Tampere University

SOFIA STARTCEVA

Regulation of Single-Cell Bacterial Gene Expression at the Stage of Transcription Initiation

Tampere University Dissertations 211

Tampere University Dissertations 211

SOFIA STARTCEVA

Regulation of Single-Cell Bacterial Gene Expression at the Stage of Transcription Initiation

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the auditorium F115
of the Arvo building, Arvo Ylpön katu 34, Tampere,
on the 17th of February 2020, at 13 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
Finland

Responsible supervisor and Custos	Professor Andre S. Ribeiro Tampere University Finland	
Supervisor	Professor Ari Visa Tampere University Finland	
Pre-examiners	Assoc. Professor Oleg Igoshin Rice University the United States of America	Assist. Professor Andreas Hilfinger University of Toronto Mississauga Canada
Opponent	Professor Heinz Köppl Technische Universität Darmstadt Germany	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-1439-2 (print)
ISBN 978-952-03-1440-8 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1440-8

PunaMusta Oy – Yliopistopaino
Tampere 2020

Abstract

One of the qualities that allow bacterial cells to survive in diverse, fluctuating environments is phenotypic plasticity, which is the ability to exhibit different phenotypes depending on the environmental conditions. Phenotypic plasticity arises via coordinated work of small genetic circuits that provide the cell with the means for decision-making. The behavior of these circuits depends, among other factors, on the ability of protein numbers to cross certain thresholds for a sufficient amount of time. In bacteria, RNA numbers largely define protein numbers and thus can be used to study the decision-making processes.

Previous research outlined the effects of mean and variance in RNA or protein numbers on the behavior of small genetic circuits. However, noise in gene expression is often highly asymmetric. This could impact the threshold-crossing abilities of molecular numbers in a way that is not detectable by considering only their mean and variance.

The focus of this thesis is to study the regulation of multi-step kinetics of bacterial gene expression in live bacteria and its effects on the shape of the distribution of RNA or protein levels. In particular, the thesis investigates how the rate-limiting steps in bacterial transcription, such as closed and open complex formation, intermittent inactive states, and promoter escape contribute to the dynamics of RNA numbers, and how this dynamics propagates to the distribution of protein levels in a cell population. This study made use of already existing techniques such as measurements at the single-RNA level and dynamically accurate stochastic modeling, complemented by the novel methodology developed in this work.

First, the thesis introduced a new method for estimating the numbers of fluorescently tagged molecules present in a cell from time series data obtained by microscopy. This method allows improving the accuracy of the estimation when fluorescently tagged molecules are absent from the cell image for time intervals comparable with cell lifetime. Second, the new methodology for dissecting *in vivo* kinetics of rate-limiting steps in transcription initiation was proposed. Applying this methodology to study initiation kinetics at lac/ara-1 promoter provided insights on the amount, duration, and reversibility of the rate-limiting steps in this process. Further, the thesis investigated the kinetics of transcription activation of lac/ara-1 promoter at various temperatures. The results indicate that additional rate-limiting steps emerge in inducer intake kinetics as temperature decreases from optimal (37 °C). Finally, the focus was shifted specifically to quantifying the asymmetry and tailedness in RNA and protein level distributions, since these features are relevant for determining threshold crossing propensities. Here, these features were found to depend both on promoter sequence and on regulatory molecules, thus being evolvable and adaptable.

Overall, the work conducted in this thesis suggests that asymmetries in RNA and protein numbers may be crucial for decision-making in bacteria, since they can be regulated

by promoter sequence, regulatory molecules levels, and temperature shifts. The thesis also contributes to the pool of existing methodology for studying *in vivo* bacterial gene expression using single-cell biology approach. These findings should be of use both for better understanding of natural systems and for fine-tuning behavior of synthetic gene circuits.

Preface

> Identifying causes ... [is] the very essence of thinking, and by this act alone sensations turn into realizations and are not lost, but become entities and start to emit like rays of light what is inside of them.
>
> Hermann Hesse, *Siddhartha*

This study was carried out at the Laboratory of Biosystem Dynamics (LBD) of the BioMediTech Institute and at the Department of Signal Processing at Tampere University (prior to 1.1.2019, at Tampere University of Technology (TUT), which became a part of Tampere University). The study was supervised by Professor Andre S. Ribeiro and co-supervised by Professor Ari Visa.

First, I am deeply grateful to Professor Andre S. Ribeiro, whose considerate and dedicated guidance enriched my doctoral studies with valuable learning experiences. Working in the multi-disciplinary, collaborative and creative environment of the LBD provided me with multiple opportunities for developing as a researcher and as a person.

Next, I would like to express my sincere gratitude to Professor Ari Visa, who kindly supported my doctoral studies during these years. I am also truly grateful to the TUT Graduate School for supporting my doctoral studies with a four-year grant.

Further, I am thankful to the members of LBD, including alumni, for partaking in this exciting journey towards uncovering the mysteries of bacterial gene expression. In particular, I would like to thank Vinodh Kandavalli and Nadia Goncalves for bringing excellent molecular biology expertise to our common projects that made this thesis possible, and Jason Lloyd-Price for sharing his knowledge in data analysis and modeling during my first year in LBD.

I also would like to thank my friends and family members whose precious emotional presence kept reminding me about life with even higher levels of complexity than that of bacterial cells. Finally, this study would not be possible without all the scientists, philosophers, artists and other remarkable individuals whose diverse works served as a foundation, support and inspiration for this thesis.

Tampere, October 2019

Sofia Startceva

Contents

Acronyms

AIC Akaike information criterion

BIC Bayesian information criterion

bp base pairs

CC closed complex

CME chemical master equation

DNA deoxyribonucleic acid

EC elongation complex

GFP green fluorescent protein

HILO highly inclined and laminated optical sheet

IPTG isopropyl β-D-1-thiogalactopyranoside

ITC initial transcribing complex

MLE maximum likelihood estimation

mRNA messenger RNA

OC open complex

PCR polymerase chain reaction

qPCR quantitative PCR

RBS ribosome binding site

RNA ribonucleic acid

RNAP RNA polymerase

[RNAP] RNAP concentration

SSA stochastic simulation algorithm

Δt distribution distribution of Δt intervals

Δt intervals time intervals between consecutive RNA production events in individual cells

$\mathbf{t_0}$ distribution distribution of time intervals between adding an inducer to the media and the first RNA production event

τ_c the average duration of the closed complex formation

τ_o the average duration of the open complex formation

τ_{prior} the average time spent in transcription initiation prior to commitment to the open complex formation

τ_{after}	the average time spent in transcription initiation after successful commitment to the open complex formation

List of Publications

This thesis is a compilation of the four original publications. In the text, these are referred to as **Publication I**, **Publication II**, **Publication III**, and **Publication IV**. The publications are reproduced with permissions from the publishers.

I **S. Startceva**, J.G. Chandraseelan, A. Visa, and A.S. Ribeiro (2016) "Quantitative Estimation of Long-living Fluorescent Molecules from Temporal Fluorescence Intensity Data Corrupted by Nonzero-mean Noise." *In Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)*, Vol.4: BIOSIGNALS, p. 17 – 24, Feb 21 – 23, Rome, Italy. DOI: 10.5220/0005605900170024.

II J. Lloyd-Price, **S. Startceva**, V. Kandavalli, J.G. Chandraseelan, N. Goncalves, S.M.D. Oliveira, A. Häkkinen, and A.S. Ribeiro (2016) "Dissecting the stochastic transcription initiation process in live *Escherichia coli*". *DNA Research* 23(3): 203 – 214. DOI: 10.1093/dnares/dsw009.

III N.S.M. Goncalves, **S. Startceva**, C.S.D. Palma, M.N.M. Bahrudeen, S.M.D. Oliveira and A.S. Ribeiro (2018) "Temperature-dependence of the single-cell variability in the kinetics of transcription activation in *Escherichia coli*". *Physical Biology* 15(2): 026007. DOI: 10.1088/1478-3975/aa9ddf.

IV **S. Startceva**, V.K. Kandavalli, A. Visa, and A.S. Ribeiro (2019) "Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression". *BBA - Gene Regulatory Mechanisms* 1862(2): 119 – 128. DOI: 10.1016/j.bbagrm.2018.12.005.

Below is a brief description of these publications, along with the information on the contributions of the author of this thesis.

Publication I presents a new method for quantitative estimation of fluorescent molecules from single-cell time-lapse microscopy data. We demonstrated that, when applied to the data affected by transient nonzero-mean noise, this method provides higher accuracy compared to the already existing methodology. Thus, it enables a more reliable analysis of the time-lapse microscopy data where fluorescent molecules can be temporarily absent from the focal plane. The author of this thesis conceived the study with A.S. Ribeiro, developed and implemented the method, and performed all stochastic simulations of *in silico* data. Following this, the author conceived and implemented the strategy for estimating the optimal values of parameters used in the method. Next, the author performed all image and data analysis of the *in vivo* microscopy measurements conducted by J.G. Chandraseelan. In addition, the author performed visual analysis of these data,

assisted by A. Visa. Finally, assisted by all authors, the author wrote the manuscript, which was revised by A.S. Ribeiro.

In **Publication II**, we proposed a methodology for characterizing the *in vivo* kinetics of the rate-limiting steps in bacterial transcription initiation and illustrated it by dissecting this kinetics for lac/ara-1 promoter in *Escherichia coli*. This methodology is similar to the existing steady-state assays for dissecting the rate-limiting steps *in vitro*. We justified the applicability of a similar approach *in vivo* by establishing that, within a certain range of cell growth media richness, while the concentration of RNA polymerases differs between conditions, the fraction of RNA polymerases available for transcription initiation does not change significantly. The author of this thesis performed image and data analysis with J. Lloyd-Price. Further, the author assisted J. Lloyd-Price in choosing methods for the image analysis and in developing the best-fitting model selection, which was used for dissecting the kinetics of transcription initiation. The author also assisted in writing the manuscript.

In **Publication III**, we estimated the mean and variability of the distribution of inducer intake times in *E. coli* following temperature shifts. For this, we developed a novel methodology based on deconvolution of empirically obtained distributions of time intervals defined largely by transcription activation and initiation kinetics. We found that the mean inducer intake time increases and variability diminishes following a shift to sub-optimal temperature, which is likely due to the emergence of additional rate-limiting steps in the intake process. The author of this thesis proposed a novel methodology for dissecting inducer intake kinetics. Experimental measurements were performed by N.S.M. Goncalves. The author of this thesis performed modeling and analysis of the measurements, assisted by C.S.D. Palma and M.N.M. Bahrudeen. Finally, the author participated in writing the manuscript as the main responsible for the sections related to modeling and computational analysis.

Publication III has been used by N.S.M. Goncalves in her Ph.D. dissertation.

Publication IV introduces the original idea of observing not only the mean and variability of gene expression kinetics but also its asymmetry and tailedness, so as to quantify the propensities for threshold crossing in RNA and protein numbers. In this study, from empirical distributions of intervals between transcription events, we found that when such thresholds are high, they are best reached by increasing the asymmetry and tailedness of this distribution, rather than the variance. We also demonstrated that these asymmetry and tailedness are regulated by the rate-limiting steps in transcription initiation, both for conditions that differ only in promoter sequence and for conditions that differ only in regulatory factors. Further, these asymmetry and tailedness are negatively correlated with the asymmetry and tailedness of the distribution of protein expression levels. We thus concluded that these asymmetry and tailedness may be the key regulatory variables that assist decision-making in *E. coli*. The author of this thesis designed the study with A.S. Ribeiro, assisted by V.K. Kandavalli. Experimental measurements were conducted by V.K. Kandavalli. The author performed modeling and analysis of the measurements. Finally, the author wrote the manuscript with A.S. Ribeiro and V.K. Kandavalli.

1 Introduction

1.1 Motivation

Bacteria survive in diverse, fluctuating environments (Kussell and Leibler, 2005) due to their cell-to-cell variability and phenotypic plasticity (Casadesús and Low, 2006, 2013; Healy and Schulte, 2015; Huang and Agrawal, 2016; Mitchell et al., 2009; Rao et al., 2002; Smits et al., 2006). These properties are based on stochastic processes (Acar et al., 2008; Süel et al., 2006) and on genetic mechanisms of decision-making using environment sensing (Arkin et al., 1998; Golding, 2011; Ribeiro, 2008; Smits et al., 2006; Wolf and Arkin, 2002). These mechanisms are implemented in cells based on small genetic circuits, known as 'motifs' (Milo et al., 2002), that are capable of, e.g., time counting, noise filtering, information storage, and binary decision-making (Alon, 2007; De Lay and Gottesman, 2012; Porter et al., 2012; Shen-Orr et al., 2002; Wolf and Arkin, 2002, 2003).

Qualitative changes in the behavior of a motif occur when numbers of one or more of its component proteins cross certain thresholds (Alon, 2007; Arkin et al., 1998; McAdams and Arkin, 1997; Panovska-Griffiths et al., 2013). The mechanisms underlying the regulation of the threshold crossing propensities of protein numbers are yet to be understood. Since proteins are translated from RNAs, one of the key factors that define protein abundances in organisms from bacteria to mammals is RNA levels (Li et al., 2014; Liu et al., 2016; Vogel and Marcotte, 2012). Thus, to understand gene network motifs behavior, it is crucial to study the regulation of RNA levels of the component genes.

RNA numbers are determined by the rates of RNA transcription from DNA template strand and RNA degradation. In bacteria, and in particular in *E. coli*, RNA degradation is known to be independent of RNA abundance, gene sequence, and metabolic function of the corresponding protein (Bernstein et al., 2002; Chen et al., 2015; Deutscher, 2006; Vogel and Marcotte, 2012), suggesting that the regulation of RNA levels occurs largely during transcription. Since not only the average rate but also noise in gene expression is able to affect the behavior of genetic circuits (Arkin et al., 1998; Kærn et al., 2005; McAdams and Arkin, 1997; Raj et al., 2006; Raser and O'Shea, 2005), it is crucial to understand how RNA production kinetics is regulated.

Previous research has shown that most of the regulation in transcription occurs at the stage of initiation (Browning and Busby, 2004, 2016; McLeod and Johnson, 2001; Ruff et al., 2015a). Moreover, transcription initiation includes several rate-limiting steps that can be affected by such factors as changes in promoter sequence, binding of regulatory molecules, σ factor competition, and transient topological constraints due to DNA supercoiling build-up, among others (Chong et al., 2014; deHaseth et al., 1998; Duchi et al., 2016; Kandavalli et al., 2016; Kærn et al., 2005; Lutz et al., 2001; McClure, 1985).

While these rate-limiting steps were originally dissected *in vitro* using steady-state assays

(Bertrand-Burggraf et al., 1984; McClure, 1980, 1985), a more recent method known as MS2-GFP tagging allows measuring transcription initiation kinetics in live cells over time (Fusco et al., 2003; Golding et al., 2005). Namely, it allows obtaining a distribution of time intervals between consecutive transcription events, collected from individual cells. Meanwhile, analytical and stochastic models of gene expression based on master equations were used to explore the possibilities and limitations of this biological system, given the existing knowledge (Arkin et al., 1998; Kepler and Elston, 2001; Rajala et al., 2010; Ribeiro, 2010; Sanchez et al., 2011a). By combining these experimental and modeling techniques, this thesis aims to study the regulation of gene expression at the stage of transcription initiation.

1.2 Aims of the study

This thesis examines the regulation of the kinetics of bacterial gene expression at the stage of transcription initiation. We hypothesized that the rate-limiting steps in transcription initiation can be studied *in vivo* using an approach analogous to the *in vitro* steady-state abortive initiation assay (McClure, 1980). This hypothesis is based on the fact that a relatively recent technology for *in vivo* ribonucleic acid (RNA) detection allows observing the kinetics of transcription initiation *in vivo* at the level of detail that was not previously possible (Golding and Cox, 2004). Further, we hypothesized that not only mean and noise in transcription initiation rate but also asymmetry and tailedness of this process are sensitive to environmental factors and depend on a promoter sequence, and that the differences in these features between conditions can affect the decision-making processes in living cells. This hypothesis is based on the past works that demonstrated that the rate-limiting steps in transcription initiation can vary independently from each other between conditions (Browning and Busby, 2016), potentially resulting in multifarious kinetics of transcription initiation. The thesis has four primary aims that had to be reached in order to test these hypotheses.

First, we had to develop a methodology for increasing the accuracy of estimating transcription initiation kinetics from measurements. Currently, the most direct method to measure *in vivo* transcription initiation kinetics in individual cells over time is single-cell time-lapse microscopy of fluorescently tagged RNA molecules (Golding and Cox, 2004). As these tagged RNAs do not degrade on the time-scale of the measurements (Tran et al., 2015), knowing their total intensities in individual cells at each time moment allows estimating the time intervals between consecutive RNA production events in individual cells (Δt intervals). However, although these molecules do not degrade during the measurement time, they can temporarily disappear from the image, e.g. due to moving out of the focal plane. While the already existing methodology accounts for this noise (Häkkinen and Ribeiro, 2015), it does not consider the cases where negative noise in intensity values appears transiently, for time lengths comparable to a cell lifetime, as observed in our microscopy measurements. Thus, we aimed to develop a new method that allows a more accurate quantitative estimation of fluorescent molecule numbers from temporal fluorescence intensity data corrupted by transient nonzero-mean noise. These goals were achieved in **Publication I**.

Achieving the first aim allowed obtaining a distribution of Δt intervals (Δt distribution) so as to next estimate the number, durations, and order of the rate-limiting steps in transcription initiation. Thus, the second aim was to dissect *in vivo* transcription initiation kinetics. This was previously done *in vitro* using the concept of the Lineweaver-Burk plot

(often called a τ plot in this context) (Bertrand-Burggraf et al., 1984; Buc and McClure, 1985; McClure, 1980). To proceed in analogy to this *in vitro* methodology, we aimed to find conditions where the closed complex formation time would vary, while the kinetics of other rate constants in transcription initiation would not differ significantly. Further, from empirical data measured in such conditions, we aimed to infer the best-fitting model of transcription initiation. The novel *in vivo* methodology that allows achieving these goals was presented in **Publication II**, along with an example application to an *E. coli* promoter.

