariate shift adaptation". *Annals of the Institute of Statistical Mathematics* 60.4 (2008), pp. 699–746.

Takahashi, K. and Yamanaka, S. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors". *Cell* 126.4 (2006), pp. 663–676.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. "Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells". *Science* 329.5991 (2010), pp. 533–538.

Tokunaga, M., Imamoto, N., and Sakata-Sogawa, K. "Highly inclined thin illumination enables clear single-molecule imaging in cells". *Nature Methods* 5.5 (2008), p. 455.

Tran, H., Oliveira, S. M. D., Goncalves, N., and Ribeiro, A. S. "Kinetics of the cellular intake of a gene expression inducer at high concentrations". *Molecular Biosystems* 11.9 (2015), pp. 2579–2587.

Tsien, R. Y. "The green fluorescent protein". *Annual Review of Biochemistry* 67 (1998), pp. 509–544.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. "An evolutionarily conserved mechanism for controlling the efficiency of protein translation". *Cell* 141.2 (2010), pp. 344–354.

Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D. J., and Kapanidis, A. N. "Single-molecule DNA repair in live bacteria". *Proceedings of the National Academy of Sciences of the United States of America* 110.20 (2013), pp. 8063–8068.

Uphoff, S., Lord, N. D., Okumus, B., Potvin-trottier, L., Sherratt, D. J., and Paulsson, J. "Mutation Rate Variation". *Science* 351.6277 (2016), pp. 1094–1098.

Valegard, K., Murray, J., Stockley, P., Stonehouse, N., and Liljas, L. "Crystal structure of an RNA bacteriophage coat protein-operator complex". *Nature* 371 (1994), pp. 623–626.

Wade, J. T. and Grainger, D. C. "Pervasive transcription: illuminating the dark matter of bacterial transcriptomes". *Nature Reviews Microbiology* 12 (2014), pp. 647–653.

Walter, G, Zillig, W, Palm, P, and Fuchs, E. "Initiation of DNA-Dependent RNA synthesis and the effect of heparin on RNA polymerase". *Eur. J. Biochem.* 3 (1967), pp. 194–201.

Walter, N. G., Huang, C. Y., Manzo, A. J., and Sobhy, M. A. "Do-it-yourself guide: how to use the modern single-molecule toolkit". *Nature Methods* 5 (2008), pp. 475–489.

Wang, F, Redding, S, Finkelstein, I. J., Gorman, J, Reichman, D. R., and Greene, E. C. "The promoter search mechanism of *E. coli* RNA polymerase is dominated by three–dimensional diffusion". *Nature Structural and Molecular Biology* 20.2 (2013), pp. 174–181.

Wang, G. Z., Lercher, M. J., and Hurst, L. D. "Transcriptional coupling of neighboring genes and gene expression noise: Evidence that gene orientation and noncoding transcripts are modulators of noise". *Genome Biology and Evolution* 3.1 (2011), pp. 320–331.

Ward, D. F. and Murray, N. E. "Convergent transcription in bacteriophage lambda: interference with gene expression". *Journal of Molecular Biology* 133 (1979), pp. 249–266.

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S. H., Noller, H. F., Bustamante, C., and Tinoco, I. "Following translation by single ribosomes one codon at a time." *Nature* 452.7187 (2008), pp. 598–603.

Wilson, D. N. "Ribosome-targeting antibiotics and mechanisms of bacterial resistance." *Nature Reviews Microbiology* 12.1 (2014), pp. 35–48.

Wu, B., Chen, J., and Singer, R. H. "Background free imaging of single mRNAs in live cells using split fluorescent proteins." *Scientific Reports* 4 (2014), p. 3615.

Wu, B., Chao, J. A., and Singer, R. H. "Fluorescence fluctuation spectroscopy enables quantitative imaging of single mRNAs in living cells". *Biophysical Journal* 102.12 (2012), pp. 2936–2944.

Wu, B., Piatkevich, K. D., Lionnet, T., Singer, R. H., and Verkhusha, V. V. "Modern fluorescent proteins and imaging technologies to study gene expression, nuclear localization, and dynamics". *Current Opinion in Cell Biology* 23.3 (2011), pp. 310–317.

Xie, X. S., Choi, P. J., Li, G.-W., Lee, N. K., and Lia, G. "Single-molecule approach to molecular biology in living bacterial cells." *Annual Review of Biophysics* 37 (2008), pp. 417–444.

Yang, S., Kim, S., Lim, Y. R., Kim, C., An, H. J., Kim, J.-h., Sung, J., and Lee, N. K. "Contribution of RNA polymerase concentration variation to protein expression noise". *Nature Communications* 5 (2014), pp. 1–9.

Yanofsky, C. "Attenuation in the control of expression of bacterial operons". *Nature* 289 (1981), pp. 751–758.

Yanofsky, C. "The different roles of tryptophan transfer RNA in regulating trp operon expression in *E. coli* versus B. subtilis". *Trends in Genetics* 20.8 (2004), pp. 367–374.

Yarchuk, O., Guillerez, J., and Dreyfus, M. "Interdependence of Translation, Transcription and mRNA Degradation in the lacZ Gene". *Journal of Molecular Biology* 226 (1992), pp. 581–596.

Young, B. A., Gruber, T. M., Gross, C. A., and Francisco, S. "Views of Transcription Initiation". *Cell* 109 (2002), pp. 417–420.

Young, J. W., Locke, J. C. W., and Elowitz, M. B. "Rate of environmental change determines stress response specificity." *Proceedings of the National Academy of Sciences of the United States of America* 110.10 (2013), pp. 4140–5.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X. S. "Probing gene expression in live cells, one protein molecule at a time." *Science* 311.5767 (2006), pp. 1600–1603.

Zafar, M. A., Carabetta, V. J., Mandel, M. J., and Silhavy, T. J. "Transcriptional occlusion caused by overlapping promoters." *Proceedings of the National Academy of Sciences* 111.4 (2014), pp. 1557–1561.

Zeng, L., Skinner, S. O., Zong, C., Sippy, J., Feiss, M., and Golding, I. "Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection". *Cell* 141.4 (2010), pp. 682–691.

Zhu, R., Ribeiro, A. S., Salahub, D., and Kauffman, S. A. "Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models". *Journal of Theoretical Biology* 246.4 (2007), pp. 725–745.

Publications

Publication I

J. Mäkelä, H. Huttunen, M. Kandhavelu, O. Yli-Harja, and A.S. Ribeiro, "Automatic detection of changes in the dynamics of delayed stochastic gene networks and *in vivo* production of RNA molecules in *Escherichia coli*", *Bioinformatics*, 27(19):2714-2720, 2011.

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 19 2011, pages 2714–2720
doi:10.1093/bioinformatics/btr471

Systems biology

Advance Access publication August 10, 2011

Automatic detection of changes in the dynamics of delayed stochastic gene networks and *in vivo* production of RNA molecules in *Escherichia coli*

Jarno Mäkelä[1], Heikki Huttunen[1], Meenakshisundaram Kandhavelu[1], Olli Yli-Harja[1,2] and Andre S. Ribeiro[1,*]

[1]Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland and [2]Institute for Systems Biology, Seattle, WA 98103-8904, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Production and degradation of RNA and proteins are stochastic processes, difficulting the distinction between spurious fluctuations in their numbers and changes in the dynamics of a genetic circuit. An accurate method of change detection is key to analyze plasticity and robustness of stochastic genetic circuits.

Results: We use automatic change point detection methods to detect non-spurious changes in the dynamics of delayed stochastic models of gene networks at run time. We test the methods in detecting changes in mean and noise of protein numbers, and in the switching frequency of a genetic switch. We also detect changes, following genes' silencing, in the dynamics of a model of the core gene regulatory network of *Saccharomyces cerevisiae* with 328 genes. Finally, from images, we determine when RNA molecules tagged with fluorescent proteins are first produced in *Escherichia coli*. Provided prior knowledge on the time scale of the changes, the methods detect them accurately and are robust to fluctuations in protein and RNA levels.

Availability: Simulator: www.cs.tut.fi/∼sanchesr/SGN/SGNSim.html

Contact: andre.ribeiro@tut.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 27, 2011; revised on August 3, 2011; accepted on August 5, 2011

1 INTRODUCTION

Gene regulatory networks (GRNs) are stochastic. However, their behavior is, to some extent, robust, e.g. when responding to environmental changes. The behavior is determined by the structure of the genetic circuits. Thus, when structural changes occur, in many cases there are changes in the dynamics of RNA and protein numbers of some genes. Some such structural changes (e.g. a mutation) can be rare, occurring once in a cell's lifetime. It is thus important to develop robust methods for detecting permanent changes in the dynamics of genetic circuits, and distinguish these from spurious fluctuations in RNA and protein numbers.

We apply automatic change detection methods to simulated and real gene expression data, to recognize candidate change points in RNA and protein numbers' dynamics. Automatic detection of change points is the discovering of points in time where the properties of the time series change. Earliest approaches were based on the Behrens–Fisher problem, a statistical hypothesis test of equal means (Belloni and Didier, 2008; Fisher, 1939). A widely used approximation to solve this problem is the Welch's *t*-test. However, these approaches assume normal distributions, typically causing them to be too sensitive for heavy-tailed distributions. The dynamics of RNA and protein production are usually not normal like, especially if a structural change occurs in the GRN during the observations.

We use two recent change point detection methods, namely, the density ratio method and the kernel change point analysis (Harchaoui *et al.*, 2009; Kawahara and Sugiyama, 2009). Our choice is based on their reported good performance compared with alternative methods. These methods use different approaches: the density ratio method has its roots in statistics and density estimation, while the kernel change point analysis is based on the theory of kernel machines and classification. In our understanding, they represent state-of-the-art approaches to the problem.

To assess the accuracy of the methods, knowing the ground truth signal is needed. For this, we require realistic simulations of RNA and protein expression dynamics. The dynamics of the models ought to be as realistic as possible so as to mimic accurately the temporal dynamics of RNA and protein numbers in real cells.

Recently, a delayed stochastic modeling strategy of gene expression and GRNs was proposed (Ribeiro *et al.*, 2006). It is based on the delayed stochastic simulation algorithm (delayed SSA) (Zhu *et al.*, 2007), and thus it accounts for the key dynamical features of real GRNs, namely, the stochasticity of the chemical kinetics (Arkin *et al.*, 1998), and the duration of events such as the promoter complex formation (McClure, 1980) and transcription elongation (Zhu *et al.*, 2007). This modeling strategy was shown to match the dynamics of RNA and protein production at the single molecule level (Yu *et al.*, 2006; Zhu *et al.*, 2007). Delayed stochastic models of GRNs can be simulated by SGNSim (Ribeiro and Lloyd-Price, 2007), which also allows introducing changes in the structure of the GRN at run time, needed to test the change point detection methods.

We apply and test the accuracy of the automatic detection of change points methods to model GRNs subject to a permanent

*To whom correspondence should be addressed.

change at run time in mean level of a protein, noise strength of a protein's time series and in the frequency of switching of a genetic switch. Further, to verify the applicability of the methods to large-scale clusters of interconnected genes, we test the ability to detect a change in the dynamics of a model GRN with 328 genes, subject to the silencing of a randomly selected gene at run time. Finally, we apply the methods to determine when new RNA molecules are produced, from our temporal measurements by confocal microscopy of RNA tagged with MS2d-GFP in *Escherichia coli*.

2 METHODS

2.1 *In vivo* detection of RNA molecules in *E.coli*

RNA detection and quantification *in vivo* in *E. coli* cells DH5α-PRO uses the ability of the coat protein of bacteriophage MS2 to tightly bind specific RNA sequences (Peabody and Lim, 1996). High-resolution detection of single RNA transcripts with 96 tandem repeats of MS2 binding sites in *E.coli* is possible by using dimeric MS2 fused to GFP (MS2d-GFP fusion protein) as a detection tag (Golding *et al.*, 2005). The method uses two genetic constructs. The first is a medium-copy vector expressing the MS2d-GFP fusion protein, whose promoter (P_{tetO}) is regulated by tetracycline repressor. The second is a single copy F-based vector, with a $P_{lac/ara}$ promoter controlling production of the transcript target, specifically mRFP1 followed by a 96 MS2 binding site array. Constructs were generously provided by I. Golding (University of Illinois). Experimental procedures of induction of the target RNA, confocal microscopy and cell and RNA spots segmentation from images are described in Supplementary Material.

2.2 Models of GRNs

We follow the modeling strategy of delayed stochastic GRNs proposed in Ribeiro *et al.* (2006). The models are implemented in the simulator SGNSim (Ribeiro and Lloyd-Price, 2007), and the dynamics is based on the delayed SSA (Zhu *et al.*, 2007), that unlike the SSA (Gillespie, 1977), uses a waiting list to store delayed output events. The algorithm of delayed SSA is presented in Supplementary Material. Delayed reactions are represented as: $A \rightarrow B + C(\tau_1) + D(\tau_2)$. In this reaction, B is instantaneously produced, while C and D are placed on the waitlist until they are released, after τ_1 and τ_2 seconds, respectively. This strategy accounts for the stochastic nature of chemical reactions and for the fact that transcription and translation are multistep processes that take non-negligible time to be completed once initiated. The strategy was validated in Zhu *et al.* (2007) by matching temporal measurements of expression of individual proteins (Yu *et al.*, 2006).

We implement four model GRNs, named models 1, 2, 3 and 4. Models 1 and 2 are identical, and consist of a two-gene network, where Gene 1 represses Gene 2. These models differ in the change at run time. In Model 1, mean protein levels change at run time, while in Model 2 it is the strength of fluctuations in protein levels that changes. Model 3 is a genetic switch whose switching frequency is changed at run time.

We also test if changes in the dynamics of larger GRNs can be detected. Gene networks consist of hundreds to thousands of genes, usually organized in clusters of dozens to hundreds, that are involved in specific tasks in development, metabolism, etc. Changes known to occur in the dynamics of these networks may be caused by mutations, deletions or duplications, or as a response to external signals or stress. Usually, such events cause one to a few genes, along with several of its neighbor genes, to alter the expression level (e.g. from high to low). The models and how the changes at run time are implemented are described in Supplementary Material.

To test if the algorithms of change point detection are successful for large genetic circuits, we apply them to a model of the core gene network of *Saccharomyces cerevisiae* inferred from microarray measurements following gene deletions and overexpressions (Chowdhury *et al.*, 2010). This network contains 328 genes. Inferred connections were verified by gene enrichment. The perturbations consist of selecting genes randomly (see Supplementary Material) and subject them to silencing at run time, one per simulation.

2.3 Methods of change point detection

Formally, the problem of change point detection can be stated as follows. Given a multidimensional time series $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathrm{R}^n$, which time points K represent change in some sense, given the data samples in the M-point backward window $\boldsymbol{X}_B = (\boldsymbol{x}_{K-M}, \ldots, \boldsymbol{x}_{K-1})$ and the M-point forward window $\boldsymbol{X}_F = (\boldsymbol{x}_{K+1}, \ldots, \boldsymbol{x}_{K+M})$. To define the dissimilarity of the two windows one can pose the question as a hypothesis testing problem:

$$\begin{cases} H_0 : p_{\boldsymbol{X}_F}(\boldsymbol{x}) = p_{\boldsymbol{X}_B}(\boldsymbol{x}) \\ H_1 : p_{\boldsymbol{X}_F}(\boldsymbol{x}) \neq p_{\boldsymbol{X}_B}(\boldsymbol{x}) \end{cases} \quad (1)$$

where $p_{\boldsymbol{X}_F}(\boldsymbol{x})$ and $p_{\boldsymbol{X}_B}(\boldsymbol{x})$ denote the probability density functions of the forward and backward windows, respectively.

Detection in non-parametric cases is still, in general, an open problem. We apply two recent change point detection methods proposed for the non-parametric case. Namely, we apply a direct density ratio test (uLSIF) (Kawahara and Sugiyama, 2009) and a kernel change point analysis method (KCpA) (Harchaoui *et al.*, 2009), described in Supplementary Material.

3 RESULTS AND DISCUSSION

3.1 Selecting a proper window size

Most parameters of change point detection algorithms can be inferred from the data by cross-validation. However, the detection window length cannot be determined from training data, since changes appear at multiple scales. Thus, the algorithm cannot decide which time scale is biologically relevant. Below, we choose the window size from knowledge of the scale of the biological phenomena studied. Before, we study the performance of the detectors for a wide range of window sizes to search for differences in robustness and to compare performances with larger number of test cases.

We first apply KCpA and uLSIF algorithms to three models. We change at run time mean protein levels in Model 1, noise in protein numbers in Model 2 and frequency of switching between noisy attractors in Model 3 (Section 2.2). Examples of the time series of protein numbers in these models, prior and after a change are shown in Supplementary Material. The performance is assessed by the area under the receiver operating characteristics (ROC) curve or the AUC criterion (Kay, 1998). The ROC curve and the corresponding AUC value were based on the specificity–sensitivity coordinates obtained by varying the detection threshold.

To improve the detection performance, the data are appended by auxiliary variables. This attempts to cast the change in, e.g. variance into a change in the mean of the auxiliary variable. For example, in Model 2, the change is in the degree of the fluctuations in protein numbers. One can convert this into a change in 'local variance' (that is, variance within a small time interval). Appending the local variance estimates for both proteins to mean levels significantly improves the detection of changes in noise in protein numbers.

We also include auxiliary variables in Model 3. The added feature is the average absolute difference of consecutive samples. This improves performance because of the nature of the change (in switching frequency) and because the levels of the two protein are dynamically coupled. The change in switching frequency from low to high is reflected in the difference between consecutive samples of $|P_1 - P_2|$ (similarly one could have used the sum of P_1 and P_2).

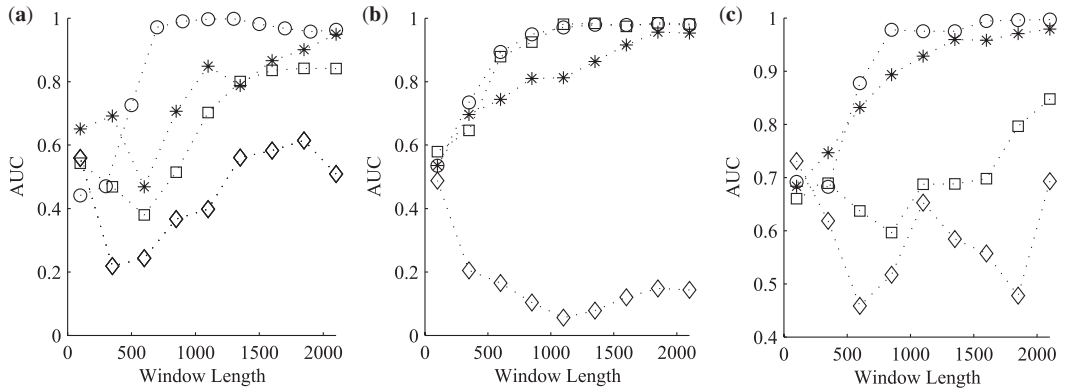


Fig. 1. Effect of window length on change point detection methods. KCpA (linear kernel) (open circle), KCpA (polynomial kernel) (open square), KCpA (RBF kernel) (asterisk) and uLSIF (diamond) detection results for each model. Horizontal axis is the size of forward and backward windows (window length size 1000 uses 1000 + 1000 = 2000 samples for detection). Vertical axis is the area under the ROC curves (AUC) for different window lengths when applied to Models 1 (**a**), 2 (**b**) and 3 (**c**).

The performance of detection of change points for various window sizes is summarized in Figure 1. For all models, KCpA seems more effective (higher AUC values for most window lengths). Also, uLSIF is more sensitive to the choice of parameters, but their manual adjustment can improve the performance to comparable levels. However, the cross-validated automatic parameter setup does not provide good results. We acknowledge that the *ad hoc* addition of auxiliary variables may favor KCpA over uLSIF. Without it, the two detectors exhibited similar (poor) performances, although for Model 2, uLSIF performed slightly better. This is likely because uLSIF models more extensively the characteristics of the distribution, while KCpA assumes Gaussian densities with equal covariances. However, for our practical purposes, neither method was robust enough, so we decided to improve the detection by adding auxiliary variables.

In general, widening the window improves performance. The best robustness of detection is obtained by KCpA with the linear kernel. This is not surprising, since simple linear models typically exhibit small variance (e.g. repeated experiments tend to have similar results). The drawback is a high bias if the model is not complex enough. However, according to Figure 1, the linear model with KCpA seems sufficient for our data.

3.2 Detection of change points in genetic circuits

Examples of the application of KCpA with linear kernel to one realization of the time series of each model are shown in Figure 2a–c, and the results for uLSIF in Figure 2d–f. The inputs for change point detection are the time series in Supplementary Figure S3a–S3c. In all cases, the true change point occurs at midpoint of the time series. KCpA outperforms uLSIF, as it is less sensitive to spurious, transient changes. We used in all cases a window length of 1100 samples as it enhances detection when compared with smaller lengths, and further increases in length did not improve the detection significantly (Fig. 1). This length corresponds to 10^5 s simulation time, allowing spurious transient fluctuations to be recognized as such. The length is realistic given the time scales of transient fluctuations in protein numbers in bacteria and eukaryotic cells. In the models, the effects of fluctuations in protein numbers last for long periods of time ($10^3 - 10^4$ s). Therefore, it is expected the need of using a window size longer by one order of magnitude. In cells, since proteins have lifetimes of the order of tenths of minutes, fluctuations last by a similar order of magnitude. For example, oscillations in protein P53 numbers in Human cells have a period of 5 h (Geva-Zatorsky *et al.*, 2006) and the period of oscillation of a repressilator engineered in *E.coli* is $\sim 10^4$ s (Elowitz and Leibler, 2000).

For the KCpA method (Fig. 2a–c), we performed multiple tests on independent time series, all of which with identical initial conditions. In the figures, we show the results of three of such independent runs for the three models. The aim is to test if KCpA is robust to the stochasticity of the time series, which will cause different spurious, transient fluctuations in each independent simulation. Visibly, the algorithm is highly robust in the first two models (Fig. 2a and b), in the sense that the change is accurately detected in all independent simulations and at the moment following occurrence of the structural change.

The results for Model 3 are more complex. Namely, while in all cases the change is detected (as depicted in Fig. 2c), the detection takes place at different points in time following the change. This is explained by the nature of the change. What changes is the frequency of switching between noisy attractors. For such a change to be detected (even by a human observer), switches between the two protein levels must take place (so that the algorithm can 'measure' the frequency). In some simulations, switches will take place shortly after the change, while in others it takes longer time. The duration of switches follows approximately an exponential distribution (Ribeiro and Lloyd-Price, 2007) causing the interval between switches to vary widely. Due to that, it is expected that the algorithm, from different runs, will detect the change in the dynamics in different moments following the structural change in the genetic circuit. Relevantly,

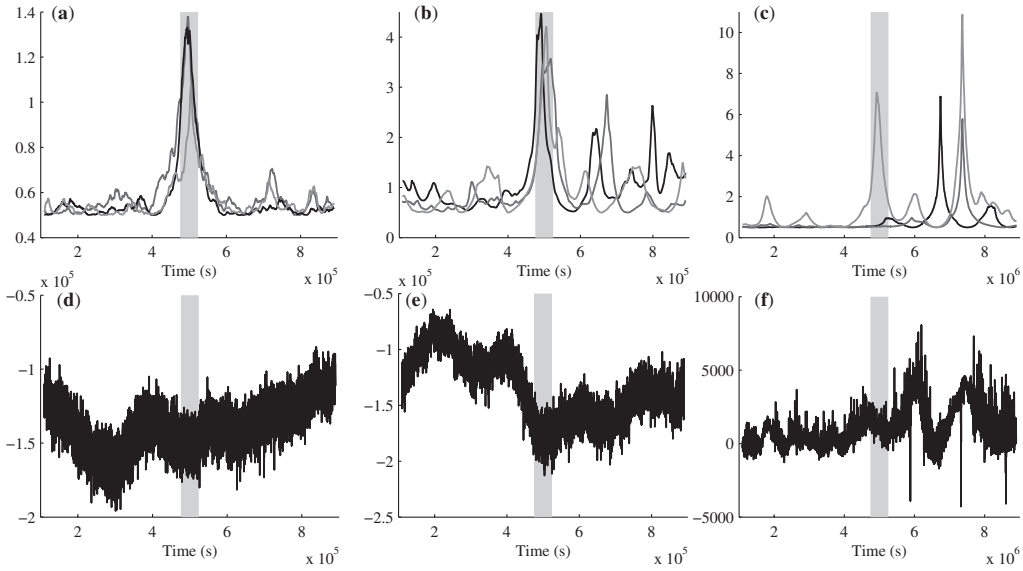


Fig. 2. Results of change point detection for KCpA (linear kernel) in models 1 (**a**), 2 (**b**) and 3 (**c**) and uLSIF in models 1 (**d**), 2 (**e**) and 3 (**f**). Vertical axis are the KCpA and uLSIF indicator outputs. In all cases, the window length is 1100 samples. Ground truth (in gray) used to compute the ROC curves. uLSIF parameters were selected by 10-fold cross-validation. In (a), (b) and (c) the results of applying KCpA to the time series from three independent runs are shown for assessing robustness to the stochastic fluctuations in protein numbers.

even for this case, in all runs KCpA detected the change, while at different moments.

We now compare the results of KCpA and uLSIF in each model. Figure 2a and d illustrate the change point detection results for Model 1. KCpA largely outperforms uLSIF. As mentioned, KCpA determines accurately the exact moment when the protein levels start changing for this model. After, the two protein levels only fluctuate around a mean level, and KCpA does not detect any significant changes. We conclude that KCpA is appropriate to detect changes in mean expression levels in highly stochastic time series, since fluctuations due to noise in the chemical kinetics are not confused with the change in mean expression levels.

Next, we test the ability of detecting changes in the degree of fluctuations in a protein's level (Model 2). In our example, the noise strength in protein numbers changes from 0.63 to 0.73, as measured by the square of the coefficient of variation (SD over the mean). The results of the detection are shown in Figure 2b and e. Again KCpA outperforms uLSIF, indicating the true change shortly after the midpoint of the time series. In comparison to the first case, the results are not as clear as there are a few false matches after the true change point. Nevertheless, the moment at which the structure changed was correctly identified. In addition, the highest peak occurs at the true change point. Thus, we conclude that KCpA detects changes in the noise strength of temporal expression levels even from time series of protein numbers that are highly stochastic both before and after the change.

We now compare KCpA and uLSIF in detecting a change in the frequency of switching of a genetic switch (Model 3). The decrease in switching frequency is due to weaker fluctuations in the protein numbers and leads to a slight increase in mean number of proteins since the decrease in fluctuations is not symmetric in relation to the mean protein level (see Supplementary Material). Thus, there are two changes, in mean and in fluctuations, which also have different time scales to be completed once initiated. The results of the detection are shown in Figure 2c and f. In this case, both the uLSIF and the KCpA have a poor performance. The KCpA does detect the true change point from one of the realizations of the data, but this might only be a coincidence. However, both methods are able to detect the strong changes in the mean levels (due to the switching dynamics) that occur at time 6×10^6 and 7.5×10^6 s (see Supplementary Fig. S1c). Thus, we conclude that for networks with switching dynamics, the detection of change of frequency requires complex analysis of the results, namely, there actually was a detection of a change point (but not of the frequency of switching), from which one can infer that the structure of the switch changed at run time. From this point of view, the uLSIF detector outperforms KCpA in this case.

3.3 Detecting change points in a complex genetic circuit

We simulate the dynamics of models from a core genetic network model of *S.cerevisiae* with 328 genes (Chowdhury *et al.*, 2010). Both topology as well as genes' transfer functions were inferred from microarray measurements following deletions and overexpressions of a gene, in optimal environmental conditions. Perhaps due to this, it was observed that unless the inferred model network is perturbed,

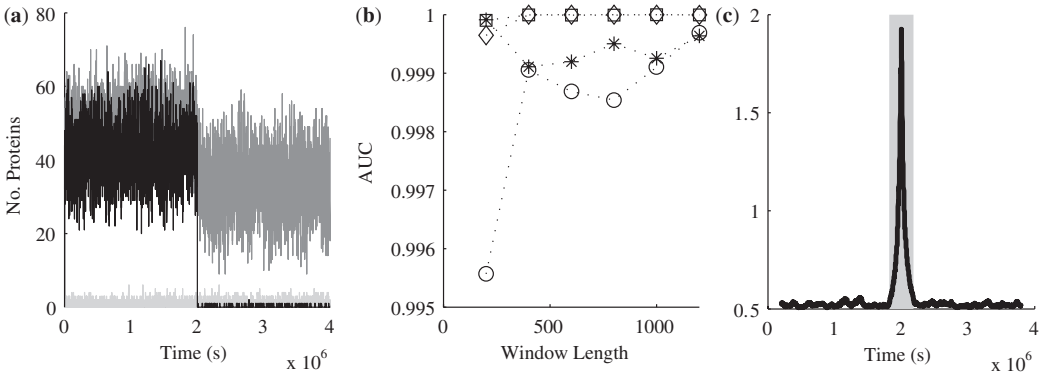


Fig. 3. (**a**) Example of protein numbers of 3 out of the 328 genes from a simulation (two were affected by the perturbation, one was not). (**b**) Results of detection. The vertical axis is the AUCs for all window sizes and all genes (the symbols 'open circle', 'asterisk', 'square' and 'diamond' are results when silencing different genes). Window size is 200 in all cases. (**c**) Example of detection results of KCpA for window size 200. Vertical axis is KCpA indicator output.

its dynamics remain relatively stable (Chowdhury *et al.*, 2010) even when modeled with the stochastic modeling strategy.

Since the topology is highly clustered and the mean connectivity $\simeq$5, and this was inferred from observing how many genes' expression level was affected by the deletion or overexpression of another gene (Chowdhury *et al.*, 2010), we can expect that when perturbing the model network by gene silencing, only a few genes' expression level will be affected, on average.

Due to that, and given our prior knowledge of the topology of this network, it is possible, for simplicity, to provide the algorithm with the data comprising only the perturbed gene and its near neighbors (between 20 and 30 genes' expression levels, selected based on smallest path length to the randomly perturbed node). In general, adding non-informative data can only decrease the performance of any detection algorithm. Therefore, attempting to detect from the expression of all 328 genes, is likely harder than when using a subset of genes. Nevertheless, this decrease ought to be minimal in this case, since even most of the selected genes were found to also be non-informative (no clear change was observed in the time series of protein numbers following the perturbation).

The dynamics is simulated for a period of time and one gene, chosen at random (see Supplementary Material), is silenced at midpoint. Examples of the protein numbers of a few genes of the network are shown in Figure 3a. Note that the mean expression levels (30–50 when gene expression is active, and close to 0 when repressed) are within realistic intervals for *S.cerevisiae* (Bar-Even *et al.*, 2006). As expected, we observed that following a perturbation, only a small fraction of genes was dynamically perturbed.

Figure 3b shows the results of detection by KCpA for varying window sizes. We use only KCpA as it exhibited the best results so far. The detection is highly accurate and robust to varying window size. In all, we ran four experiments. In each case, a different gene from the GRN was silenced. In two cases, the silenced gene was included in the simulated measurement data, while for the two other cases, we only included measurements of non-silenced genes. As one can see, the AUC's are in all case very close to 1, and it seems irrelevant whether the particular silenced gene is included or not. The accuracy in the four cases differs slightly. This is expected, since different genes have different number of outputs and thus its silencing will have differing range of effects in the GRN's dynamics. Figure 3c shows the detection results of KCpA in one case.

3.4 Time series of images of *E.coli* cells expressing RNA target for MS2d-GFP

We now apply the methods to detect changes in the number of tagged RNA molecules in cells over time from series of images taken by confocal microscopy of *E.coli* cells expressing RNA target for MS2d-GFP. The time series of total fluorescence intensity of MS2d-GFP-RNA spots in a cell is shown in Figure 4a (extracted from the images of the cell shown in Supplementary Material). Note that some time points are missing due to the microscope getting out-of-focus. From the images, one can see that in that period the objects become blurred and determining the appearance of RNA molecules becomes impossible even by a human observer. Inclusion of these outliers in the data would only result in detection of the problems of robustness of the images acquisition process.

Instead of compensating this by acquiring new data or by interpolating the existing data, we marked the out-of-focus time points as missing data and apply the detection algorithms to the remaining points. After all, missing data are common in biological measurements, and any practical algorithm should tolerate it. In our case, missing data are treated in a natural manner: both methods define a forward and a backward window, but do not restrict them to be of equal size. Thus, their use with missing data are straightforward, and it is interesting to their response to missing data.

Figure 4b and c show the results of the detection. From Figure 4a, it is visible that the fluorescence intensity of RNA-MS2d-GFP spots over time is a very noisy signal, although once the target RNA is tagged by MS2d-GFP it does not degrade (Golding *et al.*, 2005). This implies that decreases in fluorescence are not due to RNA degradation. One cause is the movements in and out of focus

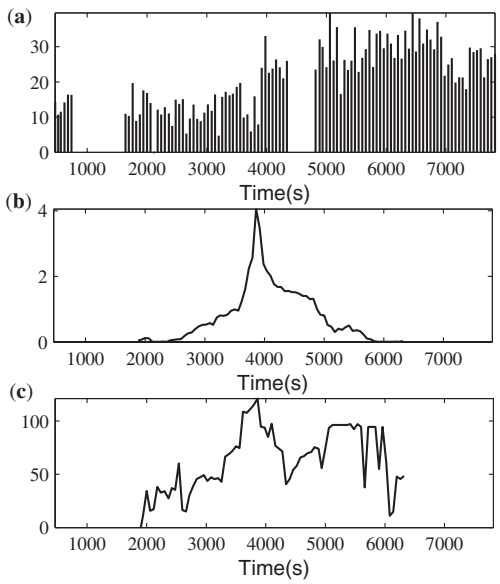


Fig. 4. (**a**) Time series of fluorescence intensity of RNA spots from images of an *E.coli* cell. Images taken for 2 h, one per minute (vertical axis is total spot intensity in arbitrary units). (**b**) Result of KCpA (vertical axis is the KCpA indicator output). (**c**) Result of uLSIF (vertical axis is the uLSIF indicator output). In both cases, window size is 25 (1500 s).

of the RNA spots along the z-axis scanned. Another source is endogenous to the tagging method, namely the number of MS2d-GFP molecules bound to the target RNA varies from 40 to 120 over time. Nevertheless, the method was found reliable in detecting, within $\leq$30 s, the appearance of new RNA molecules in the cell both empirically as well as using semiautomated methods (Golding and Cox, 2004; Golding *et al.*, 2005). From the images (Supplementary Material), a new RNA spot is detected to appear at 4100 s both empirically and by semiautomated methods. This moment should be identified by the point detection method as the one when the most significant change takes place.