The kinetics of transcription initiation can also be regulated, directly or indirectly, by transiently present regulatory molecules such as activators, repressors and other transcription factors (Mäkelä et al., 2017; Tran et al., 2015), and it is not impervious to environmental changes such as acidity and temperature (Muthukrishnan et al., 2014; Oliveira et al., 2016a). Give this, the third aim was to study the kinetics of transcription activation by an inducer introduced to the media, as a function of temperature. For this, we planned to conduct single-cell time-lapse microscopy measurements of transcription initiation kinetics (i) following the addition of inducer to the media and (ii) at the constant intracellular inducer concentration, at various temperatures. We then aimed to develop a novel methodology for estimating inducer intake times from these measurements and dissecting the inducer intake kinetics. These goals were achieved in **Publication III**.

Finally, given that cells make decisions by crossing thresholds in protein numbers (Alon, 2007), the fourth aim was to explore the role of transcription initiation kinetics on cellular decision-making processes. For this, we investigated how transcription initiation kinetics can affect this threshold crossing process. Recent studies demonstrated that *in vivo* kinetics of the rate-limiting steps in transcription initiation defines not only its mean rate but also variability (Häkkinen and Ribeiro, 2016; Mäkelä et al., 2017). Consequently, the author and colleagues aimed to investigate whether it is possible to tune asymmetries in transcription initiation kinetics independently from its mean rate and variability, and whether changes in these asymmetries have a significant effect on threshold crossing in RNA and protein numbers. For this, we used single-cell, time-lapse microscopy measurements (population and time series) of transcription initiation kinetics in various conditions differing in promoter sequence, induction schemes and media composition, along with tailored computational and statistical analysis. We also tested whether the asymmetries in transcription initiation kinetics correspond to the asymmetries in protein expression levels in individual cells. This study was presented in **Publication IV**.

1.3 Thesis outline

This thesis is organized as follows. Chapter 2 provides background information, with the emphasis on transcription initiation in *E. coli* and stochastic models of this process. Chapters 3 covers the methodology used for empirical data collection and initial data processing along with the statistical analysis techniques used in this thesis. Chapter 4 presents a summary of the results, and Chapter 5 contains conclusions and discussion.

2 Background Review

2.1 *Escherichia coli* as a model organism

Escherichia coli, a gram-negative rod-shaped bacterium, is one of the most studied living organisms and arguably is the standard model bacterium. This bacterium was discovered by Theodor Escherich in 1886 and sequenced in 1997 (Blattner et al., 1997). Most of the established concepts in molecular biology were derived from investigating *E. coli*. Further, studies of this relatively simple and widespread bacterium have contributed to the understanding of evolutionary processes. *E. coli* is usually selected as a model organism for the development of new genetic engineering techniques and is used in pharmaceutics for *in vivo* synthesis of chemicals relevant to the treatments of human diseases (Blount, 2015; Cooper, 2000).

2.2 Central dogma of molecular biology

Even the simplest of living cells have evolved intricate machinery that allows surviving, reproducing, and performing other functions under various conditions. The information utilized in such self-sufficient systems is encoded in genes, the elements that define characteristics of the species and of the individual belonging to it, and stored in cells in the form of deoxyribonucleic acid (DNA). The basic principles of preserving and extracting this information can be described by a central dogma of molecular biology (Figure 2.1). General transfers occur in all known living cells and involve producing copies of DNA by replication, information flow from DNA to RNA by transcription, and from RNA to proteins by translation. In special cases, such as viral infection or controlled laboratory setting, RNA can be replicated or can transfer information to DNA by reverse transcription, and a protein can be translated directly from DNA. These special cases do not occur in *E. coli* model that is not infected with a virus. In general, the information cannot be transferred from protein to RNA or DNA. (Alberts et al., 2008; Crick, 1970)

DNA is a double-stranded polymer where each strand is comprised of four unit types called nucleotide bases: adenine, cytosine, guanine, and thymine. The strands are complementary to each other in the sense that they form only adenine-thymine and cytosine-guanine base pairs (bp). *E. coli* DNA is a circular chromosome, a single molecule that includes about 4.6 million nucleotide pairs, whereas human DNA in haploid cells, for comparison, consists of 23 chromosomes and includes about 3.2 milliard nucleotide pairs. However, the size of the genome should not be used as a sole indicator of the organism complexity, since organisms differ in fractions of non-coding DNA sequence, even between the organisms of similar complexity (Alberts et al., 2008; Gil and Latorre, 2012). DNA is replicated once per cell cycle by DNA polymerase unwinding and pulling apart the DNA double helix and using each strand as a template to synthesize a new complementary

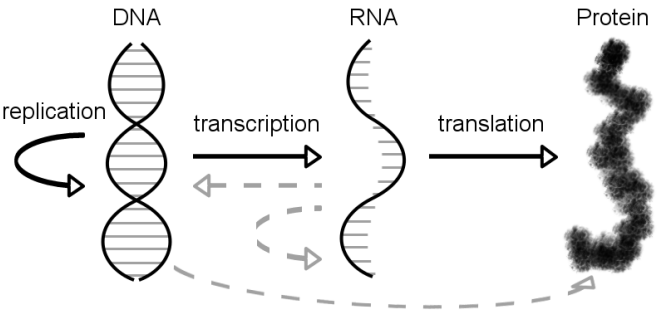


Figure 2.1: The central dogma of molecular biology. General transfers are shown as solid arrows. DNA stores genetic information and can be duplicated by replication. RNAs are transcribed from DNA, and proteins are translated from RNA. Special transfers are shown as dashed arrows. These are rare and occur only *in vitro* or in certain virus-infected cells.

strand (Alberts et al., 2008; Nielsen and Løbner-Olesen, 2008). Since DNA strands are asymmetric, namely they possess a 5′ end and a 3′ end, these polymers can be synthesized *in vivo* only in the 5′-to-3′ direction.

Meanwhile, RNA is a single-stranded polynucleotide, which differs from the DNA composition in that a nucleotide base uracil is used instead of thymine. In living cells, RNA is usually produced by transcription. This process starts with RNA polymerase (RNAP) recognizing and binding, with the help of transcription factor(s), to a DNA region called promoter, which is located upstream of the coding sequence and indicates a starting point for RNA synthesis. Next, RNAP opens the double helix and proceeds with the synthesis of a complementary RNA. Upon encountering the terminator sequence, RNAP halts and releases both the newly formed RNA and the DNA template (Alberts et al., 2008). RNA can either directly participate in cellular machinery or be used as a template for synthesizing proteins. The genetic code in the latter, called messenger RNA (mRNA), is written in triplets of nucleotides called codons, with each codon either standing for a specific amino acid or signaling the end of the region encoding a specific protein. This genetic code is redundant, with different codons able to correspond to the same amino acid, which contributes to the random mutations resistance. In eukaryotes, the transcribed RNA should undergo the process of non-coding sequence removal, splicing, before it is available for translation.

Proteins are polypeptide molecules that are usually built from up to 20 common amino acids in the process of translation from mRNA. This process is initiated by a complex molecular machine called ribosome assembling around the mRNA. In case of bacteria, this happens at the ribosome binding site (RBS), a segment of an mRNA that is located closely upstream of the start codon. While in eukaryotes the mRNA should be fully produced before it moves to the cytoplasm where ribosomes are available, in prokaryotes translation can be initiated before the completion of transcription. Amino acids are brought to the translation site by a transfer RNAs, which can recognize the corresponding mRNA codons and bind to them. When the start codon is bound by the transfer RNA, the translation elongation can begin. As the following mRNA codon is bound, the ribosome moves along the strand, binds the newly brought amino acid to the previous one and frees the transfer RNA. This process repeats until the ribosome encounters a stop codon, where translation termination occurs. During the termination, the ribosome releases the nascent

polypeptide and the mRNA. Finally, translation is followed by protein folding, during which the polypeptide chain attains the three-dimensional structure that is required for the protein to perform its functions. (Alberts et al., 2008; Lodish et al., 2016)

It is noteworthy that a given promoter does not necessarily precede only one gene. Instead, it can control an operon, which is a cluster of co-regulated genes. Operons occur frequently in a prokaryotic genome, but are more uncommon in eukaryotes (Osbourn and Field, 2009). Transcription of an operon produces a single mRNA that encodes a set of proteins usually required for a common task. Each gene encoded on such mRNA is called an open reading frame. These genes are translated separately and with rates differing up to 100 times within the same operon (Burkhardt et al., 2017). Further, promoters of some genes are closely spaced, which allows for additional means of gene expression regulation (Beck and Warren, 1988; Chen et al., 2016; Zafar et al., 2014).

2.3 Regulation of bacterial transcription initiation

While the commonalities in the regulation of transcription initiation spawn from non-cellular life to eukaryotes, the specificities of this process differ between the domains of life (Grohmann and Werner, 2011; Ptashne, 2005; Travers and Muskhelishvili, 2007; Werner and Weinzierl, 2002). At the same time, transcription initiation in bacteria, in general, employs similar regulatory mechanisms across the species (Browning and Busby, 2004, 2016; Hochschild and Dove, 1998; Rojo, 1999; Roy et al., 1998; Ruff et al., 2015b; Saecker et al., 2011). One example of this can be bacterial RNAP, which is known to be highly conserved within the domain while differing from viral, eukaryote and archaea RNAPs (nevertheless being related to these) (Grohmann and Werner, 2011; Ishihama and Nagata, 1988; Lane and Darst, 2010; Lemon and Tjian, 2000; Roeder and Rutter, 1969; Werner and Weinzierl, 2002).

Transcription initiation is arguably the most regulated stage in bacterial gene expression, likely due to the fact that it is more energy-efficient to exert the regulation before the assembling of mRNA and proteins starts (Browning and Busby, 2004, 2016; deHaseth et al., 1998; McLeod and Johnson, 2001; Ruff et al., 2015b; Thattai and van Oudenaarden, 2001). This regulation can be exerted by direct or indirect tuning of the concentrations of molecules that are required for transcription initiation to proceed, as well as of other regulatory molecules (Browning and Busby, 2004; Errington, 2003; Yanofsky, 2004), and by transient topological constraints in promoter DNA (Bryant et al., 2014; Chong et al., 2014). Further, transcription initiation kinetics can be altered by promoter modifications, such as DNA methylation (Casadesús and Low, 2006, 2013) and changes in a promoter DNA sequence (Brewster et al., 2012; Browning and Busby, 2016; Garcia et al., 2012; Ruff et al., 2015b; Wisniewski-Dyé and Vial, 2008). Since spontaneous mutations in DNA sequence occur at a slower rate than e.g. environmental changes, gene expression is, in general, regulated by other means (Wisniewski-Dyé and Vial, 2008).

Note that transcription initiation is by no means the only significant contributor to regulation in gene expression (Van Assche et al., 2015). Transcription elongation (Dobrzyński and Bruggeman, 2009; Mironov et al., 2002; von Hippel and Pasman, 2002), mRNA degradation (Bernstein et al., 2002; Chen et al., 2015; Deutscher, 2006), translation (Byrgazov et al., 2013; Wilson et al., 2016), and protein degradation (Dougan et al., 2002; Goldberg, 2003) are also able to introduce layers of global and local regulation.

2.3.1 Transcription initiation as a multi-step process

Transcription initiation is composed of multiple steps that can become rate-limiting depending on their durations and interactions with other steps (deHaseth et al., 1998; Henderson et al., 2017; McClure, 1985; Record et al., 1996; von Hippel et al., 1984). Although this process involves multiple conformational changes, transcription initiation of an active promoter is usually represented as a sequence of the following steps (Figure 2.2A). First, RNAP recognizes and binds to a promoter, forming a closed complex (CC). When bound, RNAP unwinds approximately 14 bp surrounding the transcription site, forming an open complex (OC) (Murakami and Darst, 2003; Record et al., 1996; Young et al., 2002). Further, RNAP begins synthesizing an RNA, forming an initial transcribing complex (ITC). While in ITC, the RNAP-promoter complex is likely to abort the production of nascent RNA and reverse to the OC stage, releasing the abortive RNA. As the nascent RNA reaches a certain length (approximately 10±3 nucleotides, depending on the promoter), the ITC stabilizes into an elongation complex (EC) and can no longer be reversed to OC (Henderson et al., 2017; Hsu, 2008; Murakami and Darst, 2003; Revyakin et al., 2006). This stabilization is likely caused by the accumulation of scrunching and other stresses that drive the promoter escape (Henderson et al., 2017; Kapanidis et al., 2006; Revyakin et al., 2006).

Several *in vitro* and *in vivo* studies demonstrated that, at least for some promoters, the branched pathway of transcription initiation explains the experimental observations better than the sequential one (Henderson et al., 2017; Kubori and Shimamoto, 1996; Susa et al., 2002, 2006). This pathway (Figure 2.2B) differs from the sequential pathway in that, from a CC, the system can either move to a productive OC (OC_P) or to a non-productive one (OC_{NP}, also known as a "moribund OC"). The OC_P results in a relatively fast EC formation and production of a full-length RNA transcript, whereas the OC_{NP} results only in abortive cycling accompanied by the release of a short abortive RNA (Henderson et al., 2017). Both of these complexes can be reversed to the CC by RNA cleavage factors GreA and GreB (Sen et al., 2001; Susa et al., 2006). The branched pathway is hypothesized to provide additional regulatory flexibility, not only by varying the kinetics of the rate-limiting processes in transcription initiation but also by amplifying effects of repressors or activators (Susa et al., 2002). In addition, the regulatory role of abortive RNA transcripts remains unclear, and further research is needed in order to understand the role of the branched pathway *in vivo* (Henderson et al., 2017).

Given the multi-step nature of transcription initiation, it is possible to selectively affect some of those steps while not altering other steps significantly. For example, GreA and GreB factors and Mg^{2+} ions affect only the duration spent in transcription initiation after the closed complex is formed (deHaseth et al., 1998; Hochschild, 2007; Suh et al., 1992; Susa et al., 2006). Meanwhile, according to Figure 2.2, RNAP concentration affects only the time spent in closed complex formation. Moreover, the total time spent in closed complex formation changes linearly with the inverse of RNAP concentration, and closed complex formation is assumed to be infinitely fast when the inverse of RNAP concentration approaches zero. This was demonstrated *in vitro* in (Bertrand-Burggraf et al., 1984; Buc and McClure, 1985; McClure, 1980) and *in vivo* in **Publication II**. Given the above, and since RNAP is required for transcription of any gene, RNAP concentration can be considered a global regulator of the time spent in closed complex formation. Note that the genes which spend a longer time in closed complex formation would be affected stronger by the changes in RNAP concentration.

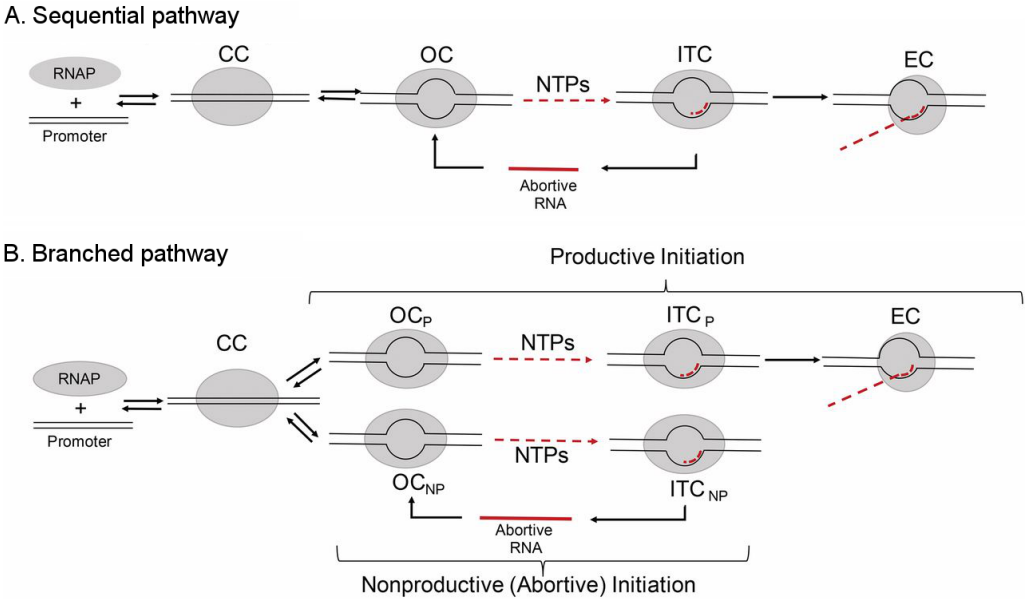


Figure 2.2: Possible pathways of transcription initiation, where RNAP is an RNA polymerase, CC is closed complex, OC is open complex, ITC is initial transcribing complex, EC is escaped (elongation) complex, and NTP represents nucleoside triphosphate molecules. (A) Sequential pathway, where abortive synthesis occurs on the path to promoter escape; (B) Branched pathway, where productive complexes (subscripted P) lead to promoter escape without abortive synthesis, while productive complexes (subscripted NP) lead to abortive synthesis but not to promoter escape. Adapted from (Henderson et al., 2017).

2.3.2 RNA polymerase composition

The main body of bacterial RNAP, called core RNAP complex or RNAP core enzyme, consists of five subunits: α (two copies), β, β', and ω (Haugen et al., 2008). The α dimer acts as a scaffold to which β and β' subunits bind with the help of ω subunit (Minakhin et al., 2001). As a part of a functional core enzyme, α subunits engage in both sequence-specific and sequence-non-specific interactions with promoter DNA, and also interact with transcription regulatory factors (Finney et al., 2002; Gourse et al., 2000; Ross et al., 1993). Two largest subunits, β and β', form the two pincers of the claw-shaped machinery responsible for RNA synthesis (Darst, 2001; Haugen et al., 2008; Mekler et al., 2002; Naryshkin et al., 2000; Zhang et al., 1999). The smallest of these components, the ω subunit, primarily facilitates RNAP assembly and enhances association between the other subunits (Gunnelius et al., 2014; Mathew and Chatterji, 2006). Core RNAP complex is capable of transcription elongation and termination but unable to recognize a promoter and initiate transcription (Browning and Busby, 2004).

A dissociable σ subunit (also known as a σ factor) is responsible for RNAP promoter recognition (Feklístov et al., 2014; Murakami, 2015). The σ factor can bind to the core RNAP complex, forming an RNAP holoenzyme, an RNAP configuration that is able to successfully initiate transcription (Figure 2.3A, the holoenzyme structure is described in detail in (Murakami et al., 2002)). While core enzyme has a fixed composition, a holoenzyme can form with one of the several σ factors available in the cell, acquiring a promoter sequence affinity defined by the σ factor (Mauri and Klumpp, 2014; Mooney

et al., 2005; Murakami and Darst, 2003). Thus, although both α and σ subunits can engage in sequence-specific interactions with promoter DNA (Figure 2.3B), most of the sequence-specific RNAP regulation is implemented via σ factor competition (Browning and Busby, 2004; Haugen et al., 2008; Mauri and Klumpp, 2014).

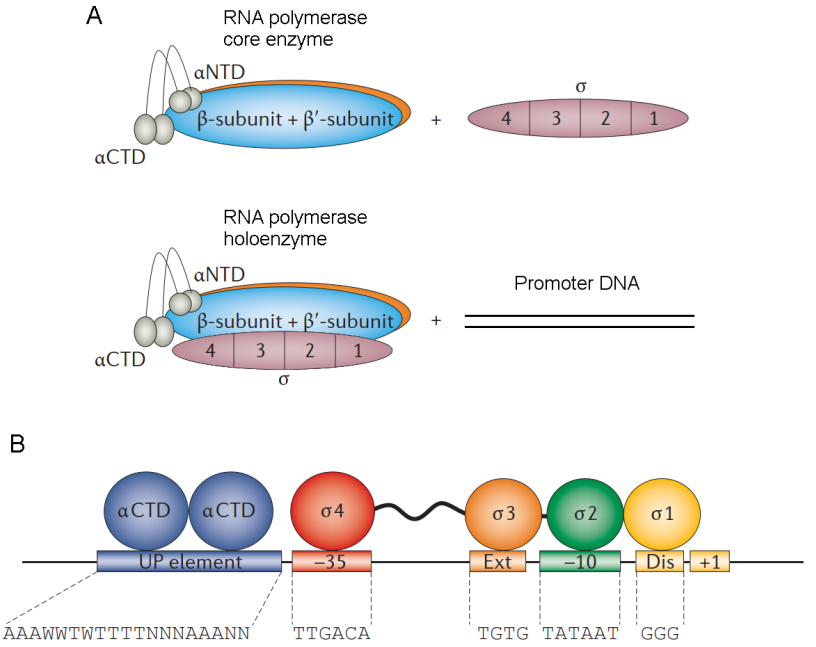


Figure 2.3: Schematic representation of RNA polymerase (RNAP) and its interaction with promoter DNA. (A) RNAP core enzyme is bound by a σ factor in order to form an RNAP holoenzyme. The holoenzyme can specifically bind to promoter DNA. RNAP subunits are represented as follows: α subunits in grey, each composed of an N-terminal domain, αNTD, and a C-terminal domain, αCTD, connected by a linker (grey line); β subunit in blue; β' subunit in orange; and σ factor, or σ subunit, in pink. Although not visible on the figure, the ω subunit is present both in the RNAP core enzyme and in the RNAP holoenzyme. (B) RNAP elements that contribute to sequence-specific binding to promoter DNA, along with their binding sites and their consensus sequences. Given that the transcriptional start cite is denoted as +1, the binding sites are positioned as follows: UP element is -37 to -58; -35 element is -35 to -30; the extended -10 element, Ext, is -17 to -14; the -10 element is -12 to -7; and the discriminator element, Dis, is -6 to -4. In the consensus sequences, W stands for A or T, and N is any base. Adapted by permission from Springer Nature Customer Service Centre GmbH: Nature Reviews Microbiology (Browning and Busby, 2016), copyright (2016).

2.3.3 Regulation by σ factor competition

Regulation by σ factor competition is enabled by the fact that the numbers of free RNAP available for transcription is limited, causing high competition for RNAP between genes (Browning and Busby, 2004; Grigorova et al., 2006; Ishihama, 2000; Maeda et al., 2000). Since different σ factors, in general, have differing affinities to a given gene, changes in the concentrations of holoenzymes containing these σ factors can strongly affect transcription initiation kinetics of this gene. These concentrations can be expressed as a function of the concentration of the holoenzymes containing other sigma factors,

along with the dissociation constants and total numbers of all σ factors (Mauri and Klumpp, 2014). Figure 2.4 gives an example of a competition between two σ factors, σ^{70} and σ^{Alt}. Interestingly, the σ factor competition mechanism provides an additional level of gene expression control. Namely, it allows for stress-responses that cannot be fully duplicated by adding activators to the system with only one type of σ factors (Gross et al., 1998).

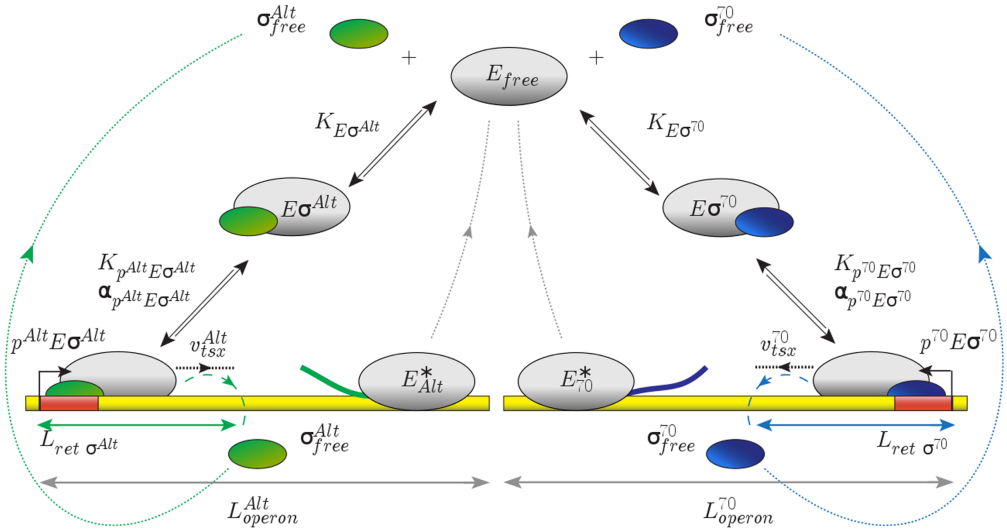


Figure 2.4: An example of σ factor competition for RNAP core enzyme, E_{free} (grey oval), given that two types of σ factors are present in the system: a housekeeping σ factor, σ^{70} (blue oval), and a generic alternative σ factor, σ^{Alt} (green oval). The σ factors σ^{70} and σ^{Alt} compete for binding to E_{free} and forming an RNAP holoenzyme, $E\sigma^{70}$ or $E\sigma^{Alt}$, respectively. Dissociation of this holoenzyme competes with reversible binding to a promoter specific to $E\sigma^{70}$ or $E\sigma^{Alt}$, respectively. Successful transcription initiation is followed by transcription elongation, during which the σ factor is released after RNAP moves forward $L_{ret}\sigma^{70}$ or $L_{ret}\sigma^{\mathrm{Alt}}$ nucleotides, respectively. The released σ factor enters the pool of the σ factors available for the competition. When transcription is complete, the RNAP core enzyme, denoted E^*_{70} or E^*_{Alt}, respectively, after the gene it transcribes, is also released and enters the pool of E_{free}. Adapted from (Mauri and Klumpp, 2014).

The change in RNAP holoenzyme concentration can also be achieved without altering relative σ factor concentrations, by changing the concentration of the core RNAP complex instead. This strategy was employed in **Publications II** and **IV**. In these studies, in most cases, we deliberately avoided experimental conditions where the σ factor competition mechanism could significantly affect the data, since this would obscure the results of interest.

2.3.4 Regulation by transcription factors

Transcription factors are the proteins that can bind to promoters, either independently or in cooperation with each other, and upregulate or downregulate transcription (Babu and Teichmann, 2003; Browning and Busby, 2004; Pérez-Rueda and Collado-Vides, 2000). Some of those target only a small amount of genes, whereas others act as global regulators (Babu and Teichmann, 2003; Hochschild and Dove, 1998; Martínez-Antonio and Collado-

Vides, 2003). The most common mechanisms of repression and activation by transcription factors in transcription initiation are shown in Figure 2.5.

One widespread repression mechanism is a steric hindrance, where repressors bind to promoter operator sites and their physical presence spatially obscures access to -10 or -35 elements, thus blocking the recruitment of RNAP (Garcia and Phillips, 2011; Oehler et al., 1990). This mechanism can be seen as a competition between the repressor and RNAP for binding to the promoter. Another repression mechanism is based on the transcription factor binding to the operators located near the RNAP recruitment site and at some distance from one another, thus bending DNA into a loop. This topological constraint prevents RNAP recruitment (Müller et al., 1996; Oehler et al., 1990; Swint-Kruse and Matthews, 2009). The less widespread mechanism involves "anti-activator" repressors that bind to a promoter near the activator binding site and interact with the adjacent activator (Valentin-Hansen et al., 1996). In addition, the concentration of free RNAP holoenzymes can be affected by small RNAs which can mimic the target promoters and block an RNAP by binding to it (Browning and Busby, 2016; Burenina et al., 2015; Wassarman and Storz, 2000). While the above mechanisms mostly affect the steps before a closed complex is formed, repressors can also hinder an open complex formation or promoter clearance (Browning and Busby, 2016; Heltzel et al., 1990; Monsalve et al., 1996; Sanchez et al., 2011b).

Transcription activation at many promoters is relatively simple and involves only a single activator, with three general mechanisms being distinguished (Browning and Busby, 2004, 2016; Lee et al., 2012). In class I activation, an operator upstream of the -35 element of the promoter is bound by a transcription factor that aids in recruiting RNAP by interacting with the α subunits (Browning and Busby, 2004; Ebright, 1993). The binding location of the activator can vary in a certain range, since α subunit has a flexible linker that allows adjusting the distance at which the interaction can occur (Browning and Busby, 2004). In class II activation, a transcription factor binds to the DNA region that overlaps with the -35 element and interacts with the RNAP holoenzyme, usually with the domain 4 of the σ subunit (Browning and Busby, 2004; Dove et al., 2003). While both class I and class II activation facilitate promoter recruitment, class II activation can also affect other steps in transcription initiation. In addition, class I and II activation can work in cooperation, which is the case at many bacterial promoters that are regulated by two input signals (Browning and Busby, 2004). The third mechanism, activation by conformational change, occurs at the promoters that usually have a non-optimal spacing between -35 and -10 elements, by distorting the promoter DNA in a way that positions the -35 and -10 elements better for RNAP binding (Browning and Busby, 2016; Philips et al., 2015).

In many cases, activity of a bacterial promoter depends on several factors simultaneously, thus many promoters are subject to cooperative regulation by several transcription factors (Browning and Busby, 2004). It is possible for the same transcription factor to act as an activator or repressor depending on the promoter it is bound to (Pérez-Rueda and Collado-Vides, 2000). Moreover, transcription factors themselves are subject to various regulatory mechanisms, e.g. their DNA binding affinity can be altered by small ligands whose concentration depends on the environmental conditions (Browning and Busby, 2004). Also, small ligands can directly affect transcription initiation rate, e.g. the global regulatory nucleotide ppGpp destabilizes the open complex at some promoters by binding to RNAP (Ross et al., 2013). In addition, bacterial CRISPR-dCas9 system, a synthetic tool for sequence-specific regulation of gene expression, is able to affect binding affinities

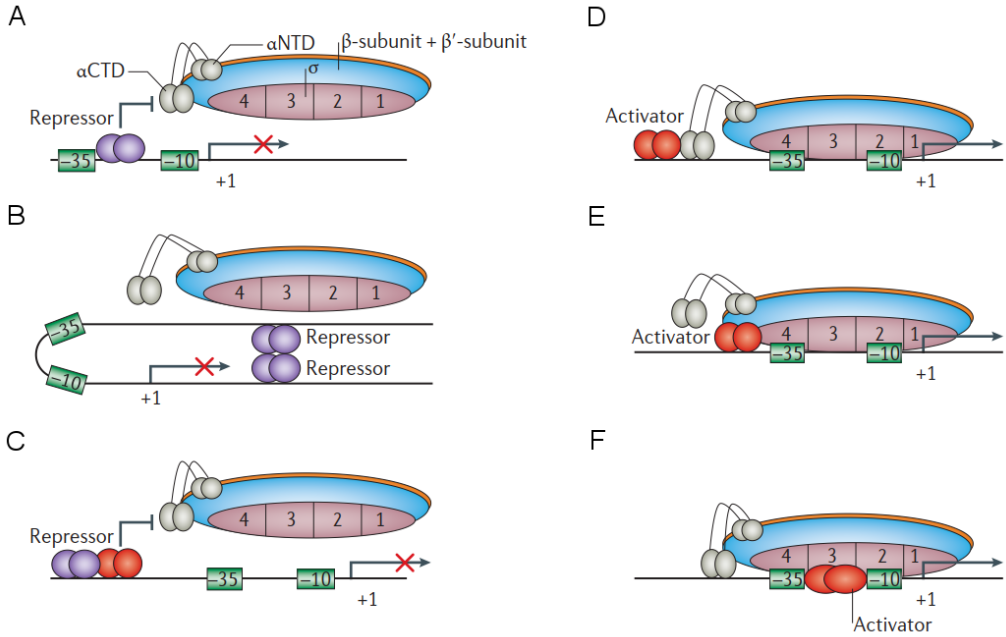


Figure 2.5: Schematic representation of common repression and activation mechanisms in bacterial transcription initiation. (A) Repression by steric hindrance; (B) Repression by looping; (C) Repression by modulation of an activator; (D) Class I activation; (E) Class II activation; (F) Activation by a promoter conformation change. RNAP composition is described in Figure 2.3. Repressor molecules are shown in purple, and activator molecules are shown in red. -35 and -10 promoter elements are highlighted with green boxes. Transcriptional starting cite is marked with +1. Adapted by permission from Springer Nature Customer Service Centre GmbH: Nature Reviews Microbiology (Browning and Busby, 2016), copyright (2016).

of transcription factors at targeted promoters (Dominguez et al., 2015).

2.3.5 Regulation by promoter modifications

Transcription initiation can also be regulated by chemical modification of nucleotide bases or changes in their sequence in the promoter region (Browning and Busby, 2016). The most widespread modification mechanism is DNA methylation, which is able to strongly affect the affinities with which transcription factors bind to their operator sites. For example, when DNA adenine methylase binds to GATC regions in the promoter region of *pap* or *agn43 E. coli* gene, repressor molecules lose the ability to bind to this area. Since methylation of this region does not prevent RNAP form proceeding with transcription, the gene becomes activated (Figure 2.6A). While DNA methylation is often overlooked in eukaryotes, it has the potential to cause significant changes in bacterial gene expression (Browning and Busby, 2016; Casadesús and Low, 2006; Sánchez-Romero et al., 2015). In some cases, modification by DNA methylation can be inherited by daughter cells, even though no actual changes in the DNA sequence take place (Veening et al., 2008).

Meanwhile, the chemical modification mechanism that leads to the most drastic effect at the promoter region is a site-specific DNA inversion (Figure 2.6B). It is performed by recombinases that recognize short inverted repeat sequences located upstream and

downstream of the DNA segment to be inverted. This segment usually contains a promoter (Wisniewski-Dyé and Vial, 2008). The site-specific inversion mechanism sets the state of a gene to either active or inactive, depending on whether the gene in upstream or downstream of the promoter. This provides a different mode of regulation from transcription factors, where the effect is usually proportional to the concentration of these factors (Browning and Busby, 2016). In *E. coli*, this mechanism is used to control the production of the major subunit of type I pili (Browning and Busby, 2016; Wisniewski-Dyé and Vial, 2008; Wolf and Arkin, 2002).

A more subtle mechanism for regulation by promoter modification is local sequence variation (Figure 2.6C). The mechanism is based on differences in the length of a DNA tract that consists of single nucleotide (or dinucleotide) repeats and is often located near the -35 element of a promoter. The length of this tract can affect transcription initiation rate, e.g. by tuning RNAP binding affinity. This length is randomly determined during DNA replication and thus differs between cell generations and between cells of the same population. Such variation allows at least a fraction of cells in a given population to have the optimal transcriptional activity at the regulated promoter (Browning and Busby, 2016).

All the three mechanisms described above are able to produce phase variation, the intra-population diversity that is crucial for survival in rapidly changing environments. This phase variation is based on the switching of bacterial phenotype at a much higher frequency than random mutations could occur (Casadesús and Low, 2013; Wisniewski-Dyé and Vial, 2008). While local sequence variation happens by chance, the DNA methylation and site-specific inversion are often regulated based on sensing the environment (Browning and Busby, 2016).

2.3.6 Example case of lac/ara-1 promoter

A synthetic hybrid promoter lac/ara-1 ($P_{lac/ara\text{-}1}$) includes (i) the activation operator site from the araBAD promoter (P_{araBAD}) and (ii) repression operator sites from the lac promoter (P_{lac}) (Lutz and Bujard, 1997; Stricker et al., 2008). The first one allows $P_{lac/ara\text{-}1}$ to be activated by AraC protein in the presence of arabinose, a monosaccharide (Stricker et al., 2008). The second one enables $P_{lac/ara\text{-}1}$ to be repressed by LacI protein in the absence of isopropyl β-D-1-thiogalactopyranoside (IPTG) (Stricker et al., 2008).