The result of KPcA is shown in Figure 4b and of uLSIF in Figure 4c. For the 1D case, all kernels for KCpA are equivalent, giving the same result. In both cases, the window length was selected arbitrarily to 50 frames (i.e. 25 frames in both the backward and forward windows). Again, KPcA performed better, although both methods detect the change in the same location (at $\sim$4000 s). Note that the window size is much smaller than the size used for the models. From the experimental data, we aim to detect the appearance of individual RNA molecules, whose effect on the number of RNAs in the cell is readily observable (given the small number produced by a cell). On the other hand, in the models we detected changes in mean levels of the order of tenths of new proteins, which is a change that requires much longer time to be completed once initiated, and thus wider windows to be detected.

The accuracy of the detection in this case is of relevance for studies of gene expression dynamics from measurements. So far, the MS2d-GFP tagging system is the only method available to detect the appearance of individual RNAs *in vivo*. The analysis of the images is, unfortunately, cumbersome (see movie in Supplementary Material). Our results are promising as they show that these methods can be used to detect in an automated fashion the moments when new RNAs appear, which will provide greater confidence in the results and allow the analysis of much larger samples of cells, making the analysis of the dynamics of gene expression, one molecule at a time, more robust. An automated unbiased analysis will also facilitate comparative studies of transcription activity of different promoters.

4 CONCLUSION

Genetic networks are subject to various structural changes and external signals, which alter their dynamics in various degrees. The detection of changes requires observing the dynamics of gene expression at the single cell, single molecule level. So far, very few direct or indirect methods allow this observation (Fusco *et al.*, 2003; Golding *et al.*, 2005; Yu *et al.*, 2006), and usually the extraction of the data from the measurements is cumbersome. Further, the ground truth signal is commonly unknown, further enhancing the need of using models to develop new methods for detecting changes in the dynamics of genetic circuits.

Changes in gene expression can be complex and diverse, e.g. in time scale. To detect them it is necessary to combine the use of adequate algorithms to particular changes, and provide information of the nature of the change one wishes to detect, which requires prior knowledge of the dynamics of gene expression at the molecular level combined with the development of new data analysis methods to distinguish real changes in signals from spurious fluctuations.

We applied recently developed point change detection methods to this problem. We tested their ability in detecting changes at run time in the dynamics of stochastic models of GRNs. The changes implemented mimic naturally occurring ones. A change in the mean expression level of a gene can occur, e.g. due to gene duplication or to silencing of a gene expressing a repressor. A change in noise in protein levels can occur, e.g. due to changes in the rate of RNA degradation. A change in the switching frequency of a two-gene switch can occur, e.g. due to a change in the number or binding affinity of the repressor proteins. Finally, the silencing of a gene that is part a large gene network can occur as a response to an external signal, and will affect the expression levels of multiple genes in the network.

In most of our test cases, KCpA outperforms uLSIF. This is likely a result of the nature of the changes that we aimed to detect and of the dynamics of protein and RNA levels. uLSIF has problems with cross-validated parameter selection and its results suffer from the sensitivity to the choice of parameter. The best kernel for detection is the linear kernel. This is probably because the changes in our examples are simple changes in mean levels or, for the more complicated cases (Models 2 and 3), the data can be cast to a change in mean level. In theory, the other kernels may detect more complex changes, which is probably the cause for multiple false matches in our case. Future studies may determine which algorithms are more appropriate to which type of change.

When applying KCpA to a model of an inferred core network of 328 genes of *S.cerevisiae* (Chowdhury *et al.*, 2010), we found

it to be very accurate in detecting changes in the overall gene expression dynamics of the network, following the silencing of randomly selected genes. The size of the network did not seem to be problematic. This example is of relevance in that it shows that the method can be applied to the analysis of time series of complex gene networks, when affected by a change in either the network's structure or by an external signal or perturbation.

Finally, we applied the methods to detect, from time series of fluorescence intensity of RNA tagged with MS2d-GFP, single transcription events in live cells. Tagging RNA with MS2d-GFP proteins is, so far, the only method for detecting *in vivo* individual RNA molecules, thus, the correct extraction of information from the measurements, such as when new RNA molecules appear, is of relevance. Also in this case, the best detector appears to be KCpA, which produces a distinctive peak at the true change point.

We note that, to detect changes in the expression level of strongly expressing genes, it may be possible to assume ergodicity of the expression dynamics (similar temporal and ensemble averages), as several strongly expressing genes in, e.g. bacteria and yeast, usually exhibit fast temporal fluctuations. Provided this assumption, the changes can be detected from measurements at several time moments of expression levels across an isogenic cell population, rather than using our method since, while it is also valid, it would be more fastidious. However, generally, this assumption may not be valid. For example, studies *in vivo* and *in vitro* in *E.coli* show that most genes are rarely expressed (Bernstein *et al.*, 2002; Taniguchi *et al.*, 2010). Our measurements in Figure 4a are in agreement, since the promoter has a very slow dynamics, expressing on average only once every 700 s.

We believe that the results are promising. While the signals analyzed were poised with noise from multiple sources, the information of dynamical changes was, to a great extent, successfully extracted, both when detecting changes in the dynamics of model GRNs as well as when detecting when RNA molecules were produced in *E.coli*. Information on the nature of the changes that one wants to detect needs to be provided, to some extent. Particularly, prior knowledge is needed on the expected time length that a change takes to provoke a tangible change in the protein numbers. This is likely to be necessary regardless of the method used, as the dynamics of GRNs is extremely 'rich' in that a variety of mechanisms can affect the system in different ways, and the change may take different time lengths to be completed. In the future, we aim to further develop these methods and use them to analyze fluorescence measurements of expression of genes within genetic circuits in live cells.

Funding: Academy of Finland and Finnish Funding Agency for Technology and Innovation.

Conflict of Interest: none declared.

REFERENCES

Arkin,A. *et al.* (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, **149**, 1633–1648.

Bar-Even,A. *et al.* (2006) Noise in protein expression scales with natural protein abundance. *Nat. Genet.*, **38**, 636–643.

Belloni,A. and Didier,G. (2008) On the Behrens-Fisher problem: a globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *Ann. Stat.*, **36**, 2377–2408.

Bernstein,J.A. *et al.* (2002) 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*, **99**, 9697–9702.

Chowdhury,S. *et al.* (2010) Information propagation within the genetic network of Saccharomyces cerevisiae. *BMC Syst. Biol.*, **4**, 143.

Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.

Fisher,R. (1939) The comparison of samples with possibly unequal variances. *Ann. Eugenics*, **9**, 174–180.

Fusco,D. *et al.* (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Curr. Biol.*, **13**, 161–167.

Geva-Zatorsky,N. *et al.* (2006) Oscillations and variability in the p53 system. *Mol. Syst. Biol.*, **2**, 2006.0033.

Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

Golding,I. and Cox,E.C. (2004) RNA dynamics in live Escherichia coli cells. *Proc. Natl Acad. Sci. USA*, **101**, 11310–11315.

Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Harchaoui,Z. *et al.* (2009) Kernel change-point analysis. *Adv. Neural Inform. Proc. Syst*.

Kawahara,Y. and Sugiyama,M. (2009) Change-point detection in time-series data by direct density-ratio estimation. In *Proceedings of 9th SIAM International Conference on Data Mining*, Vol. 1, SIAM, Nevada, USA, pp. 389–400.

Kay,S. (1998) *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall, PTR, New Jersey, USA.

McClure,W.R. (1980) Rate-limiting steps in rna chain initiation. *Proc. Natl Acad. Sci. USA*, **77**, 5634–5638.

Peabody,D.S. and Lim,F. (1996) Complementation of rna binding site mutations in ms2 coat protein heterodimers. *Nucleic Acids Res.*, **24**, 2352–2359.

Ribeiro,A. *et al.* (2006) A general modeling strategy for gene regulatory networks with stochastic dynamics. *J. Comput. Biol.*, **13**, 1630–1639.

Ribeiro,A.S. and Lloyd-Price,J. (2007) SGN sim, a stochastic genetic networks simulator. *Bioinformatics*, **23**, 777–779.

Taniguchi,Y. *et al.* (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.

Yu,J. *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.

Zhu,R. *et al.* (2007) Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. *J. Theor. Biol.*, **246**, 725–745.

Supplementary material for:

Automatic detection of changes in the dynamics of delayed stochastic gene networks and *in vivo* production of RNA molecules in *Escherichia coli*

Jarno Mäkelä [1], Heikki Huttunen [1], Meenakshisundaram Kandhavelu [1], Olli Yli-Harja [1,2] and Andre S. Ribeiro [1,*]
[1]Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland
[2]Institute for Systems Biology, 1441N 34th St, Seattle, WA, 98103-8904, USA
{andre.ribeiro}@tut.fi

August 3, 2011

1 Introduction

In this supplement we describe the method of induction of the expression of the RNA targets for MS2d-GFP as well as the details regarding microscopy and cell segmentation (Section 2). The details of the image analysis for *in vivo* detection of RNA molecules are described in Section 3. After that, the delayed stochastic simulation algorithm used to simulate models of gene expression is described (Section 4). Also, the parameters for small (Section 5) and large (Section 6) genetic circuits are described. Finally, we review the two change point detection methods used, namely the Direct Density Ratio Estimation (Section 7) and the Kernel Change Point Analysis (Section 8).

2 Inducing expression of the RNA target for MS2d-GFP, microscopy and cells segmentation

Cells with both MS2d-GFP and transcript target plasmids were grown overnight at 37^{o}C in LB supplemented by the appropriate antibiotics. The following day, cells were diluted in fresh medium plus antibiotics. To induce production of MS2d-GFP, 100 ng/mL of anhydrotetracycline (IBA GmbH, Göttingen, Germany) was added to the diluted bacterial culture. Expression of the RNA target is induced by adding IPTG (1 mM, Fermentas, Finland) and L-arabinose (6.7 mM, Sigma-Aldrich, Schnelldorf, Germany). Cells are subsequently incubated with the inducers at 37^{o}C for 1 hour with shaking to a final optical density (600 nm) of 0.4.

Following induction, cells are placed on a microscopic slide between a cover slip and 0.8% LB-agarose gel pad set. Cells are visualized by fluorescence microscopy, using a Nikon Eclipse (TE2000-U, Nikon, Tokyo, Japan) inverted C1 confocal laser-scanning system with a 100x Apo TIRF (1.49 NA, oil) objective. GFP fluorescence is measured using a 488 nm laser (Melles-Griot) and a 515/30 nm detection filter. Images of cells are taken from each slide using C1 with Nikon software EZ-C1, approximately 7 min after induction, one per minute, for approximately 2 hours. Measurements under the microscope were made in room temperature ($\sim 24^oC$).

We detect cells from raw images according to the method in [Wang *et al*., 2010], that divides a gray-scale image in three classes: background, cell border and cell region (Fig. S1). An iterative cell segmentation process identifies and segments clumped cells based on size and edge information. The performance of detection of cells degrades in regions where several cells are clumped together. This can be avoided by applying a threshold based on cell size and discarding the cells whose size goes beyond the threshold.

3 Detection in vivo of individual RNA molecules in *E. coli*

The automatic spot detection method segments the MS2d-GFP-RNA spots with the kernel density estimation method for spot detection proposed in [Chen *et al*., 2008]. This method estimates the probability density function over the image from local information, and processes an image f by filtering it with a kernel:

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

where h is the smoothing parameter or bandwidth, (k,l) represents pixel location in the kernel, card is the cardinality of the set, and $K(u)$ is the kernel. We used a Gaussian kernel [Devroye *et al*., 1996], and then applied Otsu's threshold [Otsu *et al*., 1979] to segment spots from the kernel density estimated image, highlighting the spots.

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

$$I = FGI - FGA\frac{BGI - FGI}{BGA - FGA} \quad (2)$$

Finally, the number of RNA molecules in each spot is quantified using the spot intensity distribution slicing approach [Golding *et al*., 2005], that assumes that the first peak of the distribution of intensities of many RNA spots from cells on the same slide correspond to individual RNA molecules. Subsequent peaks in the distribution of intensities correspond to spots of multiple RNA molecules.

We observed that in liquid culture each cell transcribes, on average, 3-4 RNA molecules per hour [Golding *et al*., 2005] (confirmed by qPCR). In the measurements under the microscope, the average time between the productions of consecutive RNA molecules was 700 s.

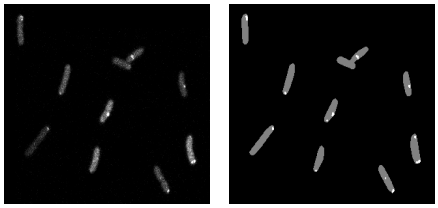
4 Delayed stochastic simulation algorithm

The delayed stochastic simulation algorithm [Roussel *et al*., 2006] (delayed SSA) differs from the original SSA [Gillespie *et al*.,1977] in that it allows the release of products of a reaction to be delayed by a specified time interval, which can be constant or a random variable. For this, a wait list of delayed events is necessary, that stores delayed products and the time when they should be released into the vessel of reactions [Roussel *et al*., 2006]. The delayed SSA proceeds as follows (t denotes time):

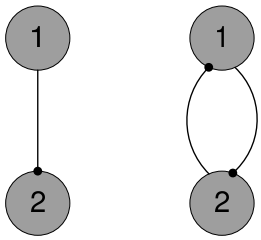
1. Set t = 0, t_{stop} = stop time, set initial number of molecules and reactions, and create empty waitlist *L*.
2. Generate an SSA step for reacting events to get the next reacting event R_1 and its time of occurrence, t_1.
3. Compare t_1 with the least time in *L*, t_{min}. If $t_1 < t_{min}$ or *L*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

5 Models of small genetic circuits

We implement three model GRNs, named 1, 2, and 3. In all, parameter values are within realistic intervals for *E. coli* [Ribeiro *et al*., 2010]. Each model is used to test the ability of the algorithms in detecting a different type of change in the dynamics of the GRN at runtime. All models are built from the set of reactions (3) to (8) [Ribeiro *et al*., 2008], where i = 1, 2 (when only the index i is present), and i, j = 1, 2



Supplementary Figure S1: Unprocessed image of MS2d-GFP-tagged RNA molecules in *E. coli* cells (left) and the corresponding segmented image showing the detected cells (grey) and the spots (white) within (right).



Supplementary Figure S2: The small genetic network used in models 1 and 2 (Left). The small genetic network used in model 3 (Right). The line ending with a dot represents repression of a gene by the other gene.

with $i \neq j$ (when both indices are present):

$$Pro_i + Rp \xrightarrow{k_t} Pro_i(\tau_1) + Rp(\tau_2) + R_i(\tau_1) \quad (3)$$

$$R_i + Rib \xrightarrow{k_{tr}} R_i(\tau_3) + Rib(\tau_4) + P_i(\tau_5, \tau_{5std}) \quad (4)$$

$$R_i \xrightarrow{rbsd} \emptyset \quad (5)$$

$$P_i \xrightarrow{k_d} \emptyset \quad (6)$$

$$Pro_i + P_j \underset{kunrep}{\overset{krep}{\rightleftharpoons}} Pro_i P_j \quad (7)$$

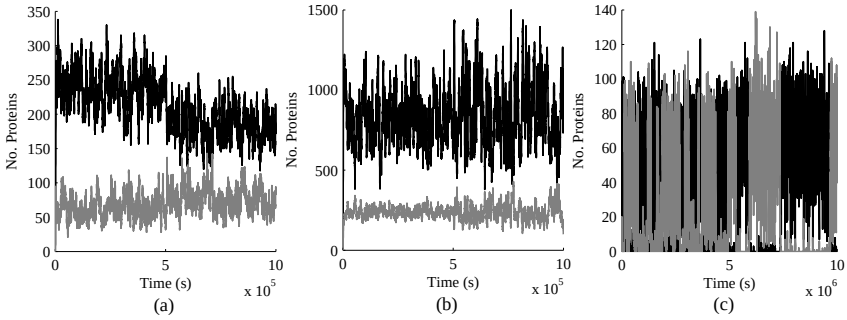
$$Pro_i P_j \xrightarrow{k_d} Pro_i \quad (8)$$

Gene expression is modeled by the multiple time-delayed reactions for transcription (3) and translation (4), where Pro_i is the promoter of gene i, Rp is an RNA polymerase, Rib is a ribosome, and R_i is the ribosome binding site of each RNA. The delays (τ_1 to τ_5) account for the duration of the processes in transcription and translation. When a product X has a delay τ, represented by X(τ), it implies that when the reaction occurs, it takes τ seconds after that for X to appear in the system.

Reaction (4) for translation accounts for the variability of the time to complete a functional protein (translation, folding, activation, etc.), given that the delay of P_i follows a normal distribution, with a mean of τ_5 and a standard deviation of $\tau_{5,std}$ [Zhu *et al.*, 2007]. Reactions (8) model the binding and unbinding of a repressor protein to a gene's transcription factor binding site. Reaction (5) models the degradation of RNAs, while reactions (6) and (8) model the degradation of proteins, when free and when bound to a promoter region, respectively.

Unless stated otherwise, the rates (in s^{-1}) of these reactions are k_t = 0.01, k_{tr} = 0.00042, d_{rbs} = 0.01, k_{rep} = 0.1, k_{unrep} = 0.1, and k_d = 0.0012. Time delays (in seconds) are τ_1 = 40, τ_2 = 90, τ_3 = 2, τ_4 = 58, τ_5 = 420, and τ_{5std} = 140. Each 'cell' is initialized with P_i = 0 and R_i = 0, for all i, and with one promoter of each gene (Pro_1 = 1, Pro_2 = 1), 40 RNA polymerases (Rp = 40), and 100 ribosomes (Rib = 100).

Model 1 is a 2-gene network, where the protein expressed by gene 1 represses the expression of gene 2. It is used to model a change in the mean level of proteins at runtime. For that, at moment t = 5x10^5 s of a simulation, the repression on gene 2 is increased by increasing the expression of P_1. This decreases



Supplementary Figure S3: (a) Time series of proteins P_1 (grey) and P_2 (black) of model 1, prior and after the change of mean levels at runtime (t = 5x10^5 s). (b) Time series of proteins P_1 (grey) and P_2 (black) of model 2, prior and after increasing the amplitude of fluctuations (t = 5x10^5 s). (c) Time series of proteins P_1 (grey) and P_2 (black) of model 3, prior and after increasing the toggling frequency of the genetic switch (at t = 5x10^6 s).

the mean rate of production of P_2. An example of a time series of the number of proteins of both genes, prior and after the change point, is shown in Fig. 1(a). The schematic illustration of network is shown in Figure S2 (left).

Model 2 is identical to model 1. Also, initially, all parameter values are identical to those of model 1. In Model 2 occurs a different change at run time, which allows changing the noise strength of the time series of the proteins. For this, at t = 5x10^5 s, both the transcription initiation rate of gene 1, as well as the degradation rate of P_1, are set to 10 times less their initial value (Fig. 1 (b)). Since the release of the promoter for new transcription events is delayed, this causes the noise strength of the time series of RNA and, thus, protein numbers to increase [Ribeiro *et al*., 2010]. The schematic illustration of network is shown in Figure S2 (left).

Model 3 consists of a genetic toggle switch (the protein expressed by each gene represses the expression of the other). The change at runtime time is a change in the noise strength of the protein levels of both genes (Fig. 12 (c)). Due to this, the switching frequency changes at runtime. To do this, at t = 5x10^6 s, we decrease in the two genes, both the rate of transcription initiation (k_t) as well as the degradation rate of the proteins to a fraction of their initial values [Ribeiro *et al*., 2009]. The schematic illustration of network is shown in Figure S2 (right).

Examples of the time series of proteins numbers in these models, prior and after a change are shown in Figure S3.

6 Model of a large scale genetic circuit

We also apply the methods to detect change points in a large scale genetic network. We model the inferred core gene regulatory of Yeast [Chowdhury *et al*., 2010]. This network consists of 328 genes, each of which expressing a protein and connected by repressive and activating interactions. At a given moment in time we delete a random gene, and then detect when this occurs, from the time series of a subset of the 328 genes. In general, the subset was defined to be the closer neighbors and of size of 20 to 30 genes. We do not inform the algorithms of when the deletion takes place, neither of what gene was randomly selected for deletion.

7 Direct density ratio estimation for change point detection

Formally, the problem of change point detection can be stated as follows. Given a multidimensional time series $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbf{R}^n$, which time points K represent change in some sense, given the data samples in the M-point backward window $\mathbf{X}_B = (\mathbf{x}_{K-M}, \ldots, \mathbf{x}_{K-1})$ and the M-point forward window $\mathbf{X}_F = (\mathbf{x}_{K+1}, \ldots, \mathbf{x}_{K+M})$.

The first detection method used is based on kernel density estimation [Kawahara *et al*., 2009] and attempts to model the densities based on the data. In the case of change point detection, one could find the kernel density estimates of both the backward and forward window directly, and judge their similarity using, for example, the Kolmogorov-Smirnov test. Instead, this method considers the likelihood

of change using density ratios directly, thus requiring the inference of only one function (the density ratio) instead of two (the densities of the forward and backward windows), which aids in avoiding the curse of dimensionality as the required data does not increase as rapidly with the dimension.

There are two direct density ratio estimation methods: *Kullback-Leibler importance estimation procedure* (KLIEP) [Sugiyama *et al*., 2008] and *Unconstrained least-squares importance fitting* (uLSIF) [Kanamori *et al*., 2009]. We use the latter due to the lower computational cost of the least squares approach, since the two methods are similar in accuracy [Kanamori *et al*., 2009]. The goal of the uLSIF estimation is to estimate the density ratio (called the importance function) $w(\mathbf{x}) : \mathbf{R}^n \to \mathbf{R}$:

$$w(\mathbf{x}) = \frac{p_{\mathbf{X}_F}(\mathbf{x})}{p_{\mathbf{X}_B}(\mathbf{x})}, \tag{9}$$

where $p_{\mathbf{X}_F}$ and $p_{\mathbf{X}_B}$ are the probability densities of the forward and backward windows, respectively. The least squares approach to estimate $w(\mathbf{x})$ uses the following linear model:

$$\hat{w}(\mathbf{x}) = \sum_{k=1}^{M} \alpha_k \varphi_k(\mathbf{x}), \tag{10}$$

where the weights $\alpha_k \in \mathbf{R}$ are learned from the data, and serve as coefficients for the basis functions $\varphi_k(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbf{R}^n$. A common choice for the basis function is the Gaussian kernel of width σ centered at the forward window points $\mathbf{X}_F = (\mathbf{x}_{K+1}, \ldots, \mathbf{x}_{K+M})$:

$$\varphi_k(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_{K+k}\|^2}{2\sigma^2}\right). \tag{11}$$

The centers are chosen as the forward window points, because the forward window density $p_{\mathbf{X}_F}(\mathbf{x})$ is the numerator of (9), and in this way the centers are at points where $w(\mathbf{x})$ is likely to have large values. With these definitions, the direct density ratio estimation problem can be formulated as a minimization problem "*Least Squares Importance Fitting (LSIF)*":

$$\min_{\boldsymbol{\alpha} \in \mathbf{R}^b} \left(\frac{1}{2}\boldsymbol{\alpha}^T \hat{\mathbf{H}} \boldsymbol{\alpha} - \hat{\mathbf{h}}^T \boldsymbol{\alpha} + \lambda \mathbf{1}^T \boldsymbol{\alpha}\right) \tag{12}$$

$$\text{subject to } \boldsymbol{\alpha} \geq 0, \tag{13}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ and $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbf{R}^M$. Moreover, the matrices $\hat{\mathbf{H}}$ and $\hat{\mathbf{h}}$ are defined through their (i, j)-th and i-th elements as

$$\left[\hat{\mathbf{H}}\right]_{i,j} = \frac{1}{M} \sum_{k=1}^{M} \varphi_i\left(\mathbf{x}_{K-k}\right) \varphi_j\left(\mathbf{x}_{K-k}\right) \tag{14}$$

and

$$[\hat{\mathbf{h}}]_i = \frac{1}{M} \sum_{k=1}^{M} \varphi_i\left(\mathbf{x}_{K+k}\right) \tag{15}$$

Finally, the LS procedure is regularized using the regularization parameter $\lambda > 0$. Equations (12) and (13) represent a convex quadratic programming problem, which can be solved [Kanamori *et al*., 2009]. It turns out, that the nonnegativity constraint (13) can be discarded without a significant loss in accuracy. However, this makes the problem ill-posed, and the linear penalty of Equation (12) has to be replaced by a quadratic one [Kanamori *et al*., 2009]. This results in the *Unconstrained* LSIF (uLSIF) optimization problem:

$$\min_{\boldsymbol{\alpha} \in \mathbf{R}^b} \left(\frac{1}{2}\boldsymbol{\alpha}^T \hat{\mathbf{H}} \boldsymbol{\alpha} - \hat{\mathbf{h}}^T \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^T \boldsymbol{\alpha}\right). \tag{16}$$

The (regularized) solution for such a quadratic minimization problem is given by:

$$\hat{\boldsymbol{\alpha}} = \left(\hat{\mathbf{H}} + \lambda \mathbf{I}\right)^{-1} \hat{\mathbf{h}}. \tag{17}$$

Having estimated the density ratio, we only need to evaluate it at the forward window points. The estimated likelihood ratio for the existence of the change-point can be shown to be [Kawahara *et al*., 2009]:

$$\mathrm{LR} = \sum_{k=1}^{M} \log(\hat{w}(\mathbf{x}_{K+k})) = \sum_{k=1}^{M} \log\left(\sum_{j=1}^{M} \alpha_j \varphi_j(\mathbf{x}_{K+k})\right). \tag{18}$$

There remains two open parameters, the kernel width σ and the regularization parameter λ. Their values can be estimated using cross-validation (CV), carried out at 100 randomly selected time points of the entire time series. The final value for the actual change point detection is the mean of these 100 CV tests. Note, that the selection of exactly 100 CV tests is arbitrary, which was selected based on experimentation with our data.

8 Kernel change point analysis for change detection

Kernel methods are a class of algorithms originally developed for pattern recognition, but their use has spread to other areas [Scholkopf *et al*., 2002]. Among the kernel methods, *support vector machines* (SVM) have gained the widest popularity. All kernel methods are based on mapping the input data into a higher dimensional feature space, and the calculations can be done effectively even in infinite dimensional spaces using the *kernel trick*. The kernel trick replaces all dot products between vectors $\mathbf{x}$ and $\mathbf{y}$ in the algorithm by a Mercer kernel, such as the Gaussian RBF kernel,

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right). \tag{19}$$

Kernel methods have been applied to the problem of change point detection in [Harchaoui *et al*., 2009, Desobry *et al*., 2005]. The *Kernel Change Detection* (KCD) [Desobry *et al*., 2005] uses a single class SVM, where the data in the backward and forward windows are clustered using the SVM and the similarity metric is calculated from the common area of the two clusters. Harchaoui *et al.* [Harchaoui *et al*., 2009] used a kernel-based binary classifier to separate the forward and backward windows from each other. The success of the classification is the dissimilarity of the two sets. Their separability can be directly measured when using the *Kernel Fisher Discriminant*, which is the basis of this method, named *Kernel Change-point Analysis* (KCpA). Since KCpA outperforms KCD [Desobry *et al*., 2005] and is conceptually simpler, we use it as a representative of kernel methods for our change-point detection problem.

The *Kernel Fisher Discriminant Ratio* (KFDR) is the ratio of the between-class-variance and the within-class-variance. The optimal classifier is the one that maximizes this ratio, but here we are more interested on using the KFDR itself as a measure of dissimilarity of the two sets (classes). It can be shown that the (maximum) KFDR is proportional to

$$\mathrm{KFDR} = \left\| \left(\hat{\mathbf{\Sigma}}_{\mathbf{X}_F} + \hat{\mathbf{\Sigma}}_{\mathbf{X}_B} + \lambda \mathbf{I}\right)^{-1/2} \left(\hat{\boldsymbol{\mu}}_{\mathbf{X}_F} - \hat{\boldsymbol{\mu}}_{\mathbf{X}_B}\right) \right\|^2, \tag{20}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{X}_F}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{X}_B}$ denote the sample covariance matrices and $\hat{\boldsymbol{\mu}}_{\mathbf{X}_F}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{X}_B}$ denote the sample means of the forward and backward windows calculated in the feature space, respectively. Moreover, the term $\lambda\mathbf{I}$ is a diagonal matrix with $\lambda > 0$, and the purpose is to regularize the solution. More specifically,

$$\hat{\boldsymbol{\mu}}_{\mathbf{X}_F} = \frac{1}{M}\sum_{k=1}^{M} \kappa\left(\mathbf{x}_{K+k}, \cdot\right), \tag{21}$$

and

$$\hat{\mathbf{\Sigma}}_{\mathbf{X}_F} = \frac{1}{M-1}\sum_{k=1}^{M} \left(\kappa\left(\mathbf{x}_{K+k}, \cdot\right) - \hat{\boldsymbol{\mu}}_{\mathbf{X}_F}\right) \otimes \left(\kappa\left(\mathbf{x}_{K+k}, \cdot\right) - \hat{\boldsymbol{\mu}}_{\mathbf{X}_F}\right), \tag{22}$$

where symbol $\otimes$ denotes outer product in the kernel space. The backward mean and covariance are defined in a similar manner. Note that, although the explicit calculation of the above quantities may be impossible due to the structure of the kernel space, the evaluation of the KFDR can be made in the feature space [Mika *et al*., 1999].

The KFDA based point of change indicator is obtained after a normalization procedure known as *studentization* [Harchaoui *et al*., 2009]. However, the studentization does not improve the detection's accuracy, as it is only used to normalize the expected mean and variance of the indicator.

In [Harchaoui *et al*., 2009], a running maximum partition strategy is used, meaning that the entire sequence $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is partitioned into two sets at all time points $K = 1, \ldots, N-1$. Our implementation uses the same online sliding window method as the direct density ratio approach. Although this choice prohibits the use of the quick computational strategy of [Harchaoui *et al*., 2009], it is essential for comparison of the two methods.

The KFDA detector has a few parameters, which affect the performance of detection. However, there are guidelines for inferring suitable values [Harchaoui *et al*., 2009]. The regularization term is proposed

to have a fixed value of $\lambda = 10^{-5}$, which we use. Another significant parameter when using the RBF kernel of Equation (19) is the kernel bandwidth σ. Harchaoui et al. propose using the Silverman's Rule of Thumb (ROT) for inferring a suitable bandwidth from the data, which we follow. The Silverman's ROT is used in kernel density estimation and is defined by $\hat{\sigma}_{\mathrm{ROT}} = 1.06\hat{\sigma}_x N^{-1/5}$, where $\hat{\sigma}_x$ is the variance of the data and N is the number of samples [Silverman *et al*., 1986].

References

[Chen *et al*., 2008] Chen, T., Lu, H. H., Lee, Y. and Lan, H. (2008) Segmentation of cDNA microarray images by kernel density estimation, *Physical Biology*, **41(6)**,1021-1027

[Chowdhury *et al*., 2010] Chowdhury, S., Lloyd-Price, J., Smolander, O.-P., Baici, W., Hughes, T., Yli-Harja, O., Chua, G., and Ribeiro, A. (2010). Information propagation within the genetic network of saccharomyces cerevisiae. *BMC Systems Biology*, **2010**, 4.

[Desobry *et al*., 2005] Desobry, F., Davy, M. and Doncarli, C. (2005) An online Kernel change detection algorithm, *IEEE Transactions on Signal Processing*, **53(8)**, 2961-2974

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

[Gillespie *et al*.,1977] Gillespie, D. T. (1977) Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry*, **81(25)**,2340-2361

[Golding *et al*., 2005] Golding, I. and Paulsson, J. and Zawilski, S. M. and Cox, E. C. (2005) Real-time kinetics of gene activity in individual bacteria, *Cell*, **123(6)**,1025-1036

[Harchaoui *et al*., 2009] Harchaoui, Z., Bach, F. and Moulines, E. (2009) Kernel Change-point Analysis, *Advances in Neural Information Processing Systems (NIPS)*

[Kanamori *et al*., 2009] Kanamori, T., Hido,S. and Sugiyama, M. (2009) A least-squares approach to direct importance estimation, *Journal of Machine Learning Research*, **10**,1391-1445

[Kawahara *et al*., 2009] Kawahara, Y. and Sugiyama, M. (2009) Change-point detection in time-series data by direct density-ratio estimation, *Proc of 9th SIAM Int Conf on Data Mining*, **1**, 385-396

[Mika *et al*., 1999] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Muller, K - R. (1999) Fisher discriminant analysis with kernels, Neural Networks for Signal Processing, 1999, 41-48

[Otsu *et al*., 1979] Otsu, N. (1979) Threshold Selection Method from Gray-level Histograms, *IEEE Trans Syst Man Cybern*, **SMC-9(1)**,62-66

[Ribeiro *et al*., 2008] Ribeiro, A. S. (2008) Dynamics and evolution of stochastic bistable gene networks with sensing in fluctuating environments, *Phys Rev E*, **78(6)**, 061902

[Ribeiro *et al*., 2009] Ribeiro, A. S., Dai, X. and Yli-Harja, O. (2009) Variability of the distribution of differentiation pathway choices regulated by a multipotent delayed stochastic switch, *J of Theor Bio*, **260(1)**, 66-76

[Ribeiro *et al*., 2010] Ribeiro, A. S. (2010) Stochastic and delayed stochastic models of gene expression and regulation, *Math. Biosciences*, **223(1)**, 1-11

[Roussel *et al*., 2006] Roussel, M. R. and Zhu, R. (2006) Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression, *Physical Biology*, **3(4)**, 274-284

[Scholkopf *et al*., 2002] Schölkopf, B. and Smola, A. (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002

[Silverman *et al*., 1986] Silverman, B. (1986) Density Estimation for Statistics and Data Analysis, Chapman-Hall, 1986

[Sugiyama *et al*., 2008] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P. and Kawanabe, M. (2008) Direct importance estimation for covariate shift adaptation, *Ann. of the Inst. of Stat. Math.*, **60(4)**, 699-746

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

[Zhu *et al*., 2007] Zhu, R., Ribeiro, A. S., Salahub, D. and Kauffman, S. A. (2007) Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models, *J. of Theor. Bio.*, **246(4)**, 725-745

Publication II

J. Mäkelä, M. Kandhavelu, S.M.D. Oliveira, J.G. Chandraseelan, J. Lloyd-Price, J. Peltonen, O. Yli-Harja, and A.S. Ribeiro, "*In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter", *Nucleic Acids Research*, 41(13):6544-6552, 2013.