The activation system by arabinose exemplifies cooperative behavior. For the promoter to become activated, two AraC proteins have to be bound upstream of the RNAP binding site, and arabinose molecules have to be bound to these AraC proteins in order to cause the conformational change that would allow AraC to interact with RNAP (Hendrickson and Schleif, 1984; Schleif, 2010). The regulation at repression operator sites derived from P_{lac} exemplifies cooperative repression and indirect activation. Namely, Lac repressor (LacI) forms a tetramer that is bound to an operator site upstream and an operator site downstream the RNAP binding site, forming a loop that blocks transcription initiation (Lutz and Bujard, 1997). When IPTG molecules are present in the cell, they bind to LacI molecules, which significantly reduces the ability of LacI to bind to its operator, thus indirectly activating the promoter. In addition, it also has been shown that the presence of IPTG interferes with induction by AraC (Lee et al., 2007; Schleif, 2010).

While P_{araBAD} and P_{lac} are considered to have at most intermediate regulatory flexibility (Deuschle et al., 1986; Lutz and Bujard, 1997), expression of a gene that is controlled by $P_{lac/ara\text{-}1}$ can be tuned in a significantly wider range compared to its predecessors, as

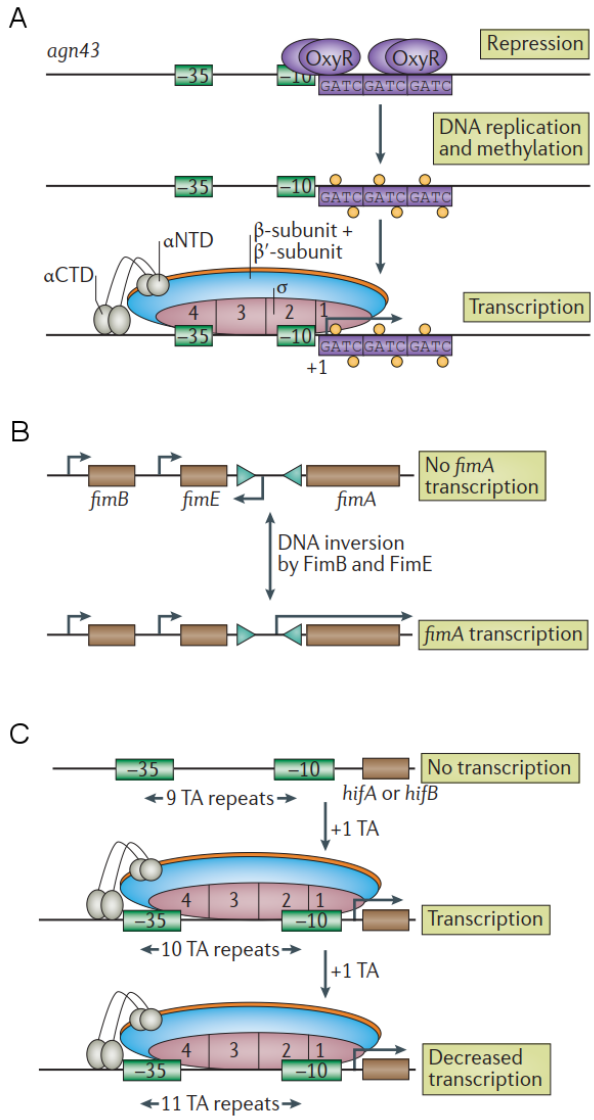


Figure 2.6: Schematic representation of common mechanisms of transcription initiation regulation by DNA modification. (A) DNA methylation. When the GATC regions of *ang43* gene is not methylated, it is bound by transcription factors OxyR (purple ovals) which block transcription initiation. Presence of DNA adenine methylase (orange circles) at the GATC regions prevents OxyR from binding, allowing transcription initiation to commence; (B) DNA inversion. The region between blue triangles can be inverted by FimB and FimE recombinases. *fimA* transcription can be initiated only when the promoter is oriented in the direction of the *fimA* gene. The elements on the DNA region shown are not at the realistic relative scale; (C) Local sequence variation. Promoters of *hifA* and *hifB* genes of *Haemophilus influenzae* differ between generations in sequence located between -35 and -10 promoter elements (the elements are highlighted with green boxes). Namely, they differ in the number of repeats of TA dinucleotides. The number of repeats affects RNAP binding rate, and thus regulates transcription initiation. RNAP composition is described in Figure 2.3. Genes are highlighted with brown rectangles and are not shown at the right relative scale. Adapted by permission from Springer Nature Customer Service Centre GmbH: Nature Reviews Microbiology (Browning and Busby, 2016), copyright (2016).

a result of the combination of the two different regulatory mechanisms described above (Lutz and Bujard, 1997). This suggests potential advantage of using synthetic promoters when aiming to develop tightly controlled genetic circuits, which in turn could have a vast array of applications (see e.g. Stricker et al., 2008).

In **Publication II**, we studied transcription initiation kinetics of $P_{lac/ara\text{-}1}$. In **Publication III**, we used $P_{lac/ara\text{-}1}$ to dissect the kinetics of inducer intake process. In **Publication IV**, we observed transcription initiation kinetics of $P_{lac/ara\text{-}1}$ in various conditions, along with transcription initiation kinetics of its mutants and other promoters.

2.4 Modeling chemical reaction systems

In classical chemical kinetics, a well-stirred system at temperature equilibrium is usually modeled using coupled first-order differential equations (Gillespie, 2007; Goutsias, 2007). Given the system with N reactant species, these reaction-rate equations describe how the number of molecules X_i of each chemical species S_i evolves in time as defined by the functions f_i $(i = 1, ..., N)$:

$$\frac{dX_i}{dt} = f_i(X_1, ..., X_N) \tag{2.1}$$

Reaction-rate equations do not provide information on the individual behavior of each molecule. Instead, they inform on the average behavior of molecules in each species. This approach works well for the systems where the number of molecules belonging to each reactant species is several orders of magnitude higher than one, which is the case in many *in vitro* experiments (Gillespie, 2007). In such systems, the number of times that a given reaction would occur in a small time interval is well-predicted from the rate of the reaction and the numbers of the reactants. However, when even one of the reactant species is present in low numbers, these predictions become inadequate (Gillespie, 2007; McAdams and Arkin, 1997).

The processes that take place in living cells are often based on the interaction of molecular species with low copy numbers (McAdams and Arkin, 1999). In the context of gene expression and its regulation, genes, RNAs, and regulatory molecules are often present in a cell in low copy numbers, and even small changes in these numbers can significantly affect behavior of gene regulatory networks (Mileyko et al., 2008; Taniguchi et al., 2010). Since continuous and deterministic reaction-rate equations cannot account for this biochemical noise, an approach that operates discrete and stochastic variables was introduced to model such systems more accurately (Gillespie, 2007; Munsky and Khammash, 2008).

The most precise way of describing these stochastic processes would be to explicitly model the state of the system at each moment by tracking the exact position and momentum of each molecule, and thus being able to calculate when and at which point in space the collisions between the molecules would occur. While this approach is technically correct, it is usually computationally demanding to the point of being impractical (Gillespie, 2007). Instead, given the assumption that the reactants are well-mixed, it is possible to omit the direct modeling of the space, and use a probabilistic approach described in the following subsections 2.4.1 and 2.4.2 (Gillespie, 1977, 1992, 2007).

2.4.1 Chemical master equation

Let us assume a system of molecules that belong to N chemical species $S_i(t)$ $(i = 1, ..., N)$, where the number of molecules of each species at a given time moment is denoted as $X_i(t) = x_i$. Then the state of the system is represented by an N-dimensional vector $\mathbf{x} = (x_1, ..., x_N)$. The molecules of these species interact with each other via M chemical reactions R_j $(j = 1, ..., M)$. These molecules are well-mixed and confined to a constant volume at thermal equilibrium, which allows disregarding the trajectories of these molecules in space, and instead consider only the molecular events that alter the population sizes of the species, stored in the state vector $\mathbf{x}$. (Gillespie, 1977, 2007)

The effect of each chemical reaction is defined by two quantities. The first one is the change in the state vector $\mathbf{x}$ that would be caused by this reaction, represented by a state-change vector $\mathbf{v}_j$. The second one is the propensity function a_j, which is defined so that, given the state vector $\mathbf{x}$ and an infinitesimal time interval $[t, t + dt)$, $a_j(\mathbf{x})dt$ would represent the probability that one reaction R_j will occur in this time interval. (Gillespie, 2007)

The rationale for the existence of the propensity function a_j can be summarised as follows. If the reaction R_j is unimolecular, represented as $S_a \rightarrow$ product(s), it can be assumed that this reaction is caused by an internal process described by quantum mechanics, similarly to the decay of a radioactive nucleus (Gillespie, 1992). From this, there exists such constant c_j that $c_j dt$ is the probability that any given molecule of the species S_a would partake in reaction R_j in the next infinitesimal time interval dt. Then, from the laws of probability, the propensity that a unimolecular reaction R_j would occur can be written as follows (Gillespie, 2007):

$$a_j(\mathbf{x}) = c_j x_a \tag{2.2}$$

where x_a is the number of molecules belonging to the species S_a present in the system. In case when the reaction R_j is bimolecular and takes the form $S_a + S_b \rightarrow$ product(s), it can be described by the kinetic molecular theory. From this, given that the system is well-stirred, there exists such constant c_j that $c_j dt$ is the probability that a randomly chosen pair of molecules belonging to the species S_a and S_b would react in the next infinitesimal time interval dt. Thus, the propensity that a bimolecular reaction R_j would occur can be written as follows (Gillespie, 2007):

$$a_j(\mathbf{x}) = c_j x_a x_b \tag{2.3}$$

where x_a and x_b are the numbers of molecules belonging to the species S_a and S_b, respectively. Note that, in case of a bimolecular reaction $S_a + S_a \rightarrow$ product(s), the number of distinct molecular pairs that can react equals $\frac{1}{2}x_a(x_a - 1)$, and the propensity function takes the following form:

$$a_j(\mathbf{x}) = c_j \frac{1}{2} x_a(x_a - 1) \tag{2.4}$$

Meanwhile, trimolecular reactions (and those of higher order), written as $S_a + S_b + S_c + ... \rightarrow$ product(s), arguably do not occur as "elementary events", and thus should not be considered as such (Gillespie, 1992). Instead, these processes should be divided into simpler components and modeled as sets of unimolecular and bimolecular reactions.

Assuming that the state-change vector $\mathbf{v}_j$ and the propensity function a_j are defined for each chemical reaction R_j, it is possible to derive an equation that describes a time-evolution of $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, the probability of the system being in a state $\mathbf{x}$ at the time point t given the initial state $\mathbf{x}_0$ and initial time t_0. The result, known as chemical master equation (CME), is a system of first-order differential equations (Gillespie, 2007):

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(\mathbf{x} - \mathbf{v}_j)P(\mathbf{x} - \mathbf{v}_j, t|\mathbf{x}_0, t_0) - a_j(\mathbf{x})P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \tag{2.5}$$

Although CME determines the function $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, it is hard to solve since the system can be highly multi-dimensional. Namely, the amount of possible states equals to the number of all possible combinations of population sizes of the reactant species. Presently, the field has accumulated a diverse set of methodologies that allow either an exact solution of CME or its approximation (see e.g. Cao et al., 2016; Kazeev et al., 2014; Lee and Kim, 2012; Munsky and Khammash, 2006; Wolf et al., 2010). For instance, the finite state projection algorithm provides a useful and intuitive framework for obtaining a direct solution for a finite number of states, or an approximation with known precision when an infinite (or an extremely large) number of states is truncated. In this approach, a finite subset of states in the state space is appropriately chosen, and the remaining states are projected onto a single space (Munsky and Khammash, 2006).

2.4.2 Stochastic simulation algorithm

The behavior of the stochastic system introduced in the previous subsection can also be studied by sampling the time evolution of the state vector $\mathbf{x}$ using the stochastic simulation algorithm (SSA). This approach generates trajectories of $\mathbf{x}$ over time that are exact numerical realizations of the process described by CME. Thus, a combination of the infinite number of such samples is logically equivalent to the exact numerical solution to the CME (Gillespie, 2007).

The SSA is based on iteratively answering two questions, given the defined set of reactions and the initial populations of the reactant species: which of the reactions will occur next, and when will it happen (Gillespie, 1977). To answer these questions, let us introduce the reaction probability density function $p(\tau, \mu)dt$, a probability that, given the current state $\mathbf{x}$ at the time t, the next reaction will occur in an infinitesimal interval $(t + \tau, t + \tau + d\tau)$, and it will be the R_μ reaction. Given the definition of $a_j(\mathbf{x})$ and by applying the laws of probability, this function takes the following form (Gillespie, 1976, 2007):

$$\begin{aligned} p(\tau, \mu|\mathbf{x}, t)dt &= a_\mu(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau), \\ a_0(\mathbf{x}) &= \sum_{j=1}^{M} a_j(\mathbf{x}) \end{aligned} \tag{2.6}$$

The equation 2.6 is central for the SSA approach. Here, the time until the next reaction will occur, τ, is an exponentially distributed random variable with a mean of $1/a_0(\mathbf{x})$, and it is equivalent to the minimum of the times it would take for each reaction to occur next given that no other reaction would occur before it. Meanwhile, the number of the next reaction, j, is a categorical random variable that holds the integer number of a reaction R_μ with the probability $a_\mu(\mathbf{x})/a_0(\mathbf{x})$. Note that these two variables are statistically independent of each other.

The samples of τ and μ can be generated using one of the existing Monte Carlo procedures, such as the direct method, which employs the standard inverse transform sampling (Gillespie, 2007). Given that r_1 and r_2 are two independent uniformly distributed random numbers in the interval (0, 1), the samples are generated as follows:

$$\tau = \frac{-\ln(r_1)}{a_0(\mathbf{x})}, \tag{2.7}$$

while μ is set to the value that makes the statement true:

$$\sum_{j=1}^{\mu-1} a_k(\mathbf{x}) \leq r_2 a_0(\mathbf{x}) < \sum_{j=1}^{\mu} a_k(\mathbf{x}) \tag{2.8}$$

Once τ and μ are obtained, the state vector $\mathbf{x}$ is updated based on the values stored in the state-change vector $\mathbf{v}_\mu$. Next, one has to decide whether the simulation should stop or the number and waiting time of the next reaction should be generated. This decision can be based on the current time, on the population levels of the reactant species, or on the number of reactions that occurred so far. Also, the algorithm should be terminated in case the value of $a_0(\mathbf{x})$ ever reaches 0. The stepwise procedure of the SSA is presented in Algorithm 1, given the start time t_0, the stop time t_{stop}, and the initial state of the system $\mathbf{x}$ (Gillespie, 1977).

Algorithm 1 : Stochastic simulation algorithm

1: $t \leftarrow t_0$; $\mathbf{x} \leftarrow \mathbf{x_0}$
2: evaluate all $a_j(\mathbf{x})$ values (section 2.4.1) and the $a_0(\mathbf{x})$ value (equation 2.6)
3: **while** $(t < t_{stop})$ and $(a_0(\mathbf{x}) > 0)$ **do**
4: generate a random pair (τ, μ), e.g. using equations 2.6 and 2.7
5: $t \leftarrow t + \tau$; $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v}_\mu$
6: save the values of t and $\mathbf{x}$
7: evaluate all $a_j(\mathbf{x})$ values (section 2.4.1) and the $a_0(\mathbf{x})$ value (equation 2.6)
8: **end while**

2.4.3 Delay stochastic simulation algorithm

The SSA provides a framework for describing the events that occur with a given propensity and are completed at the same moment as they are initiated. However, gene expression involves many complex processes that do not happen in an instant, but instead, once initiated, require certain time to be completed. This is the case when these processes, chemically, are not elementary reactions but instead are also composed of multiple steps, with each step being a simpler reaction. One strategy that would allow accounting for this is modeling each of these steps explicitly, as in (Mäkelä et al., 2011). This strategy is preferable when the dynamics of these steps is relevant for the study. However, considering the additional reactions and species can significantly slow down the simulation times. Another strategy is to introduce a delayed release of the reaction products, which allows accounting for the dynamics of the intermediate steps without explicitly introducing them into the model (Bratsun et al., 2005; Gibson and Bruck, 2000; Roussel and Zhu, 2006).

Following the second strategy, the delay SSA allows a reaction product to be released into the system only a defined time interval after the reaction occurs. The duration of the

delay is drawn from an arbitrary distribution, and the distributions from which the delay is drawn may vary between reactions as well as between the products of the same reaction. This approach is able to support more realistic modeling at reasonable computational costs. (Roussel and Zhu, 2006)

To implement the delay SSA based on the SSA, a delayed reaction should be separated into reacting events, which include instantaneous consumption of reactants, and generating events, which are the release of the possibly delayed reaction products. The time when the next reaction will occur and the reaction that will occur next are determined in the same way as in SSA, and the non-delayed generating events are performed at the same moment as the reacting events. Meanwhile, the generating events with non-zero delays are stored in a waiting list L, sorted by the time of their future release. After the next time point at which the next reaction should occur is drawn, it is compared with the smallest time value stored in the waiting list L (unless the list is empty). If the next reaction should happen before the release of the product stored in the waiting list, then this reaction occurs. Otherwise, the release of the product occurs. Following either of these events, a new time for the next reaction is drawn. (Roussel and Zhu, 2006; Zhu et al., 2007)

The procedure for the delay SSA is shown in Algorithm 2, given that t_{min} is the earliest release time among the generation events stored in the waiting list, $\mathbf{g}_{min}$ is the state change of this generation event, and $\mathbf{v}_\mu$ accounts only for the production and non-delayed generation events, while $\mathbf{g}_\mu$ accounts for the delayed generation events. Note that if all delays equal zero, the delay SSA behaves exactly as the regular SSA, and thus can be considered its generalization (Roussel and Zhu, 2006; Zhu et al., 2007).