6544–6552 *Nucleic Acids Research, 2013, Vol. 41, No. 13* *Published online 3 May 2013*
doi:10.1093/nar/gkt350

In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter

Jarno Mäkelä[1], Meenakshisundaram Kandhavelu[1], Samuel M. D. Oliveira[1], Jerome G. Chandraseelan[1], Jason Lloyd-Price[1], Juha Peltonen[1], Olli Yli-Harja[1,2] and Andre S. Ribeiro[1,*]

[1]Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland and [2]Institute for Systems Biology, 1441N 34th Street, Seattle, WA 98103-8904, USA

Received March 21, 2013; Revised April 11, 2013; Accepted April 13, 2013

ABSTRACT

Using a single-RNA detection technique in live *Escherichia coli* cells, we measure, for each cell, the waiting time for the production of the first RNA under the control of P_{BAD} promoter after induction by arabinose, and subsequent intervals between transcription events. We find that the kinetics of the arabinose intake system affect mean and diversity in RNA numbers, long after induction. We observed the same effect on $P_{lac/ara-1}$ promoter, which is inducible by arabinose or by IPTG. Importantly, the distribution of waiting times of $P_{lac/ara-1}$ is indistinguishable from that of P_{BAD}, if and only if induced by arabinose alone. Finally, RNA production under the control of P_{BAD} is found to be a sub-Poissonian process. We conclude that inducer-dependent waiting times affect mean and cell-to-cell diversity in RNA numbers long after induction, suggesting that intake mechanisms have non-negligible effects on the phenotypic diversity of cell populations in natural, fluctuating environments.

INTRODUCTION

Transcription in *E. coli* is, at a genome-wide scale, a relatively rare stochastic event (1–3). Further, many genes only become active in response to external stimuli (4–7), via processes that are also stochastic (7). Although much is known on the noise in gene expression at the single-cell level (1–3,7–10), most of our present knowledge concerning the kinetics of response, in terms of gene activity, to external signals concerns the average behaviour of cell populations alone (11). However, to characterize the dynamics and the underlying steps of intake processes, it is necessary to observe their effects in individual live cells (12). This observation should inform also on the robustness of cellular response mechanisms by informing on the degree of change in the responses of a single cell to multiple occurrences of the same stimulus, as well as the difference in responses to different stimuli.

One of the most well-known gene activation mechanisms is the arabinose utilization system of *E. coli*. This system imports arabinose into the cell by AraFGH, an arabinose-specific high-affinity ABC transporter (11,13–15), and by a low-affinity transporter, AraE, which binds to the inner membrane and makes use of electrochemical potential to intake the arabinose (11,16,17). This system exhibits wide variability in the timing of activation and in the rates of accumulation of inducer molecules (18). It has been hypothesized that this is due to the cell-to-cell variability in the numbers of proteins responsible for the intake of arabinose (18). Interestingly, if the intake gene *araE* is placed under the control of a constitutive promoter the intake rates become more homogenous (19–21), suggesting that the diversity in the number of intake proteins is a non-negligible source of cell-to-cell variability in the kinetics of the arabinose utilization system (12).

Evidence suggests that when the intracellular concentration of arabinose exceeds a threshold, the dimeric AraC protein activates the genes that code for the proteins responsible for the intake (AraE and AraFGH) and for the catabolism of arabinose (*araBAD*) (11,22). In the absence of arabinose, AraC binds two half-sites on the DNA (I_1 and O_2) and promotes the formation of a DNA loop that prevents access of RNA polymerases to the promoters in that region (P_C and P_{BAD}). When bound by arabinose, AraC binds instead to the adjacent I_1 and I_2 half-sites. The resulting configuration promotes transcription initiation at P_{BAD} (11).

Transcription initiation is a complex, multi-stepped process (23,24). *In vitro* measurements suggest that this process has at least two to three rate limiting steps (25,26). It starts when the RNA polymerase binds to the

*To whom correspondence should be addressed. Tel: +358408490736; Fax: +358331154989; Email: andre.ribeiro@tut.fi

promoter region of the DNA molecule, forming the closed complex, which is followed by the open complex formation and promoter escape (27,28). The RNA polymerase then elongates the nascent RNA (28). Evidence suggests that, in general, initiation is much longer in duration than elongation (26,29). Recent *in vivo* measurements of the kinetics of initiation of $P_{lac/ara-1}$ and P_{tetA} promoters have shown that RNA production under the control of these promoters is a sub-Poissonian process (8–10). These studies also support the existence of multiple steps at the stage of initiation, significantly limiting the rate of RNA production, as suggested by *in vitro* measurements (30).

Here, we investigate the degree of contribution of the process of intake of arabinose and of the process of transcription under the control of P_{BAD} to the cell-to-cell diversity in RNA production. Namely, we report measurements of the *in vivo* kinetics of induction and transcript production of P_{BAD} with single-molecule sensitivity, making use of the MS2d-GFP tagging of RNA in *E. coli* (31). For that, in each cell, we measure the waiting time until the first RNA is produced after induction and the subsequent intervals between consecutive transcript productions. For comparison, we conduct the same measurements for $P_{lac/ara-1}$ when induced by either of its two inducers, arabinose and IPTG.

MATERIALS AND METHODS

Strains and plasmids

Escherichia coli strain DH5α-PRO was generously provided by I. Golding, University of Illinois and contains the construct PROTET-K133, carrying $P_{LtetO-1}$-MS2d-GFP (31), along with a new construct, pMK-BAC (P_{BAD}-mRFP1-MS2-96bs), which is a single-copy F-based vector carrying a sequence coding for a monomeric red fluorescent protein (mRFP1) followed by a 96 binding site array under the control of P_{BAD} (cloning information provided in Supplementary Methods) (see also Supplementary Figures S1 and S2). The strain with plasmids $P_{LtetO-1}$-MS2d-GFP and pIG-BAC ($P_{lac/ara-1}$-mRFP1-MS2-96bs) (32) was used as well. The DH5α-PRO strain [identical to Z1 (31)] is a genuine producer of AraC (33). No modifications were made to the chromosome of this strain in our experiments.

Media and growth conditions

Cells were grown overnight at 30°C with aeration and shaking in Luria-Bertani (LB) medium, supplemented with antibiotics according to the plasmids. The cells were diluted in fresh M63 medium and allowed to grow until an optical density of $OD_{600} \approx 0.3$–0.5. To attain full induction of the MS2d-GFP reporter, cells were pre-incubated for 40 min with 100 ng/ml anhydrotetracycline (aTc, IBA GmbH). The same protocol was used for each strain.

Microscopy

For microscopy measurements, cells were pelleted and re-suspended in ~50 µl of fresh M63 medium. Afterwards, few microlitres of cells were placed between a 3% agarose gel pad made with medium and a glass coverslip before assembling the imaging chamber (FCS2, Bioptechs). Before the starting of the experiment, the chamber was heated to 37°C.

Cells were visualized in a Nikon Eclipse (TE2000-U, Nikon, Japan) inverted microscope with C1 confocal laser-scanning system using a ×100 Apo TIRF objective. A flow of fresh, pre-warmed M63 medium containing the inducer was provided with a peristaltic pump at a rate of 1 ml/min. Images were taken once per minute for 2 h, and the laser shutter was open only during the exposure time to minimize photobleaching. The peristaltic pump was initialized at the same time as the collection of the time series. For image acquisition, we used Nikon EZ-C1 software. GFP fluorescence was measured using a 488 nm argon ion laser (Melles-Griot), 515/30 nm emission filter and a pixel dwell time of 3.36 µs (total image acquisition time of 3.5 s).

An interacting multiple model filter-based autofocus strategy (34) was used to correct focus drift in time series acquisitions. The method estimates the focal drift using an interacting multiple model filter algorithm to predict the focal drift at time t based on the measurement at t-1. It allows reducing the number of required images at different z-planes for drift correction, thus minimizing photobleaching.

Data and image analysis

Data and images were analysed using custom software written in MATLAB 2011b (MathWorks). Cells were detected from fluorescence images by a semi-automatic method described previously (8). In time series, the area occupied by each cell was manually masked. Principal component analysis was used to obtain the dimensions and orientation of the cells within each mask. Fluorescent spots in the cells were automatically segmented using density estimation with a Gaussian kernel (35) and Otsu's thresholding (36). Finally, background-corrected spot intensities were calculated and summed to produce the total spot intensity in each cell.

Moments of appearance of novel target RNA molecules in each cell were obtained from time-lapse fluorescence images by fitting the corrected total spots intensity over time in each cell to a monotone piecewise-constant function by least squares (37). The number of terms was selected using the F-test with a P-value of 0.01. Each jump corresponds to the production of a single RNA molecule (37). An example of this procedure is shown in Figure 1D. For more details on the image analysis see (8). Note that, in cells that do not contain target RNA molecules at the start of the measurements, the number of novel RNA molecules detected since the start of the measurements until a given moment equals the total number of RNA molecules in the cell at that moment.

Because some cells already contained target RNA molecules at the start of the measurement, the total RNA numbers within cells at a given moment in time is obtained using a different method. Specifically, when

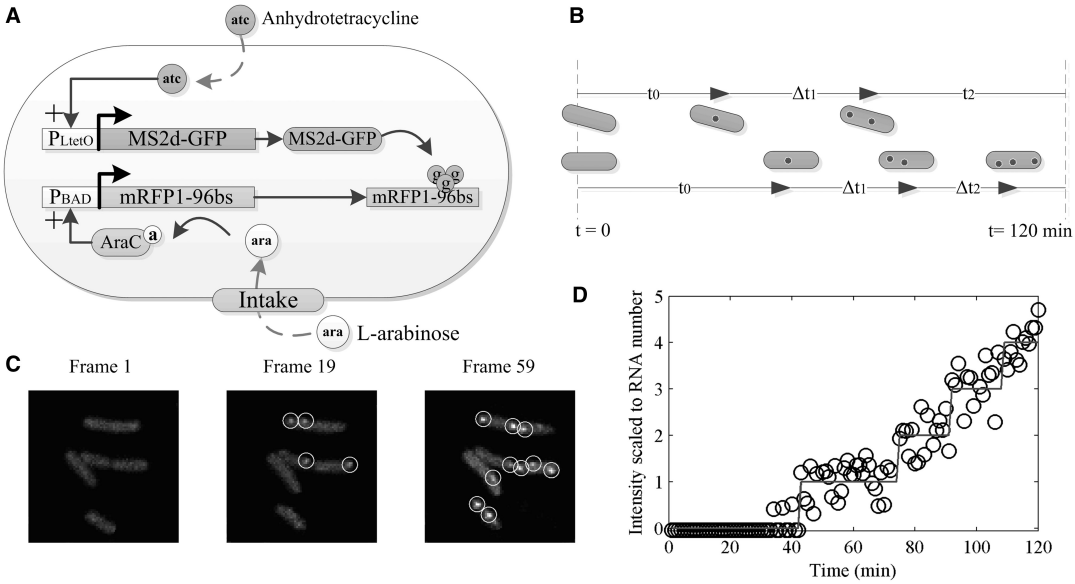


Figure 1. Measurement system. (**A**) Components of the detection system. The expression of the tagging protein, MS2d-GFP, is controlled by P_{LtetO} (33) and is inducible by anhydrotetracycline (aTc). The target RNA contains an mRFP1 coding region, followed by an array of 96 MS2d-binding sites. Expression of the target RNA is controlled by P_{BAD} whose activity is regulated by AraC and the inducer L-arabinose. The target construct is on a single-copy F-plasmid. The tagging construct is on a medium-copy vector. (**B**) Figurative description of the waiting time for the first RNA production (t_0) and intervals between subsequent productions (Δt). Images are taken once per minute for 2 h. (**C**) Example of *E. coli* cells expressing MS2d-GFP and target RNA. GFP-tagged RNA molecules are marked by circles. (**D**) Time course of total intensity of spots in a cell (circles) and monotone piecewise-constant fit (line).

comparing measurements using MS2d-GFP tagging and using plate reader (Supplementary Figure S4), the total number of MS2d-GFP–tagged RNA molecules was extracted from the total spot intensity distribution, obtained from all cells in an image obtained at a given moment after induction. For this, the first peak of the obtained distribution is set to correspond to the intensity of a single-RNA molecule. The number of tagged RNAs in each spot can be estimated by dividing its intensity by that of the first peak (32).

RESULTS

Experimental design

To study the kinetics of expression of P_{BAD}, we detect individual RNA molecules, as these are produced in live cells and register when these events occur. For this, we placed the P_{BAD} promoter on a single-copy F-plasmid, followed by a coding region for mRFP1 and an array of 96 binding sites for MS2d-GFP–tagging proteins (32) (Figure 1A). The expression of MS2d-GFP is controlled by P_{TetO}, which is activated before the gene of interest so that sufficient MS2d-GFP proteins are present when target RNA molecules appear. Induction of P_{BAD} and image acquisitions is initialized simultaneously (Figure 1B). For this, we use a temperature-controlled imaging chamber and a peristaltic pump for introducing inducers and fresh media. From the fluorescence images, using semi-automated cell segmentation and tracking (Figure 1C) (8), we measure in each cell the time for the first RNA to appear (named 'waiting time', t_0), as well as the subsequent intervals between consecutive RNA productions, Δt, until cell division occurs or until the end of the measurement period (Figure 1D).

Given that values of t_0 can only be obtained from cells of the first generation (i.e. cells already on the slide when the measurement begins), and as cells that do not divide in the first 2 h will not, in general, divide afterwards, we limited the measurement period to 2 h for simplicity. This was possible, as this period also proved to be sufficient to acquire enough samples of Δt.

From cells born during the measurement period, we only extract intervals between consecutive RNA productions, not waiting times, as these contain inducers by inheritance. We detected no difference in the distributions of intervals obtained from such cells and cells already present when induction is initiated. Finally, we observed ~0.2 RNA molecules per cell, at the moment preceding induction, because of spurious transcription events. Cells where a target RNA was already present at the start of the measurement were also not used to obtain values of t_0.

First, we compared by quantitative polymerase chain reaction the RNA production from the F-plasmid and from the native gene under the control of P_{BAD} (Supplementary Methods). Using 16S rRNA as reference

gene, we observe similar trend in activity over time in the native promoter and in the one on the F-plasmid (Supplementary Figure S3).

We next compare expression levels of the target gene, when assessed by independent methods, for two induction levels, namely, 0.1 and 1% L-arabinose (Supplementary Methods). In Supplementary Figure S4A and B, we show the temporal variation after induction in mean numbers of MS2d-GFP–tagged RNA molecules in cell populations and in the fluorescence intensity of RFP measured by plate reader, respectively.

The plate reader measurements of mRFP1 levels, 2 h after induction in liquid culture, show a fold change of 1.67 times when L-arabinose is increased from 0.1 to 1%. The MS2d-GFP *in vivo* detection method shows a fold change of 1.74 between these same conditions, showing that the results from the two methods are in accordance. From this and the previous experiment, we also conclude that the MS2d-GFP tagging method accurately detects RNA production of the target gene, and that the target gene behaves similarly to the natural system.

We also assessed for what range of inducer concentrations is the target gene under full induction. We measured with the plate reader its expression for varying inducer concentration, 2 h after induction. From Supplementary Figure S5, maximum induction is achieved for 1% arabinose. Here onwards, unless stated otherwise, we use this concentration to assess the kinetics of RNA production under the control of P_{BAD}.

First RNA and intervals between consecutive RNA molecules in individual cells

From the time-lapse images acquired with confocal microscopy, after induction, we measure in each cell both t_0 and subsequent values of Δt. t_0 is expected to include the time for arabinose to enter the cell via the intake mechanism, the time to find the promoter and release the repressor and also the time for the recruitment of the RNA polymerase and subsequent production of the first target RNA. The latter process includes events such as the closed and the open complex formation at the promoter region, as well as the elongation time. Both the elongation time and the time for MS2d-GFP to bind to a target RNA are expected to be negligible in comparison with the duration of the intake and of transcription initiation (8,12,31). Meanwhile, Δt should depend only on the events in transcription initiation (37).

The distribution of values of the waiting times, t_0, is shown in Figure 2A. Cells were induced in the gel with fresh media and 1% arabinose. The distribution is broad, as the waiting times spread through the measurement time and has a mean of 3071 s.

The distribution of intervals between consecutive productions of target RNA molecules (Δt) is shown in Figure 2B. This production is a sub-Poissonian process, as the normalized variance (σ^2/μ^2) of the distribution is 0.37. Similar conclusions were obtained from measurements of the *in vivo* kinetics of RNA production under the control of $P_{lac/ara\text{-}1}$ and P_{tetA} (9,10).

The distributions in Figure 2A and B differ significantly. We verified this with a statistical testing of equality of two empirical distributions, the Kolmogorov–Smirnov (K–S) test. We obtained a *P*-value of 2.8×10^{-18}, much smaller than 0.05, which allows rejecting the null hypothesis of similarity. We conclude that in the case of P_{BAD} and the arabinose intake mechanism, the time of intake of inducers affects significantly both mean and standard deviation of RNA numbers in individual cells, long after induction. Finally, note that the difference between the distributions of t_0 and Δt provides evidence that the activity of P_{BAD} changes significantly with induction. Otherwise, these two distributions should not differ significantly, as they would both result, e.g. from spurious transcription events alone.

One recent study (12) also focuses on the *in vivo* induction kinetics of P_{BAD}. This study uses measurements of GFP levels in cell populations, whose expression is controlled by P_{BAD} (inserted into a medium-copy vector) and a model to extrapolate the mean activation time of the promoter, after induction. Assuming a threshold for GFP levels to consider the promoter as active, the mean appearance time of GFP after induction was ~960 s. By considering several features of the measurement system, including the mean maturation time of GFP, a value was then extrapolated for the expected activation time of the promoter, namely, ~250 s. This does not include the time for transcription to be completed, once the closed complex is formed. This study thus predicts a faster mean initiation time than what our direct measurements indicate (~3000 s). Two main reasons exist for this difference. First, in the mutant used previously (12), the chromosomal *araBAD* operon is deleted, avoiding the negative feedback mechanism, which likely fastens the response time significantly. Additionally, gene expression was assessed from a medium-copy vector, which should respond much faster than the single-copy vector system used here, as its response time depends on the fastest of the response times of several promoter copies. Thus, we find that the results reported previously (12) and ours are in agreement. For example, while observing mean waiting times one order of magnitude longer, we do observe RNA molecules appearing in some of the cells within a time scale of 200–400 s after induction. Therefore, provided the usage of a multi-copy vector instead of the single-copy vector used here, we expect mean waiting times one order of magnitude smaller and thus in agreement with the measurements described previously (12).

Correlations between consecutive processes

To study whether the durations of the processes of intake and of transcription initiation are correlated, we first assessed whether consecutive intervals of Δt in individual cells are correlated. We measured the Pearson correlation from 101 pairs of consecutive intervals, and found it to be 0.16. We obtained a *P*-value of 0.11, assuming no correlation as the null hypothesis, which implies that we cannot prove that the correlation is significant. This is in agreement with previous studies of $P_{lac/ara\text{-}1}$ kinetics, which also

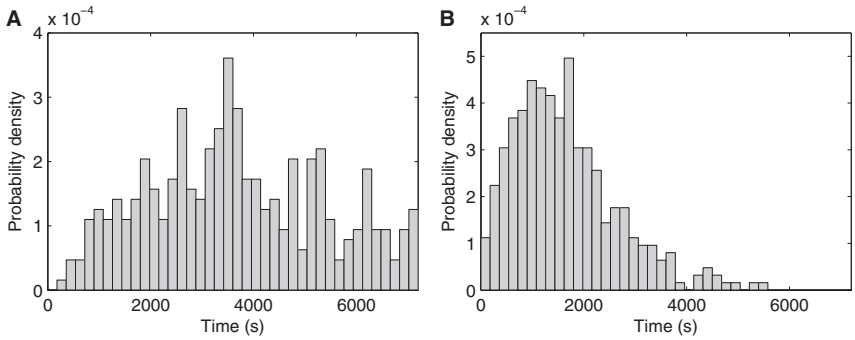


Figure 2. Kinetics of the intake and production. (**A**) Probability density distribution of waiting times ($\mu = 3071$ s, $\sigma = 1711$ s) for the first RNA to be produced in cells induced by 1% L-arabinose (354 data points). (**B**) Probability density distribution of intervals between transcription events for P_{BAD} when induced by 1% L-arabinose ($\mu = 1672$ s, $\sigma = 1012$ s) (347 data points).

indicate inexistence of correlation between durations of consecutive intervals between RNA productions (8).

We next assessed whether the distributions of t_0 and values of Δt (Figure 2A and B) are correlated. Note that t_0 and the Δt are of similar order of magnitude as the measurement period. This introduces artificial correlations between t_0 and Δt of individual cells, as, e.g. a cell with a large t_0 is expected to exhibit smaller than average Δt values, as larger intervals would not be detected during the measurement period as likely as in cells with smaller values of t_0. To remove these artificial correlations between t_0 and Δt of individual cells, in this assessment, we only considered RNA productions for a certain window size (Supplementary Methods and Supplementary Figure S6). This window is set so as to maximize the number of data points that can be extracted from the measurements.

From the windowed data, we calculated the Pearson correlation between t_0 and Δt values in individual cells to be -0.15. We calculated a *P*-value of 0.18 assuming no correlation as the null hypothesis, which implies that we cannot prove that the correlation is significant. This result is in line with (12), which reports a lack of correlation between initiation of protein expression and subsequent rate of protein synthesis in individual cells.

Dynamics of induction and of transcription initiation under different induction schemes

The distinctiveness of the distributions of t_0 and Δt of P_{BAD}, as assessed by the K–S test, suggests that they are, partially, the result of different processes. Although t_0 ought to depend on the kinetics of intake of arabinose and on the first transcription initiation event, Δt values ought to depend mostly on the kinetics of transcription initiation events alone.

These assumptions arise from the following. First, *in vitro* and *in vivo* measurements (26,38) suggest that transcription initiation (including closed and open complex formation) is a long-duration, multi-step process, usually taking 10^2–10^3 s in bacterial promoters (10,25,26,37,38). Other events that need to occur before the appearance of a target RNA because of the tagging of the MS2d-GFP are not expected to affect Δt significantly. These are transcription elongation and the tagging by multiple MS2d-GFP. Elongation of the target RNA was measured to take only tens of seconds (31). Also, the tagging occurs at a rate that makes the RNA visible during elongation or shortly after (31).

To test the two assumptions, we measured the distributions of t_0 and Δt for another promoter, $P_{lac/ara\text{-}1}$, in two conditions. $P_{lac/ara\text{-}1}$ can be induced either by IPTG or by arabinose (as P_{BAD}), or by both inducers simultaneously (9). According to our assumption, the distribution of t_0 of P_{BAD} is expected to be similar to that of $P_{lac/ara\text{-}1}$ when the latter is induced by arabinose, because of depending on the same intake mechanism, whereas it should differ significantly when $P_{lac/ara\text{-}1}$ is induced by IPTG, given the different intake mechanisms of IPTG.

We measured the distributions of t_0 and Δt for $P_{lac/ara\text{-}1}$ when induced by IPTG alone and when induced by arabinose alone (Table 1). We used the same concentration of arabinose as when inducing P_{BAD}. The IPTG concentration used is the one required for maximum induction of $P_{lac/ara\text{-}1}$ (33). Results in Table 1 follow the windowing procedure described earlier in the text. The table shows mean, standard deviation and square of the coefficient of variation (μ^2/σ^2) of t_0 and of Δt for the two promoters, each of which in two induction schemes.

We first assessed the distinctiveness of the distributions of t_0 and Δt by the K–S test, for each promoter in each condition (Table 2). In all cases, these two distributions differ in a statistical sense. This is in agreement with the assumption that although both Δt and t_0 depend on the kinetics of initiation at the promoter, only t_0 depends on the kinetics of intake of inducers.

We next performed statistical tests to assess the distinctiveness between the induction kinetics (t_0) of the two promoters (Table 3), when subject to the same inducer and when subject to different inducers. Also, we compared the effects of a different inducer concentration in the case of P_{BAD}. From Table 3, when P_{BAD} and $P_{lac/ara\text{-}1}$ are induced with 1% arabinose, they exhibit distributions of t_0 that cannot be distinguished. However, when $P_{lac/ara\text{-}1}$ is

Table 1. Measurements of t_0 and Δt

Promoter	Inducer	No. of samples (Δt)	$\mu_{\Delta t}$ (s)	$\sigma_{\Delta t}$ (s)	σ^2/μ^2	No. samples (t_0)	μ_{t0} (s)	σ_{t0} (s)	σ^2/μ^2
P_{BAD}	1% arabinose	102	1440.6	532.8	0.14	84	2885.0	1159.8	0.16
P_{BAD}	0.1% arabinose	78	1475.4	481.2	0.11	70	3519.4	1236.2	0.12
$P_{lac/ara\text{-}1}$	1% arabinose	149	1516.5	516.0	0.12	125	2832.5	1184.6	0.17
$P_{lac/ara\text{-}1}$	1 mM IPTG	485	1314.4	576.0	0.19	286	2697.0	913.6	0.11

The table shows the mean (μ), the standard deviation (σ) and the normalized variance (σ^2/μ^2) of the measured distributions of t_0 and Δt.

Table 2. *P*-values of the Kolmogorov–Smirnov test between t_0 and Δt distributions for each promoter and induction condition

Promoter	Inducer	*P*-value
P_{BAD}	1% arabinose	2.83×10^{-18}
P_{BAD}	0.1% arabinose	4.06×10^{-21}
$P_{lac/ara\text{-}1}$	1% arabinose	2.48×10^{-26}
$P_{lac/ara\text{-}1}$	1 mM IPTG	3.32×10^{-72}

For $P < 0.05$, it is generally accepted that the hypothesis that the two distributions are the same should be rejected.

Table 3. *P*-values of the Kolmogorov–Smirnov test between t_0 distributions for each promoter and induction condition

	P_{BAD} 1% arab	P_{BAD} 0.1% arab	$P_{lac/ara\text{-}1}$ 1% arab	$P_{lac/ara\text{-}1}$ IPTG
P_{BAD} 1% arab	1			
P_{BAD} 0.1% arab	5.93×10^{-4}	1		
$P_{lac/ara\text{-}1}$ 1% arab	0.8533	1.10×10^{-4}	1	
$P_{lac/ara\text{-}1}$ IPTG	0.0126	4.49×10^{-12}	0.0049	1

For $P < 0.05$, it is generally accepted that the hypothesis that the two distributions are the same should be rejected.

induced with IPTG, the resulting t_0 distribution is statistically distinguishable from that of P_{BAD}, when induced by either 0.1 or 1% arabinose. It is also distinct from its own t_0 distribution when induced by 1% arabinose. This statistically significant difference supports the hypothesis that the distributions of t_0 are dependent on the kinetics of the intake system of the inducers, and that these differ for IPTG and arabinose.

Finally, we observed that the distributions of t_0 of P_{BAD}, when induced by 0.1% and by 1% arabinose, are distinct. This is expected as the time for inducers to 'first reach the promoter' ought to depend on the inducer's concentration.

Kinetics of the intake process

The intake time of an inducer, here named 't_{diff}', differs from t_0 in that it does not include the time for the first transcription initiation event to occur. Because of this, t_{diff} cannot be measured directly with the MS2-GFP–tagging method. We thus estimate the mean and variance of the distribution of values of t_{diff} by subtracting the means and variances of the Δt distribution from the t_0 distribution. This method is based on the fact that we were unable to establish the existence of a correlation between the values of t_0 and Δt. Given this, and as they are, at most, weakly correlated (Pearson correlation of −0.15), we assume that they are independent so as to be able to estimate the standard deviation of the duration of the intake process alone (note that the mean of this quantity can be estimated as described later in the text, regardless of the existence of dependence).

The estimated mean and a standard deviation of t_{diff} are similar for P_{BAD} and for $P_{lac/ara\text{-}1}$, when induced with 1% arabinose. Namely, in both cases, we obtained a mean of ∼1400 s and a standard deviation of ∼1100 s. This is expected, given the usage of the same intake mechanism and inducer concentration. Importantly, when $P_{lac/ara\text{-}1}$ is induced by IPTG, the standard deviation of t_{diff} is much smaller (∼700 s), whereas the mean is similar to when induced by arabinose (∼1400 s). This suggests that the intake of arabinose is a noisier process (concerning the uncertainty of the intake time) than the intake of IPTG. Finally, we find that in the case of P_{BAD}, the concentration of arabinose affects the mean of t_{diff} significantly, as it equals ∼2000 s for 0.1% arabinose.

Effect of the intake process on the temporal cell-to-cell diversity in RNA numbers

Because of being stochastic and thus variable in duration from one event to the next (i.e. it differs from one cell to the next), the intake process impacts on the diversity in RNA numbers of a cell population. This impact should decrease with time, after induction. We estimated the time during which the effect is tangible for each measurement condition. For this, we assume that values of t_0 depend mostly on the intake of arabinose and on the first transcription initiation event at the start site of P_{BAD}. Meanwhile, the distribution of intervals between consecutive RNAs is assumed to depend solely on the kinetics of transcription initiation (8,10,37).

The events determining Δt as well as t_0 are modelled as *d*-step processes, each step with an exponentially distributed duration (Supplementary Methods) (37). From this assumption, it is possible, for a given number of steps, to find the duration of each step that best fits the measurements. We assume transcription initiation to be a three-step process, namely, the closed complex formation, the open complex formation and promoter escape (27,38), as evidence suggests that these are the most rate-limiting steps in normal conditions, i.e. the ones most contributing to the intervals between production of consecutive RNA molecules (26). This assumption also relies on recent

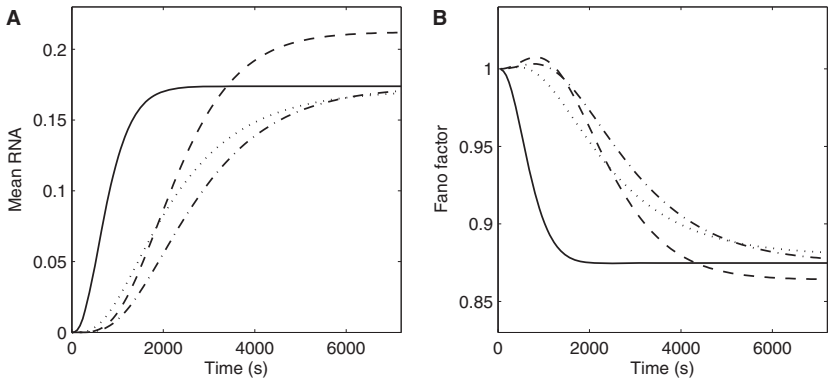


Figure 3. Mean and Fano factor of transient times for different models of intake and subsequent RNA production kinetics. Mean (**A**) and Fano factor (**B**) of RNA numbers as obtained by CME models of activation and expression. The models shown are that of $P_{lac/ara\text{-}1}$ with 1 mM IPTG (dashed line), $P_{lac/ara\text{-}1}$ with 1% arabinose (dotted line), P_{BAD} with 1% arabinose (dash-dotted line) and P_{BAD} with 1% arabinose and infinitely fast intake (solid line).

studies (37) that indicate that assuming this number of steps suffices to generate distributions that cannot be distinguished, in a statistical sense, from measurements with accuracy and quantity of data similar to the measurements reported here. Finally, we assume the intake to be a two-step process, namely, the binding of extracellular arabinose to an uptake protein and, once bound, its translocation to the cytoplasm (12). The combination of the two processes (intake followed by transcription initiation) is, consequently, assumed to be a five-step process.

Assuming these numbers of steps and stable conditions (e.g. induction level), we searched for models that fit the distributions accurately enough so that the K–S test does not find differences between model and measurements. The *P*-values of these tests are shown in Supplementary Table S1 and show that in all but one case, it is possible to find a model that cannot be distinguished from the empirical distribution, in a statistical sense.

The case for which we could not find a model that fits the measurements is that of P_{BAD} at 0.1% arabinose induction. This may be due to lack of sufficient data or because the model is unsuitable. Future studies are required to assert this. One explanation may be that, in this case, the distribution of intake times results from two distinct kinetics, one being the productions under induction and the other being spurious productions by promoters in the 'non-induced' state.

Given the models aforementioned and provided a rate of RNA degradation, it is possible to estimate the time it takes for the mean RNA numbers of a model cell population to reach equilibrium, as this time depends solely on the rate of degradation of RNAs and t_0. We do not have measurements of the degradation rate of the target RNA, as the tagging with MS2d-GFP 'immortalizes' it for the duration of the measurements (32). Instead, the models in Figure 3 assume an RNA degradation rate of $5\,min^{-1}$, which is within realistic intervals for *E. coli* (1).

From all of the aforementioned data, we estimated the mean times for RNA numbers to reach near-equilibrium, as well as the Fano factor of this quantity since the start of the simulations. Results are shown in Figure 3, as estimated for each of the models. Also shown is an estimation that assumes the model of transcription initiation of P_{BAD} when induced by 1% arabinose, coupled with an infinitely fast intake.

In all cases, reaching equilibrium in mean RNA numbers takes >1 h, except when assuming infinitely fast intake, in which case the time to reach equilibrium is <0.5 h. Thus, for a time length as long as 1–2 h, the intake process has a non-negligible contribution on the mean and the on the cell-to-cell diversity in RNA numbers of the cell populations. From Figure 3A, one also observes different shapes in the curves of $P_{lac/ara\text{-}1}$ when induced by IPTG (dashed line) and when induced by arabinose (dotted line), because of differing intake kinetics.

From Figure 3B, the contribution of the intake kinetics on the cell-to-cell variability in RNA numbers is also significant. For example, the kinetics of intake causes an increase in the Fano factor in the initial moments not observable in the case of infinitely fast intake.

We also tested models of P_{BAD} induced by 1% arabinose (normal and infinitely fast intake) with other RNA degradation rates (Supplementary Figure S7), within realistic intervals (1). Aside from assessing the degree of dependency on the intake time and degradation rate, one also observes from the figure that although the latter determines the rate at which the system reaches equilibrium, the former acts as a delay towards reaching the numbers at equilibrium. Further, one can see that the intake step adds diversity to the RNA numbers in the cells, during the transient to reach equilibrium.

DISCUSSION

We measured, at the single-cell level, how long it takes for the first RNA under the control of P_{BAD} to be produced, followed the introduction of the inducer in the media. Also, we measured the subsequent intervals between

consecutive RNA productions. From the intervals between transcription events, we determined that RNA production under the control of P_{BAD} is a sub-Poissonian process. Two recent studies reached similar conclusions for $P_{lac/ara-1}$ and P_{tetA}, for all induction conditions tested (9,10). We hypothesize that this may be a common phenomenon because of the kinetic properties of the process of transcription initiation in bacteria, in particular, because of its multi-stepped nature.

From the distributions of the time, it takes for the appearance of the first RNA in each cell when under the control of P_{BAD} and of $P_{lac/ara-1}$, for different induction conditions, we assessed the effect of the kinetics of the intake process on the mean and cell-to-cell diversity in RNA numbers of cell populations. Relevantly, this effect was found to be tangible for a long period after induction. Also, we verified that different intake mechanisms differ significantly not only in mean but also in the degree of variability of the intake time, and that this has a non-negligible effect on RNA population statistics.

Given the aforementioned data, and considering that natural environments are fluctuating, we expect the kinetics of cellular intake mechanisms to have a significant effect on the degree of phenotypic diversity of cell populations. Finally, we expect the methodology used here to assess the *in vivo* kinetics of intake of arabinose and of IPTG to be applicable to any gene of interest. Such studies should provide valuable insight into the adaptability of prokaryotic organisms to environmental changes and stress. They should also provide a better understanding of the observed cell-to-cell phenotypic diversity in *E. coli* when in fluctuating environments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–7, Supplementary Methods and Supplementary References [39–43].

FUNDING

The Academy of Finland [257603 to A.S.R.]; the Finnish Funding Agency for Technology and Innovation [40226/12 to O.Y.-H.]; the Foundation for Science and Technology, Portugal [PTDC/BBB-MET/1084/2012 to A.S.R.]. Funding for open access charge: Academy of Finland [257603, 2012 to A.S.R.].

Conflict of interest statement. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Bernstein,J.A., Khodursky,A.B., Lin-Chao,S. and Cohen,S.N. (2002) 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*, **99**, 9697–9702.
2. Yu,J., Xiao,J., Ren,X., Lao,K. and Xie,X.S. (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.
3. Taniguchi,Y., Choi,P.J., Li,G.-W., Chen,H., Babu,M., Hearn,J., Emili,A. and Xie,X.S. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.
4. Weickert,M.J. and Adhya,S. (1993) The galactose regulon of *Escherichia coli*. *Mol. Microbiol.*, **10**, 245–251.
5. Skerra,A. (1994) Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in *Escherichia coli*. *Gene*, **151**, 131–135.
6. Schleif,R. (2000) Regulation of the L-arabinose operon of *Escherichia coli*. *Trends. Genet.*, **16**, 559–565.
7. Choi,P., Cai,L., Frieda,K. and Xie,X. (2008) A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, **322**, 442–446.
8. Kandhavelu,M. and Häkkinen,A. (2012) Single-molecule dynamics of transcription of the lar promoter. *Phys. Biol.*, **9**, 026004.
9. Kandhavelu,M., Lloyd-Price,J., Gupta,A., Muthukrishnan,A.B., Yli-Harja,O. and Ribeiro,A.S. (2012) Regulation of mean and noise of the *in vivo* kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett.*, **586**, 3870–3875.
10. Muthukrishnan,A.B., Kandhavelu,M., Lloyd-Price,J., Kudasov,F., Chowdhury,S., Yli-Harja,O. and Ribeiro,A.S. (2012) Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucleic Acids Res.*, **40**, 8472–8483.
11. Schleif,R. (2010) AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiol. Rev.*, **34**, 779–796.
12. Megerle,J.A., Fritz,G., Gerland,U., Jung,K. and Rädler,J.O. (2008) Timing and dynamics of single cell Gene expression in the arabinose utilization system. *Biophys. J.*, **95**, 2103–2115.
13. Hogg,R.W. and Englesberg,E. (1969) L-arabinose binding protein from *Escherichia coli* B/r. *J. Bacteriol.*, **100**, 423–432.
14. Schleif,R. (1969) An l-arabinose binding protein and arabinose permeation in *Escherichia coli*. *J. Mol. Biol.*, **46**, 185–196.
15. Horazdovsky,B.F. and Hogg,R.W. (1989) Genetic reconstitution of the high-affinity L-arabinose transport system. *J. Bacteriol.*, **171**, 3053–3059.
16. Lee,J.H., Al-Zarban,S. and Wilcox,G. (1981) Genetic characterization of the araE gene in Salmonella typhimurium LT2. *J. Bacteriol.*, **146**, 298–304.
17. Stoner,C. and Schleif,R. (1983) The araE low affinity L-arabinose transport promoter. Cloning, sequence, transcription start site and DNA binding sites of regulatory proteins. *J. Mol. Biol.*, **171**, 369–381.
18. Siegele,D.A. and Hu,J.C. (1997) Gene expression from plasmids containing the araBAD promoter at subsaturating inducer concentrations represents mixed populations. *Proc. Natl Acad. Sci. USA*, **94**, 8168–8172.
19. Khlebnikov,A., Risa,O., Skaug,T., Carrier,T.A. and Keasling,J.D. (2000) Regulatable arabinose-inducible gene expression system with consistent control in all cells of a culture. *J. Bacteriol.*, **182**, 7029–7034.
20. Khlebnikov,A., Datsenko,K.A., Skaug,T., Wanner,B.L. and Keasling,J.D. (2001) Homogeneous expression of the P-BAD promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology*, **147**, 3241–3247.
21. Morgan-Kiss,R.M., Wadler,C. and Cronan,J.E. (2002) Long-term and homogeneous regulation of the *Escherichia coli* araBAD promoter by use of a lactose transporter of relaxed specificity. *Proc. Natl Acad. Sci. USA*, **99**, 7373–7377.
22. Johnson,C.M. and Schleif,R.F. (1995) In vivo induction kinetics of the arabinose promoters in *Escherichia coli*. *J. Bacteriol.*, **177**, 3438–3442.
23. Walter,G., Zillig,W., Palm,P. and Fuchs,E. (1967) Initiation of DNA-Dependent RNA synthesis and the effect of heparin on RNA polymerase. *Eur. J. Biochem.*, **3**, 194–201.
24. Chamberlin,M.J. (1974) The selectivity of transcription. *Annu. Rev. Biochem.*, **43**, 721–775.
25. Buc,H. and McClure,W.R. (1985) Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac

UV5 promoter. Evidence for a sequential mechanism involving three steps. *Biochemistry*, **24**, 2712–2723.
26. Lutz,R., Lozinski,T., Ellinger,T. and Bujard,H. (2001) Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Res.*, **29**, 3873–3881.
27. Hsu,L.M. (2002) Promoter clearance and escape in prokaryotes. *Biochim. Biophys. Acta*, **1577**, 191–207.
28. DeHaseth,P.L., Zupancic,M.L. and Record,M.T. (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.*, **180**, 3019–3025.
29. Greive,S.J. and Von Hippel,P.H. (2005) Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.*, **6**, 221–232.
30. McClure,W.R. (1985) Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, **54**, 171–204.
31. Golding,I. and Cox,E.C. (2004) RNA dynamics in live *Escherichia coli* cells. *Proc. Natl Acad. Sci. USA*, **101**, 11310–11315.
32. Golding,I., Paulsson,J., Zawilski,S.M. and Cox,E.C. (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.
33. Lutz,R. and Bujard,H. (1997) Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.*, **25**, 1203–1210.
34. Chowdhury,S., Kandhavelu,M., Yli-Harja,O. and Ribeiro,A.S. (2012) An interacting multiple model filter-based autofocus strategy for confocal time-lapse microscopy. *J. Microscopy*, **245**, 265–275.
35. Chen,T.B., Lu,H.H., Lee,Y.S. and Lan,H.J. (2008) Segmentation of cDNA microarray images by kernel density estimation. *J. Biomed. Inform.*, **41**, 1021–1027.
36. Otsu,N. (1979) A threshold selection method from gray-level histograms. *IEEE Trans. Sys. Man Cybern.*, **9**, 62–66.
37. Kandhavelu,M., Mannerström,H., Gupta,A., Häkkinen,A., Lloyd-Price,J., Yli-Harja,O. and Ribeiro,A.S. (2011) *In vivo* kinetics of transcription initiation of the lar promoter in *Escherichia coli*. Evidence for a sequential mechanism with two rate-limiting steps. *BMC Syst. Biol.*, **5**, 149.
38. McClure,W.R. (1980) Rate-limiting steps in RNA chain initiation. *Proc. Natl Acad. Sci. USA*, **77**, 5634–5638.
39. Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.
40. Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, **25**, 402–408.
41. Daruwalla,K.R., Paxton,A.T. and Henderson,P.J.F. (1981) Energization of the transport systems for arabinose and comparison with galactose transport in *Escherichia coli*. *Biochem. J.*, **200**, 611–627.
42. Gillespie,D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403–434.
43. Munsky,B. and Khammash,M. (2006) The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, **124**, 044104.

Supplementary Information

In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter

Jarno Mäkelä[1], Meenakshisundaram Kandhavelu[1], Samuel M.D. Oliveira[1], Jerome G. Chandraseelan[1], Jason Lloyd-Price[1], Juha Peltonen[1], Olli Yli-Harja[1,2] and Andre S. Ribeiro[1,*]

[1] Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland.
[2] Institute for Systems Biology, 1441N 34th St, Seattle, WA, 98103-8904, USA

* Corresponding author: andre.ribeiro@tut.fi (Andre S. Ribeiro)

Supplementary Methods

Chemicals

Bacterial cell cultures were grown in two media, namely Luria-Bertani (LB) and M63. The chemical components of LB broth (Tryptone, Yeast extract and NaCl) were purchased from LabM (UK). For M63 media, the following components were used: 2 mM $MgSO_4.7H_20$ (Sigma-Aldrich, USA), 7.6 mM $(NH_4)_2SO_4$ (Sigma Life Science, USA), 30µM $FeSO_4.7H_2O$ (Sigma Life Science, USA), 1 mM EDTA (Sigma Life Science, USA), 60 mM KH_2PO_4 (Sigma Life Science, USA), Glycerol 0.5 % (Sigma Life Science, USA), and Casaminoacids 0.1 % (Fluka Analytical, USA). Isopropyl β-$_D$-1-thiogalactopyranoside (IPTG), $_L$-(+)-Arabinose and anhydrotetracycline (aTc) used for induction of the cells and the antibiotics (100 mg/ml kanamycin and 35 mg/ml chloramphenicol) were purchased from Sigma-Aldrich (USA). Agarose (Sigma Life Science, USA) was used for the microscopic slide gel preparation.

Bacterial Strain

Cloning and expression experiments were performed in *E. coli* DH5α-PRO strain (Clontech; identical to DH5α-Z1 (31). The strain information is: deoR, endA1, gyrA96, hsdR17($r_{k-}m_{k+}$), recA1, relA1, supE44, thi-1, Δ(lacZYA-argF)U169, Φ80δlacZΔM15, F-, λ-, P_{N25}/tetR, P_{lacIq}/lacI, and Sp^R. Frag1A: F-, rha-, thi, gal, $lacZ_{am}$, ΔacrAB::kan^R, P_{N25}/tetR, P_{lacIq}/lacI, and Sp^R. Frag1B: F-, rha-, thi, gal, $lacZ_{am}$, P_{N25}/tetR, P_{laciq}/lacI, and Sp^R. The P_{N25}/tetR, P_{laciq}/lacI, Sp^R cassette was transferred from *DH5αPRO* to Frag1 to generate Frag1B by P1 transduction. The ΔacrAB:kan^R cassette was transferred from KZM120 to Frag1B, so as to generate Frag1A.

Construction of the pMK-BAC vector

To construct the pMK-BAC (P_{BAD}-mRFP1-96 binding site (96 BS) array, the following plasmids were used: a plasmid with mRFP1 plus 96bs array region in the BAC vector, originally designed and generously provided by Prof. Ido Golding ($P_{lac/ara-1}$- mRFP1-96 bs) (32). To amplify the construct containing the AraC and pBAD promoter region from the pGLO vector (Biorad), a primer set was designed as follows:

Ara_AatII-Fw-5´CCTAAGACGTCATCGATGCATAATGTGCC 3´
Ara_AatII-Rv-5´CCTTGATGACGTCATGTATATCTCCTTCTTAAAGTTA3´

The target BAD promoter region along with AraC coding region from the pGLO vector was amplified and inserted into the pIG-BAC vector by standard molecular biology techniques. The construct was verified by sequencing with the appropriate primers and transformed into the *E. coli* DH5α-PRO strain carrying the bacterial expression vector pPROTET.E (Clontech) coding for MS2d–GFP. For more details see Supplementary Figures 1 and 2.

Plate reader experiment

The mean fluorescence of RFP under the control of P_{BAD} was measured with a microplate fluorometer (Fluoroskan Ascent, Thermo Scientific). 200 ml of cells at OD600 ≈ 0.5 were induced with 0.1 % or 1 % L-arabinose and placed on 96 well microplate. From this, cells were measured for 2 hours for relative fluorescence levels of mRFP1 protein (excitation and emission wavelengths were 584 nm and 607 nm, respectively). The cell density was kept identical in all wells of the plate for all conditions.

Quantitative PCR for mean mRNA quantification

The change in the rate of transcription of genes araB and mRFP was studied using qPCR. *E. coli* DH5α-PRO cells containing the constructs were grown as described in the section describing the microscopy measurements. Cells were grown overnight at 30°C with aeration, diluted into fresh medium and allowed to grow at the appropriate temperature of the experiment until an optical density of OD_{600} ≈ 0.3-0.5 was reached. For the experiment, 5 ml of cells were pre-incubated with 100 ng/ml of aTc to induce the expression of MS2d-GFP. 1 % $_L$-arabinose was used for induction of the BAD promoter, 30 minutes after induction, the first sample was taken. From then onwards, samples were taken at an interval of 60 minutes. Rifampicin was added to the samples immediately, so as to prevent further transcription and the cells were fixed with RNA protect reagent immediately followed by enzymatic lysis using Tris-EDTA lysozyme buffer (pH 8.3). RNA was purified from each sample by RNeasy mini-kit (Qiagen). The total RNA was separated by electrophoresis through a 1 % agarose gel and stained with SYBR Safe DNA Gel Stain. The RNA was found intact with discreet bands for 16 S and 23 S ribosomal RNAs. To ensure purity of the RNA samples, they were subject to treatment with DNase free of RNase, to remove residual DNA. The yield of RNA obtained was 0.4 – 0.6 mg/ml. Approximately 40 ng of RNA was used for cDNA synthesis using iSCRIPT reverse transcription super mix (Biorad) according to the manufacturer's instructions.

Quantification of cDNA was performed by real-time PCR using SYBR-green supermix with primers for the amplification of target and reference genes at a concentration of 200nM. Primers specific to AraB (Forward: 5' GGTACTTCCACCTGCGACAT 3', Reverse: 5' CAACCTGACCGCAAATACCT 3') and mRFP genes (Forward: 5' TACGAC GCCGAGGTCAAG 3' and Reverse: 5' TTGTGGGAGGTGATGTCCA 3') were designed using PRIMER3 (39), the length of the amplicon for the target and reference were maintained at 90bp. The sequence of the primers for the reference gene 16S rRNA (EcoCyc Accession Number: EG30090) (Forward: 5' CGTCAGCTCGTGTTGTGAA 3' and Reverse: 5' GGACCGCTGGCAACAAAG 3') and the primers were obtained from Thermo Scientific. The level of 16s rRNA was used to normalize the expression data of each target gene. 10 ng of cDNA was used as a template. The cycling protocol used was 94 °C for 15 s, 51 °C for 30 s, and 72 °C for 30 s, up to 39 cycles. The amplification was monitored in real time by measuring the fluorescence intensities at the end of each cycle. The experiment was performed in triplicates along with the No-RT and no template controls. The volume used for each reaction was 25 µl in low-profile tube strips in a MiniOpticon Real time PCR system (Biorad). The Cq values were obtained from the CFX ManagerTM Software and the fold change of expression of the target gene was analysed by normalizing against the reference gene according to the Livak method (40). See Supplementary Figure 3 for the results.

Normalization between samples of the distributions of time intervals

The observation time for the production of RNAs is two hours. In some cells, the intervals between transcription events (Δt) are of this order of magnitude. This causes shorter intervals to be 'favored'. This is more likely to occur in cells where the waiting time for the first RNA to be produced (t_0) is longer, since the remaining observation time is shorter. This introduces an artificial anti-correlation between t0 and Δt in individual cells. Similar correlations are introduced by different division times as well, i.e., shorter division times hamper the collection of longer Δt samples.

Thus, prior to determine if any real correlation exists between t_0 and Δt in individual cells, it is necessary to remove these artificial sources of anti-correlation due to the limits in the measurement period. For this, in all cells, all intervals between consecutive RNAs were collected only for a time window of size t_c after the previous production. The value of t_c is identical in all cells. This causes the probability of appearance of the next RNA molecule during that period to be uniform for all cells, if the underlying process is in fact identical in all cells.

This restriction in the collection of values of Δt is made when assessing correlations between t_0 and Δt and when comparing these two distributions between conditions. When imposing the restriction, we thus consider only cells that produce at least 2 RNA molecules during their life time and measurement period. The value of t_c was selected so as to maximize the number of data points collectable from the data sets. Here, t_c was set to 39 minutes (see Supplementary Figure S6).

Fitting the empirical distributions to a sum of *d*-exponential variates

The arabinose intake mechanism can be described by a single Michaelis-Menten function (41). Since the backward reaction of the intake process is slower than the forward reaction (12), the intake process is modeled, roughly, by a sequence of non-reversible reactions. Interestingly, we found from the measurements and the inference procedure, evidence of two steps at this stage (exponential in duration), which is in agreement with the number of forward steps assumed in other studies for this process (12). Finally, transcription initiation, which follows the intake process, can also be modeled by a 3-step exponential model according recent in vivo measurements (9, 10). Thus, we fit the measured distributions of t_0 to a 5-step exponential model.

To fit the empirical distribution with a sum of *d*-exponential variates (of possibly unequal rates), we select the exponential rate parameters $\lambda_1, \ldots, \lambda_d$ such that the Kolmogorov-Smirnov (K-S) statistic is minimized. That is, parameters are selected as $\hat{\theta} = \arg\max_{\theta=\lambda_1,\ldots,\lambda_d} \sup_x |F_\theta(x) - G(x)|$, where $F_\theta(x)$ is the cumulative distribution function (CDF) of a sum of d exponentials with parameters $\theta = (\lambda_1, \ldots, \lambda_d)$, and $G(x)$ is the CDF of the empirical distribution.

$$F_{\theta=L_1,\ldots,L_d}(X) := \sum_{i=1}^{d} \left((1 - e^{-L_i x}) \prod_{\substack{j=1 \\ j \neq i}}^{d} \frac{L_j}{L_j - L_i} \right)$$

The parameter values θ are found using a nonlinear numerical optimizer. This method is convenient, since if the K-S test was rejected for the parameters $\hat{\theta}$, such a test would also be rejected for any other set of parameters θ in this family of fitted distributions, indicating that these distributions are inappropriate models of the data. The results of the fitting are shown in Table S1.

As a final note, the model assumed above can be considered as the simplest possible, i.e., each step is an elementary reaction of the form $A \xrightarrow{c} B$, with a constant probability of occurring per unit time. This entails that the distributions of intervals between steps are exponential (42). Notably, the inferred distributions and the experimental data are statistically indistinguishable by the K-S test, which implies that there is no evidence to assume that the model is wrong (see Table S1).

CME solution

To estimate the effect of the intake on the cell-to-cell diversity in RNA numbers we made use of direct integration of the Chemical Master Equation (CME) of the model described in the previous section, using the Finite State Projection algorithm (43). This method truncates the infinite state space of the CME such that the amount of probability outside the truncated region is negligible. In all cases, we truncated the state space at 20 RNA molecules. This number sufficed for this space to contain virtually all of the total probability in the system. The probability mass vector at each time moment is then solved by numerically integrating the truncated CME. From this distribution over time, we calculate mean, variance, and Fano factor of RNA molecules of a model at each moment.

Supplementary Figures

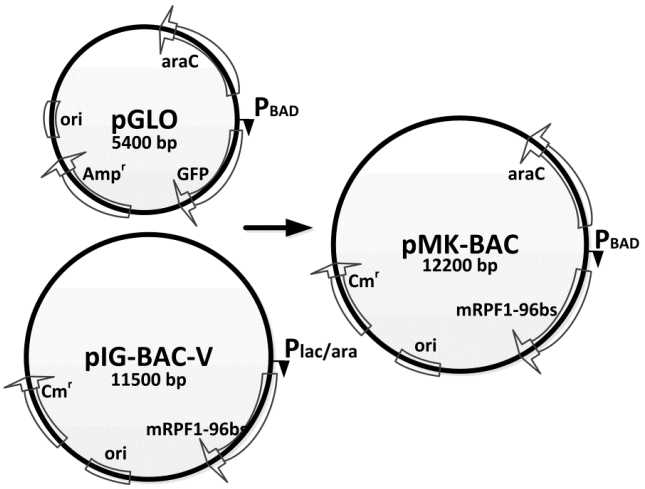


Figure S1. Plasmids used for the pMK-BAC construction. The pMK-BAC(P_{BAD}-mRFP1-96bs) plasmid was engineered by linking the amplified region, containing the P_{BAD} promoter and the araC gene, obtained from pGLO, to the pIG-BAC expression vector, without the lac/ara-1 promoter, obtained from pIG-BAC($P_{lac/ara-1}$- mRFP1-96 bs)-V.

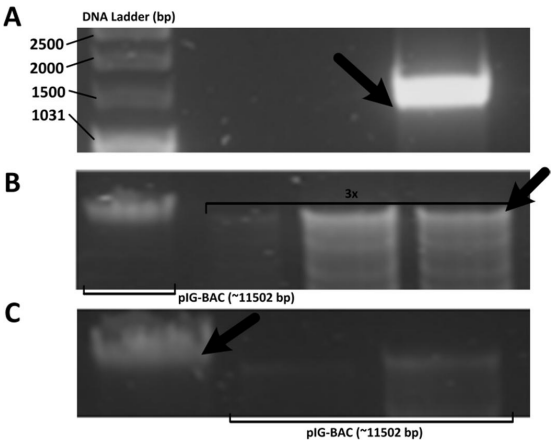


Figure S2. Split gels of the plasmid construction. (A) The PCR fragment of 1347bp amplified from pGLO with the appropriate primers. (B) Lanes containing pIG-BAC-V without the $P_{lac/ara-1}$ promoter region (10849bp), and the pIG-BAC-V expression plasmid (11502bp. (C) The pMK-BAC plasmid (12196bp) containing the araC-P_{BAD} amplified fragment inserted to the BAC expression vector, and the pIG-BAC-V (11502bp). Note the black arrows indicating the bands.

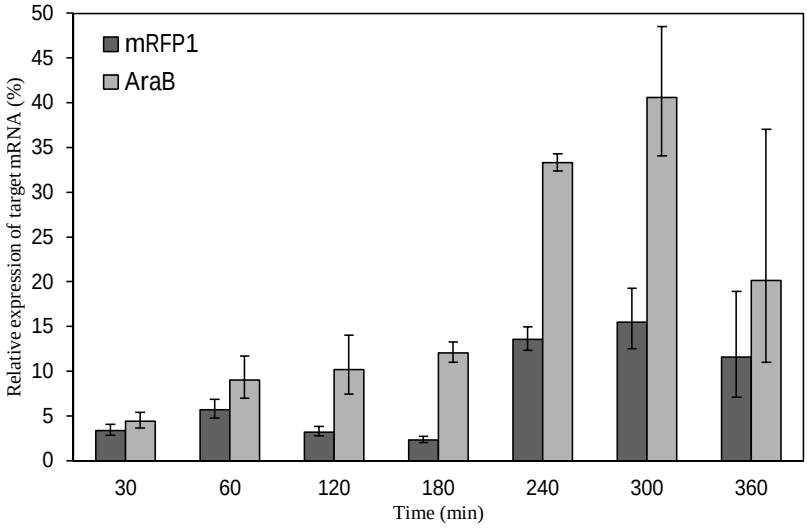


Figure S3. Q-PCR of the native and of the target gene. Q-PCR of RNA expression of the native, integrated AraB gene and of the mRFP1 probe in the F-plasmid, as a function of time, when subject to induction by 1% L-arabinose in liquid culture. The standard deviation bars are from three independent experiments.

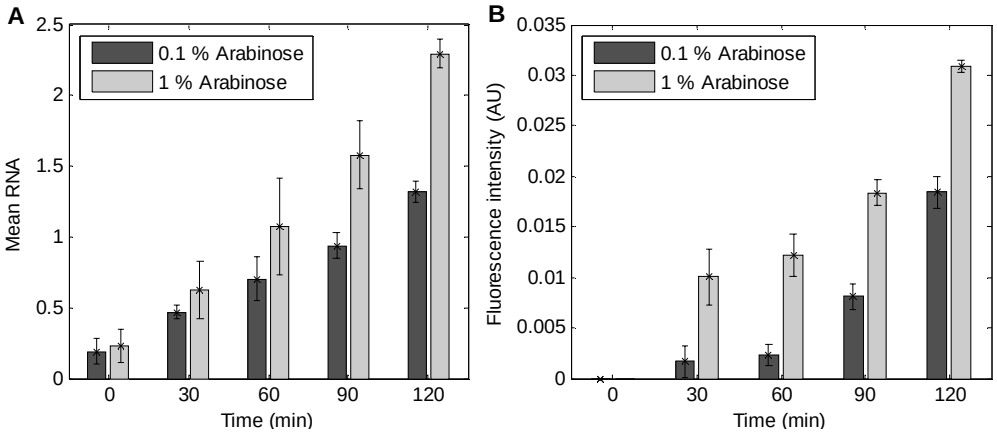


Figure S4. MS2-GFP measurement of RNA numbers compared with Plate reader results. (A) RNA numbers over time measured in vivo with the MS2-GFP method for 0.1 % and 1 % L-arabinose. Mean and standard deviation of RNA numbers in individual cells were calculated for each sample separately. Error bars show the standard error of the mean from independent measurements (3 measurements) (B) Fluorescent intensity of RFP over time for 0.1 % and 1 % L-arabinose as measured by Plate reader. Error bars show the standard error of the mean obtained from 8 wells.

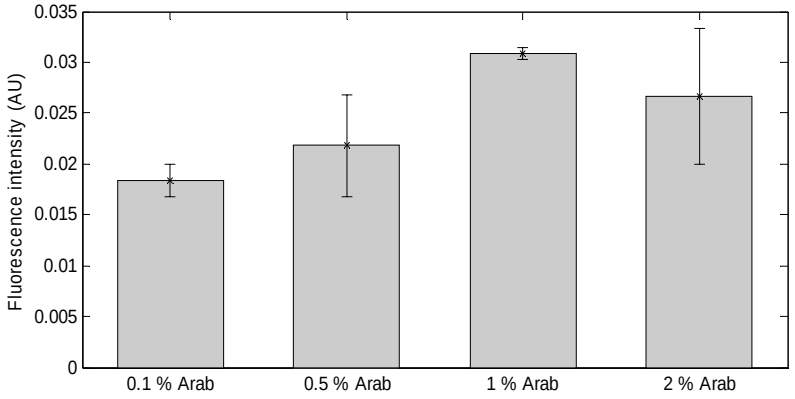


Figure S5. Gene expression as measured by Plate Reader. Comparison of different inducer concentrations by plate reader measurements, 2 hours following induction. Maximum induction is achieved with 1 % L-arabinose. Error bars show the standard error of the mean obtained from 8 wells.

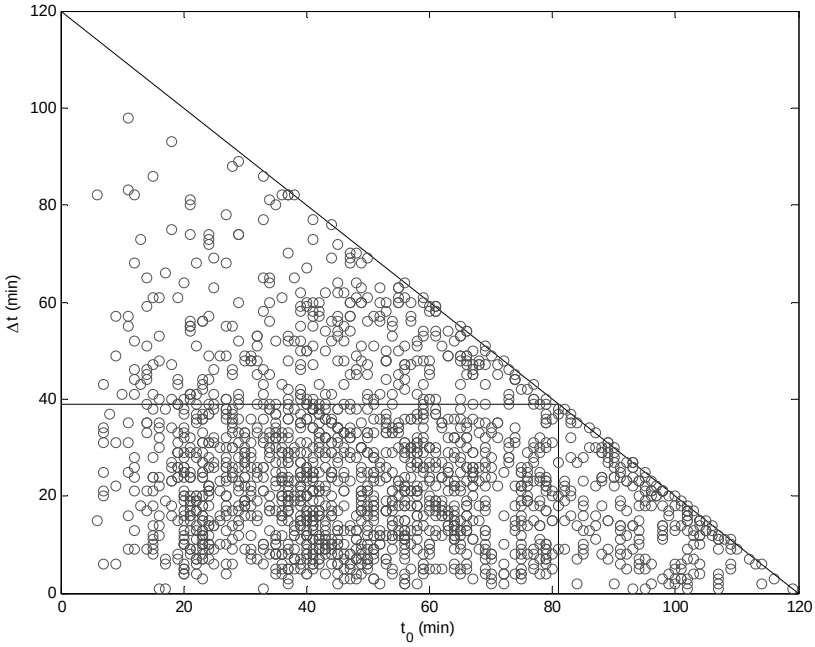


Figure S6. Normalization of the data. The values of t_0 and the corresponding values of the first Δt in each cell. The diagonal line is the total observation time (120 min). Vertical and horizontal (t_c = 39 min) lines define the intervals that meet the requirements for un-biasedness.

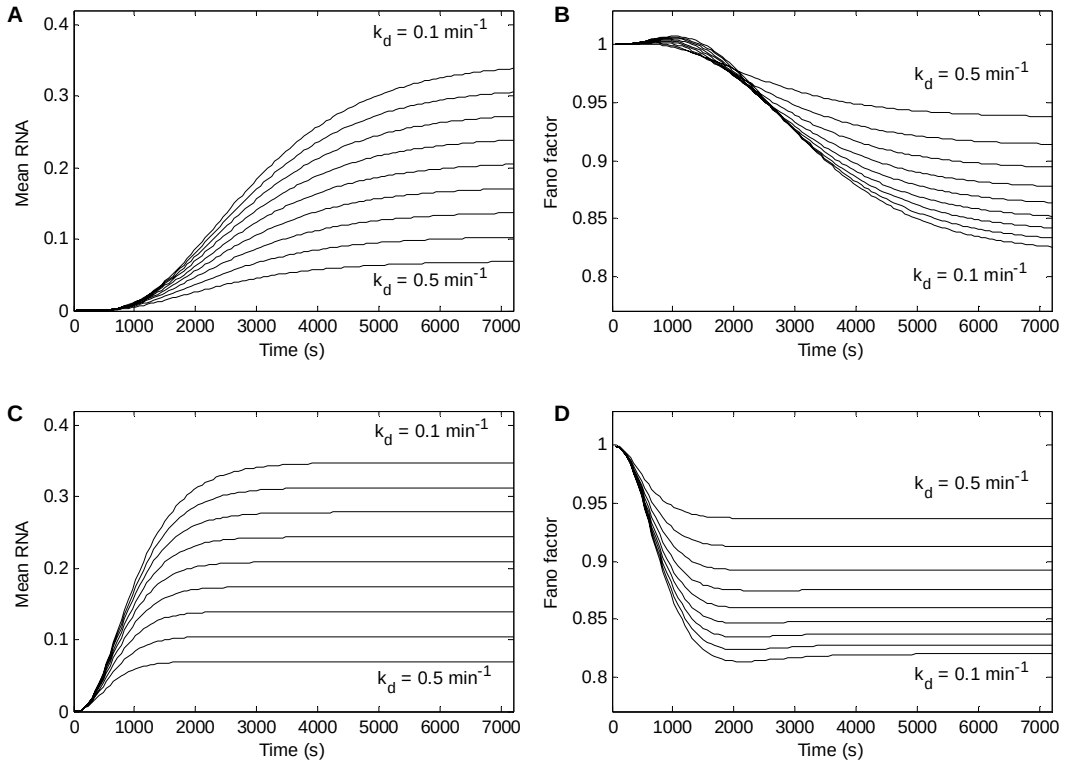


Figure S7. Models with different degradation rates. The degradation rate was set to the following values: 0.1 min^{-1}, 0.111 min^{-1}, 0.125 min^{-1}, 0.143 min^{-1}, 0.167 min^{-1}, 0.2 min^{-1}, 0.25 min^{-1}, 0.333 min^{-1}, 0.5 min^{-1}. In the figures, only the highest and the lowest values are marked. Mean RNA numbers shown for (A) P_{BAD} with 1 % Arabinose and for (C) P_{BAD} with 1 % Arabinose and infinitely fast intake. Fano factors of RNA numbers are shown for (B) P_{BAD} with 1 % Arabinose and, (D) P_{BAD} with 1 % Arabinose and infinitely fast intake.

Supplementary Table

	p-value for t_0	p-value for Δt
P_{BAD} 1 % arabinose	0.2613	0.8930
P_{BAD} 0.1 % arabinose	0.0020	0.5728
$P_{lac/ara-1}$ 1 % arabinose	0.1759	0.3826
$P_{lac/ara-1}$ 1 mM IPTG	0.1155	0.2413

Table S1. Results of the K-S fitting. Asymptotic p-values of the Kolmogorov-Smirnov goodness-of-fit test when fitting the empirical distribution with a sum of 5-exponential variates in the case of t_0 and of 3-exponential variates in the case of Δt. We compare these p-values with a standard value of 0.05.