Algorithm 2 : Delay stochastic simulation algorithm

1: $t \leftarrow t_0$; $\mathbf{x} \leftarrow \mathbf{x_0}$; $L \leftarrow$ empty waiting list
2: evaluate all $a_j(\mathbf{x})$ values (section 2.4.1) and the $a_0(\mathbf{x})$ value (equation 2.6)
3: **while** $(t < t_{stop})$ and $(a_0(\mathbf{x}) > 0)$ **do**
4: generate a random pair (τ, μ), e.g. using equations 2.6 and 2.7
5: **if** L is empty **then**
6: $t_{min} \leftarrow \infty$
7: **else**
8: $t_{min}, \mathbf{g}_{min} \leftarrow$ the time and state change of the earliest event in L
9: **end if**
10: **if** $\tau < t_{min}$ **then**
11: $t \leftarrow t + \tau$; $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v}_\mu$
12: if the reaction μ includes $\mathbf{g}_\mu$, add those and their delay times to L
13: **else**
14: $t \leftarrow t + t_{min}$; $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{g}_{min}$
15: remove the earliest event from L
16: **end if**
17: save the values of t and $\mathbf{x}$
18: evaluate all $a_j(\mathbf{x})$ values (section 2.4.1) and the $a_0(\mathbf{x})$ value (equation 2.6)
19: **end while**

2.5 Stochastic models of bacterial gene expression

2.5.1 Transcription

Given stochastic nature of bacterial gene expression, it can be well-modeled by a set of stochastic chemical reactions (Arkin et al., 1998; McAdams and Arkin, 1997; Ribeiro, 2010; Zhu et al., 2007). The model presented here offers the level of detail required for studying how the regulation of rate-limiting steps in transcription initiation affects the kinetics of RNA production.

$$\mathrm{P_{ON}} + \mathrm{R} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP}_c \xrightarrow{k_2} \mathrm{RP}_o \xrightarrow{k_3} \mathrm{P_{ON}} + \mathrm{R} + \mathrm{RNA} \tag{2.9}$$

$$\mathrm{P_{ON}} \underset{k_{\mathrm{ON}}}{\overset{k_{\mathrm{OFF}}}{\rightleftarrows}} \mathrm{P_{OFF}} \tag{2.10}$$

Transcription initiation starts when an RNAP molecule, R, binds to an active promoter, $\mathrm{P_{ON}}$, at the rate k_1. This process is in competition with the promoter entering an inactive state, $\mathrm{P_{OFF}}$, at the rate k_{OFF}, e.g. due to the activity of transcription factors or as a result of transient topological constraints (Browning and Busby, 2016; Chong et al., 2014). The promoter returns to the active state at the rate k_{ON}. Thus, as long as neither k_{ON} or k_1 equals zero and RNAP is present in the system, following the laws of probability and given enough time, the promoter and RNAP will eventually form a reversible closed complex, RP_c. At this stage, the reaction with the rate k_2 that describes commitment to the formation of an open complex, RP_o, is in competition with the reaction that describes the dissociation of the RNAP from the promoter at the rate k_{-1}. Finally, the reaction with the rate k_3 represents several successive steps that include promoter escape, promoter clearance, transcription elongation and its termination, followed by a release of the product RNA and the RNAP. As a result of this reaction, the active promoter becomes available for the next transcription initiation attempt, new RNA is produced, and the RNAP re-enters the pool of available RNAPs.

The reaction rates of this model depend on various factors such as promoter sequence, activities of regulatory molecules and topological location of the promoter, among others (for more detail, see section 2.3). Moreover, the processes represented by these reactions are not necessarily elementary, but instead can involve multiple steps, including e.g. reversibility and additional transient isomerization complexes (see section 2.3.1 and Figure 2.2).

Also, reaction 2.10 might take a different form, depending on the cause of the transient promoter inactivation. For example, if the inactive state is caused by binding of a transcription factor to the promoter, then this repressor, Rep, either can be modeled explicitly (reaction 2.11) or, if its concentration is in equilibrium, could be accounted for in the rate of promoter inactivation in reactions 2.10. Namely, k_{OFF} should be defined as follows: $k_{\mathrm{OFF}} = \mathrm{Rep}\, k_{\mathrm{rep}}$, where k_{rep} is the rate of the reaction given that only one repressor molecule is present in the system.

$$\mathrm{P_{ON}} + \mathrm{Rep} \underset{k_{\mathrm{ON}}}{\overset{k_{\mathrm{rep}}}{\rightleftarrows}} \mathrm{P_{OFF}} \tag{2.11}$$

The model of transcription described in reactions 2.9 and 2.10 was considered when interpreting experimental data in **Publications II** and **IV**. In **Publication III**, to

model transcription from an active promoter, a simplified version of reactions 2.9 is assumed based on the properties of genetic constructs employed in the study. Meanwhile, promoter repression in the **Publication III** is modeled as in reactions 2.11, since explicitly introducing repressor molecule species was required to describe promoter activation by inducer (see section 2.5.3).

2.5.2 Coupled transcription and translation

In bacteria, transcription and translation are coupled (Belogurov and Artsimovitch, 2015). Translation of an RNA can start before its transcription is completed, as soon as the RBS is produced. Thus, one feature required to couple the models of transcription and translation is the explicit production of an RBS. Meanwhile, transient promoter inactivation can be modeled in the same way as in the previous section, by reactions 2.10 or 2.11, since it is assumed to not interfere with translation. Given the above, transcription from an active promoter in this model is described as follows:

$$\mathrm{P_{ON}} + \mathrm{R} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP}_c \xrightarrow{k_2} \mathrm{RP}_o \xrightarrow{k_e} \mathrm{P_{ON}}(\tau_1) + \mathrm{RBS}(\tau_1) + \mathrm{R}(\tau_2) + \mathrm{RNA}(\tau_2) \tag{2.12}$$

In reactions 2.12, in addition to the rate constants described in the section 2.5.1, k_e is the rate of promoter escape. Utilizing the strategy presented in section 2.4.3, this model has delays on product release times. Promoter clearance delay, τ_1, accounts for the time that RNAP takes to move approximately 30–60 bp downstream of the transcription starting side after it has escaped the promoter. Since RBS is located near the transcription starting site, upstream of the start codon, and given that its length usually does not exceed ~40 bp (Ringquist et al., 1992; Shultzaberger et al., 2001; Zhu et al., 2007), RBS release happens at approximately the same time as promoter clearance. It is worth noting that, given transcription elongation rate of 12 bp/s (Adelman et al., 2002), τ_1 duration is about 3–4 s, which is considered fast compared to the rate-limiting steps in transcription.

Meanwhile, τ_2 represents the mean duration of transcription elongation (since the time taken by termination is negligibly small in comparison (Ray-Soni and Landick, 2016)). Being a sequence of multiple short single-nucleotide events, the total duration of transcription elongation is usually well-approximated by a constant (Adelman et al., 2002). Nevertheless, in the case of some genes, rare long sequence-specific pauses can occur and might become rate-limiting. To account for this additional complexity, a single-nucleotide model of elongation can be used (Mäkelä et al., 2011).

As soon as the RBS is released, it is available for translation initiation. A single-reaction model of translation that incorporates delayed product release captures the main rate-limiting steps:

$$\mathrm{RBS} + \mathrm{Rib} \xrightarrow{k_{\mathrm{tr}}} \mathrm{RBS}(\tau_3) + \mathrm{Rib}(\tau_4) + \mathrm{Protein}(\tau_5) \tag{2.13}$$

In reaction 2.13, a ribosome, Rib, binds to the RBS and initiates translation with the rate k_{tr}. As the ribosome proceeds with translation, it frees the RBS with the delay τ_3 after the successful initiation event. Translation elongation and termination are completed with the delay τ_4, at which point the ribosome releases the RNA (note that the RNA molecule is not modeled explicitly). The protein release delay, τ_5, equals to the sum of the τ_4 duration and the time it takes for the protein to fold.

Note that the model described here does not account for RNAP and ribosome traffic events, which in some cases are able to significantly affect gene expression kinetics (Belogurov and Artsimovitch, 2015; Lesnik et al., 2000; Mäkelä et al., 2011).

The model described by reactions 2.10, 2.12 and 2.13 was used in **Publication IV** for interpreting experimental data. This model was selected because it provides a detailed insight into the rate-limiting steps in transcription initiation, the only stage of gene expression that was intended to differ between the experimental conditions used in the study.

2.5.3 Inducer intake and transcription activation by inducer binding to a repressor

When an activator molecule is placed in the media, it must pass through the cell wall in order to become available for transcription activation. This process, called inducer intake, is stochastic and can introduce additional noise to gene expression (Megerle et al., 2008). In gram-negative bacteria, the cell wall consists of an outer membrane, a thin inner membrane, and a gel-like layer called periplasm located between those (Beveridge, 1999). As such, when transcription is activated externally and given that the process is diffusive-like (i.e. no active intake mechanism is involved, or its contribution is negligible compared to the intake by diffusion), the inducer intake kinetics can be modeled as follows:

$$\mathrm{I_{env}} \xrightarrow{int_1} \mathrm{I_{peri}} \xrightarrow{int_2} \mathrm{I} \tag{2.14}$$

In reactions 2.14, an inducer molecule located outside of the cell wall, $\mathrm{I_{env}}$, can cross the outer membrane with the rate int_1 and become an inducer molecule located in the periplasm, $\mathrm{I_{peri}}$. Meanwhile, $\mathrm{I_{peri}}$ can cross the inner membrane with the rate int_2 and become an active inducer molecule, I, that is able to bind the repressor molecules, Rep. Note that reactions 2.14 are not necessarily elementary. As has been shown in **Publication III**, while in normal conditions the inducer intake process appears to have only two rate-limiting steps, at suboptimal temperatures the existence of additional rate-limiting steps can be inferred.

The general case of promoter inactivation by repressor molecules is modeled by reactions 2.11. Given that the inducer can bind both to free repressor molecules and to those that are bound to the promoter, transcription activation can be modeled as follows:

$$\mathrm{Rep} + \mathrm{I} \underset{k_{\mathrm{inact}}}{\overset{k_{\mathrm{act}}}{\rightleftarrows}} \mathrm{Rep.I} \tag{2.15}$$

$$\mathrm{P_{OFF}} + \mathrm{I} \xrightarrow{k_{\mathrm{act}}} \mathrm{P_{ON}} + \mathrm{Rep.I} \tag{2.16}$$

The reversible reaction 2.15 describes the binding of an inducer to the repressor molecule with the rate k_{act}, which reduces the repressor concentration in the system and thus indirectly activates transcription initiation. The resulting complex Rep.I can dissociate into the inducer and repressor molecules with the rate k_{inact}. Meanwhile, reaction 2.16 shows that an inducer molecule can bind to a repressor molecule that occupies the promoter with the same rate as to a free repressor molecule, directly activating the promoter. This reaction is irreversible since the formed Rep.I complex leaves the promoter and is indistinguishable from the complexes formed in reaction 2.15.

3 Materials and Methods

3.1 Fluorescent proteins

The first fluorescent proteins – aequorin, which emits blue light when reacting with Ca^{2+} ions, and a wild-type green fluorescent protein (GFP), which absorbs blue and emits green light – were isolated from Aequorea, a bioluminescent jellyfish (Morise et al., 1974; Shimomura et al., 1962; Ward et al., 1980). About two decades later, GFP was shown to produce stable fluorescence in live *E. coli* cells, expressed under the control of a T7 promoter (Chalfie et al., 1994). In the following years, new types of fluorescent proteins were engineered, providing a vast array of tools for *in vivo* and *in situ* visualization of cellular components and for visual quantification of gene expression activity (Day and Davidson, 2009; Shaner et al., 2005; Tsien, 1998).

The key advantage of using fluorescent proteins over other fluorescent labeling techniques (Hayashi-Takanaka et al., 2014; Schneider and Hackenberger, 2017) is that they allow fusion-tagging of the molecules of interest. This can be employed for observing gene expression dynamics in real time (Golding and Cox, 2004; Yu et al., 2006). However, there are certain limitations that must be considered when selecting a fluorescent protein for a particular experimental design (Shaner et al., 2005). The main issues are toxicity of the fluorescent proteins, detectability of their fluorescence, and robustness of the proteins to environmental conditions. First, many wild-type fluorescent proteins form dimers or trimers. This oligomerization can be toxic to cells, and thus monomeric variants of the fluorescent proteins are usually used instead (Shaner et al., 2005). Moreover, toxicity can originate not only from the protein itself but also from the light wavelengths used to excite the protein or emitted by it. For example, exposure to near-UV light is known to affect the physiology of some microorganisms (Jagger, 1976; Kramer and Ames, 1987). To ensure cell health, tests for toxicity should be performed when introducing a fluorescent protein to a new cell strain or cell line. Second, the fluorescence signal must be sufficiently above the autofluorescence of the cell and of the background (e.g. cell growth media). The techniques chosen for acquiring the cell images can affect the signal-to-noise ratio in fluorescence detection. Further, photostability and protein maturation time should be considered, although the degree to which these parameters are relevant depends on the experimental design. Finally, protein folding often depends on temperature, and performance of many fluorescent proteins is sensitive to acidity of the environment (Shaner et al., 2005).

In **Publication II**, *E. coli* strain with fluorescently tagged RNAP molecules that are functionally identical to their non-tagged version (Bratton et al., 2011; Cabrera and Jin, 2003) was used to validate the values of the relative change in RNAP concentration ([RNAP]) between experimental conditions (first measured by quantitative PCR (qPCR), see section 3.4). In **Publication IV**, the same strain was used to demonstrate that

skewness of the distribution of [RNAP] measured from individual cells is not correlated with the skewness in RNA production kinetics. The tagging method that is crucial for the work presented in this thesis is described below.

3.1.1 MS2-GFP tagging method

The method for tagging RNAs with MS2-GFP molecules was first introduced in living yeast (Bertrand et al., 1998), then adapted for usage in live mammalian cells (Fusco et al., 2003), and in live *E. coli* (Golding and Cox, 2004; Golding et al., 2005). In this method, GFP is fused with MS2, which is the coat protein of a bacteriophage MS2, an RNA virus that can infect *E. coli* and other similar bacteria. This protein binds to a specific hairpin loop formed by a segment of the viral RNA (Peabody, 1993). These hairpin loops, called MS2 binding sites, can be added to the sequence of the gene controlled by the target promoter, which allows MS2-GFP tagging of this RNA.

The MS2-GFP tagging method presented in (Golding et al., 2005) (Figure 3.1A) has an advantage over its earlier version (Golding and Cox, 2004) in that it also allows observing the proteins translated from the target RNA. In this tagging method (Golding et al., 2005), the promoter of interest, $P_{lac/ara\text{-}1}$, is placed on a single-copy plasmid and controls the expression of the target RNA that includes two major segments. The segment that is located immediately after the RBS encodes for a red fluorescent protein, and the following segment encodes for 96 MS2 binding sites (each site can be bound by one MS2 dimer). The expression of an RNA that encodes for MS2-GFP proteins is controlled by a reporter promoter, $P_{LtetO\text{-}1}$. The reporter RNA sequence includes the region encoding a fusion of two MS2 proteins that form a dimer, MS2d, followed by the region encoding a green fluorescent protein, GFP. Thus, by the time the MS2-GFP complex becomes fluorescent, it is expected to be fully produced. The reporter is carried by a plasmid with high copy number, to ensure sufficient concentration of MS2-GFP molecules in the cell.

The abundance of the reporter protein guarantees that the background fluorescence intensity formed by diffusing MS2-GFP molecules does not change strongly when a new target RNA is produced and 96 MS2-GFP proteins bind to it. The newly formed MS2-GFP-RNA complex appears as a bright spot in the cell on a fluorescence microscopy image (Figure 3.1B). Note that multiple MS2 binding sites at the target RNA allow for a brighter spot to form, and thus a higher signal-to-noise ratio on the microscopy images. Further, the RNA molecule is assumed to be fully tagged when first detected under the microscope, given that transcription elongation of the target RNA segment with the MS2 binding sites and the MS2-GFP binding to the RNA are fast compared to the time interval between the consecutive microscopy images (which is usually about 1 min) (Tran et al., 2015).

In order to reach a sufficient concentration of MS2-GFP molecules in the cell by the time the first target RNA is produced, the reporter is activated prior to activation of the target promoter. It is crucial that the activation of the reporter would not affect the expression from the target promoter, and it is in general undesirable to affect the reporter when regulating the expression from target promoter. Thus, when selecting the promoter that controls the reporter, the crosstalk between the target and the reporter promoters should be minimized.

In this thesis, in **Publications I-IV**, the MS2-GFP tagging method described above was utilized for *in vivo* quantification of integer-valued RNA numbers in *E. coli*. These numbers were then used to estimate time intervals between consecutive RNA production

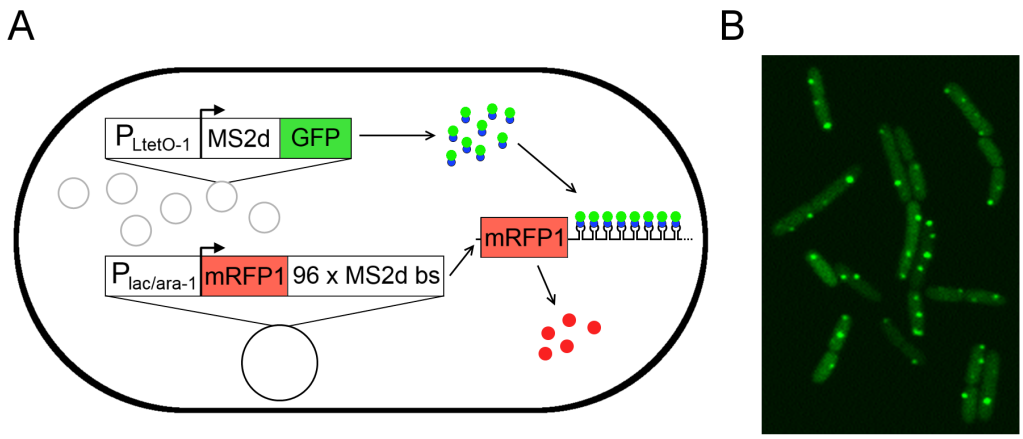


Figure 3.1: *In vivo* detection of RNA molecules using the MS2-GFP tagging method. (A) A schematic representation of the MS2-GFP tagging method used for *in vivo* RNA detection *E. coli*. The target gene is carried on a single-copy plasmid (large black circle) and produces the target RNA under the control of $P_{lac/ara-1}$ promoter. The reporter gene, controlled by $P_{LtetO-1}$ promoter, is carried on a multicopy plasmid (small gray circles). The MS2-GFP molecules (merged blue and green circles) are expressed from the reporter plasmids and bind to the MS2d binding sites located on the target RNA. Target RNA expresses the red fluorescent protein, mRFP1 (red circles). (B) An example image obtained by fluorescence microscopy. The uniform background fluorescence inside the cells is caused by freely diffusing MS2-GFP molecules, whereas bright spots correspond to the tagged RNA molecules.

events. In **Publication III**, these numbers were also used to estimate the time between the addition of the target promoter activator into the media and the first target RNA being produced. In addition, **Publication IV** made use of the MS2-GFP tagging method with different pairs of target and reporter promoters.