Publication III

L. Martins*, J. Mäkelä*, A. Häkkinen, M. Kandhavelu, O. Yli Harja, J.M. Fonseca and A.S. Ribeiro, "Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time", *Journal of Theoretical Biology*, 301:83 – 94, 2012. (*equal contribution)

Journal of Theoretical Biology 301 (2012) 83–94

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time

Leonardo Martins [a,1], Jarno Mäkelä [b,1], Antti Häkkinen [b], Meenakshisundaram Kandhavelu [b], Olli Yli-Harja [b,c], José M. Fonseca [a], Andre S. Ribeiro [b,*]

[a] Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa, Monte da Caparica, 2829-516 Caparica, Portugal
[b] Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, P.O Box 553, FI-33101 Tampere, Finland
[c] Institute for Systems Biology, 1441N 34th St, Seattle, WA 98103-8904, USA

ARTICLE INFO

Article history:
Received 24 October 2011
Received in revised form
8 February 2012
Accepted 13 February 2012
Available online 20 February 2012

Keywords:
Closely spaced promoters
Stochastic dynamics
Transcription start sites
Open complex formation

ABSTRACT

Many pairs of genes in *Escherichia coli* are driven by closely spaced promoters. We study the dynamics of expression of such pairs of genes driven by a model at the molecule and nucleotide level with delayed stochastic dynamics as a function of the binding affinity of the RNA polymerase to the promoter region, of the geometry of the promoter, of the distance between transcription start sites (TSSs) and of the repression mechanism. We find that the rate limiting steps of transcription at the TSS, the closed and open complex formations, strongly affect the kinetics of RNA production for all promoter configurations. Beyond a certain rate of transcription initiation events, we find that the interference between polymerases correlates the dynamics of production of the two RNA molecules from the two TSS and affects the distribution of intervals between consecutive productions of RNA molecules. The degree of correlation depends on the geometry, the distance between TSSs and repressors. Small changes in the distance between TSSs can cause abrupt changes in behavior patterns, suggesting that the sequence between adjacent promoters may be subject to strong selective pressure. The results provide better understanding on the sequence level mechanisms of transcription regulation in bacteria and may aid in the genetic engineering of artificial circuits based on closely spaced promoters.

1. Introduction

Genes of *Escherichia coli* differ widely in expression kinetics (Taniguchi et al., 2010) due to, among other factors, the diversity of promoter sequences driving their expression. The regulation of expression levels is usually exerted during transcription initiation, a highly complex, multi-stepped process that starts with the binding of the RNA polymerase (RNAp) to the promoter region, followed by DNA unwinding and stabilization of the closed complex, assembly of the clamp/jaw on downstream DNA, formation of the open complex, and promoter escape (Browning and Busby, 2004; Saecker et al., 2011; Hsu, 2002). Only after, can RNA synthesis begin.

In vitro studies of the kinetics of initiation of several promoters in *E. coli* suggest that this process can have up to three sequential steps: formation of a closed complex, isomerization, and formation of the open complex (Saecker et al., 2011; Buc and McClure, 1985; McClure, 1985). Their duration varies between promoters, even when the sequences only differ slightly (Lutz et al., 2001; Singh et al., 2011), and are tightly regulated by repressing and activating molecules, whose bindings are, in general, in the promoter region.

The genome of *E. coli* contains various sites with closely spaced transcription start sites (TSSs). The geometry of these promoters with closely spaced TSSs can be tandem (same direction of elongation), divergent (directions of elongation are opposite, in a back-to-back fashion), or convergent (directions of elongation are opposite, in a front-to-front fashion) (McClure, 1985; Beck and Warren, 1988). Other sources of diversity between these promoters with closely spaced TSSs are the distance, in number of nucleotides, between the two TSSs and the location of the transcription factor binding sites (TFBS). Here, isolated TSSs are referred to as unidirectional promoters.

A recent survey suggests that approximately 15% of the promoters in *E. coli* are closely spaced (Gama-Castro et al., 2010). The same configurations have been found in quantity in other organisms, also exhibiting structural diversity at various levels (Beck and Warren, 1988; Häkkinen et al., 2011).

The kinetics of expression of genes driven by closely spaced promoters remains relatively unstudied, particularly in prokaryotes. In these organisms, it is yet unknown to what extent is the

* Corresponding author. Tel.: +358 408 490 736; fax: +358 331 154 989.
E-mail addresses: andre.sanchesribeiro@tut.fi, andre.ribeiro@tut.fi (A.S. Ribeiro).
[1] Equal contributions.

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2012.02.015

production of RNA molecules affected by possible interference between RNA polymerases (RNAps). In Sneppen et al. (2005) an analytical model was proposed to study the impact of transcriptional interference on mean expression levels. The model includes three mechanisms of interference, namely, occlusion (passing RNAps block access to the promoter), collisions between elongating RNAps, and "sitting duck" interference. Given the features of their model, the analysis focused on study of mean expression rates and, regarding this feature of the kinetics, the results agreed with measurements from convergent promoters (Callen et al., 2004).

While there is little study of the effects of proximity of TSSs in prokaryotes, there is some information available from studies in eukaryotes. In Wang et al. (2011) the dynamics of a stochastic model of closely spaced promoters was analyzed. The model used accounted for chromatin remodeling by switching the promoter state between ON and OFF via stochastic, first order reactions. Gene expression was modeled as a single step event. The results suggest that the orientation and distance between TSSs affect noise in RNA and proteins numbers. It is noted that this model assumes *a priori* that distance and other topological features (not modeled explicitly) have effects on the correlation between the expression of the two genes, thus on both genes' mean expression and noise, rather than determining these effects from the structure.

Another work studied the effects of genetic and epigenetic properties of promoters on expression variability in budding yeast (Woo and Li, 2011) from genome-wide datasets of gene expression and nucleosome occupancy. The authors suggest that, for this organism, divergent TSSs tend to have lower expression variability than tandem TSSs and that this variability, for both configurations, tend to decrease with decreasing distance. These results are somewhat in disagreement with those of Wang et al. (2011) for distances between TSSs shorter than 300 nucleotides. Finally, in Ebisuya et al. (2008) experimental evidence was reported that, in eukaryotes, transcription initiation appears to exert a "ripple effect", that is, the induction of expression of a gene tends to stimulate the expression of neighbor genes. The authors suggest that this mechanism may be advantageous for coordinated expression of genes participating in similar functions. Since chromatin dynamics and DNA methylation may play a role in this effect, and given several other differences in the mechanisms of gene expression, it is unknown if the effects of TSSs proximity is similar in prokaryotes and eukaryotes, although it is likely that the proximity does play a role in the expression kinetics of closely spaced genes in both cases. Due to that, models must be made at the nucleotide level so as to study, among other features, interferences in the expression dynamics.

Here, using the delayed stochastic simulation algorithm (delayed SSA) (Roussel and Zhu, 2006) to drive the dynamics, we model promoters at the nucleotide level and simulate the kinetics of transcription, one RNAp at a time. One of the novelties of this study, allowed by the model used, is the quantification of the effects of changes in promoters, at the nucleotide level, on the kinetics of RNA production. Also, we account for the duration of rate limiting steps at the TSS, such as isomerization and the open complex formation, which varies from one event to the next.

Following the description of the model and comparison of its predictions to measurements, we first study the kinetics of the binding of RNAps to the promoter sequence. Next, we study the dynamics of RNA production as well as the degree of correlation between consecutive choices of directions of elongation as a function of the geometry. Finally, we study the kinetics of expression and its regulation by repression by occlusion as a function of the positioning of the TFBS, among other variables. In the end, we present our conclusions and address the following questions: what are the effects of the rate limiting steps at the TSS, the closed and open complex formations, on the kinetics of RNA production for various promoter configurations? Are there abrupt changes in the kinetics of RNA production with nucleotide distance between TSSs? To what extent does the proximity between two TSSs correlate the dynamics of RNA production under their control?

2. Methods

Transcription in prokaryotes is both a stochastic process (Arkin et al., 1998) and sparse in time (Taniguchi et al., 2010; McClure, 1985), which imposes the use of Monte Carlo methods to simulate it, such as the stochastic simulation algorithm (SSA) (Gillespie, 1977). Additionally, the process of initiation contains several rate limiting steps. Usually, there appear to be two major rate limiting steps, the closed and the open complex formations (McClure, 1985; Lutz et al., 2001). The first includes the finding of the promoter region and diffusion of the RNAp along the DNA template until reaching the TSS and forming of the closed complex. The second includes a few isomerization steps until the open complex is formed (deHaseth et al., 1998).

Given the above, we use the delayed SSA (Roussel and Zhu, 2006) to drive the kinetics of the models since, unlike the original SSA (Gillespie, 1977), it allows delayed events. In these, once the reaction occurs and the reactants are removed from the vessel of reactions, the products are kept on a wait list for a predetermined amount of time, and only after are made available for reactive events. To implement the delayed SSA and simulate the models described below we use the simulator SGNSim (Ribeiro and Lloyd-Price, 2007). An example of the implementation of one the models that can be simulated by SGNSim is provided in supplementary material. All models are described in detail in the supplementary material as well.

The models of promoters, namely, their nucleotide structure, as well as various kinetic rate constants, are extracted from measurements. We extracted the sequences of known and predicted divergent and convergent closely spaced promoters in *E. coli* from the RegulonDB database (version 7.0) (Gama-Castro et al., 2010). The distributions of nucleotide length between TSSs are shown in Fig. 1. The bulk of the distribution is below 200 nucleotides in length (88.8% for convergent and 61.8% for divergent). Mean distances are 108.4 and 225.7 for convergent and divergent promoters, respectively. In all models below, the distances between TSS are set within these realistic intervals.

The model of transcription, the set of possible reactions and events, along with the stochastic rate constants are described in supplementary material. The first step towards the production of an RNA molecule is the binding of the RNAp to the DNA template. The RNAp can then diffuse one nucleotide at a time in a direction chosen initially at random. Provided long time intervals, it is believed that diffusion RNAps can change direction. However, these changes are not common events, that is, for short distances (of tenths of nucleotides) evidence suggests that the direction of diffusion does not change (Sakata-Sogawa and Shimamoto, 2004; Gorman and Greene, 2008). Nevertheless, we note that this assumption (if wrong) likely does not affect the results significantly, due to the much higher speed of diffusion (600 nuc/s) and disassociation ($0.3\ s^{-1}$) in comparison with the other possible events, such as those at the TSS and during elongation.

The RNAp can unbind from the template at any step. If there are multiple RNAps on the template, there can be collisions. In that case, one of the RNAps falls off the template (randomly chosen). In convergent promoters, elongating and diffusing RNAps can collide (Callen et al., 2004). Since the elongating RNAp is more

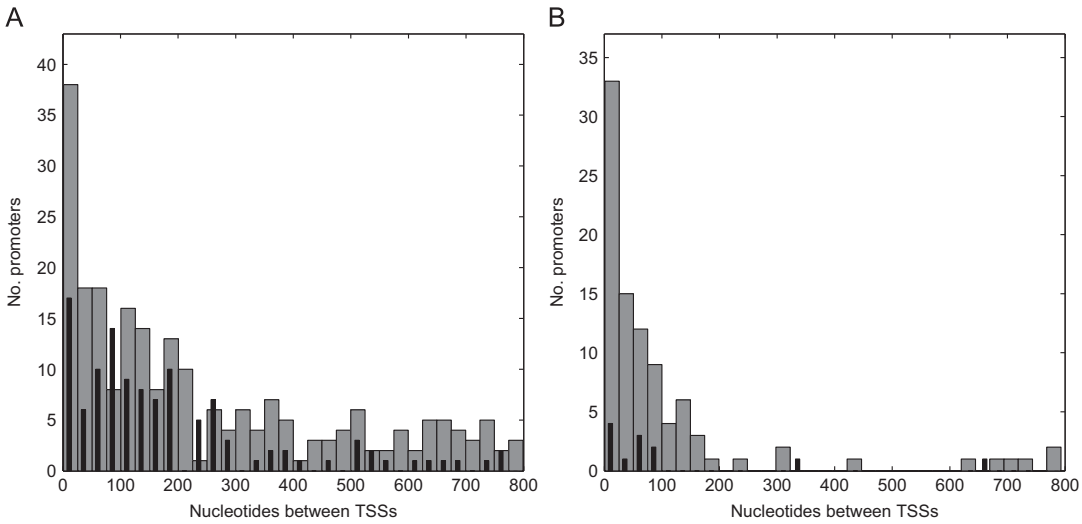


Fig. 1. Number of closely spaced promoters in *E. coli*. Number of closely spaced promoters with various numbers of nucleotides between the two TSSs for (A) divergent promoters and (B) convergent promoters. Black bars are experimentally verified promoters; wider gray bars are predicted promoters (Gama-Castro et al., 2010).

tightly bound to the template (Callen et al., 2004), the diffusing RNAp (or an RNAp at the TSS) is removed from the template. If two elongating RNAps collide, either one or both are removed from the sequence (50% chance of each case).

Once the diffusing RNAp finds the TSS oriented in the same direction as its movement, several events take place. First, the closed complex is formed. This is followed by isomerization, and then the open complex formation (deHaseth et al., 1998). The duration of isomerization and of the open complex formation are randomly generated at each event from an exponential distribution whose means are set in accordance with in vitro measurements (Buc and McClure, 1985)(see supplementary material). Once the open complex is formed, either elongation begins and the TSS is released or a short, incomplete, transcript is produced and the RNAp returns to the TSS (Hsu, 2002). Once the TSS is cleared another diffusing RNAp can occupy it. As a side note, the assumption that the RNAp only engages the TSS oriented in the same direction as its direction of diffusion has not been experimentally validated. Instead, it is a hypothesis made here that relies on the fact that the RNAp molecule is not symmetric and thus it is reasonable to assume that it can only recognize a TSS sequence in one direction.

The last stage is elongation. Within the promoter sequence, elongation is modeled at the nucleotide level to account for the interference between RNAps. Once this region is cleared, the rest of the process is modeled by a single step delayed event. The time length of this process is generated at each occurrence from a Gamma distribution whose mean is given by the product between the expected time for the RNAp to elongate from one nucleotide to the next and the number of nucleotides of the sequence.

The model allows repression by one or more TFs, which can bind to existing TFBSs. Depending on the DNA region occupied by the repressor; one can model repression by steric occlusion or by DNA looping (i.e. preventing binding of RNAps). When a diffusing RNAp collides with a repressor, it is released from the template, decreasing the rate of production of transcripts. Elongating RNAps are not released by such collisions instead they remain paused until the repressor unbinds (Lopez et al., 1998).

In Fig. 2 we represent two models of closely spaced promoters. Locations are designated by the position relative to the TSS at position $+1$: positions to the left are negative and to the right are positive (position 0 does not exist, by convention). In Fig. 2A, a divergent promoter is represented. The TSSs are at -151 and $+1$. The gene to the left can only be transcribed by RNAps diffusing in the left direction, and the one to the right can only be transcribed by RNAps diffusing to the right. Regions of elongation (which the RNAp can also percolate by diffusion) are represented in black, while regions where only diffusion occurs are in gray. Elongating RNAps are represented with an elongating RNA chain, not present in diffusing RNAps. If a repressor is bound (in this case the TFBS is between -140 and -120), it blocks diffusing and elongating RNAps. It also prevents binding of RNAps in that region. In Fig. 2B is shown a convergent promoter with overlapping elongation regions. TSSs are at $+152$ and $+1$.

3. Results

3.1. Mean RNA numbers at near-equilibrium and the rate of binding to the promoter region

All parameters values of the model, including RNA degradation rates (Bernstein et al., 2002), are from in vitro measurements except for the rate constant of association of the RNAp to the promoter region (k_{bind}). There is evidence that the RNAp can bind to any nucleotide in the promoter region and that the binding rate to this region is higher than in other regions, but exact values are unknown (Singer and Wu, 1987).

To test with the model what values for k_{bind} result in realistic mean RNA numbers at near-equilibrium obtained from genome wide measurements (Taniguchi et al., 2010), we vary k_{bind} and measure mean RNA numbers at near-equilibrium for convergent, divergent, and unidirectional promoters. In all models, the length of nucleotides of the binding region of the RNAp is 200 nucleotides. In the two former models, the two TSSs are 150 nucleotides apart from each other.

Results are shown in Fig. 3 and each data point is calculated from 50 concatenated time series, each 10^5 s long, sampled every 10 s. There are 28 RNAps available (Bremer et al., 2003) and RNA degradation rate is 0.36 min^{-1} (Bernstein et al., 2002). Simulations

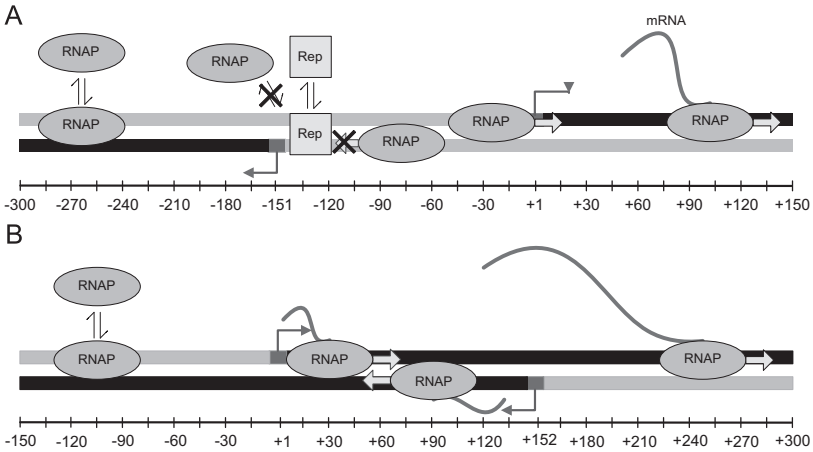


Fig. 2. Models of promoters. Representation of models of (A) divergent and (B) convergent promoters. Elongation regions depicted in black. RNAps can bind to any nucleotide. Angled arrows represent TSSs. Harpoons represent binding and unbinding of RNAps and repressors to the DNA. RNA sequences are depicted in elongating RNAps.

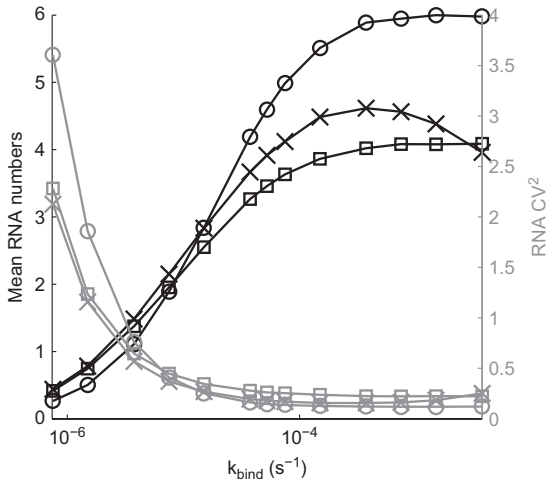


Fig. 3. Mean and square of coefficient of variation of RNA numbers at near-equilibrium. Mean RNA numbers at near-equilibrium as a function of k_{bind} rate (black solid lines): (○) unidirectional promoters, (x) divergent promoters and (□) convergent promoters. Also shown are the CV^2 (square of the standard deviation over the square of the mean) of RNA numbers from the statistics over time (gray solid lines) for each case. Note the the y-scale of CV^2 is shown in the vertical axis at the right side, while y-scale for mean numbers is shown in the vertical axis at the left side.

are initialized without RNA molecules but the transient to reach near-equilibrium is negligible, given the length of the series.

Mean RNA numbers at near-equilibrium are within realistic intervals (Taniguchi et al., 2010) for all values of k_{bind} and promoter arrangements (Fig. 3). Lower values correspond to weakly expressed genes, while higher values correspond to more strongly expressed ones. However, expression is not subject to repression in these simulations, while in *E. coli*, weakly expressing genes usually have such behavior due to the presence of repressor molecules. Due to this, we consider the most realistic values of k_{bind} are likely to values beyond $10^{-5}\,s^{-1}$. Interestingly, the rate of saturation (due to the existence of the rate limiting steps, the open and closed complex formations) is $k_{bind} \sim 10^{-4}\,s^{-1}$. We thus estimate realistic values to be between 10^{-5} and $10^{-4}\,s^{-1}$.

Singer and Wu, (1987) estimated the in vitro rate of non-specific association of RNAp to circular DNA to be $4.6 \times 10^4\,M^{-1}\,s^{-1}$ (per nucleotide). From this, one can estimate the value for k_{bind} by dividing the measured value by the expected volume of *E. coli* (10^{-15} L) (Sundararaj et al., 2004) and the Avogadro constant. It results in a value for k_{bind} of $0.75 \times 10^{-4}\,s^{-1}$ per nucleotide, thus within the interval of the estimation (between 10^{-5} and $10^{-4}\,s^{-1}$). Here onwards we set k_{bind} to $0.75 \times 10^{-4}\,s^{-1}$, unless stated otherwise.

Also, the kinetics of RNA numbers differs with the geometry of the promoter (Fig. 3). Provided identical kinetic rate constants for the various models of promoters including, e.g., binding regions for RNAps of the same length, same duration for processes such as the open complex formation, etc., we find that unidirectional promoters have higher mean RNA numbers, as the production is not affected by occlusion and collisions between RNAps traveling in opposite directions. The decrease in mean RNA numbers for increases in k_{bind} beyond $\sim 10^{-4}\,s^{-1}$ in divergent promoters is due to the increased interference between diffusing RNAps. This is less frequent in convergent promoters as elongating RNAps remove many diffusing RNAps while percolating the DNA template in the outer regions.

3.2. Binding kinetics of RNA polymerases to promoter regions

To study the dynamics of binding of RNAps to promoters, we measured the fraction of times each of the nucleotides of a divergent promoter is bound by the center of a previously free RNAp (50 simulations, each 10^5 s long) (Fig. 4), when k_{bind} is set to $0.75 \times 10^{-4}\,s^{-1}$ and when is set to 10 and 100 times smaller. Simulations are long enough so that increasing the duration of the binding would not alter the results significantly.

From Fig. 4, the fractions of bindings per nucleotide are not uniform across the template, except for $k_{bind}/100$. This is due to the rate limiting steps at the TSSs (closed complex formation, isomerization, open complex formation, and abortive initiations) and the non-negligible footprint of the RNAp. The intermediate regions between the TSSs are those most available for new RNAps to bind to. The discontinuities are less pronounced as k_{bind} is decreased, since the regions where the rate limiting steps occur are not so often occupied.

To verify this, we next set the rates of closed and open complex formations and of isomerization to values identical to the rate of diffusion of the RNAp. The rate of abortive initiation is set to zero. The resulting distribution of fraction of binds to each nucleotide for standard values of k_{bind} (data not shown) becomes identical to that of $k_{bind}/100$ shown in Fig. 4. Additionally, to show the dependence of the discontinuities in the distributions on the footprint of the RNAp, in Fig. 5 we show the results for divergent promoters with increasing distances between TSSs.

From Fig. 5 it is visible how the shapes of the distributions depend on the ratio between the footprints of the RNAps when diffusing and the length of the binding region. These distributions only become uniform-like if this ratio is very large or if the promoter region is smaller than twice the footprint length. The same conclusions are valid for convergent promoters (data not shown).

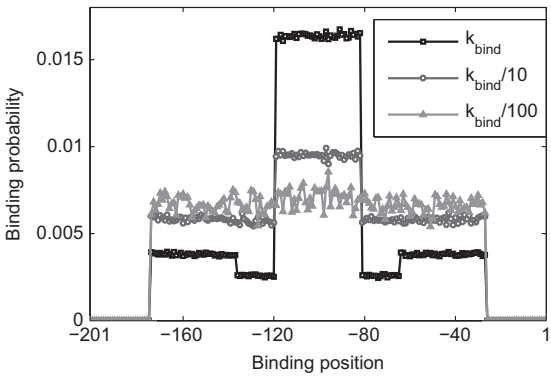


Fig. 4. Promoter binding kinetics. Fraction of times that a free RNAp bound to each of the nucleotides of a divergent promoter with a binding region of 200 nucleotides located between the two TSSs. When the closed complex forms at a TSS the RNAp occupies nucleotides +1 to −53 (right) and −147 to −201 (left).

3.3. Dynamics of RNA production in closely spaced promoters of different geometries

To determine how the geometry of a promoter affects the dynamics of RNA production, we now compare the distributions of intervals between the productions of consecutive RNAs from one of two TSSs for differing geometries (in all models this distribution is identical for the two RNAs, unless stated otherwise). In all models the binding region is 200 nucleotides. Results are shown in Fig. 6 and the tails of the distributions are shown in inset for each case (except Fig. 6F, where it is within the range of 120 s) and differ significantly in length.

Models A–C are divergent, differing in the distance between the two TSS. In A the distance is 200, in B is 150 and in C is 65 (can only contain one RNAp at a time). As the distance increases, the mean and standard deviation of the intervals decrease due to the decrease in the number of collisions between elongating and diffusing RNAps, and the consequent reduction of the width of the distribution. Model D is also divergent, identical to A, but without the rate limiting steps at the two TSS and all steps, including

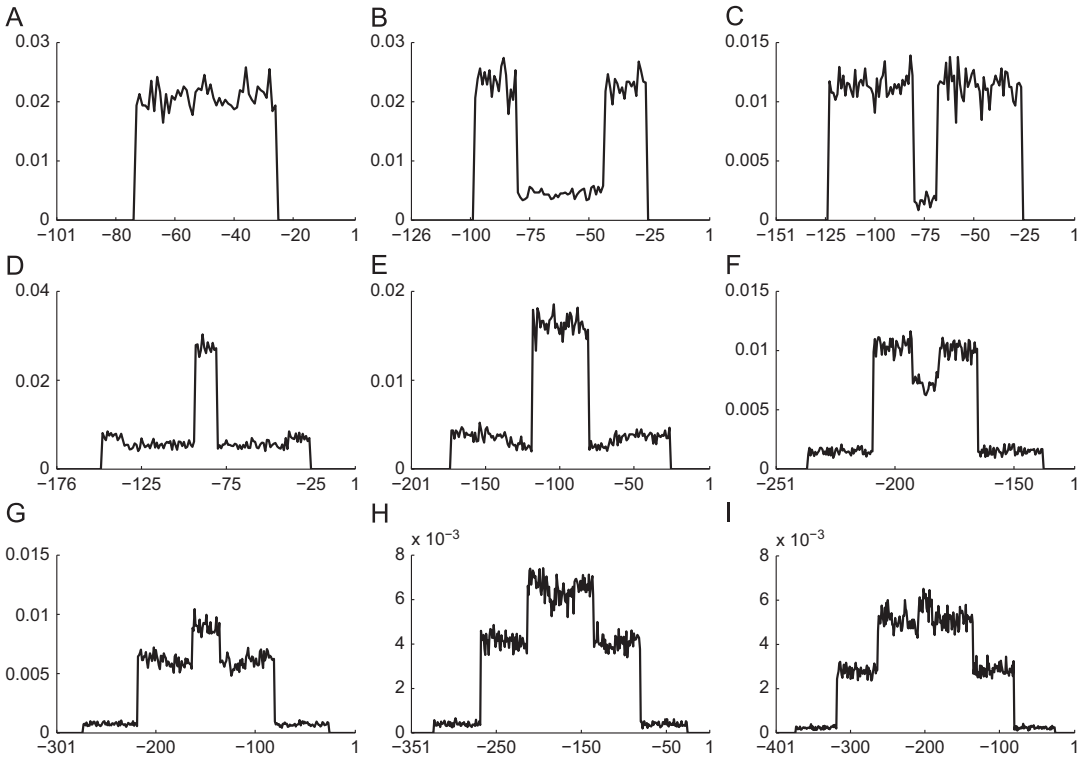


Fig. 5. Promoter binding kinetics for varying promoter lengths. Fraction of times free RNAps bound to each of the nucleotides for varying number of nucleotides between the two TSSs in divergent promoters. In all cases, the two TSSs are at the extreme positions represented in the *x*-axis.

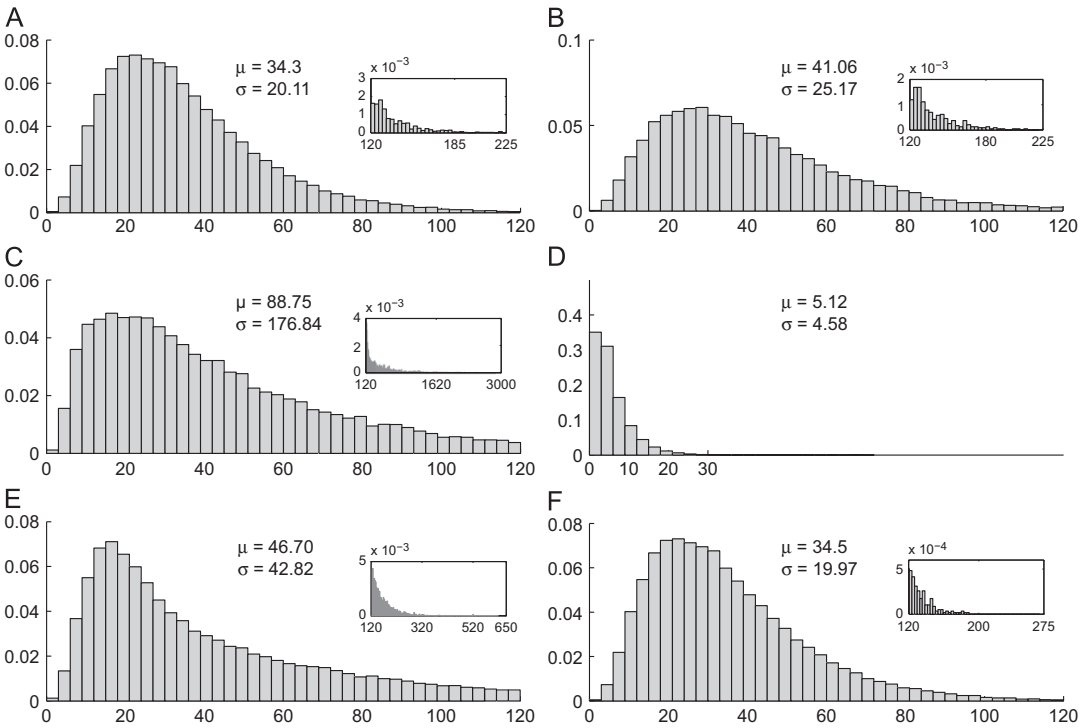


Fig. 6. Distributions of intervals between the productions of consecutive RNAs from one TSS. In all cases the bins size is 3 s: (A) divergent promoter with 200 nucleotides between TSSs, (B) divergent promoter with 150 nucleotides between TSSs, (C) divergent promoter with 65 nucleotides between TSSs, (D) same as (A) but no rate limiting steps at the TSSs, (E) convergent promoter with 100 nucleotides between the two TSSs and (F) unidirectional promoter with a binding region of 200 nucleotides. The tails of the distributions are shown in inset. Each figure also shows the mean (μ) and standard deviation (σ) of each distribution.

elongation, closed and open complex formation and isomerization, occur at the same speed as diffusion (abortive initiation is not modeled). In comparison to A, the mean of the intervals is much smaller as the distribution becomes exponential-like, due to the absence of the rate limiting steps, as predicted in Ribeiro et al. (2010).

Model E is a convergent promoter with 100 nucleotides between the two TSSs. In comparison to A, RNA production is reduced and noisier. The distribution of intervals increases in mean and standard deviation due to the interference between elongating RNAps traveling in opposite directions. Finally, model F, unidirectional, behaves as a divergent promoter with a configuration such that there are no collisions between elongating and diffusing RNAps (i.e. model A).

In general, the kinetics of transcript production is similar in all closely spaced promoters. The steps that most shape the distributions are the rate limiting steps at the two TSS. However, the mean and standard deviation of the distributions of intervals between productions of RNAs depends, to some extent, on the relative positions of the two TSSs and the promoter's geometry.

In closely spaced promoters the dynamics of transcripts production from the two TSSs are dynamically correlated by interferences between diffusing and elongating RNAps (i.e. collisions). This can be verified by calculating the autocorrelation of the time series of RNAs' production, although collisions may not be the only source of correlations. If the autocorrelation is null, there is no effect of the interferences between RNAps on the production of transcripts. If the autocorrelation is positive, it implies that once one of the two types RNAs is produced, the promoter is biased by the interferences to produce the same type of RNA in the next event. If the autocorrelation is negative, the opposite is more likely. The autocorrelations were calculated for models A–F, and for an additional model of two, non-interacting, unidirectional promoters. The calculations are done for several lags (lag 1 corresponds to correlations between the present choice of RNA produced and the next one) and confirm the above hypothesis (Fig. 7).

First, from Fig. 7A, there is a negative correlation between consecutive production events, but this correlation is not propagated for longer lags. The correlation depends on the existence of rate limiting steps, as seen by comparing with Fig. 7D (model without rate limiting steps). When a TSS is occupied by an RNAp, it remains occupied until the RNAp begins a successful elongation event. Due to that, it is more likely that the next RNA to be produced will be from the other TSS (since this TSS may or not be occupied at the same time, the negative autocorrelation is not "perfect", i.e. equal to -1). The same reason explains the small positive correlation between choices for lag of 2. On the other hand, if there are no rate-limiting steps (model D), there are no correlations between consecutive choices of which RNA to produce, since the system is virtually memoryless.

If the distance between TSSs is decreased (Fig. 7B) both the autocorrelations at lag 1 and at lag 2 decrease, since the number of RNAps that can be bound to it at any moment is smaller, thus decreasing the probability that one successful transcription event at a TSS will be followed by another event at the other TSS.

As the distance between the two TSSs is even further decreased so that only one RNAp can be between the two TSS at any moment, there is an abrupt change in the kinetics of transcription (Fig. 7C). In this configuration, if one of the two TSS loaded with RNAp, going

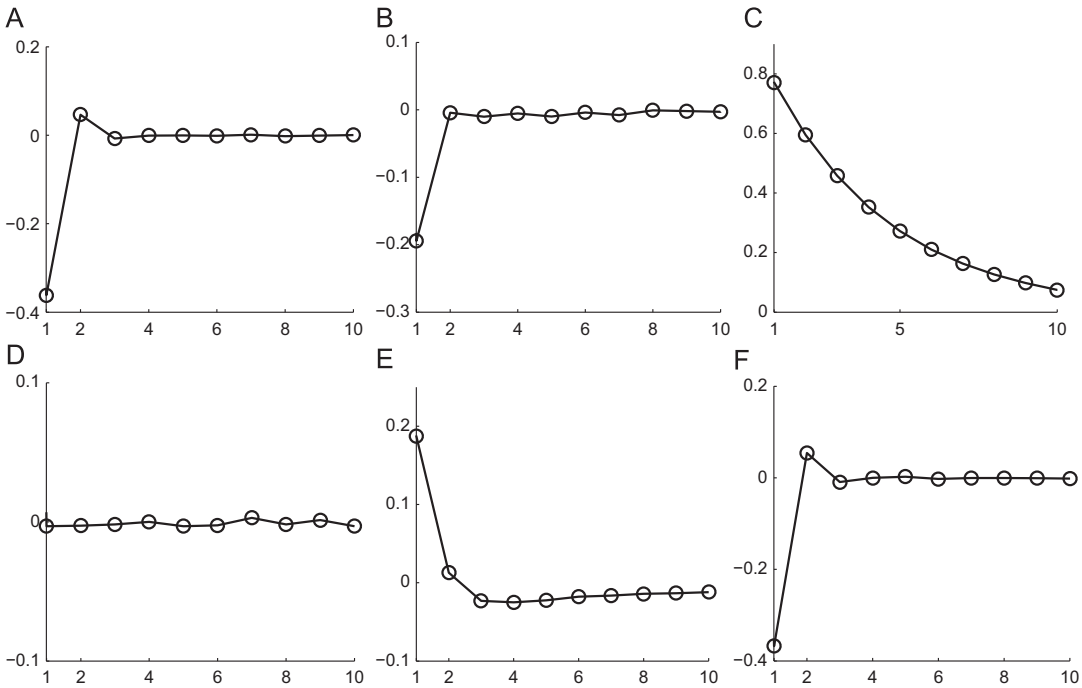


Fig. 7. Correlation between consecutive choices of RNA production. Correlation between consecutive choices of which of the two RNAs are transcribed for increasing lag: (A) divergent promoter with 200 nucleotides between TSSs, (B) divergent promoter with 150 nucleotides between TSSs, (C) divergent promoter with 65 nucleotides between TSSs, (D) same as (A) but no rate limiting steps at the TSSs, (E) convergent promoter with 100 nucleotides between the two TSSs and (F) two independent unidirectional promoters each with a binding region of 200 nucleotides. Self-correlation (lag 0) not shown.

through the open complex formation, the other TSS cannot be loaded by an RNAp, since the region between the two TSS is too small to allow diffusing RNAps to reach that TSS. Once the open complex is completed, the RNAp elongates along the template and removes from the template any RNAp diffusing in the opposite direction, making it more likely that the next TSS to be occupied to be the one from where the elongating RNAp is coming from. This leads to a very strong positive correlation of consecutive choices (Fig. 7C). In the next section, it is shown how strongly dependent this phenomena is of the distance between the two TSS. Finally, Fig. 7D shows that, in the absence of rate limiting steps and overlapping divergent configuration, there are no correlations between consecutive choices.

In Fig. 7E, we observe a phenomenon similar to that in Fig. 7C. In convergent geometries it is more likely that one elongation event from one TSS is followed by another such event from the same TSS. The effect is not as strong as in model C because here another elongating event can start at the other TSS while the RNAp coming from the first TSS is elongating. When two elongating RNAps collide, they have identical probabilities of being removed from the sequence, aborting RNA production.

Finally, in Fig. 7F we show the correlations between consecutive transcripts production from two independent unidirectional promoters. The similarity between this figure and Fig. 7A shows how negative correlation emerges due to the rate limiting steps, regardless of the geometry of the promoter.

3.4. Distance between TSSs

We now study how changing the distance between TSSs may alter the dynamics of RNA production. Fig. 8 shows the degree of correlation between consecutive choices of direction of elongation

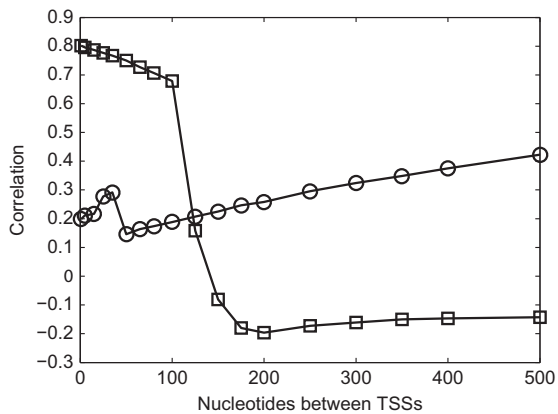


Fig. 8. Correlation between consecutive choices of RNA production with varying promoter lengths. Correlation in consecutive choices (lag 1) of directionality of elongation events in convergent (○) and divergent (□) promoters for varying nucleotide length between the two TSSs.

(at lag 1, i.e. between consecutive events) for convergent (○) and divergent (□) promoters with varying nucleotide length between the two TSSs. In all cases, the binding region of the RNAp is 300 nucleotides long.

Results from Fig. 8 show that the distance between the TSSs strongly affects the correlation between consecutive choices of direction of elongation in divergent promoters. For distances smaller than ~110 nucleotides there is a strong positive correlation between

consecutive choices. As the distance is further increased, there is an abrupt change and the choices become anti-correlated. This transition corresponds to the change in the structure from overlapping to not overlapping. When overlapping (<110 nucleotides), the RNAp at a TSS, when elongating, clears the other TSS from any RNAp. For larger distances, the correlation is negative because of the rate limiting steps at the TSS and because no longer does an elongating RNAp interfere with the activity of the other TSS.

In convergent promoters, there is interference between the activities of the two TSSs for all distances, as elongating RNAps can clear the other promoter from any RNAp. This interference increases with distance because the longer the time that it takes for the elongating RNAp to pass by the other TSS, the longer will be the interval during which no successful transcription event can arise from this other TSS. The small peak at the 35 nucleotides between TSSs is due to the fact that at such a distance, the RNAp at one TSS impedes RNAps to reach the other TSS. When the distance becomes large enough, both TSSs can have an open complex event at the same time (one initiated before the other). In this event, it is possible for the open complex that initiated at a later stage is completed first, and, in that case, it will clear the other TSS once elongation begins.

The correlations in Fig. 8 affect the mean RNA numbers at near equilibrium (Fig. 9). In divergent promoters, in general, the higher is the positive correlation, the smaller is the mean number of RNAs at near-equilibrium. In convergent promoters, the relationship between correlation and mean RNA numbers is the opposite for small distances between the two TSSs. In both geometries, the stronger is k_{bind}, the stronger are the correlations (both positive and negative ones) (data not shown). Finally, beyond a certain length, further increases in length no longer change mean RNA levels significantly. This is due to other rate limiting steps, such as the open complex, that limit further increases in RNA production.

In convergent promoters, as the distances between the two TSS increase, there is strong increase in mean RNA numbers when the distance becomes large enough for having the two TSS simultaneously occupied by an RNAp. Further increases in distance between the TSSs decrease mean RNA numbers due to the increase in number of interferences and collisions between elongating RNAps.

3.5. Repression by occlusion

The most common mechanism of repression of transcription is steric occlusion, which blocks the access of RNAps to a specific region of the promoter (McClure, 1985). Depending on the location of the binding site, it affects different stages of initiation, from closed complex to open complex formation, preventing elongation initiation (Sanchez et al., 2011; Garcia et al., 2010). Different repressors occupy different number of nucleotides. Blocking large portions of the DNA usually requires DNA looping (Carey et al., 1991; Lewis et al., 1996; Horton et al., 1997).

Steric occlusion can, in theory, block transcription completely since, provided a very large number of repressors the expected time for one of them to bind to the DNA is virtually zero, hampering transcription events. The only case where complete repression is not achievable is if there is sufficient space between the region occupied by the repressor and the TSS for an RNAp to bind. In this scenario, as the number of repressor molecules increase, the rate of RNA production would decrease only until a plateau of minimum expression rate.

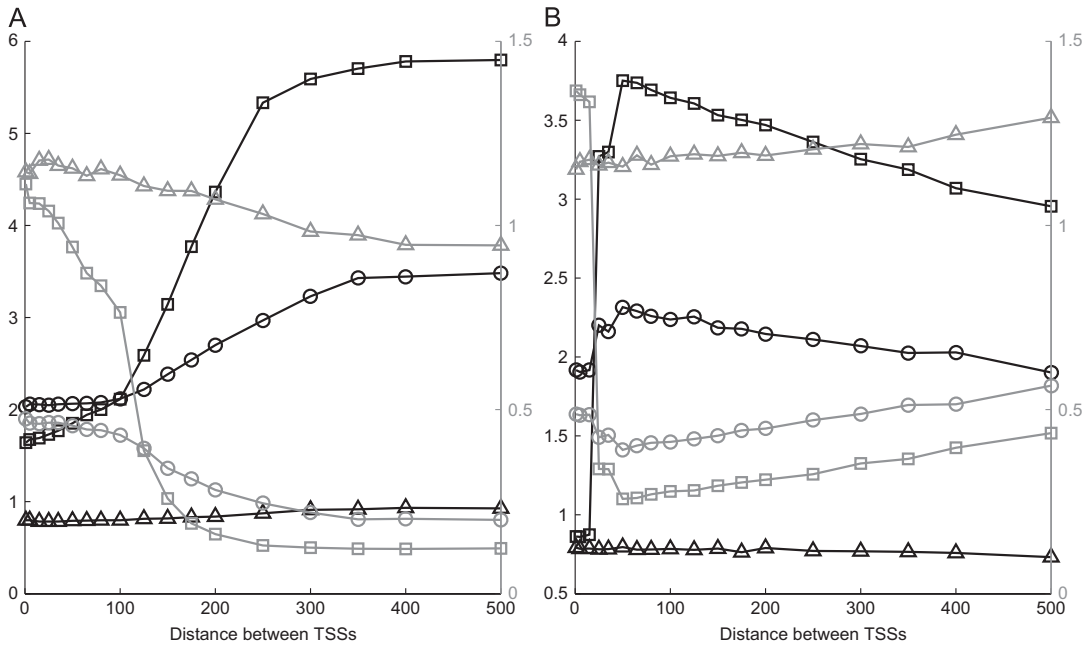


Fig. 9. Mean and square of coefficient of variation of number of RNAs at near-equilibrium from one of the two TSSs with varying promoter lengths. Mean number of RNAs at near-equilibrium from one of the two TSSs in (A) divergent and (B) convergent promoters for varying nucleotide length between the TSSs. (□) represents standard k_{bind}, (○) represents $k_{bind}/10$ and (Δ) represents $k_{bind}/100$. Also shown are the CV^2 (square of the standard deviation over the square of the mean) of RNA numbers from the statistics over time (gray solid lines) for each case. Note the the y-scale of CV^2 is shown in the vertical axis at the right side, while y-scale for mean numbers is shown in the vertical axis at the left side.

Here, we first investigate the kinetics of transcription of unidirectional promoters subject to a repressor as a function of the number of repressors and the location of the binding site. We model promoters with the repressor's binding site centered at positions +1, +12, and +37. In all cases, the repressor occupies 21 nucleotides centered at the binding position.

When diffusing, an RNAp occupies 55 nucleotides (McClure, 1985). The first rate limiting step begins when the RNAp reaches the TSS. During isomerization, the footprint of the RNAp increases to 75 in the downstream direction implying that it now occupies the 20 nucleotides following the TSS. After promoter escape, the release of the σ factor reduces the footprint to 25 nucleotides. Thus, when repressors bind at +1 it blocks the closed complex formation, at +12 it allows the closed complex but blocks the open complex formation, and at +37, it allows initiation but blocks elongation.

To model repression we introduce in the model the reactions for binding and unbinding of repressors. The ratio between the rates of these two reactions has been estimated for several repressor molecules (So et al., 2011). Here, on average, we set these rates' values so that a repressor is bound to its binding site approximately 80% of the time (see supplementary material).

The model assumes that RNAps cannot, by any means, dislodge repressors, i.e., the kinetics of unbinding of the repressor from the DNA only depends on the kinetic rate of unbinding. If an RNAp is occupying the binding region of the repressor, the repressor cannot bind. To assess the strength of repression we define a repression factor as the ratio between mean RNA numbers at near-equilibrium when no repressors are present and when a certain number of repressors are present. Fig. 10 shows how this quantity varies with the position of the binding site and with the number of repressors. The rates of binding and unbinding are identical in all cases.

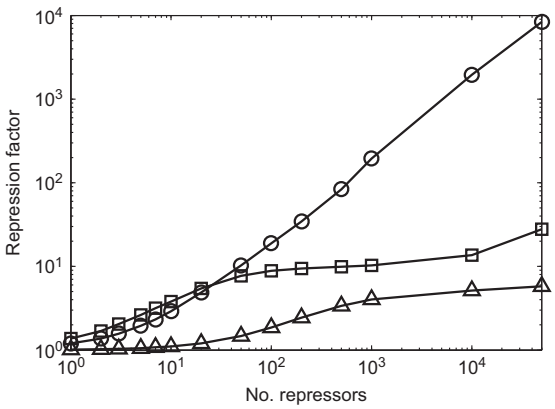


Fig. 10. Repression factor of unidirectional promoters. Repression factor of unidirectional promoters as a function of the position of binding site and of the number of repressors. The *y*-axis is the repression factor (ratio between mean RNA numbers at near-equilibrium in the absence and in the presence of repressors) and *x*-axis is the number of repressor molecules in the cell. Repression by occlusion with a binding site for the repressor centered at positions +1 (Δ), +12 (○), and +37 (□).

From Fig. 10, in all cases, increasing the number of repressors increases the repression factor, which also depends on the location of the binding site. Binding sites at the TSS or right after it (at +12 or +37) provide equally efficient repression for small number of repressors. For large number of repressors, repression is stronger if the open complex is blocked (at +12). Repressing the closed complex is the least efficient since binding of RNAps to the template is a fast process, and thus able to compete with the binding of repressors.

When blocking the open complex formation at +12, repressors only compete with isomerization. In these conditions, increasing the number of repressors steadily decreases the rate of RNA production. Increasing the number of repressors blocking promoter escape leads to more complex changes. In small amounts, repressors only delay the movement of elongating RNAps but do not actually prevent elongation; thus, they have a limited effect in RNA production. Only when the speed of binding of repressors (due to increased number of repressors) overcomes the speed of elongation do further increases in repressors numbers lead to additional decreases in the production of RNA molecules. In this regime, RNAps are prevented from leaving the TSS as the template is virtually always occupied by a repressor.

Next, we test how the positions of the binding sites of the repressors determine the effects on the kinetics of transcription of convergent and divergent promoters. Results are shown in Table 1, where it is visible that by placing the repressor closer to one of the two TSS it is possible to bias to kinetics of RNAs production in both convergent and divergent promoters. For example, placing a repressor at −65 in a convergent promoter (TSSs at +1 and +150) only reduces the expression from the TSS at +1 (right). The overall production of both RNA molecules can also be affected, as in unidirectional promoters. For example, a repressor at +36 in convergent promoters or at −35 in divergent promoters decreases the overall expression by approximately 60% and 40%, respectively. Finally, decreasing the overall expression without biasing the production of the two RNAs is also possible. For that, the binding site ought to be located close to the midpoint between the two TSSs, or two repressors can be placed at symmetric positions from one another.

It has been suggested that the relatively small distance between TSSs may facilitate the co-regulation of gene expression in both directions (McClure, 1985). This would imply, for example, facilitating the simultaneous repression or activation of transcription from the two genes, which in other words, implies that the expression levels of the two genes ought to be correlated.

Table 1
Repression in closely spaced promoters.

Convergent	Left	Right	Divergent	Left	Right
No repressor	1	1	No repressor	1	1
−65	1 (elong.)	0.3 (closed)	+15	0.8 (diff.)	0.38 (open)
+15	0.8 (elong.)	0.26 (open)	−35	0.9 (diff.)	0.6 (closed)
+36	0.6 (elong.)	0.17 (esc.)	−75	0.56 (diff.)	0.55 (diff.)
+75	0.25 (elong.)	0.25 (elong.)	−35, −115	0.4 (closed)	0.4 (closed)

Mean RNA numbers relative to the basal level of overlapping promoters with 150 nucleotides between TSSs. Binding regions are 300 nucleotides long and repressors occupy 21 nucleotides. The type of repression, determined by the location of the binding site, is indicated. In convergent promoters, TSSs are at +1 (controlling the gene at the right side) and +150 (controlling the gene at the left). In divergent, TSSs are at −150 (left) and +1 (right). In all cases, there is only one repressor in the cell except in the last case for divergent promoters, where there are two repressors, since there are two binding sites.

We next study how repression may correlate the two time series of RNA numbers under the control of the two TSSs.

We start by modeling two identical unidirectional promoters in the same cell and under the control of same repressor. The number of repressors is set to 100 and, thus, the correlation is null for all lags (data not shown). This implies that any correlations in closely spaced promoters are not originated by the rate limiting steps. If the number of repressors was from one to a few, a spurious anti-correlation would appear, as the repression of one TSS would diminish the chance of repression of the other at the same time.

Next, we model pairs of closely spaced promoters. We simulate the models and compute the correlation between temporal choices for all lags. In all cases, the binding site of the repressor is at midpoint between the two TSS, so as to not generate spurious correlations due to biases in the expression levels. One model is a divergent (□ in Fig. 11) and the other is a convergent promoter (○ in Fig. 11). The models were simulated both with (in black in Fig. 11) and without repressors (in gray in Fig. 11).

From Fig. 11, comparing the models with and without repressors it is visible that, for both the divergent and the convergent promoter, the mutual repression mechanism correlates, in a positive fashion, the consecutive choices of production of RNAs. In the convergent, the correlation goes from almost null to positive, while in the overlapping divergent promoter it decreases the inherent negative correlation.

In non-overlapped divergent promoters, repression also increases the inherent negative correlation (data not shown). This occurs if there are two binding sites and two distinct repressor molecules (one for each TSS) each of which with a binding site, or if the binding sites overlap, causing the repression of one to hamper the repression of the other TSS. This suggests that complex repression mechanisms, via the use of multiple repressors and binding sites configurations, likely allow various degrees of correlation between the transcription kinetics of adjacent genes driven by closely spaced promoters.

3.6. Comparison between the model's predictions and measurements of cell–cell heterogeneity

From the simulations described above, particularly the intervals between the productions of consecutive RNA molecules, one can conclude that unidirectional and divergent promoters with a large distance between TSSs (i.e., non-overlapping) will generate the least noisy time series of RNA numbers. The amount of noise in the RNA numbers over time can be assessed by the square of the coefficient of variation, CV^2, of the time intervals between consecutive RNA productions (Pedraza and Paulsson, 2008) whose distributions are shown in Fig. 6. Specifically, the distributions from non-overlapping divergent promoters, namely A and B, have CV^2 values equal to 0.34 and 0.38, respectively, while the unidirectional promoter F has a CV^2 of 0.34. On the other hand, the divergent overlapping promoter C and the convergent promoter E have CV^2 of 3.97 and 0.84, indicating much higher noise in the kinetics of RNA production.

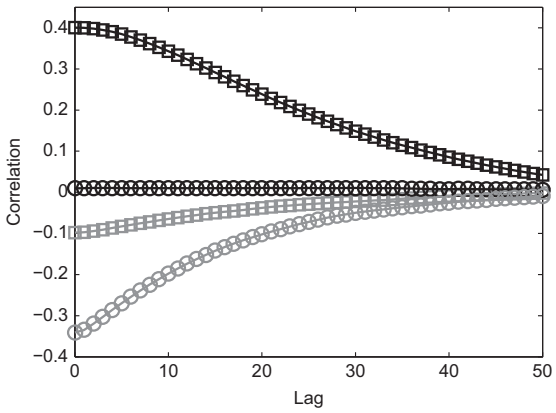


Fig. 11. Correlations in consecutive RNAs at different lags. Correlations in consecutive RNAs produced at different lags for different a divergent promoter with 65 nucleotides between TSSs (○) and an overlapped divergent promoter with 150 nucleotides between TSSs (□). Lines are gray in the cases of no repression and black when there are repressors in the system.

These results can, to some extent, be compared to measurements of gene expression dynamics in *E. coli*. Recently, a quantitative system-wide analysis of protein and mRNA expression was carried out in individual cells with single-molecule sensitivity using a yellow fluorescent protein fusion library for *E. coli* (Taniguchi et al., 2010). From this data, we flagged the divergent and convergent promoters for comparison. In particular, we compared our predictions regarding noise in transcript production between closely spaced promoters (divergent and convergent) versus unidirectional TSSs with the measurements of the cell-to-cell diversity in RNA numbers from these two sets of promoters. We found no significant differences in the kinetics of RNA production between these two sets of genes from the data in Taniguchi et al. (2010). Further, we found no correlation between the distance between TSSs and the noise levels in bidirectional promoters from the same data. In supplementary material we show the calculations that were made to determine the possible differences in the kinetics of RNA production as well as to determine possible correlations between distances and noise levels.

However, we note that the measurements of cell-to-cell diversity in RNA numbers in Taniguchi et al. (2010) are not the most informative of noise in transcription. Evidence suggests that complex mechanisms of RNA degradation (see e.g. Yarchuk et al., 1992; Taniguchi et al., 2010) may significantly affect the cell-to-cell diversity in RNA numbers. Another mechanism that may affect the cell-to-cell diversity in RNA numbers in a population is errors in the partitioning of RNA molecules in cell division. Stochasticity in this process will enhance the diversity even when the partitioning is unbiased (Huh and Paulsson, 2011a, 2011b). If biases exist, this effect will have an even stronger impact on the levels of diversity of RNA numbers (Lloyd-Price et al., 2012). Such biases are likely to exist and may vary from one RNA sequence to the next, particularly given the recent evidence that the location of the RNA molecules in *E. coli* is far from arbitrary (Llopis et al., 2010).

Finally, it is of relevance to note that the mean RNA levels in the measurements reported in Taniguchi et al. (2010) are between $\sim 10^{-3}$ and ~ 5 molecules per cell, as assessed by RNA-seq, and between $\sim 10^{-2}$ and ~ 5 per cell, as assessed by FISH. This suggests very small rates of RNA production in optimal growth conditions for most genes (the genes were not subject to any artificial induction). From our model predictions, for such rates of RNA production, one does not expect observable differences in the kinetics of RNA production between bidirectional and unidirectional promoters or with different distances between TSSs (see Fig. 9).

To test whether promoter geometry affects, to some degree, the fluctuations in RNA numbers, it is necessary to measure time intervals between the consecutive productions of RNA molecules, e.g. using an MS2-GFP-based RNA tagging technique (Fusco et al., 2003; Golding et al., 2005). Such measurements have already been reported for the *lar* promoter in Kandhavelu et al. (2011). Using this technique, we must engineer genes with promoters with different geometries followed by an elongation region that codes for MS2-GFP binding sites. These genes ought to be driven

by promoters strong enough, when induced, so that the production rates are sufficient to allow detection of differences in the stochasticity of the production mechanism. Note that, with this method, one can detect RNA molecules as soon as they are produced (Golding and Cox, 2004), implying that production rates of the order of 4–10 RNAs per hour may suffice to detect differences in the kinetics of production of RNAs under the control of promoters with different geometries.

4. Conclusions

We studied the dynamics of expression of pairs of genes driven by closely spaced promoters within realistic intervals of parameter values for *E. coli*. For that, we used a delayed stochastic model that mimics transcription initiation in *E. coli*, one RNAp and one nucleotide at a time. From the simulations, in general, we find that changing the sequence between the two TSS and the kinetics of the closed and open complex formation at each TSS allows pairs of genes with overlapping promoters to have widely diverse kinetics of RNA production, and complex dynamics of RNA production not easily achievable by sets of genes that interact via transcription factors alone.

In general, the rate limiting steps at the two TSSs are responsible for a degree of anti-correlation between consecutive choices of which of the two RNAs are transcribed next. The blocking of a TSS by an RNAp during the rate limiting steps (isomerization and open complex formation) makes more likely the choice of the other RNA to be produced next. However, if the distance between the two TSS is below a certain number of nucleotides, the opposite occurs.

The simulations showed that the sequence of the promoter (nucleotide length, kinetics of rate limiting steps, etc.) significantly affects the dynamics of RNA production. In that sense, this dynamics is sequence dependent and thus subject to selection. If the kinetics of RNA production is subject to selection, the arrangement of closely spaced promoters is also subject to selection, since promoters with different geometries were shown to have widely diverse kinetics of RNA production. Further, given the observed ranges of variability in the distributions of intervals between the productions of consecutive RNA molecules as a function of the configuration of the promoter, it is expected, provided first order degradation rates of RNA molecules, that cell-to-cell diversity in RNA numbers may range from sub- to supra-poissonian as a function of the kinetics of transcription initiation. We verified this (data not shown) by calculating the Fano factor in RNA numbers in different cells at near equilibrium.

One interesting outcome of the results is that they provide a means to, from the behavior analysis, define the meaning of 'closely spaced promoter' which, at the moment, still has a rather loose definition. From Figs. 8 and 9, regarding divergent promoters, we observe a sharp behavioral change (in mean and correlation) when the distance between TSSs increases from 100 to 200 nucleotides. Provided future experimental validation of this result, this may allow define as 'closely spaced' promoters, for this geometry, those that are separated by less than 100 nucleotides. In the case of convergent promoters there is a strong change in the mean RNA levels (Fig. 9) when the nucleotides distance increases beyond 25 nucleotides and thus this distance could be used as a means to define closely spaced convergent promoters.

As a side note, the discrete nature of the probabilities of binding of the RNAp to the nucleotides close to the two TSS may be of significance. It suggests that the location of binding sites for repressors and activators relative to the TSSs is likely to be far from random, as the location will determine the overall probabilities of freely diffusing RNAps to reach either TSS.

The study of the effects of repressors showed that these can be a means to achieve complex patterns of behaviors, not possible otherwise. Their effect depends on the geometry of the promoter and on the length of the sequences between the two TSS. Further, it depends on the location of the binding sites. For example, placing the binding site between the two TSSs, but closer to one of them, biases the mean expression levels of the genes. Finally, depending on these parameters, repression by occlusion can correlate or anti-correlate the two RNAs levels.

Our results show that repression at different stages of transcription can lead to similar as well as distinct kinetics of RNA numbers, depending on several factors such as the number of repressor molecules in the cell. In general, for the same number of repressors and binding affinity, the effects on RNA production differ with the stage of transcription that is repressed. Recent observations are in agreement. When Schlax et al. (1995) studied the repression kinetics they concluded that the most probable down-regulation mechanism is the inhibition of closed complex formation. However, the diversity of regulatory mechanisms and broad distribution of locations of binding sites of repressors relative to the TSSs (Garcia et al., 2010) suggests that, repression occurs at different stages in different genes, including during the promoter escape.

It is of interest to compare our results regarding repression with those of Wang et al. (2011) from a single step stochastic model of eukaryotic gene expression. In this work, the authors postulate that, in divergent promoters, the expression of one gene facilitates the expression of the other by preventing chromatin compaction. In our case, a similar effect takes place, in that the expression of one of the genes tends to prevent the binding of the repressor between the TSSs, provided close proximity between the TSS and the repressor binding site.

Much effort has been given to the engineering of artificial genetic circuits (Elowitz and Leibler, 2000; Gardner et al., 2000). *E. coli* is one of the model organisms used in these studies. Most circuits engineered so far rely on commonly used promoters, such as lac, tet, and ara (McClure, 1985). These are used because their sequences are well characterized and the regulatory molecules are known (Lutz et al., 2001). So far, little attention has been given to the panoply of native promoters in *E. coli* and other organisms. A genome-wide characterization of the kinetics of the endogenous promoters in *E. coli* is likely to show that there is a much wider range of dynamical behaviors than those observed so far. A more complete survey of the state space of dynamical behaviors at the promoter level will aid the engineering of novel circuits that will be able to perform far more complex dynamical behaviors than what is presently possible. The applicability of these circuits will thus be much enhanced. Additionally, if these circuits can be engineered so as to be more tightly coupled with the native genetic networks, they can make use of the native regulatory mechanisms of the dynamics of expression in the host cells. Our study aimed to aid in this effort by assessing the range of dynamical behaviors possible by varying the geometry and structure of closely spaced promoters. Finally, we hypothesize that the multitude of regulatory steps of the dynamics of RNA production not only partially explains the observed diversity of kinetic behavior of genes in *E. coli*, but it also suggests that different mechanisms can be used to attain similar kinetics of RNA production, thereby allowing for the emergence of neutral evolutionary pathways.

Author contributions

A.S.R. conceived the study and supervised the interpretation of data. O.Y.H., J.M.F., and M.K. aided in the conception of the study

and interpretation of the data. J.M. and A.S.R. conceived the model. L.M. and AH implemented the model. L.M. and J.M. did most simulations and data acquisition. All authors performed research. A.S.R. and J.M. drafted the manuscript that was revised by all authors.

Acknowledgments

Work supported by the Academy of Finland (A.S.R., A.H., M.K.), and the FiDiPro program of Finnish Funding Agency for Technology and Innovation (J.M., A.S.R., A.H., O.Y.-H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Jason Lloyd-Price for valuable advice and discussions.

Appendix A. supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2012.02.015.

References

Arkin, A., Ross, J., McAdams, H.H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics 149 (4), 1633–1648.

Beck, C.F., Warren, R.A., 1988. Divergent promoters, a common form of gene organization. Microbiol. Rev. 52 (3), 318–326.

Bernstein, J.A., Khodursky, A.B., Lin, P., Lin-Chao, S., Cohen, S.N., 2002. 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 99 (15), 9697–9702.

Bremer, H., Dennis, P., Ehrenberg, M., 2003. Free RNA polymerase and modeling global transcription in *Escherichia coli*. Biochimie 85, 597–609.

Browning, D.F., Busby, S.J.W., 2004. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2, 57–65.

Buc, H., McClure, W.R., 1985. Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps. Biochemistry 24 (11), 2712–2723.

Callen, B.P., Shearwin, K.E., Egan, J.B., 2004. Transcriptional interference between convergent promoters caused by elongation over the promoter. Mol. Cell 14, 647–656.

Carey, J., Lewis, D.E., Lavoie, T.A., Yang, J., 1991. How does trp repressor bind to its operator? J. Biol. Chem. 266 (36), 24509–24513.

deHaseth, P.L., Zupancic, M.L., Record Jr, M.T., 1998. RNA polymerase–promoter interactions: the comings and goings of RNA polymerase. J. Bacteriol. 180 (12), 3019–3025.

Ebisuya, M., Yamamoto, T., Nakajima, M., Nishida, E., 2008. Ripples from neighbouring transcription. Nat. Cell Biol. 10, 1106–1113.

Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403, 335–338.

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S.M., Blanchard, J.-M., Singer, R.H., Bertrand, E., 2003. Single mrna molecules demonstrate probabilistic movement in living mammalian cells. Curr. Biol. 13 (2), 161–167.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J.S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A.M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., Collado-Vides, J., 2010. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res. 39, 98–105.

Gardner, T.S., Cantor, C.R., Collins, J.J., 2000. Construction of a genetic toggle switch in *Escherichia coli*. Nature 403, 339–342.

Garcia, H.G., Sanchez, A., Kuhlman, T., Kondev, J., Phillips, R., 2010. Transcription by the numbers redux: experiments and calculations that surprise. Trends Cell Biol. 20 (12), 723–733.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81 (25), 2340–2361.

Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C., 2005. Real-time kinetics of gene activity in individual bacteria. Cell 123 (6), 1025–1036.

Golding, I., Cox, E.C., 2004. Rna dynamics in live *Escherichia coli* cells. Proc. Natl. Acad. Sci. USA 101 (31), 11310–11315.

Gorman, J., Greene, E.C., 2008. Visualizing one-dimensional diffusion of proteins along DNA. Nat. Struct. Mol. Biol. 15 (8), 768–774.

Häkkinen, A., Healy, S., Jacobs, H.T., Ribeiro, A.S., 2011. Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse. J. Theor. Biol. 281 (1), 74–83.

Horton, N., Lewis, M., Lu, P., 1997. *Escherichia coli* lac repressor-lac operator interaction and the influence of allosteric effectors. J. Mol. Biol. 265 (1), 1–7.

Hsu, L.M., 2002. Promoter clearance and escape in prokaryotes. Biochim. Biophys. Acta 1577, 191–207.

Huh, D., Paulsson, J., 2011a. Random partitioning of molecules at cell division. Proc. Acad. Natl. Sci. USA 108 (36), 15004–15009.

Huh, D., Paulsson, J., 2011b. Non-genetic heterogeneity from stochastic partitioning at cell division. Nat. Genet. 43 (2), 95–100.

Kandhavelu, M., Mannerström, H., Yli-Harja, O., Ribeiro, A.S., 2011. In vivo kinetics of transcription initiation of the lar promoter in *Escherichia coli*. Evidence for a sequential mechanism with two rate limiting steps. BMC Syst. Biol. 5, 149.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., Lu, P., 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 271 (5253), 1247–1254.

Lopez, P.J., Guillerez, J., Sousa, R., Dreyfus, M., 1998. On the mechanism of inhibition of phage T7 RNA polymerase by lac repressor. J. Mol. Biol. 276 (5), 861–875.

Llopis, P.M., Jackson, A., Sliusarenko, O., Surovtsev, I., Heinritz, J., Emonet, T., Jacobs-Wagner, C., 2010. Spatial organization of the flow of genetic information in bacteria. Nature 466, 77–81.

Lloyd-Price, J., Lehtivaara, M., Kandhavelu, M., Chowdhury, S., Muthukrishnan, A.-B., Yli-Harja, O., Ribeiro, A.S., 2012. Probabilistic RNA partitioning generates transient increases in the normalized variance of RNA numbers in synchronized populations of *Escherichia coli*. Mol. BioSyst. 8 (2), 565–571.

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

McClure, W.R., 1985. Mechanism and control of transcription initiation in prokaryotes. Annu. Rev. Biochem. 54, 171–204.

Pedraza, J.M., Paulsson, J., 2008. Effects of molecular memory and bursting on fluctuations in gene expression. Science 319, 339–343.

Ribeiro, A.S., Lloyd-Price, J., 2007. SGN Sim, a stochastic genetic networks simulator. Bioinformatics 23 (6), 777–779.

Ribeiro, A.S., Häkkinen, A., Mannerström, H., Lloyd-Price, J., Yli-Harja, O., 2010. Effects of the promoter open complex formation on gene expression dynamics. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 81 (1 Part 1), 011912.

Roussel, M.R., Zhu, R., 2006. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. Phys. Biol. 3, 274–284.

Saecker, R.M., Record Jr, M.T., deHaseth, P.L., 2011. Mechanism of bacterial transcription initiation: RNA polymerase–promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J. Mol. Biol. 412 (5), 754–771.

Sakata-Sogawa, K., Shimamoto, N., 2004. RNA polymerase can track a DNA groove during promoter search. Proc. Natl. Acad. Sci. USA 101 (41), 14731–14735.

Sanchez, A., Osborne, M.L., Friedman, L.J., Kondev, J., Gelles, J., 2011. Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules. EMBO J. 30, 3940–3946.