3.2 Microscopy

Fluorescence microscopy allows observing fluorescently tagged molecules in live cells by illuminating the sample with the light of the excitation wavelength and detecting the emitted light. The most common illumination scheme, widefield, illuminates a large area and depth of the sample. As a consequence, besides the signal from the sample of interest, the obtained image includes also the signal from a large background area, which can result in a low signal-to-noise ratio. Confocal microscopy addresses this problem by using point illumination and point detection, exciting and receiving signal only from a small section of the sample at a given time. This approach improves optical resolution but increases the time required to obtain the image, since a confocal microscope can register only a small area of the sample at a single time moment. Parallel scanning, e.g. using a spinning disk, helps to reduce this time. Various other fluorescence microscopy approaches and illumination schemes allow reducing the signal-to-noise ratio via the restriction of the illuminated sample volume. Among others, total internal reflection fluorescence microscopy uses an evanescent field of light to excite only the thin top layer of a sample, whereas highly inclined and laminated optical sheet (HILO) microscopy utilizes a highly inclined thin sheet of a laser beam to visualize the illuminated slice inside

the sample. (Minsky, 1988; Stephens and Allan, 2003; Tokunaga et al., 2008)

Aside from the fluorescent molecules, it is often useful to detect the location of the elements that are not fluorescently tagged, such as cell walls or inclusion bodies. It can be challenging to reliably extract this information from fluorescence microscopy images. Instead, bright-field or phase-contrast microscopy can be used for this purpose (Chowdhury et al., 2013; Selinummi et al., 2009). The former is a simple optical microscopy technique where the sample is illuminated with white light, and the effects of the sample refractive index and light absorption then detected from the light that passes through the sample. The latter is a more elaborate optical microscopy technique that allows detecting the phase shift in the light weaves scattered by the sample and converting this phase shift into the contrast of the image (Zernike, 1942).

In **Publication I**, **Publication II**, and **Publication IV** confocal and phase contrast images of the same cells were taken approximately at the same time. A confocal laser-scanning system was used to obtain the fluorescence images of the cells, and phase-contrast microscopy to detect the cell borders. In **Publication III**, fluorescence images were obtained both by confocal and by HILO microscopy.

In all microscopy experiments conducted during the work on this thesis, the cells to be imaged were placed on a thin agarose gel pad and topped with a coverslip (Golding et al., 2005; Kandavalli et al., 2016). This agarose pad should contain nutrients and other chemicals required for the experiment. Further, it should be dense enough for the bacteria to become immobilized but not so dense as to prevent healthy growth and division, thus allowing *in vivo* time-lapse imaging.

Finally, controlling the sample temperature during the experiment can be of value (Kumar and Libchaber, 2013). To maintain the cells at a constant temperature, the sample can be placed under the microscope in a thermal chamber, which was done in **Publications I-IV**. In addition, **Publication III** employs a thermal microfluidic system that allows shifting temperature during the course of a microscopy imaging experiment.

3.3 Analysis of microscopy images

Extracting single-cell data on the kinetics of fluorescently tagged RNA production from time-lapse microscopy images involves a sequence of non-trivial data processing steps. The main stages of the data extraction from microscopy images used in this thesis are outlined in Figure 3.2. First, the imaging process often consequently covers several cell areas, called panels, at each time point, thus producing several sets of time series images during one experiment. The images obtained from each panel should be aligned over time in order to minimize the small shift in the imaging position that occurs due to mechanical iteration over the panels. This alignment can be done e.g. using cross-correlation (Gupta et al., 2014; Häkkinen et al., 2013).

The most reliable method for detecting borders of individual cells, in many cases, is still drawing the cell masks manually by a human expert. However, since this is a time-consuming process, automatic cell segmentation is often used for preliminary construction of cell masks, followed by manual correction (Chowdhury et al., 2013; Häkkinen et al., 2013). To track individual cells over time, cell lineages can be reconstructed with CellAging software (Häkkinen et al., 2013). If the masks are constructed from phase contrast or brightfield microscopy images, they must be aligned with the fluorescence microscopy images, since they can have a different resolution and a slightly different position and area

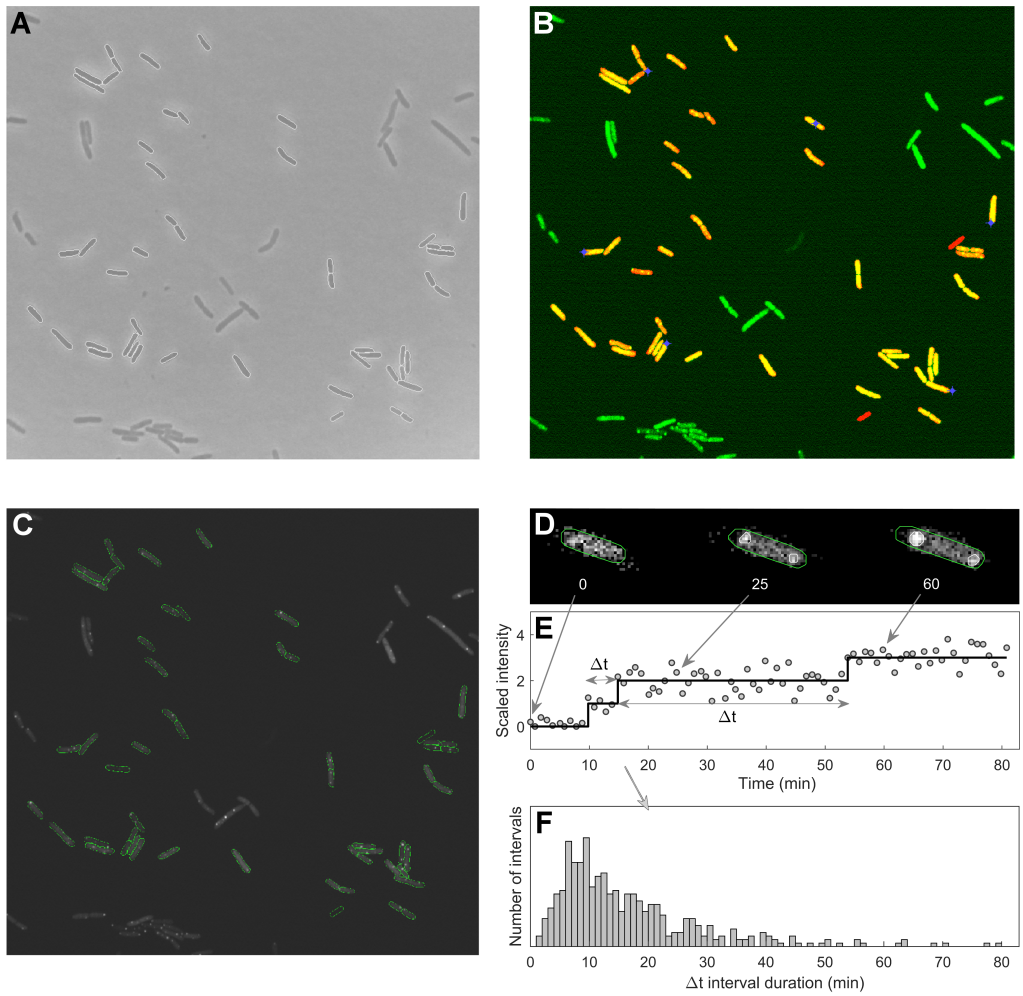


Figure 3.2: Extracting data from microscopy images. (A) Phase-contrast microscopy image with semi-automatically segmented cells. The cell masks are shown as white lines. (B) Alignment of the cell masks with a confocal microscopy image. Masks are shown in transparent red, seen as yellow when located on top of the GFP fluorescence. Landmarks are shown in blue. (C) Confocal microscopy image with segmented cells. The cell masks are shown as green lines. (D) Confocal microscopy image of one cell followed over time, after the detection of MS2-GFP-RNA complexes. Cell masks are shown as green lines, the detected MS2-GFP-RNA complexes are outlined in white. (E) Scaled total intensity of MS2-GFP-RNA complexes over time (grey circles) and the estimated number of target RNA molecules in the cell at each time point (black line). The total intensity is scaled by the estimated intensity of one MS2-GFP-RNA complex. (F) An example Δt distribution.

coverage. After this alignment, the masks applied to fluorescent images can be further improved in accuracy by adjusting their borders based on the fluorescence intensity of the nearby pixels.

Next, the total intensity of MS2-GFP-RNA complexes, visible as bright fluorescent spots, is detected from each masked cell using the procedure described in (Häkkinen et al., 2014). Two main steps of this procedure are the cell background removal and fitting a set of Gaussian surfaces to the remaining cell image in order to detect the bright fluorescent spots. Based on the cell lineage information, the time series of these intensities for each observed cell is extracted. The time points at which a new RNA is produced are usually estimated using a least-square fit of a monotonic piecewise-constant function with a constant increment (Häkkinen and Ribeiro, 2015). From these, the Δt intervals are calculated, and the obtained Δt distributions can be used to study the kinetics of transcription initiation.

In this thesis, the main results were obtained using data extracted from time-lapse microscopy images. **Publication I** focuses on processing the time series of the total intensity of MS2-GFP-RNA complexes. **Publication II** and **Publication IV** are based on analyzing the shape of Δt distributions. Finally, **Publication III** makes use of the distributions of time intervals between the target gene induction and the first RNA production, along with the Δt distributions collected in another recent study (Oliveira et al., 2016a). Note that the Δt distributions used in **Publication III** were obtained by microscopy and image analysis techniques that are similar to those used in this thesis, differing only in that all cell masks were produced manually.

In **Publication I**, **Publication II** and **Publication IV**, cell segmentation was performed from phase contrast images. In **Publication III**, cell segmentation was performed directly from fluorescent images. In all cases, the images were initially subject to automatic segmentation which was followed by a manual correction. In **Publication I** and **Publication II**, the automatic segmentation was performed using MAMLE (Chowdhury et al., 2013). In **Publication III** and **Publication IV**, the automatic segmentation was performed using CellAging (Häkkinen et al., 2013). These two segmentation tools produce results of comparable quality (Chowdhury et al., 2013; Häkkinen et al., 2013). In the later publications, CellAging was chosen for segmentation since the next steps in the image analysis pipeline are also performed using CellAging.

In **Publication I**, **Publication II** and **Publication IV**, alignment of the masks obtained from phase-contrast images with the confocal microscopy images was performed semi-automatically. In the first time frame, we manually selected 5-8 landmarks that placed the cell masks on top of the corresponding cells, which were visible due to fluorescent cell background formed by free GFP molecules. After establishing the landmarks, we used the thin-plate spline interpolation to transform the masks into the coordinate system of the confocal images and propagate the alignment to all time frames. The results were inspected by human specialists, who repeated the procedure with additional landmarks in case the outcome was not satisfactory.

In **Publication I**, the time points at which new RNAs are produced were estimated using both a new methodology developed in this thesis and the methodology proposed in (Häkkinen and Ribeiro, 2015), as the aim of **Publication I** was to propose the new method and to evaluate its accuracy relative to the method that is currently in use. In **Publication II**, **Publication III** and **Publication IV**, the performance of both the new methodology and the one from (Häkkinen and Ribeiro, 2015) was visually evaluated by human specialists. In all cases, either the method from (Häkkinen and Ribeiro, 2015)

provided better results or there was no clear difference between the performance of the two methods. Thus, for consistency, the **Publication II**, **Publication III** and **Publication IV** employed only the method from (Häkkinen and Ribeiro, 2015).

3.4 Quantitative PCR

Polymerase chain reaction (PCR) is a molecular biology method that allows creating multiple copies of a specific DNA segment. When combined with a reverse transcription technique, PCR can be used to amplify RNA numbers and thus measure relative gene expression levels (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008). Quantitative PCR (qPCR) allows measuring these gene expression levels in real time by using fluorescent labels. The amount of the amplification product can be estimated from the measured fluorescence intensity. The labeling can be performed by dyes that bind to DNA non-specifically, by probes specific to the target sequence, or by combining the dyes and the probes (Lind et al., 2006). In this thesis, a non-specific dye (SYBR Green I) was used in all qPCR experiments, which is a standard approach when quantifying gene expression levels (Livak and Schmittgen, 2001; Schmittgen and Livak, 2008).

In **Publication II**, qPCR was employed to estimate relative RNAP concentrations by measuring relative transcript levels of the rpoC gene, which encodes the β' subunit. In **Publication II** and **Publication IV**, qPCR measurements of the transcript levels produced under control of the target promoter were used to construct τ-plots (see section 3.8). In **Publication III**, qPCR measurements of relative transcript levels produced under the control of lac/ara-1 promoter at different IPTG concentrations were used to obtain an induction curve.

3.5 Western blot

Western blot is a molecular biology technique that provides means for identifying specific proteins in the sample and measure protein size and relative abundance (Burnette, 1981; Mahmood and Yang, 2012). In **Publication IV**, relative RNAP concentrations were estimated by measuring RpoC protein levels using western blot. This differs from estimating relative RNAP concentrations by qPCR in that western blot provides information on proteins, whereas qPCR provides information on RNA transcripts. Given that regulation in gene expression occurs not only at the stage of transcription but also during translation, western blot is, in general, more reliable for estimating relative RNAP concentrations.

3.6 Flow cytometry

Flow cytometry is a technique that allows collecting single-cell data in a format that is faster to process than, e.g., single-cell microscopy data. In flow cytometry, the cells are carried by a liquid stream in single file. When passing the detector, each cell is illuminated by a laser beam, with forward-scattered and side-scattered light being detected (Shapiro, 2005). To filter out the instances of cell doublets and other measurement instances that can be considered as debris, gating techniques should be applied to the measured data (Aghaeepour et al., 2013; Razo-Mejia et al., 2018). While flow cytometry allows gathering data on thousands of cells in a matter of minutes, it is limited in the sense that the same cell cannot be tracked over time, as it is possible under the microscope. In **Publication**

IV, flow cytometry was used to obtain distributions of fluorescent protein levels in cell populations. The data was gated by an automated method proposed in (Razo-Mejia et al., 2018).

3.7 Lineweaver-Burk plot

Originally, the Lineweaver-Burk plot was introduced as a graphical method that uses a double reciprocal plot for evaluation of the constants involved in the description of enzyme kinetics (Lineweaver and Burk, 1934). Such catalytic reactions usually involve reversible binding stage and mostly irreversible catalysis stage. For the purposes of our study, we apply the Lineweaver-Burk plot method to a process that is well-described by Michaelis-Menten enzyme kinetics, which can be modeled by the following stochastic reactions (Chen et al., 2010):

$$\mathrm{E} + \mathrm{S} \underset{k_r}{\overset{k_f}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P} \tag{3.1}$$

In equation 3.1, enzyme, E, binds to a substrate, S, at the rate k_f, forming a reversible active complex, ES. Dissociation of this complex at the rate k_r is in competition with a catalytic reaction of product formation at the rate k_{cat}, which results in a release of the enzyme and the product, P. Classical Michaelis-Menten equation that describes a relation between the rate of product formation, V, and the concentration of a substrate, S, takes the following form:

$$\begin{gathered} \mathrm{V} = \frac{\mathrm{V}_{\max}[\mathrm{S}]}{[\mathrm{S}] + \mathrm{K}_{\mathrm{M}}}, \\ \text{where } \mathrm{V}_{\max} = k_{\mathrm{cat}}[\mathrm{E}]_0 \text{ and } \mathrm{K}_m = \frac{k_r + k_{\mathrm{cat}}}{k_f}. \end{gathered} \tag{3.2}$$

In equation 3.2, $\mathrm{V}_{\max}$ is the maximum possible rate of the product formation, $[\mathrm{E}]_0$ is the starting concentration of the enzyme, and K_{M} is the Michaelis constant. Note that K_m equals to a subtract concentration at which $\mathrm{V} = \mathrm{V}_{\max}/2$.

The Lineweaver-Burk equation can be written as a reciprocal of equation 3.2:

$$\frac{1}{\mathrm{V}} = \frac{\mathrm{K}_m}{\mathrm{V}_{\max}} \frac{1}{[\mathrm{S}]} + \frac{1}{\mathrm{V}_{\max}} \tag{3.3}$$

This equation reveals linear relationship between the inverse rate of product formation, 1/V, and the inverse of the substrate concentration, 1/[S]. Figure 3.3 shows a Lineweaver-Burk plot, from which $\mathrm{V}_{\max}$ can be calculated as the inverse of the y-intercept, and K_{M} can be calculated as a ratio between the slope of the line and the y-intercept. In **Publication III**, the concept of Lineweaver–Burk equation was employed to estimate the intake times of an inducer.

3.8 τ plot

τ plot is a double reciprocal plot that describes a relationship between the inverse of transcription rate and the inverse of [RNAP] (McClure, 1980). A τ plot is technically a

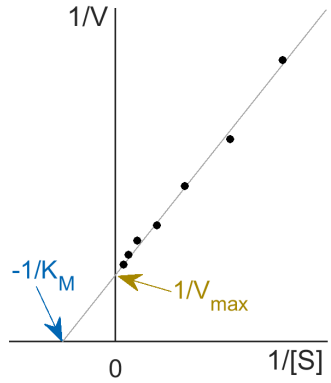


Figure 3.3: Schematic representation of a Lineweaver-Burk plot providing a graphical representation of Michaelis-Menten enzyme kinetics. The inverse of the rate of product formation, 1/V, is plotted against the inverse of the substrate concentration, 1/[S]. Shown are the example data points (black circles) and the best-fitting line to the data (grey line). The line intersects the y-axis at a point with the value of $1/\mathrm{V_{max}}$, and the x-axis at a point with the value of $-1/\mathrm{K_M}$.