Schlax, P.J., Capp, M.W., Record Jr, M.T., 1995. Inhibition of transcription initiation by lac repressor. J. Mol. Biol. 245 (4), 331–350.

Singer, P., Wu, C.-H., 1987. Promoter search by *Escherichia coli* RNA polymerase on a circular DNA template. J. Biol. Chem. 262 (29), 14178–14189.

Singh, S.S., Typas, A., Hengge, R., Grainger, D.C., 2011. Escherichia coli σ70 senses sequence and conformation of the promoter spacer region. Nucleic Acids Res. 39 (12), 5109–5118.

Sneppen, K., Dodd, I.B., Shearwin, K.E., Palmer, A.C., Schubert, R.A., Callen, B.P., Egan, J.B., 2005. A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. J. Mol. Biol. 346, 399–409.

So, L.H., Ghosh, A., Zong, C., Sepúlveda, L.A., Segev, R., Golding, I., 2011. General properties of transcriptional time series in *Escherichia coli*. Nat. Genet. 43 (6), 554–560.

Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M., Wishart, D.S., 2004. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *Escherichia coli*. Nucleic Acids Res. 32 (D293), D295.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., Xie, X.S., 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533–538.

Wang, G.-Z., Lercher, M.J., Hurst, L.D., 2011. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. Genome Biol. Evol. 3, 320–331.

Woo, Y.H., Li, W.-H., 2011. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. Proc. Natl. Acad. Sci. USA 108 (8), 3306–3311.

Yarchuk, O., Jacques, N., Guillerez, J., Dreyfus, M., 1992. Interdependence of translation, transcription and mRNA degradation in the lacZ Gene. J. Mol. Biol. 266, 581–596.

Supplementary Material for "Dynamics of transcription of closely spaced promoters in Escherichia coli, one event at a time".

Leonardo Martins, Jarno Mäkelä, Antti Häkkinen, Meenakshisundaram Kandhavelu, Olli Yli-Harja, José M. Fonseca and Andre S. Ribeiro

1. Stochastic Model of Closely Spaced Promoters at the Nucleotide Level

The delayed stochastic model of transcription initiation at the nucleotide level includes the non-specific binding of RNA polymerase (RNAp) to the DNA, search by diffusion for transcription start sites (TSSs), rate limiting steps leading to the open complex formation at the TSS, abortive initiations and productive elongation. In addition, it accounts for changes in the footprint of the RNAp while diffusing, at the TSS and when elongating [1-3]. When referring to closed complex formation, we mean the steps that occur after the finding of the TSS, but prior to the start of the open complex formation [1,4].

Most reactions in the model are instantaneous, i.e. once the two reacting molecules meet and react, the product is released instantaneously. Instantaneous reactions are represented as: $A+B \xrightarrow{k} C$. In this reaction, when A and B meet according to the rules of the stochastic simulation algorithm (SSA) [5], molecule C is produced instantaneously. The expected time for A and B to meet is determined by the propensity of this reaction at each moment, given by the product of k, the number of A molecules and the number of B molecules [5].

Some reactions need to account for the time the process takes to occur, once initiated. Such delays in the release of products are represented as follows: $A+B \xrightarrow{k} C(\tau)$. When this reaction occurs, C is placed on a waitlist and only made available for further reactions after τ seconds have elapsed. τ can be generated randomly from any desired distribution each time the reaction is chosen to occur. Such delayed events are only introduced when the time that the process takes to occur is sufficiently long to affect the kinetics of the system. Finally, note that in some reactions, not all necessary reactants for the occurrence are consumed by the reaction. Substrates that are not consumed are indicated by an (*).

The model of transcription has four main components: RNAps, repressor molecules when existing, and DNA and RNA sequences, modeled at the nucleotide level. The promoter sequence contains TSSs and binding sites (BS) for repressors. This model can be coupled to the model of transcription and translation elongations at the nucleotide and codon level proposed in [6]. For simplicity, here we only present the model of initiation. The reactions, stochastic rate constants and time delays, are shown in Table S1. Variables used in reactions are shown in Table S2.

RNAps freely diffusing in the cell are allowed to bind to any nucleotide of the promoter region, provided that it is free. This occurs via reaction (1). The strand to which it binds determines the direction of diffusion, which does not change until the RNAp unbinds the DNA template. The RNAp can unbind via reaction (2) at any stage of difusion.

The model accounts for the footprint of the RNAp at each stage. Ranges of nucleotides are denoted in square brackets, e.g. $U_{[start,end]}$. Footprint studies [3,7,8] indicate that a bound RNAp, while diffusing, occupies ~55 nucleotides. Such occupied nucleotides are named O_n, where n denotes its number, while U_n stands for the n^{th} unoccupied nucleotide. To refer to a position of an RNAp we use the nucleotide where its active center is at, while the range occupied by the RNAp is referred to as $[n-\Delta_D, n+\Delta_D]$, where $\Delta_D = 27$.

Once an RNAp binds to the DNA, it diffuses on the template one nucleotide at a time, provided that the nucleotides are available (3). If the path is blocked by another RNAp or a repressor, it will eventually dissociate from the DNA strand via (2) [4].

When the RNAp finds the specific TSS, a chain of events takes place such as the closed complex formation (4) and isomerization (5) [1,2]. At this stage, the RNAp structure changes and occupies more nucleotides (~75) [8,9]. Next, the open complex formation occurs (6) [10].

The model accounts for collisions between elongating complexes (EC) and diffusing RNAps (8) and between two ECs (9). The former causes diffusing RNAps to disassociate from the DNA and the latter disassociates one or both ECs from the DNA. In collisions between ECs and diffusing RNAps (8), the EC remains in the template and the diffusing RNAp dissociates from the template. A similar reaction models collisions between diffusing RNAps (7), where one or both RNAps dissociate from the template. Finally, we model the "sitting duck" mechanism (10) [11]. When an EC collides with a promoter complex (e.g. open complex), since the EC is tightly bound to the DNA, the complex is removed.

Reaction (11) models the formation of elongating complexes, while (12) models the initial steps in elongation, during which the RNAp "scrunches" the DNA [12,13] until enough energy is accumulated for the RNAp to escape the promoter (14). Prior to this, abortive initiation events (13) can occur, which causes the RNAp to return to the open complex state. We set the rate of abortive initiations to 4.2 s^{-1} to be within the ranges reported in [14].

For simplicity we only allow escape after the scrunching of the 12th nucleotide, although this differs from gene to gene, and from one event to the next [15]. As soon as the RNAp escapes the promoter (14), productive elongation events can occur (15). Also, the TSS becomes available for other RNAps.

During elongation, the EC (named E_n in the model) occupies 25 nucleotides [3,16]. The range is represented as $[n-\Delta_E, n+\Delta_E]$, where $\Delta_E=12$. Elongation (16) is modeled as a delayed reaction, with the delay for the production of a RNA molecule following a Gamma distribution (resultant from the composition of many sequential exponential distributions with the same mean). In (16), k equals the number nucleotides and θ, the rate of elongation, equaling 42 s^{-1} per nucleotide [17]. Finally, reaction (17) models RNA degradation as a single step reaction [18].

Repression is modeled via (18). It can compete with RNAp binding and, once the repressor is bound, it blocks the RNAp movement. Dissociation of the repressor from the template is modeled by (19). The footprint of the repressor is $[n-\Delta_{rep}, n+\Delta_{rep}]$, where Δ_{rep} is 10, within realistic intervals of footprints of repressors in *E. coli* [19]. The rate constants for the reactions associate with repression are from [20].

Table S1. Reactions modeling transcription initiation, elongation, RNA degradation, repression and unrepression. Reactions, rate constants (in s^{-1}), and delays (in *s*) used to model transcription initiation, elongation, repression, and RNA degradation. Parameter values were obtained from measurements in *E. coli* [4,10-14,17,18,20].

Event	**Reaction**	**Rate constant and delays**
Binding (1)	$RNAp + U_{[n-\Delta_D,\, n+\Delta_D]} \xrightarrow{k_b} O_n$	$k_b = 0.000075\ s^{-1}$
Unbinding (2)	$O_n \xrightarrow{k_f} RNAp + U_{[n-\Delta_D,\, n+\Delta_D]}$	$k_f = 0.3\ s^{-1}$
Diffusion (3)	$O_n + U_{n+\Delta_D+1} \xrightarrow{k_m} O_{n+1} + U_{n-\Delta_D}$	$k_m = 660\ s^{-1}$
Completion of closed complex (4)	$O_{TSS+\Delta_D} \xrightarrow{k_c} RP_c$	$k_c = 0.5\ s^{-1}$
Isomerization (5)	$RP_c + U_{[TSS+1,\, TSS+19]} \xrightarrow{k_i} RP_i$	$k_i = 0.095\ s^{-1}$
Open complex formation (6)	$RP_i \xrightarrow{k_o} RP_o$	$k_o = 2\ s^{-1}$
Collisions between diffusing RNAps (7)	$*O_{n+2\Delta_D+1} + O_n \xrightarrow{k_m} U_{[n-\Delta_D,\, n+\Delta_D]} + RNAp$	$k_m = 660\ s^{-1}$

Collisions between diffusing and elongating RNAps (8)	$^{*}E_{n+2\Delta_E+1} + O_n \xrightarrow{k_m} U_{[n-\Delta_D, n+\Delta_D]} + RNAp$	$k_m = 660\ s^{-1}$
Collisions between elongation RNAps (9)	$^{*}E_{n+2\Delta_E+1} + E_n \xrightarrow{k_{el}} U_{[n-\Delta_E,\ n+\Delta_E]} + RNAp$	$k_{el} = 42\ s^{-1}$
Collision between RNAps elongating and at the promoter (10)	$^{*}E_{TSS-\Delta_E} + RP_c \xrightarrow{k_{el}} U_{[TSS,\ TSS-2\Delta_D]}$, $^{*}E_{TSS} + RP_i/RP_o/E_{TSS-12} \xrightarrow{k_{el}} U_{[TSS+\Delta_E,\ TSS-2\Delta_D]}$	$k_{el} = 42\ s^{-1}$
Transcription complex formation (11)	$RP_o \xrightarrow{k_{el}} E_{TSS}$	$k_{el} = 42\ s^{-1}$
Initial elongation (Scrunching) (12)	$E_{TSS+n} \xrightarrow{k_{el}/4} E_{TSS+n+1}$	$k_{el} = 42\ s^{-1}$, $n \leq 12$
Abortive initiation (13)	$E_{TSS+n} \xrightarrow{k_a} RP_o$	$k_a = k_{el}/10\ s^{-1}$
TSS clearance (14)	$E_{TSS+12} + U_{TSS+\Delta_E+12} \xrightarrow{k_{el}} E_{TSS+13} + U_{[TSS+12,\ TSS+2\Delta_D+12]}$	$k_{el} = 42\ s^{-1}$
Elongation (15)	$E_n + U_{n+\Delta_E} \xrightarrow{k_{el}} E_{n+1} + U_{n-\Delta_E}$	$k_{el} = 42\ s^{-1}$, $n \geq 13$
RNA production (16)	$E_{n_{last}} \xrightarrow{k_{el}} RNA(\tau_{el}) + RNAp(\tau_{el}) + U_{[n_{last}-2\Delta_E,\ n_{last}]}$	$k_{el} = 42\ s^{-1}$ $\tau_{el} = G(e_x, k_{el}^{-1})$ e_x = no. nuc.
Degradation (17)	$RNA \xrightarrow{k_d} \varnothing$	$k_d = 0.006\ s^{-1}$
Repression (18)	$Rep + U_{[n-\Delta_{rep},\ n+\Delta_{rep}]} \xrightarrow{k_r} R_n$	$k_r = 0.0167\ s^{-1}$
Unrepression (19)	$R_n \xrightarrow{k_u} Rep + U_{[n-\Delta_{rep},\ n+\Delta_{rep}]}$	$k_u = 0.004\ s^{-1}$

Table S2. Description of the variables used in the model of gene expression.

Variable	Description
U_n	Free nucleotide at location *n*
O_n	Occupied nucleotide at location *n*
E_n	Elongation complex (EC) at location *n*
R_n	Nucleotide occupied by repressor at location *n*
$U_{[start, end]}$	Ranges of nucleotides are denoted in square brackets
Δ_D	Half-range occupied by the RNAp in diffusion. Total range is $[n-\Delta_D, n+\Delta_D]$, where $\Delta_D = 27$
Δ_E	Half-range occupied by the RNAp in elongation. Total range is $[n-\Delta_E, n+\Delta_E]$, where $\Delta_E=12$
*	Indicates substrates not consumed in the reaction
RNAp	RNA polymerase
RP_c	Closed complex
RP_i	Isomerized complex
RP_o	Open complex
TSS	Transcription start site
Rep	Repressor molecule
$RNA(\tau_{el})$	RNA substrate is released with delay (τ_{el})

The models are simulated by the SGNSim Simulator [21]. This simulator and the manual for its usage can be found in: http://www.cs.tut.fi/~sanchesr/SGN/SGNSim.html. SGNSim makes use of a reactions file, where the rate constants, the initial amount of each substance, and the chemical reactions of the model are specified. An example of such a reactions file is provided in a text file named "Reactionsfile_bidirectional.g". In this example, a divergent promoter with 150 nucleotides between TSSs is modeled and the two TSS locations are at nucleotides 51 and 200.

2. Statistical Analysis

It is possible to compare our predictions regarding noise in transcript production between closely spaced promoters (divergent and convergent) and unidirectional TSSs with measurements of the cell-to-cell diversity in RNA numbers from these two sets of promoters. Namely, we assessed from the data in (Taniguchi et al, 2010) whether the noise levels in RNA numbers, as measured by the square of the coefficient of variation (CV^2), differed between closely spaced promoters and the others. By closely spaced promoters we refer to pairs of promoters that are less than 500 nucleotides apart.

We thus assessed the null hypothesis (that the sets originate from the same distribution) by the two-sample Kolmogorov-Smirnov (K-S) test (Matlab 2007a). The samples sizes are 38 for closely spaced promoters and 99 for unidirectional promoters. Setting α to 0.01, the K-S test confirms the goodness of fit between the two sample distributions (p-value equal to 0.67). Thus, we conclude that there are no significant differences in the kinetics of RNA production between these two sets of genes from the data in (Taniguchi et al, 2010).

Next, we assessed whether there is any correlation between the distance between the two TSSs and the observed noise in RNA numbers, as measured by the CV^2 of the RNA numbers in individual cells. For this, we selected all pairs of promoters' separated by less than 500 nucleotides and calculated the Pearson correlation between the length and the CV^2 values. The Pearson correlation coefficient of -0.1004 is indicative that these measurements do not detect a strong correlation between the distance between TSSs and the CV^2.

References

1. Bai L, Santangelo TJ, Wang MD (2006) Single-molecule analysis of RNA polymerase transcription. Annu Rev Biophys Biomol Struct 35: 343–360.
2. McClure WR (1985) Mechanism and control of transcription initiation in prokaryotes. Ann Rev Biochem 54: 171-204.
3. Metzger W, Schickor P, Heumann HA (1989) Cinematographic view of Escherichia coli RNA polymerase translocation. The EMBO Journal 8(9): 2745-2754.
4. Singer P, Wu C-H (1987) Promoter Search by Escherichia coli RNA polymerase on a Circular DNA Template. The Journal of Biological Chemistry 262(29): 14178-14189.
5. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25): 2340-2361.

6. Mäkelä J, Lloyd-Price J, Yli-Harja O, Ribeiro AS (2011) Stochastic sequence-level model of coupled transcription and translation in prokaryotes. BMC Bioinformatics 12: 121.
7. Carpousis AJ, Gralla JD (1985) Interaction of RNA polymerase with lacUV5 promoter DNA during mRNA initiation and elongation. Footprinting, methylation, and rifampicin-sensitivity changes accompanying transcription initiation. J Mol Biol 183(2): 165–177.
8. Record MT Jr, Reznikoff WS, Craig ML, McQuade KL and Schlax PJ (1996) Escherichia coli RNA Polymerase (Eσ70), Promoters, and the Kinetics of the Steps of Transcription Initiation. Second Edition of Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology 2: 792-821.
9. Hsu LM (2002) Promoter clearance and escape in prokaryotes. Biochimica et Biophysica Acta 1577: 191– 207.
10. Buc H, McClure WR (1985) Kinetics of Open Complex Formation between Escherichia coli R N A Polymerase and the lac UV5 Promoter. Evidence for a Sequential Mechanism Involving Three Steps. Biochemistry 24(11): 2712-2723.
11. Callen BP, Shearwin KE, Egan JB (2004) Transcriptional Interference between Convergent Promoters Caused by Elongation over the Promoter. Molecular Cell 14: 647–656.
12. Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, Ebright RH (2006) Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science 314(5802): 1144-1147.
13. Revyakin A, Liu C, Ebright RH, Strick TR (2006) Abortive Initiation and Productive Initiation by RNA polymerase Involve DNA Scrunching. Science 17 314(5802): 1139-1143.
14. Hsu LM, Cobb IM, Ozmore JR, Khoo M, Nahm G, et al. (2006) Initial transcribed sequence mutations specifically affect promoter escape properties. Biochemistry 45(29): 8841-8854.
15. Xue X, Liu F and Ou-Yang Z (2008) A Kinetic Model of Transcription Initiation by RNA Polymerase. J Mol Biol 378: 520–529.
16. Greive SJ, von Hippel PH (2005) Thinking quantitatively about transcriptional regulation. Nature Reviews Molecular Cell Biology 6: 221-232.
17. Phroskin S, Rachid Rahmouni A, Mironov A, and Nudler E (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328(5977): 504-508.
18. Bernstein JA, Khodursky AB, Lin P, Lin-Chao S, Cohen SN (2002) 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 99(15): 9697-9702.
19. Carey J, Lewis DE, Lavoie TA, Yang J (1991) How does trp repressor bind to its

operator? J Biol Chem 266(36): 24509-24513.
20. So LH, Ghosh A, Zong C, Sepúlveda LA, Segev R, Golding I (2011) General properties of transcriptional time series in Escherichia coli. Nat Genet 43(6): 554-560.
21. Ribeiro AS and Lloyd-Price J (2007) SGN Sim, a Stochastic Genetic Networks Simulator. Bioinformatics 23(6): 777-779.

Publication IV

J. Mäkelä, J. Lloyd-Price, O. Yli-Harja, and A.S. Ribeiro, "Stochastic sequence-level model of coupled transcription and translation in prokaryotes", *BMC Bioinformatics* 12:121, 2011.

Mäkelä et al. BMC Bioinformatics 2011, 12:121
http://www.biomedcentral.com/1471-2105/12/121

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Stochastic sequence-level model of coupled transcription and translation in prokaryotes

Jarno Mäkelä[1], Jason Lloyd-Price[1], Olli Yli-Harja[1,2] and Andre S Ribeiro[1*]

Abstract

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

Results: First, we show that the model accurately matches measurements of sequence-dependent translation elongation dynamics. Next, we characterize the degree of coupling between fluctuations in RNA and protein levels, and its dependence on the rates of transcription and translation initiation. Finally, modeling sequence-specific transcriptional pauses, we find that these affect protein noise levels.

Conclusions: For parameter values within realistic intervals, transcription and translation are found to be tightly coupled in *Escherichia coli*, as the noise in protein levels is mostly determined by the underlying noise in RNA levels. Sequence-dependent events in transcription elongation, e.g. pauses, are found to cause tangible effects in the degree of fluctuations in protein levels.

Background

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

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

Two recent experiments have given a preliminary glimpse at the dynamics of production of individual proteins [13] and RNA molecules [14] *in vivo* in bacteria. However, as of yet, there is no experimental setting to simultaneously observe the production of both RNA and proteins at the molecular level. Further, in the aforementioned experiments [13,14], the rate of gene expression was kept very weak, as otherwise the number of

* Correspondence: andre.ribeiro@tut.fi
[1]Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland
Full list of author information is available at the end of the article

molecules would not be easily quantifiable. This implies that they cannot be used to study the effects of events such as the promoter open complex formation [15]. The present shortcomings of these techniques enhance the need for realistic models of gene expression in prokaryotes.

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

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

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

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

Results and discussion

Dynamics of transcript production

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

In this simulation, transcription is not repressed. Thus, provided that the promoter is available for transcription, the expected time for a transcription event to start is approximately 2.5 s, given the value of the rate constant of reaction (1) in Table 1 and that there are 28 RNAp molecules available in the system [15]. The promoter open complex formation step, with a mean duration of 40 s [25] and a standard deviation of 4 s [21] is the major limiting factor of transcription events in these conditions.

Figure 1A shows, for a time window of 400 seconds, the positions (y-axis) over time (x-axis) of several RNAp molecules on the DNA template. In real time, this simulation takes ~30 s, on an Intel Core 2 Duo processor. Transcription elongation is visibly stochastic, with events such as arrests (e.g. at t = ~450 s), ubiquitous pauses and pyrophosphorolysis. Several collisions between RNAp molecules are also visible, caused in part by these events. Note that one RNAp never overtakes another on the template.

Figure 1B shows the distribution of the time intervals between transcription initiation events, which is Gaussian-like, due to the open complex formation step. The

Table 1 Reactions modeling transcription

Event	Reaction	Rate constant	Ref.
Initiation and promoter complex formation (1)	$\text{Pro} + \text{RNAp} \xrightarrow{k_{init}} \text{RNAp} \bullet \text{Pro}(\tau_{oc})$	$k_{init} = 0.015$ $\tau_{oc} = 40 \pm 4$	[21]
Promoter clearance (2)	$\text{RNAp} \bullet \text{Pro} + \text{U}_{[1,\Delta_{\text{RNAp}}+1]} \xrightarrow{k_m} \text{O}_1 + \text{Pro}$	$k_m = 114$	[37]
Elongation (3)	$\text{A}_n + \text{U}_{n+\Delta_{\text{RNAp}}+1} \xrightarrow{k_m} \text{O}_{n+1} + \text{U}_{n-\Delta_{\text{RNAp}}} + \text{U}^{\text{R}}_{n-\Delta_{\text{RNAp}}}$	$k_m = 114$	[37]
Activation (4)	$\text{O}_n \xrightarrow{k_a} \text{A}_n$	$k_a = 114$, n>10, $k_a = 30$, n≤10	[37]
Pausing (5)	$\text{O}_n \underset{1/\tau_p}{\overset{k_p}{\rightleftarrows}} \text{O}_{n_\text{p}}$	$k_p = 0.55$ $\tau_p = 3$	[2]
Pause release due to collision (6)	$\text{O}_{n_\text{p}} + \text{A}_{n\,-\,2\Delta_{\text{RNAp}}-1} \xrightarrow{0.8k_m} \text{O}_n + \text{A}_{n\,-\,2\Delta_{\text{RNAp}}-1}i$	$k_m = 114$	[38]
Pause induced by collision (7)	$\text{O}_{n_\text{p}} + \text{A}_{n\,-\,2\Delta_{\text{RNAp}}-1} \xrightarrow{0.2k_m} \text{O}_{n_\text{p}} + \text{O}_{n\,-\,2\Delta_{\text{RNAp}}-1_\text{p}}$	$k_m = 114$	[38]
Arrests (8)	$\text{O}_n \underset{1/\tau_{ar}}{\overset{k_{ar}}{\rightleftarrows}} \text{O}_{n_{\text{ar}}}$	$k_{ar} = 0.00028$ $\tau_{ar} = 100$	[5]
Editing (9)	$\text{O}_n \underset{1/\tau_c}{\overset{k_{ec}}{\rightleftarrows}} \text{O}_{n_{\text{correcting}}}$	$k_{ec} = 0.008$ $\tau_c = 5$	[2]
Premature termination (10)	$\text{O}_n \xrightarrow{k_{pre}} \text{RNAp} + \text{U}_{[n\,-\,\Delta_{\text{RNAp}},n\,+\,\Delta_{\text{RNAp}}]}$	$k_{pre} = 0.00019$	[39]
Pyrophosphorolysis (11)	$\text{O}_n + \text{U}_{n-\Delta_{\text{RNAp}}-1} + \text{U}^{\text{R}}_{n-\Delta_{\text{RNAp}}-1} \xrightarrow{k_{pyro}} \text{O}_{n-1} + \text{U}_{n+\Delta_{\text{RNAp}}-1}$	$k_{pyro} = 0.75$	[40]
Completion (12)	$\text{A}_{n_{last}} \xrightarrow{k_f} \text{RNAp} + \text{U}_{[n_{last},n_{last}\,-\,\Delta_{\text{RNAp}}]}$	$k_f = 2$	[41]
mRNA degradation (13)	$\text{R} \xrightarrow{k_{dr}} \emptyset$	$k_{dr} = 0.011$	[13]

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

longer tail on the right side of the distribution is mainly due to the contribution of the time it takes for the RNAp to bind to the template, a bimolecular reaction whose expected time to occur follows an exponential distribution with a mean of 2.5 s [26,27].

Figure 1C shows the distribution of time intervals between transcription completion events in the same simulation as Figure 1B. This distribution is strikingly different from that of Figure 1B due to the stochastic events in transcription elongation. Pauses, arrests and other stochastic events cause the distribution to be bimodal due to the bursty dynamics (many short intervals and some long intervals). When these probabilistic events occur to some RNAp molecules, they significantly alter the distances in the strand between consecutive RNAps. For example, when one RNAp pauses, its distance to the preceding RNAp increases, while the distance to subsequent RNAps shortens, allowing completion events to be separated by intervals shorter than the promoter delay.

Dynamics of production of proteins

Figure 2A exemplifies the dynamics of ribosomes on one RNA strand. Stochastically, the transcription elongation process of this particular mRNA was halted at t = 50 s for a long period, and was thus selected to illustrate how long pauses in transcription affect the dynamics of translation of the multiple ribosomes on the RNA strand. The solid gray region in the bottom left part of the figure corresponds to the as-of-yet untranscribed sequence of the mRNA. When the RNAp pauses or is arrested (e.g. at t = 50 s), ribosomes accumulate in the region of the mRNA preceding the leading edge of transcription. Stochasticity in the translation elongation process is also visible. However, this process, modeled with realistic parameter values, appears to be less stochastic than transcription elongation, in that the stepwise elongation of ribosomes on the RNA template is more uniform than that of the RNAps on the DNA template. This is especially visible after the effects of the long arrest disappeared (at t > 230 s), at which point the

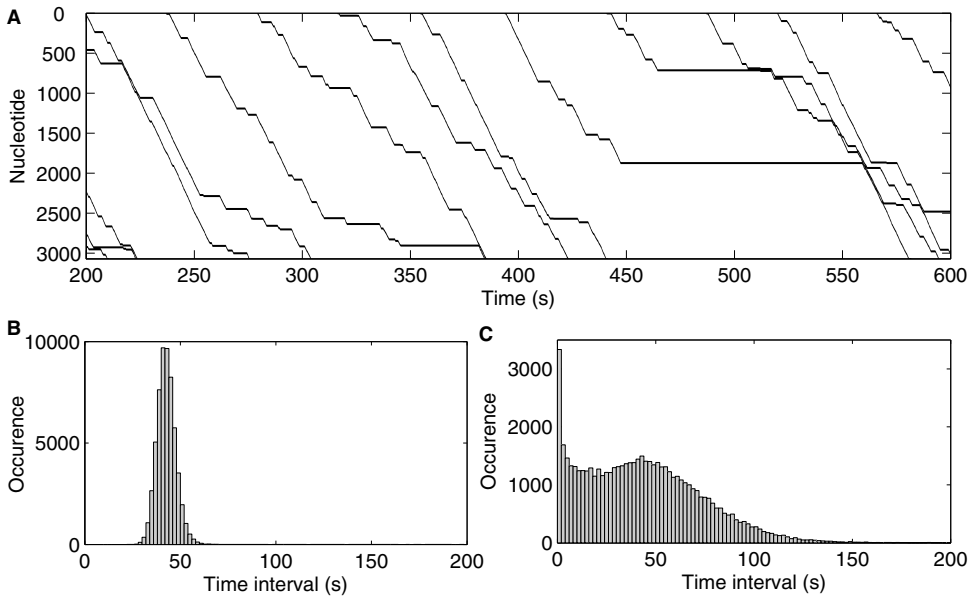


Figure 1 Kinetics of RNA polymerases on the DNA strand. (A) Example of the kinetics of multiple RNAp molecules on the DNA template over 400 s. Note that, on several occasions, the RNAp molecules pause and that one RNAp never overtakes another on the DNA template. (B) Distribution of time intervals between consecutive transcription initiation and (C) completion events. Data is from 57 000 initiation events.

distributions of time intervals between consecutive ribosomes at the start and at the end of translation elongation do not differ significantly.

Figure 2B shows the distribution of intervals between translation initiation events. Since there is no significant delay in translation initiation (as the one due to the promoter open complex formation), this distribution is exponential-like. Figure 2C shows the corresponding distribution of intervals between translation completion events (grey bars), given the presence of a sequence dependent arrest site at nucleotide 1850. This distribution, while resembling that of Figure 2B, shows more short time intervals, due to the long arrest in transcription elongation. For comparison, we also show a distribution of intervals between translation completion events drawn from cases without the sequence dependent arrest in transcription (solid black line). The difference between the two distributions illustrates how events in transcription elongation (e.g. a sequence dependent arrest site) can significantly affect the dynamics of translation.

Comparing the dynamics of the model of translation with measurements

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

We next compare the kinetics of translation in our model with measurements of the translation elongation speed in three engineered *E. coli* strains designed to enhance queue formation and traffic in translation [17]. Each strain contains a different mutant of *LacZ*. The pMAS23 strain corresponds to the wild-type *lacZ*. The other two sequences differ in that a region of slow-to-translate codons was inserted (~24 in pMAS-24GAG and ~48 in pMAS-48GAG). The speed of protein chain elongation was measured by subjecting the cells to a pulse of radioactive methionines, and then measuring

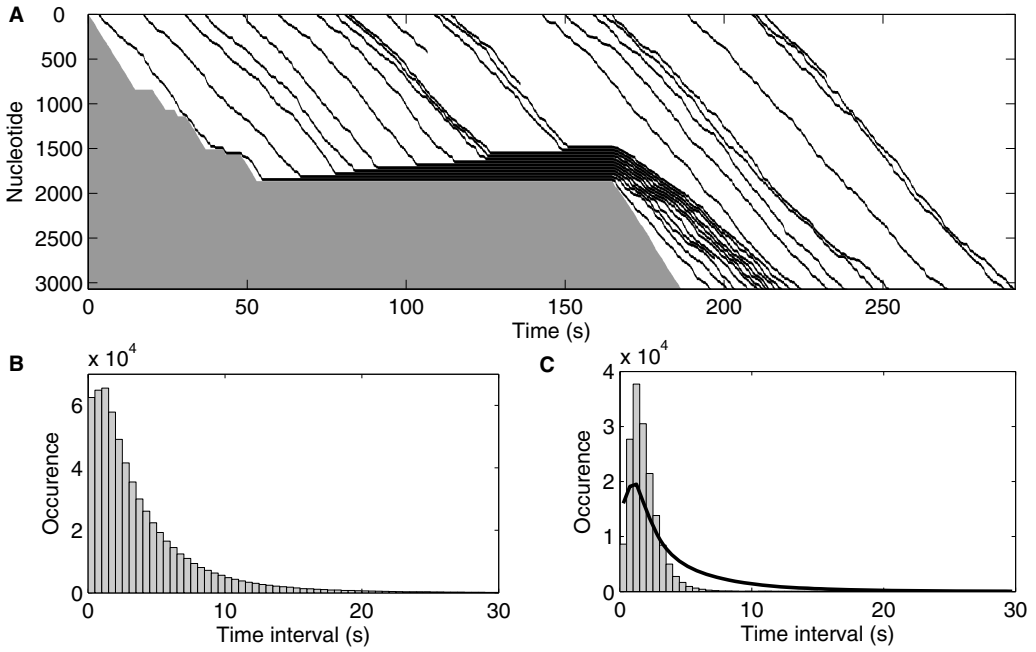


Figure 2 Kinetics of ribosomes on an RNA strand. (A) Example of the kinetics of several ribosomes along an mRNA template that suffered an arrest at nucleotide 1850, from the moment the ribosome binding site is formed to the degradation of the mRNA. The continuous gray region in the bottom left corresponds to the untranscribed sequence of the mRNA. (B) Distribution of time intervals between consecutive translation initiation events. (C) Distribution (grey bars) of time intervals between consecutive translation completion events given the presence of a sequence dependent arrest site at nucleotide 1850. The solid black line shows the distribution of time intervals between consecutive translation completion events without the sequence-dependent arrest site, normalized to the same scale. Data is from 600 000 initiation events.

the level of radioactivity in cells of each population, every 10 s after the pulse. Each strand contained 23 methionines, spread out unevenly on the DNA sequence, causing the incorporation curve to be non-linear.

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

To model this, we simulate the transcription and translation processes of the three sequences [17]. We model the incorporation of radioactive methionines at the same locations as in these sequences. The three model strands differ only in sequence, as in the measurements. During the simulations, we measure the number of incorporated radioactive methionines at the same points in time as in the experiment. Results of our simulations and of the measurements [17] are shown in Figure 3, showing good agreement between model and measurements.

Propagation of fluctuations in RNA levels to protein levels

We simulate the model for varying effective rates of transcription initiation (denoted k_{eff}). This rate is determined by the basal rate of transcription initiation (k_{init}), which sets the binding affinity of the RNAp to the transcription start site, and by the strength of repression of transcription. Thus, to vary k_{eff}, we vary the number of repressor molecules present in the system. Three sets of simulations are performed, differing in rate of translation initiation (k_{tr}). This rate is one of the kinetic parameters of the model, thus can be changed directly, and not by indirect means as k_{eff}. In *E. coli* genes, this rate is believed to be determined by the RBS sequence [28].

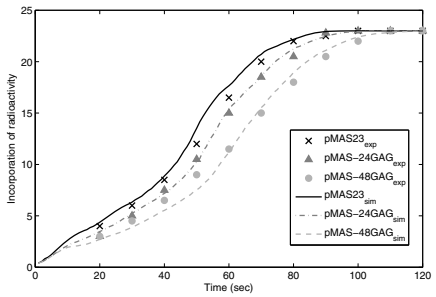


Figure 3 Appearance of radioactivity in β-galactosidase. Appearance of radioactivity incorporated from the three different mRNA strands, at different times after initiation of translation elongation in the models (lines) and in the measurements (crosses, triangles and circles) [17]. Values of radioactivity are normalized such that the maximum corresponds to 23 radioactive methionines.