Lineweaver-Burk plot used to dissect rate-limiting steps of bacterial transcription. The concept of Lineweaver-Burk plot is not specific to interactions between an enzyme and a substrate. In general, it is applicable to any biochemical process that can be described by equivalent stochastic chemical kinetics. For example, this concept was previously utilized *in vitro* in the steady-state assay methodology for studying separate steps in transcription initiation (McClure et al., 1978), allowing e.g. for a better understanding of the role of rifampicin in this multi-step process (McClure and Cech, 1978).

A τ plot, which is a graphical method for dissecting rate-limiting steps in transcription initiation, emerged as another steady-state assay application of this concept (Bertrand-Burggraf et al., 1984; McClure, 1980). In particular, a τ plot is based on the assumption that the rate of an RNAP binding to a promoter is directly proportional to the [RNAP]. This is equivalent to the inverse of the average time spent in the closed complex formation being directly proportional to the inverse of the [RNAP]. Meanwhile, the time spent in the subsequent steps remains constant with the changes in [RNAP]. Given this, the average total time spent in transcription initiation should change linearly with the inverse of [RNAP]. As such, by changing the [RNAP] between experimental conditions and measuring this average total time in each condition, it is possible to estimate the average time spent in the steps whose duration depends on [RNAP] and the average time spent in the subsequent steps.

This methodology was originally used to study *in vitro* abortive transcription initiation, the process that involves fewer stages than *in vivo* transcription modeled by reactions 2.9 and 2.10. First, no repression mechanism was present in these *in vitro* studies, which means that the reactions 2.10 are not needed. Second, transcription initiation always was abortive, releasing a short product sequence negligibly fast after successfully forming an open complex (McClure, 1980). This *in vitro* abortive transcription initiation thus can be described by the following reactions (see section 2.5.1 for a description of the reactant species and rate constants):

$$\mathrm{P_{ON}} + \mathrm{R} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP}_c \xrightarrow{k_2} \mathrm{RP}_o \xrightarrow{\infty} \mathrm{P_{ON}} + \mathrm{R} + \text{abortive product} \tag{3.4}$$

The average time between an active promoter becoming available for RNAP binding and an abortive product being released, denoted as τ, can be obtained, e.g., using a moment-generating function of the distribution of the transition times between $\mathrm{P_{ON}}$ and P_o states (Häkkinen and Ribeiro, 2016), which takes the following form:

$$M_{(in\ vitro)}(t) = \frac{k_2(k_1 - t)}{(k_{-1} + k_2 - t)(k_1 - t) - k_1 k_{-1}} + \frac{k_2 - t}{k_2} \tag{3.5}$$

$$\tau = M'_{(in\ vitro)}(0) = \underbrace{\frac{1}{[\mathrm{RNAP}]} \frac{k_{-1} + k_2}{k_1 k_2}}_{\tau_c} + \underbrace{\frac{1}{k_2}}_{\tau_o} \tag{3.6}$$

Equation 3.6 shows a linear relationship between the inverse of the rate of abortive transcription initiation (τ) and the inverse of the [RNAP]. the average duration of the closed complex formation (τ_c) is directly proportional to 1/[RNAP], whereas the average duration of the open complex formation (τ_o) remains constant. Given equation 3.6, the linear fit to the data on an *in vitro* τ plot (Figure 3.4) intersects with y-axis at the value that equals to the inverse of the production rate when [RNAP] is assumed to be infinitely large. Since, at this point, the steps that depend on [RNAP] occur infinitely fast, this value equals to τ_o. Knowing this value, one can obtain τ_c in a measured condition of interest: $\tau_c = \tau - \tau_o$.

In **Publication II**, the concept of an *in vitro* τ plot was utilized to propose a new methodology that allows to construct τ plots based on *in vivo* data. This *in vivo* τ plot methodology was then used in **Publication III** and **Publication IV** to estimate the average time spent in transcription initiation prior and after commitment to open complex formation.

3.9 Uncertainty estimation

Estimation of uncertainty in the results is essential for evaluating the significance of the conclusions. There is no universal approach to uncertainty estimation, with the choice of the method depending, e.g., on whether the variable in question follows a known distribution and on the sample size of the data set. Therefore, this thesis employed a case-based approach.

3.9.1 Delta method

The Delta method allows estimating the variance of a function of random variables by using the Taylor series approximation, given that the variances (var) and covariances (cov) of these random variables are known. Let $\mathbf{X} = \{X_1, ..., X_k\}$ be k random variables with means $\mathbf{M} = \{M_1, ..., M_k\}$. Then the variance of the function $g(\mathbf{X})$ can be written as follows (Casella and Berger, 2001):

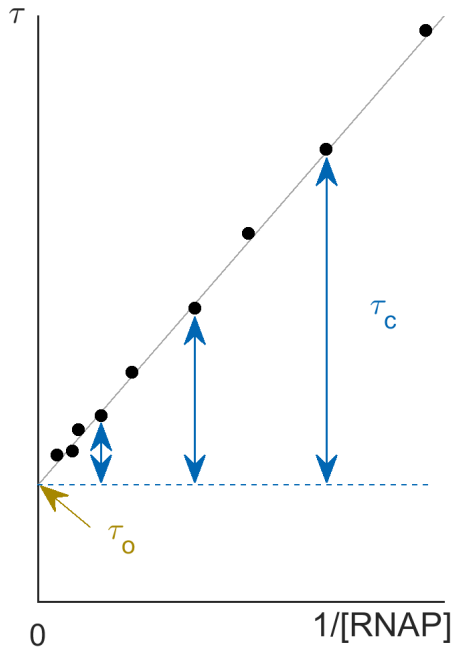


Figure 3.4: Schematic representation of an *in vitro* τ plot, showing the inverse of the rate of abortive transcription initiation, τ, against the inverse of RNAP concentration, 1/[RNAP]. The average duration of the open complex formation, τ_o, is estimated as a y-intercept of the best-fitting line to the empirical data (black dots).

$$var(g(\mathbf{X})) = \sum_{i=1}^{k} (g_i'(\mathbf{M}))^2 var(X_i) + 2 \sum_{i>j} g_i'(\mathbf{M}) g_j'(\mathbf{M}) cov(T_i, T_j),$$
$$g_i'(\mathbf{M}) = \left. \frac{\partial g(\mathbf{X})}{\partial X_i} \right|_{T_1=M_1,\ldots,T_k=M_k} \tag{3.7}$$

In **Publication II**, **Publication III**, and **Publication IV**, this method was used to estimate the standard error of the relative transcript levels obtained by qPCR. In all cases, the qPCR data provided 2 random variables in all conditions: the reference and control gene transcripts. Since the activity of these two genes is independent of each other, the covariance between the two random variables considered always equals zero. Meanwhile, the variance of the random variables is obtained by repeating the experiment several times in each condition. In **Publication II**, the Delta method was used to estimate the standard uncertainty of the time durations spent on the events prior and after closed complex formation in transcription initiation. In **Publication IV**, it was used to estimate the standard uncertainty of the relative durations spent on the events prior and after closed complex formation in transcription initiation.

3.9.2 Non-parametric bootstrap confidence intervals

Bootstrapping is a technique for estimating a distribution of some feature of the data set, e.g. its skewness, by using resampling with replacement. The resampling can be either non-parametric or parametric. The former does not require any assumptions about the distribution from which the data was drawn, contrary to the latter, where a parametric model of the data is assumed. A non-parametric bootstrap confidence interval for a given feature of the data sample is constructed as follows. First, the data is resampled with replacement B times, with the size of each bootstrapped data set usually being the same as the size of the original sample, and B being high enough to produce repeatable confidence intervals in the end, in general $B \geq 10^3$. From each bootstrapped data set, the bootstrapped feature is calculated. Finally, the $\alpha/2$ and $1-\alpha/2$ percentiles of this bootstrapped distribution of the feature values form the bootstrap confidence interval of this feature with the significance level α. (Carpenter and Bithell, 2000; DiCiccio and Efron, 1996)

In **Publication III**, non-parametric bootstrap was used to construct confidence intervals for the mean and coefficient of variation of the distributions of (i) time intervals between the inducer placement in the media and the cell producing the first target RNA and (ii) time intervals between consecutive RNA production events (Δt distributions). In **Publication IV**, non-parametric bootstrap was used to construct confidence intervals for the mean, coefficient of variation, skewness, and kurtosis of Δt distributions.

3.9.3 Simultaneous estimation of confidence bands

Let us assume a linear regression model $\hat{y} = \hat{a}x + \hat{b}$ that is best-fit to the data (x_1, y_1), ..., (x_n, y_n). In order to estimate the confidence bands on $\hat{y}$ for all possible values of x, it does not suffice to combine the confidence intervals with the desired confidence level $1-\alpha$ obtained at each x_i point, since this would usually produce a confidence band with the confidence level $1-\alpha'$, where $\alpha' > \alpha$ (Casella and Berger, 2001; Degras, 2017). One way to correct for this and obtain a good approximation of the confidence bands with the confidence level $1-\alpha$ is based on the Bonferroni inequality. In this approach, the confidence level of each confidence interval is equaled to $1-\gamma$, $\gamma = 1 - \frac{\alpha}{m}$, where m is the number of individual confidence intervals. Another approach employs the Scheffé's method to estimate the confidence bands of $\hat{y}$ simultaneously for all values of x, defining the confidence bands with the confidence level $1-\alpha$, $y_\pm$, as follows (Casella and Berger, 2001):

$$
\begin{aligned}
y_\pm &= \hat{y} \pm M_a S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}, \\
S &= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \\
S_{xx} &= \sum_{i=1}^{n}(y_i - \bar{y})^2, \\
M_a &= \sqrt{2F_{2,n-2,\alpha}},
\end{aligned}
\tag{3.8}
$$

where $F_{2,n-2,\alpha}$ is the F-statistic at the significance level α with the degrees of freedom $d_1 = 2$ and $d_2 = n-2$. In **Publication II**, the Scheffé's approach was used to estimate

confidence bands of one standard uncertainty for the linear regression models that describe (i) the inverse of RNA production rate and (ii) the inverse of red fluorescent protein production rate, as a function of the inverse RNAP concentration.

3.10 Maximum likelihood estimation

Maximum likelihood estimation (MLE) is a method for finding such parameters of a statistical model that would maximize the likelihood that the data was generated by this model. Let the data set $\mathbf{x} = \{x_1, ..., x_n\}$ consist of n independent and identically distributed random variables. Given a probability density function (or a probability mass function) $f(x|\boldsymbol{\theta})$, the likelihood that the data set $\mathbf{x}$ was generated by the model with parameters $\boldsymbol{\theta} = \{\theta_1, ..., \theta_k\}$ takes form (Casella and Berger, 2001):

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \tag{3.9}$$

The parameter set at which the value of $L(\boldsymbol{\theta}|\mathbf{x})$ is maximized is called a maximum likelihood estimator, $\hat{\boldsymbol{\theta}}(\mathbf{x})$. One common way of finding the $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is by differentiation. First, extrema of $L(\boldsymbol{\theta}|\mathbf{x})$ are found as follows:

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}|\mathbf{x}) = 0, i = 1, ..., k \tag{3.10}$$

Next, at each extremum point, the second derivative test is performed in order to determine whether the point is a maximum or a minimum. In general case, this requires construction of a Hessian matrix (which at $k = 1$ equals the second derivative of the likelihood function). Finally, $L(\boldsymbol{\theta}|\mathbf{x})$ is evaluated at the local optima along with the boundary $\boldsymbol{\theta}$ values, and the $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is set to the one that produces the highest value.

It is a common practice to replace the likelihood function with its natural logarithm, log-likelihood $\log L(\boldsymbol{\theta}|\mathbf{x})$, since it is often easier to differentiate:

$$\log L(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}) \tag{3.11}$$

This is possible because the logarithm function is monotonous, and thus $L(\boldsymbol{\theta}|\mathbf{x})$ and $\log L(\boldsymbol{\theta}|\mathbf{x})$ have a common estimator $\hat{\boldsymbol{\theta}}(\mathbf{x})$ (Casella and Berger, 2001).

In **Publication II** and **Publication III**, MLE was used to estimate parameters of statistical models of gene expression and inducer intake kinetics from single-cell microscopy data, utilizing the methodology developed in (Häkkinen and Ribeiro, 2016). Namely, in **Publication II**, the best-fitting parameters of several possible models of transcription initiation dynamics were obtained. Meanwhile, in **Publication III**, the best-fitting gamma distributions for the distributions of time intervals between the placement of the inducer in the media and the first RNA production event were obtained at various temperatures. In **Publication IV**, linear fits used in τ plots were obtained by MLE as described in (Bevington and Robinson, 2003).

3.11 Methods for model selection

It is often of interest not only which parameters of a given statistical model maximize its likelihood function, but also which of the two given models can explain data the best. A common approach for comparing nested models (i.e. a pair of models where one model is identical to the other under a certain set of constraints) is a likelihood ratio test. The test assumes a null hypothesis H_0 that the model parameters θ belong to the constrained subset Θ_0 of the parameter space Θ. The alternative hypothesis H_1 states that the model parameters θ belong to the subset Θ_0^c which is complementary to the subset Θ_0. The likelihood ratio test statistic is calculated as follows (Casella and Berger, 2001):

$$\Lambda(\mathbf{x}) = \frac{\sup L(\boldsymbol{\theta} \in \Theta_0|\mathbf{x})}{\sup L(\boldsymbol{\theta} \in \Theta|\mathbf{x})} \tag{3.12}$$

In equation 3.12, the nominator is always equal or less than the denominator, thus $\Lambda(\mathbf{x})$ is in the interval $[0; 1]$, with the higher values showing that the two models perform similarly, and the lower values showing that the general model outperforms the model with additional constraints. According to the Wilks' theorem, as the sample size of the data increases, $-2\log\Lambda$ is asymptotically approaches a chi-squared distribution (Wilks, 1938). Further, the degree of freedom of this chi-squared distribution equals to the difference in the dimensionality between Θ_0 and Θ. Chi-squared distribution evaluated at $-2\log\Lambda(\mathbf{x})$ thus produces a p-value that is then compared to the chosen significance level α of the test. If p-value $\leq \alpha$, than the H_0 is rejected in favor of H_1. If the p-value $> \alpha$, the H_1 does not provide significant improvement and H_0 cannot be rejected.

Multiple methods for comparing non-nested models are available, with the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) being the most commonly used. Both methods measure the goodness of fit of a model by taking into account the maximum value of its likelihood function, $\hat{L}$, and penalizing the model based on the number of parameters, k. The difference between AIC and BIC is in the penalty on the number of parameters, which in the case of BIC is scaled by the sample size on the data, n:

$$\begin{aligned} \mathrm{AIC} &= -2k - 2\ln(\hat{L}) \\ \mathrm{BIC} &= -2\ln(n)k - 2\ln(\hat{L}) \end{aligned} \tag{3.13}$$

The smaller value of AIC or BIC corresponds to the best-fitting model among the tested set. In general, BIC performs better than AIC when the "true model" that produced the data is assumed to be in the set of the models being compared (Burnham and Anderson, 2004).

The likelihood ratio test was used in **Publication II** and **Publication IV** to test whether the data is best-fit by a line or a polynomial of the higher order. In **Publication IV**, the linear model was also compared with the models where data is constant on x-axis or on y-axis. The maximum of the likelihood function for these likelihood ratio tests was estimated based on the approach presented in (Krystek and Anton, 2008). In **Publication III**, the likelihood ratio test was used to compare the models of d exponential steps fitted to the estimated distribution of the inducer intake times at various temperatures. Finally, in **Publication II**, BIC was used to select between several plausible models of transcription initiation at lac/ara-1 promoter.

3.12 Applying deconvolution to empirical data

In certain cases, in a process that consists of two sequential sub-processes, the distributions of completion times of the whole process, d_{full}, and of one sub-process, d_1, are known, but the distribution of completion times of the other sub-process, d_2, cannot be measured directly. This composite process can be expressed as a convolution of its sequential sub-processes: $d_{full} = d_1 * d_2$. Thus, d_2 can be obtained by deconvolving d_1 from d_{full}. This problem is usually solved numerically by converting the distributions d_{full} and d_1 to the frequency domain, by the Fourier transform, and obtaining the d_2 in the frequency domain by complex division: $D_2 = D_{full}/D_1$ (Sheu and Ratcliff, 1995; Smith, 1990). D_2 is then converted to the time domain using the inverse Fourier transform, yielding a d_2 estimate.

For this approach to produce meaningful results, the measured d_{full} and d_1 used for deconvolution should be close to the actual distributions that produced the data. For this, the measurements should contain a sufficient amount of samples, d_{full} and d_1 should not be significantly corrupted by noise, and the duration of the experimental observations should be long enough to avoid right-censoring of the data. Even if a large number of samples is used, the data is filtered from noise, and right-censoring does not pose a problem, a small number of samples in d_2 could still take negative values. This is the case since the experimentally obtained d_{full} and d_1 are not the exact representations of the distributions that produced these data. It is also common for this method to underestimate the peak value of the d_2 distribution (Sheu and Ratcliff, 1995). With these limitations in mind, this deconvolution approach can be used to obtain insights into the processes that otherwise are challenging to quantify.

In **Publication III**, deconvolution was applied to empirical distributions of (i) time intervals between the inducer placement in the media and the first RNA production event and (ii) time intervals between consecutive RNA production events from an induced gene, in order to estimate the distributions of time intervals that it takes for the inducer to enter the cell after being placed in the media, t_{int}, at various temperature conditions. To evaluate the confidence in the estimates, confidence intervals for the mean and coefficient of variation of the t_{int} distributions in various temperature conditions were constructed using the bootstrap approach (section 3.9.2).