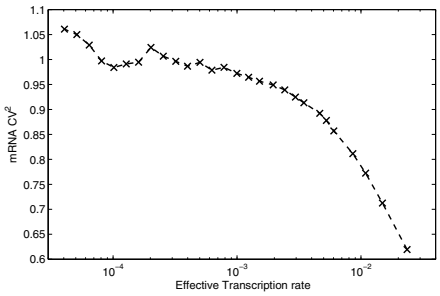


Figure 4 Noise in mRNA as a function of the transcription initiation rate. Noise (CV^2) in mRNA levels for varying effective transcription initiation rates. The mRNA degradation rate is set so that the mean mRNA levels at steady state are identical in all cases.

mRNA and protein degradation rates are set so that the mRNA and protein mean levels are identical for all cases, allowing us to study how the level of noise in mRNA and protein levels changes.

For each set of values of k_{eff} and k_{tr} we perform 100 independent simulations. Depending on these rates, the mean time to reach steady state differs. Each case is simulated for long enough to reach steady state and for an additional 100 000 s after that. The time series of the 100 simulations for each set of parameter values is concatenated into one time series, from which the noise is quantified by the square of the coefficient of variation, CV^2 (variance over the mean squared) [29]. This number of long simulations is necessary to properly sample the system due to the stochasticity of the underlying processes.

In Figure 4, we first show the CV^2 of mRNA time series for varying k_{eff}. Noise decreases as k_{eff} increases due to the promoter open complex formation step [6]. Without this event, the distribution of time intervals between transcription initiation events would be exponential, and the CV^2 would not vary. However, with this step, if the expected time for an RNAp to bind to the free promoter is faster than the duration of the promoter open complex formation, then the distribution of time intervals becomes Gaussian-like [6].

No measurements have yet been made to study experimentally the relation between the noise in mRNA levels and the corresponding protein levels. Nevertheless, it is possible to create a robust estimate, provided reasonable assumptions on the nature of the underlying processes [8]. Our model allows for a direct assessment, and it additionally includes realistic events such as RNAp and ribosome traffic, in transcription and translation elongation, which are not included in the aforementioned estimations [8]. Figure 5 shows the noise (CV^2) in protein levels, for varying k_{eff} and three values of k_{tr}. The data was obtained from the same simulations used to generate the results in Figure 4.

In general, we find that increasing k_{eff} decreases the noise in protein levels due to the decrease of noise in mRNA levels. Increasing k_{tr} increases the noise in protein levels, due to the increased size of the bursts in the protein level [8,29]. This finding has not yet been experimentally validated by direct means.

An interesting observation from Figures 4 and 5 is that, for $k_{eff} < 5 \times 10^{-4}$ s^{-1}, as k_{eff} is increased, the noise in protein levels decreases significantly, while the noise in RNA levels does not noticeably change. This is due to the decrease in mean protein burst size, i.e., the

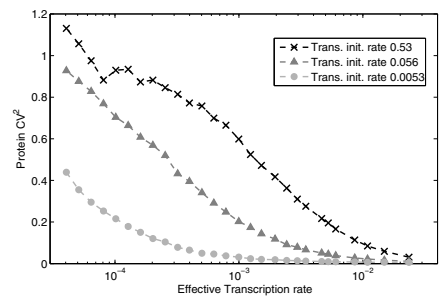


Figure 5 Noise in protein levels for varying transcription and translation initiation rates. Noise (CV^2) in protein levels for varying effective transcription initiation rates and three different rates of translation initiation. mRNA and protein degradation rates are set so that the mean mRNA and mean protein levels at steady state are identical in all cases.

mean number of proteins produced from each RNA molecule, as both k_{eff} and the degradation rate of RNA molecules are varied.

From these results, we conclude that the degree of coupling between transcription and translation is likely to be a key determining factor of the noise in protein levels. This can be verified by computing the normalized maximum correlation between time-series of protein and mRNA levels for each set of parameter values (Figure 6). Comparing Figures 5 and 6, we see that higher correlation values are obtained for the regime of higher noise in the protein levels. This implies that the principal source of this noise is the fluctuations in RNA levels.

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

Effects of transcriptional pauses on the fluctuations in protein levels

Recent work [1] reported that long transcriptional pauses enhance the noise in mRNA levels. We next investigate to what extent the fluctuations in RNA levels caused by long transcriptional pauses propagate to protein levels. Long sequence-dependent pauses [16,30,31] in transcription elongation may cause the ribosome to stall in the mRNA chain. This will likely cause subsequent ribosomes to accumulate in the preceding sequence. When the RNAp is spontaneously released from the pause [31], translation of the stalled ribosomes likely resumes but the distribution of intervals between them will differ significantly from what it would have been without the pause event. Consequently, the protein production is likely to become burstier, especially if the long pause site is located near the end of the sequence. An increase in burstiness ought to increase the noise in protein levels.

To verify this, we perform two simulations. We introduce a long-pause sequence with mean pause durations of 500 s in one case, and 100 s in the other (both values are within realistic intervals [30]). In both cases, we set the probability that an RNAp will pause at that site to 70% (identical to the value for *his* pause sites [16]).

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

The effects of several pause sites on the same strain are cumulative, namely, the higher the number of pause sites, the higher the noise in RNA levels [32]. Combined with the present results, this leads us to the conclusion that the sequence-dependent transcriptional pausing mechanism likely exists to allow a wide variation of both RNA and protein noise levels.

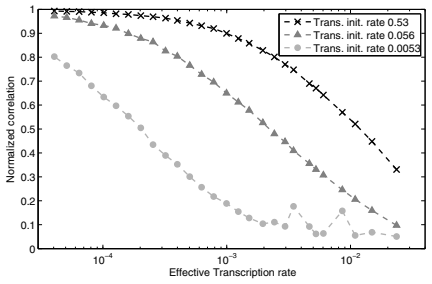


Figure 6 Normalized maximum correlation between RNA and protein time series. The higher the rate of translation initiation (and thus higher protein degradation to keep the mean the same), the more correlated the fluctuations in protein and RNA levels become, as measured by the normalized maximum correlation. This is because the protein levels follow any fluctuations in the RNA levels faster. Similarly, increasing the rate of transcription initiation, while maintaining the rate of translation initiation constant, decreases the correlation between fluctuations in protein and RNA levels.

Conclusions

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

act as generators of strong fluctuations in RNA levels [15], the propagation of these fluctuations to protein levels, and the ability of fluctuations in protein levels to affect cellular phenotype [33], suggest that these mechanisms may be evolvable.

As a previous study has suggested [8], the translation initiation rate was found to be key in determining the degree of coupling between the fluctuations in RNA and protein levels, if one assumes that the degradation rate of the proteins is changed accordingly to maintain their mean level unchanged. Varying this sequence-dependent, and thus, evolvable parameter [28] within realistic ranges gave a widely varying degree of coupling between the fluctuations in RNA and protein levels. It is therefore not necessarily true that noisy production of RNA molecules results in noisy protein levels. Interestingly, while decreasing the coupling between transcription and translation by decreasing the rate of translation initiation causes the protein levels to become less noisy, it also takes longer for a change in RNA levels to be followed by the protein levels. This suggests that to be able to change rapidly in response to, e.g., environmental changes, the levels of a protein will be necessarily noisier.

Confirming previous studies [1,5,8,19], we found that the distributions of time intervals between transcription initiation and completion events differ significantly and that the faster the rate of transcription initiation events, the more they differ. This implies that in the regime of fast transcription, both the transcription and translation elongation processes need to be modeled explicitly and coupled, if one is to match the mean and fluctuations in the protein levels at the molecular level. This is of relevance, since bursts in protein levels may trigger many processes, such as phenotypic differentiation [33,34]. A final justification for using the model proposed here is the complexity of the process of gene expression in *E. coli*, and the fact that many events therein may or may not affect the temporal RNA and protein levels significantly, depending on their specific sequence-dependent features. Such effects, due to the complexity of the system, are not easily predictable without performing explicit numerical simulations.

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

However, the present model does not yet account for known effects of ribosomes on the dynamics of transcription elongation. These might need to be included in future developments of the proposed model as recent results [27,35] suggest that the rate of translation elongation can affect the rate of transcription elongation, due to possible interactions between the ribosome that first binds to the mRNA and the RNAp transcribing it. Possible effects may include facilitating the release of paused RNAp's, which could affect the degree of the contribution of pauses to the noise in RNA and thus protein levels. We do not exclude the possibility that the contrary may occur in specific cases, that is, that the paused state of the RNAp may cause pauses in the ribosome translational dynamics, which would amplify the effect of transcriptional pauses on the fluctuations of protein levels. Whether the pause is ubiquitous or due to loop formations in the nascent RNA may affect the results of the interaction as well. Provided experimental evidence on the nature and consequences of these interactions, once included in the model, we may be able to test, among other things, whether long transcriptional pauses located in an attenuator system provide an additional layer of control over premature transcription terminations, and thus over RNA and protein noise levels.

Methods

Model of transcription, one nucleotide at a time

We model the dynamics of gene expression as in [23]. This model was shown [21] to match the dynamics of RNA and protein production at the single molecule level [13]. The dynamics of the system of chemical reactions is driven by the delayed stochastic simulation algorithm (delayed SSA [19]) so as to include events whose time of completion once initiated is non negligible, in that it affects the dynamics of production of RNA and protein molecules. Specifically, several steps in gene expression, such as the promoter open complex formation, are time consuming [36]. To include these events when simulating gene expression, the delayed SSA was proposed [19].

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

The delayed stochastic model of transcription at the nucleotide level [5] includes the promoter occupancy time, pausing, arrests, editing, premature terminations, pyrophosphorolysis, and accounts for the RNAp footprint in the DNA template [2]. Additional reactions

model the stepwise forward movement and activation of the RNAp, pausing and unpausing of the RNAp due to collisions with adjacent RNAps, release of the promoter when the RNAp begins elongation, and error correction.

The reactions, stochastic rate constants and time delays, are shown in Table 1 and described in detail in [5,37-41]. Here, Pro stands for the promoter region, RNAp for the RNA polymerase, and RNAp·Pro for the promoter region occupied by an RNAp. A_n, O_n and U_n stand for the nth nucleotide when activated, occupied, and unoccupied, respectively. Ranges of nucleotides are denoted such as $U_{[start,\ end]}$, denoting a stretch of unoccupied nucleotides from indexes *start* to *end*. O_{n_p}, $O_{n_{ar}}$ and $O_{n_{correcting}}$ are used to represent a paused, arrested, or error correcting RNAp at position n. On the template, each RNAp occupies $(2\Delta_{RNAp}+1)$ nucleotides, where $\Delta_{RNAp} = 12$. These nucleotides cannot be occupied by any other RNAp at the same time. U_n^R denotes transcribed ribonucleotides which are free (i.e., not under the RNAp's footprint). These transcribed ribonucleotides are created in a separate part of the simulation (denoted by the R superscript), one separate set per RNA strand, so that we can simulate the translation of all individual RNA molecules independently and simultaneously.

We use a delayed reaction event to model the first step in transcription, the promoter closed and open complex formation (1). These processes could instead be modeled by a set of non-delayed, consecutive, reactions [42]. We use a delayed reaction as it was shown to accurately model the dynamics of this process [19,21,23]. The duration of this step likely varies from one event to the next, but while values for the mean duration are known, as of yet, there are no exact measurements of the standard deviation. Nevertheless, it is likely small compared to the mean, given the very small standard deviations of promoter activity [25]. For these reasons, we set the promoter delay, τ_{oc}, as a random variable, following a normal distribution with a mean of 40 s and a standard deviation of 4 s, whose value is randomly drawn each time a transcription event occurs.

Once the first nucleotide is occupied via reaction (2), stepwise elongation can begin (3). Also, as soon as the promoter is released, a new transcription initiation event can occur. Following each elongation step (3), an activation step occurs (4), which is necessary for the RNAp to move along the template to the next nucleotide. The following events compete with stepwise elongation: pausing (5) and (7), released via (5) or (6), arrests and their release (8), editing (9), premature terminations (10), and pyrophosphorolysis (11).

At the end of the elongation process, the RNAp is released (12). mRNA degradation is modeled, for simplicity, as a first order reaction (13). When (13) occurs, the first few ribonucleotides of the RNA are immediately removed from the system, preventing any new translation event [43]. Thus, we model the degradation process such that it begins in the vicinity of the RBS and then gradually cuts the mRNA as it is being released from the ribosomes. This allows the translating ribosomes to complete protein production before the whole mRNA is degraded. When the final ribosome unbinds from the RNA, the rest of the RNA strand, denoted by R in reaction (13), is destroyed.

If the model of RNA degradation was such that some of the ribosomes on the RNA template fell off when degradation begins (i.e. due to endonucleatic cleavage of the RNA chain at a random position [43]), one consequence would be the reduction of the mean protein burst size as these RNAs would contribute far fewer proteins than if the ribosomes were allowed to finish translating. This would likely result in a reduction of protein noise levels. Alternatively, the ribosome occupancy of the ribosome binding site might determine mRNA longevity [28]. In this case, for the same mean burst size, the noise is expected to increase since large bursts will get larger and small bursts will get smaller, likely increasing protein noise levels. We opted not to include these additions to the degradation model since they are not yet well characterized [43].

Finally, we note that in present model we do not add an explicit reaction for abortive initiation of transcription [44]. This could be done by adding a reaction (2b) which would compete with reaction (2). Its rate, k_{ab}, would be set so as to match the fraction of abortive initiations after the formation of the promoter open complex [44]:

$$\mathrm{RNAp} \bullet \mathrm{Pro} \xrightarrow{k_{ab}} \mathrm{Pro} + \mathrm{RNAp} \tag{2b}$$

For simplicity, we opted not to include this reaction in the simulations, and instead set a value for the rate of transcription initiation that matches realistic rates of RNA production. From the point of view of RNA production, since (2b) competes with reaction (2), it would be dynamically equivalent to decrease the rate of transcription initiation in (2) to account for the fraction of abortive initiations.

The model of transcription and the reaction rates in Table 1 are described in greater detail in [5]. Parameter values were obtained from measurements in *E. coli*, mainly for *LacZ*.

Model of translation, one codon at a time

The stochastic model of translation at the codon level includes initiation (14) and stepwise translocation (codon incorporation) (15-17) followed by activation

(18). Reactions competing with translocation are back-translocation (19), drop-off (20), and trans-translation (21). The process ends with elongation completion (22), followed by protein folding and activation (23). Protein degradation (24) is included to allow us to study fluctuations in protein levels at steady state. All reactions and rate constants are presented in Table 2 [45-47]. Here, Rib denotes a free ribosome complex in the cellular medium, while Rib^R denotes a ribosome bound to a specific RNA strand. Similar to Δ_{RNAp}, Δ_{Rib} denotes the ribosome's footprint in the RNA template. Each ribosome occupies $(2\Delta_{Rib}+1)$ ribonucleotides, where Δ_{Rib} = 15 [20]. U_n^R, O_n^R and A_n^R are the ribonucleic equivalents of U_n, O_n and A_n. U_n^R denotes an unoccupied ribonucleotide, while O_n^R denotes that a translating ribosome is currently positioned at ribonucleotide n. Similarly, A_n^R denotes that a ribosome has created peptide bond for the peptide coded by the codon at position $[n\text{-}2,n]$, where n is a multiple of 3 (n = 3, 6, 9,...). Since different codons are translated at different rates, the activation reaction has a codon-specific rate [17]. Specific rates were set for four codons, while the remaining ones fall into three different classes [20], A, B and C, whose rates are denoted $k_{trans\{A,\ B,\ C\}}$.

Translation has three main phases: initiation, elongation and termination. It begins with the binding of the ribosome complex to the mRNA strand. During elongation, the amino acids, determined by the RNA sequence, are added to the elongating peptide chain. Termination is the final step, as specific release factors detach the peptide and the RNA chain from the ribosome. *E. coli* has specific translation factors for each phase: initiation factors IF1, IF2 and IF3, elongation factors EF-G, EF-Tu and EF-Ts and three release factors RF1, RF2 and RF3 [48]. These are not explicitly modeled, as they exist in abundance under normal conditions.

The binding of the ribosome to the ribosome binding site (RBS) of the RNA starts with the binding of the 30S ribosomal subunit to the nascent mRNA. After that, fMet-tRNA binds to the P-site forming a 30S complex. The 50S ribosome subunit attaches to it, forming the 70S initiation complex [48]. This process is modeled as a single step reaction (14). The next ribosome can only to bind after the preceding one has moved away from the RBS. This implies that the initiation of two consecutive translation events is separated by a non-negligible time interval.

Translation elongation occurs through successive translocation-and-pause cycles [3]. Translocation includes three steps (15-17), after which there is a pause (18), during which the bond between amino acids is formed. The time that (18) takes to occur accounts for this pause, which is much longer than the time for (15-17) to occur [3].

The genetic code contains two mechanisms for redundancy: some tRNAs can be charged with the same

Table 2 Reactions modeling translation

Event	Reaction	Rate constant	Ref.
Initiation (14)	$\mathrm{Rib} + \mathrm{U}^{\mathrm{R}}_{[1,\Delta_{\mathrm{Rib}}+1]} \xrightarrow{k_{trans\_init}} \mathrm{O}^{\mathrm{R}}_1 + \mathrm{Rib}^{\mathrm{R}}$	k_{trans_init} = 0.33	[20]
Stepwise translocation (15-17)	$\mathrm{A}^{\mathrm{R}}_{n-3} + \mathrm{U}^{\mathrm{R}}_{[n+\Delta_{\mathrm{Rib}}-3,n+\Delta_{\mathrm{Rib}}-1]} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n-2}$ $\mathrm{O}^{\mathrm{R}}_{n-2} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n-1}$ $\mathrm{O}^{\mathrm{R}}_{n-1} \xrightarrow{k_{tm}} \mathrm{O}^{\mathrm{R}}_{n} + \mathrm{U}^{\mathrm{R}}_{[n-\Delta_{\mathrm{Rib}}-2,n-\Delta_{\mathrm{Rib}}]}$	k_{tm} = 1000	[3]
Activation (18)	$\mathrm{O}^{\mathrm{R}}_{n} \xrightarrow{k_{trans\{A,B,C\}}} \mathrm{A}^{\mathrm{R}}_{n}$	k_{transA}= 35, k_{transB}= 8, k_{transC}= 4.5	[20]
Back-translocation (19)	$\mathrm{O}^{\mathrm{R}}_{n} + \mathrm{U}^{\mathrm{R}}_{[n-\Delta_{\mathrm{Rib}}-2,n-\Delta_{\mathrm{Rib}}]} \xrightarrow{k_{bt}} \mathrm{A}^{\mathrm{R}}_{n-3} + \mathrm{U}^{\mathrm{R}}_{[n+\Delta_{\mathrm{Rib}}-3,n+\Delta_{\mathrm{Rib}}-1]}$	k_{bt} = 1.5	[51]
Drop-off (20)	$\mathrm{O}^{\mathrm{R}}_{n} \xrightarrow{k_{drop}} \mathrm{Rib} + \mathrm{U}^{\mathrm{R}}_{[n\ -\ \Delta_{\mathrm{Rib}},n+\Delta_{\mathrm{Rib}}]}$	k_{drop} = 0.000114	[45]
Trans-translation (21)	$\mathrm{R} \xrightarrow{k_{tt}} [\mathrm{Rib}^{\mathrm{R}}]\mathrm{Rib}$	k_{tt} = 0.000052	[46]
Elongation completion (22)	$\mathrm{A}^{\mathrm{R}}_{n_{last}} \xrightarrow{k_{trans\_f}} \mathrm{Rib} + \mathrm{U}^{\mathrm{R}}_{[n_{last},n_{last}\ -\ \Delta_{\mathrm{Rib}}]} + \mathrm{P}_{\mathrm{prem}}$	k_{trans_f} = 2	[20]
Folding and activation (23)	$\mathrm{P}_{\mathrm{prem}} \xrightarrow{k_{fold}} \mathrm{P}$	k_{fold} = 0.0024	[47]
Protein degradation (24)	$\mathrm{P} \xrightarrow{k_{dec}} \emptyset$	k_{dec} = 0.0017	[47]

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

amino acid, and a single tRNA can recognize more than one codon due to a "wobble" effect in position three of the anti-codon [48]. The net effect is that multiple codons code for the same amino acid. These codons are called synonymous codons. Synonymous codons read by the same tRNA have been shown to translate at significantly different rates [17], implying that our model must incorporate per-codon translation rates for reaction (18), rather than per-tRNA or per-amino acid rates. Only a few of these translation rates have been measured directly [17] but indirect assessment is available [20]. In our case, we assume normal cellular conditions, including an abundance of charged tRNA, implying that we do not need to model the tRNA explicitly.

Since each codon is translated at a different rate, the codon frequency also needs to be accounted for explicitly [49]. In the model, the sequence can either be randomly generated or selected from a known gene. In the former case, the sequence is randomly generated according to the known statistical frequency of each codon in *E. coli*.

The competing reactions of stepwise translation elongation are back-translocation (19), drop-off (20) and trans-translation (21), which are explicitly modeled. Back-translocation generally occurs when the tRNA has not yet locked into the peptide chain, causing the ribosome to move backwards on the mRNA template to the position of the previous codon. While the occurrence of back-translocation has been observed and can be promoted by certain antibiotics [50-52], its exact causes remain somewhat unknown. Nevertheless, the kinetic rates for translocation and back-translocation have been measured under various conditions [51]. Alternatively, the ribosomes can randomly dissociate from the RNA, in a process called drop-off, modeled by reaction (20). The overall rate of drop-off has been measured in [45], from which we have inferred a per-codon rate.

Trans-translation is the process by which the ribosome is released from the RNA template after stalling, which can occur for a variety of reasons, such as the incorporation of an incorrect codon, premature mRNA degradation, or spontaneous frameshifting [53]. Trans-translation is executed by the tmRNA that, together with SmpB and EF-Tu, binds to the A-site of the ribosome and releases it from the mRNA [53]. Once the ribosome is released, the mRNA is degraded. In the model, stalling followed by trans-translation can occur spontaneously with a given probability at any codon via reaction (21). When this reaction occurs, the RNA strand is immediately destroyed in the simulation, and all translating ribosomes are released back into the cellular medium, denoted in reaction (21) by $[Rib^R]Rib$, where $[Rib^R]$ denotes the number of ribosomes bound to the RNA at that moment.

Translation elongation continues until the STOP codon is reached (22), after which RF1 or RF2 binds and releases the ribosome together with RF3 [48]. These are not modeled explicitly in the model. Its kinetic rate is higher than initiation, preventing queuing near the stop codon [20]. Reaction (22) is followed by folding and activation (23), modeled as a first order process for simplicity [21]. The rate of this reaction is set to model the maturation time of GFP, as most measurements of protein expression at the single cell level use this protein. P_{prem} denotes the unfolded protein, while P denotes the complete activated protein, which can then degrade via reaction (24).

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

The chemical reactions and rate constants (in s^{-1}) used to model translation initiation, elongation, and termination, as well as protein folding and activation and protein degradation are in Table 2. Parameter values were obtained from measurements in *E. coli*, mainly for *LacZ*.

Quantifying the correlation between protein and mRNA levels

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

To assess the extent to which fluctuations in RNA levels are propagated to protein levels, we compute the normalized discrete cross-correlation [54] between the time series of RNA and protein numbers. The normalized cross-correlation function r for m pairs of time series (x and y) of discrete signals of length n is given by:

$$\mathrm{r}[\tau] = \frac{\sum_{l=1}^{N}\sum_{k=1}^{n-\tau}\left(x_l[k] - m_{x1,\ldots,N[1,\ldots,n-\tau]}\right)\left(y_l[k+\tau] - m_{y1,\ldots,N[1+\tau,\ldots,n]}\right)}{((\mathrm{n}-\tau)N-1)\, s_{x1,\ldots,N[1,\ldots,n-\tau]} s_{y1,\ldots,N[1+\tau,\ldots,n]}} \quad (25)$$

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

$$\mathrm{m}_{\mathrm{w}_{1\ldots N}[i\ldots j]} \doteq \frac{1}{(j-i+1)N}\sum_{l=1}^{N}\sum_{k=i}^{j} w_l[k] \quad (26)$$

$$s_{\mathrm{w}_{1\ldots N}[i\ldots j]} \doteq \sqrt{\frac{1}{(j-i+1)N-1}\sum_{l=1}^{N}\sum_{k=i}^{j}\left(w_l[k] - m_{w_{1\ldots N}[i\ldots j]}\right)^2} \quad (27)$$

Acknowledgements
This work was supported by the Academy of Finland (JLP, ASR) and by the FiDiPro programme of Finnish Funding Agency for Technology and Innovation (JM, OYH, and ASR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details
[1]Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland. [2]Institute for Systems Biology, 1441N 34th St, Seattle, WA, 98103-8904, USA.

Authors' contributions
All authors contributed in the design of the study, data acquisition and interpretation, and participated in the drafting of the article. All authors have read and approved the final manuscript.

Received: 19 November 2010 Accepted: 26 April 2011
Published: 26 April 2011

References

1. Rajala T, Häkkinen A, Healy S, Yli-Harja O, Ribeiro AS: **Effects of transcriptional pausing on gene expression dynamics.** *PLOS Comput Biol* 2010, **6(3)**:e1000704.
2. Greive SJ, von Hippel PH: **Thinking quantitatively about transcriptional regulation.** *Nat Rev Mol Cell Biol* 2005, **6**:221-232.
3. Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, Bustamante C, Tinoco I Jr: **Following translation by single ribosomes one codon at a time.** *Nature* 2008, **452**:598-603.
4. Landick R: **The regulatory roles and mechanism of transcriptional pausing.** *Biochem Soc Trans* 2006, **34(6)**:1062-1066.
5. Ribeiro AS, Rajala T, Smolander OP, Häkkinen A, Yli-Harja O: **Delayed Stochastic Model of Transcription at the Single Nucleotide Level.** *J Comput Biol* 2009, **16**:539-553.
6. Ribeiro AS, Häkkinen A, Mannerstrom H, Lloyd-Price J, Yli-Harja O: **Effects of the promoter open complex formation on gene expression dynamics.** *Phys Rev E* 2010, **81(1)**:011912.
7. Kaern M, Elston TC, Blake WJ, Collins JJ: **Stochasticity in gene expression: from theories to phenotypes.** *Nat Rev Genet* 2005, **6**:451-464.
8. Pedraza J, Paulsson J: **Effects of Molecular Memory and Bursting on Fluctuations in Gene Expression.** *Science* 2008, **319**:339-334.
9. Murphy KF, Balazsi G, Collins JJ: **Combinatorial promoter design for engineering noisy gene expression.** *Proc Natl Acad Sci USA* 2007, **104**:12726-12731.
10. Mayr E: *What evolution is* Basic Books, NY, USA; 2001.
11. Lee HH, Molla MN, Cantor CR, Collins JJ: **Bacterial charity work leads to population-wide resistance.** *Nature* 2010, **467**:82-86.
12. Acar M, Mettetal J, van Oudenaarden A: **Stochastic switching as a survival strategy in fluctuating environments.** *Nature Genetics* 2008, **40**:471-475.
13. Yu J, Xiao J, Ren X, Lao K, Xie XS: **Probing gene expression in live cells, one protein molecule at a time.** *Science* 2006, **311**:1600-1603.
14. Golding I, Paulsson J, Zawilski SM, Cox EC: **Real-time kinetics of gene activity in individual bacteria.** *Cell* 2005, **123**:1025-1036.
15. Ribeiro AS: **Stochastic and delayed stochastic models of gene expression and regulation.** *Mathematical Biosciences* 2010, **223(1)**:1-11.
16. Herbert KM, La Porta A, Wong BJ, Mooney RA, Neuman KC, Landick R, Block SM: **Sequence-resolved detection of pausing by single RNA polymerase molecules.** *Cell* 2006, **125**:1083-1094.
17. Sorensen MA, Pedersen S: **Absolute in vivo translation rates of individual codons in Escherichia coli.** *J Mol Biol* 1991, **222**:265-280.
18. Bernstein J, Khodursky A, Lin P, Lin-Chao S, Cohen S: **Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.** *Proc Natl Acad Sci USA* 2002, **99**:9697-9702.
19. Roussel MR, Zhu R: **Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression.** *Phys Biol* 2006, **3**:274-284.
20. Mitarai N, Sneppen K, Pedersen S: **Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization.** *J Mol Biol* 2008, **382(1)**:236-245.
21. Zhu R, Ribeiro AS, Salahub D, Kauffman SA: **Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models.** *J Theor Biol* 2007, **246**:725-745.
22. Voliotis M, Cohen N, Molina-Paris C, Liverpool TB: **Fluctuations, pauses and backtracking in DNA transcription.** *Biophys J* 2008, **94**:334-348.
23. Ribeiro AS, Zhu R, Kauffman SA: **A general modeling strategy for gene regulatory networks with stochastic dynamics.** *J Comput Biol* 2006, **13**:1630-1639.
24. Ribeiro AS, Lloyd-Price J: **SGN Sim, a Stochastic Genetic Networks Simulator.** *Bioinformatics* 2007, **23(6)**:777-779.
25. Lutz R, Lozinski T, Ellinger T, Bujard H: **Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator.** *Nuc Ac Res* 2001, **29**:3873-3881.
26. Gillespie DT: **Exact stochastic simulation of coupled chemical reactions.** *J Phys Chem* 1977, **81**:2340-2361.
27. Arkin A, Ross J, McAdams H: **Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected E. coli cells.** *Genetics* 1998, **149**:1633-1648.
28. Yarchuk O, Jacques N, Guillerez J, Dreyfus M: **Interdependence of translation, transcription and mRNA degradation in the *lacZ* gene.** *J Mol Biol* 1992, **226**:581-596.
29. Paulsson J: **Models of stochastic gene expression.** *Phys Life Rev* 2005, **2(2)**:157-175.
30. Shaevitz JW, Abbondanzieri EA, Landick R, Block SM: **Backtracking by single RNA polymerase molecules observed at near-base-pair resolution.** *Nature* 2003, **426**:684-687.
31. Landick R: **Transcriptional pausing without backtracking.** *Proc Natl Acad Sci USA* 2009, **106(22)**:8797-8798.
32. Ribeiro AS, Häkkinen A, Healy S, Yli-Harja O: **Dynamical effects of transcriptional pause-prone sites.** *Comput Biol Chem* 2010, **34(3)**:143-148.
33. Choi PJ, Cai L, Frieda K, Xie XS: **A Stochastic *Single-Molecule Event* Triggers Phenotype Switching of a Bacterial Cell.** *Science* 2008, **322(5900)**:442-446.
34. Xie XS, Choi PJ, Li GW, Lee NK, Lia G: **Single-molecule approach to molecular biology in living bacterial cells.** *Annu Rev Biophys* 2008, **37**:417-444.
35. Burmann BM, Schweimer K, Luo X, Wahl MC, Stitt BL, Gottesman ME, Rösch P: **A NusE:NusG Complex Links Transcription and Translation.** *Science* 2010, **328(5977)**:501-504.
36. Ota K, Yamada T, Yamanishi Y, Goto S, Kanehisa M: **Comprehensive Analysis of Delay in Transcriptional Regulation Using Expression Profiles.** *Genome Informatics* 2003, **14**:302-303.
37. Phroskin S, Rachid Rahmouni A, Mironov A, Nudler E: **Cooperation between translating ribosomes and RNA polymerase in transcription elongation.** *Science* 2010, **328(5977)**:504-508.
38. Epshtein V, Nudler E: **Cooperation between RNA polymerase molecules in transcription elongation.** *Science* 2003, **300(5620)**:801-805.
39. Lewin B: *Genes IX* Jones and Bartlett Publishers, USA; 2008, 256-299.
40. Erie DA, Hajiseyedjavadi O, Young MC, von Hippel PH: **Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription.** *Science* 1993, **262**:867-873.

41. Greive SJ, Weitzel SE, Goodarzi JP, Main LJ, Pasman Z, von Hippel PH: **Monitoring RNA transcription in real time by using surface plasmon resonance.** *Proc Natl Acad Sci USA* 2008, **105**:3315-3320.
42. McClure WR: **Rate-limiting steps in RNA chain initiation.** *Proc Natl Acad Sci USA* 1980, **77**:5634-5638.
43. Balesco JG: **All things must pass: Contrasts and commonalities in eukaryotic and bacterial mRNA decay.** *Nat Rev Mol Cell Biol* 2010, **11(7)**:467-478.
44. Hsu LM: **Promoter clearance and escape in prokaryotes.** *Biochimica et Biophysica Acta - Gene Structure and Expression* 2002, **1577(2)**:191-207.
45. Jorgensen F, Kurland CG: **Processivity errors of gene expression in Escherichia coli.** *J Mol Biol* 1990, **215**:511-521.
46. Moore SD, Sauer RT: **Ribosome rescue: tmRNA tagging activity and capacity in Escherichia coli.** *Mol Microbiol* 2005, **58**:456-466.
47. Cormack BP, Valdivia RH, Falkow S: **FACS-optimized mutants of the green fluorescent protein (GFP).** *Gene* 1996, **173(1)**:33-38.
48. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular biology of the cell* Garland Science, USA; 2002.
49. Sorensen MA, Kurland CG, Pedersen S: **Codon usage determines translation rate in Escherichia coli.** *J Mol Biol* 1989, **207**:365-377.
50. Menninger JR: **Peptidyl transfer RNA dissociates during protein synthesis from ribosomes of Escherichia coli.** *J Biol Chem* 1976, **251**:3392-3398.
51. Shoji S, Walker SE, Fredrick K: **Ribosomal translocation: One step closer to the molecular mechanism.** *ACS Chem Biol* 2009, **4**:93-107.
52. Qin Y, Polacek N, Vesper O, Staub E, Einfeldt E, Wilson DN, Nierhaus KH: **The highly conserved LepA is a ribosomal elongation factor that back-translocates the ribosome.** *Cell* 2006, **127**:721-733.
53. Keiler KC: **Biology of trans-translation.** *Annu Rev Microbiol* 2008, **62**:133-151.
54. Bracewell R: *Pentagram Notation for Cross Correlation. The Fourier Transform and Its Applications* New York: McGraw-Hill; 1965, 46-243.

doi:10.1186/1471-2105-12-121
Cite this article as: Mäkelä *et al.*: Stochastic sequence-level model of coupled transcription and translation in prokaryotes. *BMC Bioinformatics* 2011 **12**:121.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3794-3
ISSN 1459-2045