4 Summary of the Results

In **Publication I**, a new method for quantitative estimation of fluorescent molecule numbers from temporal fluorescence intensity data corrupted by transient nonzero-mean noise was developed. The method aimed to prevent the transient disruptions in the signal from affecting the molecule number estimates at the time moments before the disruption was introduced. For this, the author and colleagues developed an algorithm that utilizes a stepwise approach, where only the information from several following time points is allowed to affect the estimated number of molecules at a given moment. This improved accuracy in the cases where fluorescent molecules were absent from the cell image for durations comparable to the cell lifetime.

This method is based on the assumption that temporal fluorescence intensity data can be described as a monotonic non-decreasing function corrupted by three types of noise. First, the imprecisions of a microscope and detector add an independent, normally distributed noise to the fluorescence intensity value at each time point (Chowdhury et al., 2012; Waters, 2009). Second, fluorescent molecules are able to move out of the focal plane and remain out of focus for durations on the order of the cell lifetime before reemerging. This causes transient, non-periodic negative noise in the fluorescent intensity signal. Third, such events as, e.g., false-positive detection of fluorescent molecules result in rarely occurring positive spikes in the fluorescent intensity values. The parameters of the algorithm developed in this thesis for detection of RNA production events can be tuned to account for various intensities of these three types of noise.

The algorithm, shown in Figure 2 of the **Publication I**, has two parameters. The first one, ω, is the number of consecutive points in the time series, starting with the current one, used to decide whether a new RNA has been produced at the current time point. The decision is made by comparing the average intensity at these time points to a specified threshold. Considering only ω points at a time allows to avoid the influence of the negative noise (type 2) that occurs before or after this time window. However, low ω values could cause false-positive detection of RNA production events based on rare spikes of positive noise (type 3) or based on random fluctuations of the local mean of the zero-mean noise (type 1). The second parameter, v, is the value used to modify the aforementioned threshold so that these random fluctuations in the local mean would not hinder detection of RNA production events.

Using stochastic simulations, **Publication I** demonstrates that the optimal value of ω depends mostly on two factors: f, the sampling frequency of the time series, and σ, the standard deviation of the consistent zero-mean noise. Namely, ω increased both with the increase in f and with the increase in σ. Meanwhile, the intensity of the transient, non-periodic negative noise (type 2) does not seem to affect this value. The optimal value of v equals 0.25 and does not respond to changes in any of these factors. However, the

optimal value of v increased when the mean of the consistent noise (type 1) was changed from zero to a negative value.

To evaluate the performance of this new algorithm in comparison with the previously existing one, both the new and the reference methods were applied to the simulated data. As Figure 5 in **Publication I** shows, the new method consistently performs better when the data is affected by the transient, non-periodic negative noise, and performs worse in the absence of this noise. The transient noise was modeled by allowing each present RNA to leave the focal plane, on average, once in 60 min, and by permitting each absent RNA to return into focus, on average, in 20 min, mimicking the behavior of the MS2-GFP-RNA complexes observed under the microscope. Finally, both the new and the reference methods were applied to empirical data corrupted by this transient noise, and the new method was found to produce more accurate results, as evaluated by human experts and illustrated in Figure 6 of **Publication I**.

In **Publication II**, we proposed a methodology for dissecting *in vivo* transcription initiation kinetics and applied it to an *E. coli* promoter $\mathrm{P_{lac/ara\text{-}1}}$. For this, we first established the concept of an *in vivo* τ plot and demonstrated its viability. The *in vivo* τ plot allows to estimate, from *in vivo* measurements of transcription initiation kinetics, the average time spent in transcription initiation prior to commitment to the open complex formation (τ_{prior}) and the average time spent in transcription initiation after successful commitment to the open complex formation (τ_{after}). Further, the concept of an *in vivo* τ plot was utilized to put constraints on the parameters of a stochastic model of transcription initiation. This allowed to infer the rate-limiting constants of *in vivo* transcription initiation in more detail than it was possible previously.

Construction of a τ plot using *in vivo* measurements of transcription initiation kinetics utilizes the similarities in behavior of τ_{prior} and τ_{after} with τ_c and τ_o, correspondingly (see section 3.8). Namely, given the stochastic model of *in vivo* transcription (reactions 2.9 and 2.10), τ_{prior} changes linearly with the inverse of [RNAP], whereas τ_{after} is not significantly affected by this concentration. To demonstrate this, a moment-generating function of the distribution of the transition times between the P_{ON} state and the release of the product RNA (see reactions 2.9 and 2.10) was derived as in (Häkkinen and Ribeiro, 2016):

$$\begin{aligned} M_{(in\ vivo)}(t) &= \frac{(k_{\mathrm{ON}} - t)}{k_{\mathrm{ON}}} \frac{Q_2(0)}{Q_2(t)} + \frac{k_3 - t}{k_3}, \\ Q_2(t) &= (k_{-1} + k_2 - t)Q_1(t) - k_1 k_{-1}(k_{\mathrm{ON}} - t), \\ Q_1(t) &= (k_{\mathrm{OFF}} + k_1 - t)(k_{\mathrm{ON}} - t) - k_{\mathrm{ON}} k_{\mathrm{OFF}}. \end{aligned} \tag{4.1}$$

Equation 4.1 allows obtaining the analytical expression for the mean of this distribution, denoted here as μ:

$$\mu = M'_{(in\ vivo)}(0) = \underbrace{\frac{1}{[\mathrm{RNAP}]} \frac{(k_{\mathrm{ON}} + k_{\mathrm{OFF}})(k_{-1} + k_2)}{k_1 k_2 k_{\mathrm{ON}}}}_{\tau_{prior}} + \underbrace{\frac{1}{k_2} + \frac{1}{k_3}}_{\tau_{after}} \tag{4.2}$$

From equation 4.2, μ changes linearly with 1/[RNAP]. In particular, τ_{prior} is directly proportional to 1/[RNAP] and τ_{after} is not affected by [RNAP]. This allows to construct a schematic *in vivo* τ plot (Figure 4.1). Note that τ_{prior} differs from τ_c by also including the time the promoter spends in a transient inactive state P_{OFF}, and τ_{after} differs from τ_o by accounting for transcription elongation.

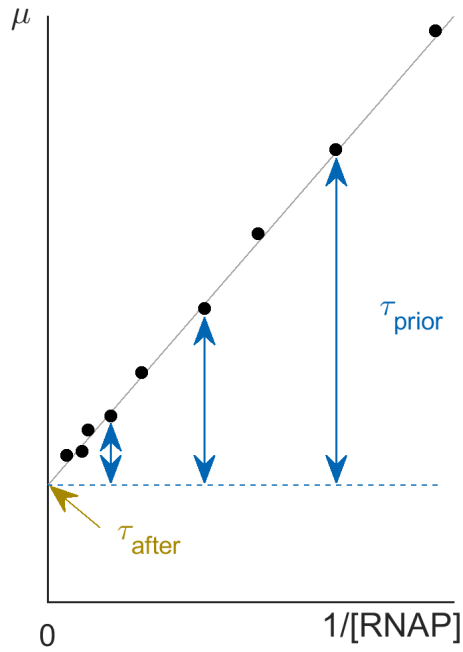


Figure 4.1: Schematic representation of an *in vivo* τ plot, showing the inverse of RNA production rate, μ, against the inverse of RNAP concentration, 1/[RNAP]. The average time spent in transcription initiation after commitment to the open complex formation, τ_{after}, is estimated as a y-intercept of the best-fitting line to the empirical data (black dots).

To construct this τ plot based on empirical data, it is essential to vary [RNAP] *in vivo* without significantly affecting other variables in the modeled system (reactions 2.9 and 2.10). Here, *in vivo* [RNAP] was varied by changing growth media composition in a specific range that was determined empirically. The criterion for this range of media concentrations was that the inverse of RNA production rate should change linearly with the inverse of [RNAP] in this range. This assumption of linearity was supported by empirical data. The inverse of RNA production rate from an *E. coli* promoter $\mathrm{P_{lac/ara\text{-}1}}$ and the inverse of [RNAP] were estimated from qPCR measurements (see section 3.4) in live *E. coli* cells. The inverse of RNA production rate was shown to change linearly with the inverse of [RNAP] within a specific range of media conditions. Additional results provided evidence that this linear relationship is strain-independent.

Next, using single-cell time-lapse microscopy (see section 3.2), we obtained the distributions of time intervals between RNA production events from $\mathrm{P_{lac/ara\text{-}1}}$ promoter in different media conditions. Using these and the RNAPs concentrations measured above, we utilized the *in vivo* τ plots methodology to estimate not only τ_{prior} and τ_{after}, but also the rate-limiting steps in transcription initiation. Based on the concept of a τ plot,

only the rate constant that depends on [RNAP], the rate of RNAP binding to an active promoter, was allowed to differ between different media conditions during the model fitting. To infer the rate-limiting steps, we considered the full model of transcription initiation (reactions 2.9 and 2.10) and simplified models that could be preferred if some steps of the full model do not influence the transcription initiation kinetics significantly. We considered the following simplifications (in various combinations): (i) the time that promoter spends in the locked state is negligible, either because it rarely gets into the locked state or because unlocking is fast; (ii) the closed complex formation is reversible, (iii) the closed complex formation is irreversible, (iv) the steps in transcription initiation that take place after the open complex is formed happen negligibly fast. By comparing the models using BIC (see section 3.11), we found that the best-fit model is the one that allows the promoter to be in an inactive state for non-negligible time and assumes that the closed complex formation is a reversible process. The best-fitting rate constants are shown in Table 2 of **Publication II**. Finally, by varying the induction scheme of the $P_{lac/ara\text{-}1}$ promoter and by applying the same model fitting and selection technique as above, we determined that the intermittent inactive states are caused by intermittent binding of LacI repressor.

In **Publication III**, we studied how kinetics of inducer intake varies with temperature shifts. For this, we developed a new method for estimating the distribution of time intervals between the moment when inducer is placed in the media and the moment when the inducer has entered the cell cytoplasm, where it activates the target gene. By applying this method to the empirical data collected at various temperature conditions, we found that the mean inducer intake time increases at suboptimal temperatures, while the cell-to-cell variability of the intake times decreases. Using the standard model-fitting procedure, we demonstrated that this is likely due to emergence of an additional step in the inducer intake process at suboptimal temperatures.

Using single-cell time-lapse microscopy, we measured the distribution of time intervals between adding an inducer to the media and the first RNA production event (t_0 distribution) under various temperature shifts. The Δt distributions under the same conditions were obtained in a previous work (Oliveira et al., 2016a). To estimate the distribution of induction times in each condition, we deconvolved each Δt distribution from the corresponding t_0 distribution using a common strategy for applying deconvolution to empirical data (see section 3.12). We then calculated the mean and the variability (the squared coefficient of variation) of the distributions of the intake times, and found that the mean increases at suboptimal temperatures, while the variability decreases. The former finding was validated by estimating the mean inducer intake time at various temperatures, from qPCR measurements (see section 3.4), using the Lineweaver-Burk equation (see section 3.7 and equation (3) in **Publication III**).

Finally, we estimated the number and durations of the rate-limiting steps in the deconvolved distributions of inducer intake times in maximum likelihood sense (see section 3.10). Namely, we assumed a model of inducer intake to involve d consecutive exponential steps, with this number of steps possibly differing between various temperature conditions. Using the likelihood ratio tests (see section 3.11), we found that the inducer intake involves two rate-limiting steps at the optimal temperature, with at least one additional rate-limiting step emerging at lower temperatures (the detailed results are shown in Table 4 of **Publication III**).

In **Publication IV**, we studied the means to control propensities of threshold crossing in RNA and protein numbers in *E. coli*. For this, we investigated whether it is possible

to tune asymmetry and tailedness in transcription initiation kinetics independently of its mean rate and variability, and whether changes in these asymmetry and tailedness can significantly affect threshold crossing in RNA and protein numbers. By applying the tests of statistical comparison to empirical data, we demonstrated that changes in these asymmetry and tailedness are independent of changes in the mean and variability (while correlated between each other), and play a key role in crossing more exclusive thresholds. Further, these asymmetry and tailedness propagate to the asymmetry and tailedness of protein levels in a cell population. Since this can be observed for the sets of conditions that differ only in a promoter sequence and only in regulatory factors, the asymmetry and tailedness in transcription initiation kinetics could play a crucial role in cellular decision making processes by controlling the threshold crossing propensities both through evolution and by regulation.

First, we performed single-cell microscopy measurements to obtain the Δt distributions in various conditions differing in promoter sequence, induction schemes, and media composition (see Table 1 of **Publication IV** for the detailed description of the conditions). From each Δt distribution, we estimated its mean (M), coefficient of variation (CV), asymmetry (assessed by skewness, S), and tailedness (assessed by kurtosis, K). The results are shown in Figure 4.2A, with the standard error of the mean estimated using non-parametric bootstrapping (see section 3.9.2). We found that S and K are correlated with each other, whereas all other pairs of these four features do not exhibit correlation. This was also found true for the set of conditions that differ only in promoter sequence and for the set of conditions that differ only in regulatory mechanisms (Table 2 of **Publication IV**). In addition, Figure 4.2B shows that S and K can differ widely between conditions even when M and CV remain nearly the same.

We then investigated whether the values of S and K contribute significantly to crossing thresholds in RNA numbers. Since these numbers are defined by the Δt intervals, we tested how the percentage of the Δt intervals that are higher than a given threshold depends on CV, S and K. To eliminate the influence of M from our results, we scaled the threshold by the mean of the Δt distribution. Figures 4.2C and 4.2D show that, at lower thresholds, the values of S and K don't seem to affect the threshold propensities and CV plays the major role in the threshold crossing. However, to cross more challenging thresholds (5M and 6M), both higher values of CV and of S and K are required.

Further, we estimated τ_{prior} and τ_{after} in each condition using the *in vivo* τ plot methodology developed in **Publication II**. The inverse of RNAP concentration was measured using the Western blot, and the inverse of RNA production rate was measured using qPCR. S and K were shown to change linearly with τ_{prior} when the number of variables that were allowed to change between conditions was restricted (only the promoter sequence or only regulatory factors were allowed to change). Interestingly, the linear relationship is positive in the set of conditions where only promoter sequence changes, and is negative in the set of conditions where only regulatory factors change. This suggests that these two sets of conditions affect τ_{prior} differently, e.g. by tuning different rate-limiting steps in transcription initiation that occur before commitment to open complex formation. In addition, no significant relationship between S and K and τ_{after} or τ_{after}/M was found.

Finally, using flow cytometry, we measured the distributions of corresponding protein expression levels in individual cells, in various conditions. We found that the skewness of this distribution is negatively correlated with the skewness of a Δt distribution. Further, the mean of the Δt distribution was not found to affect the skewness and kurtosis of the distribution of protein expression levels. **Publication IV** demonstrates that τ_{prior}

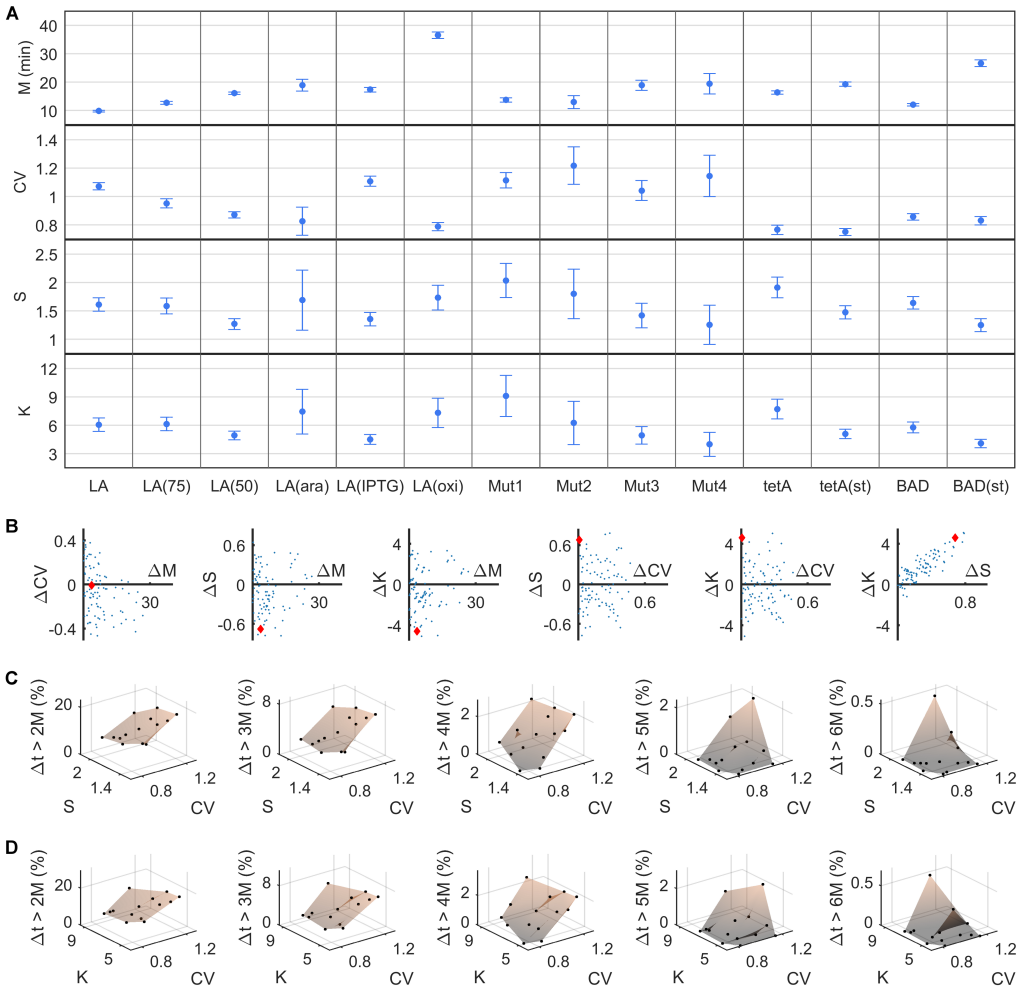


Figure 4.2: Skewness (S) and kurtosis (K) of a Δt distribution are independent of the mean (M) and the coefficient of variation (CV) of this distribution and affect the probability of crossing the upper-bound thresholds in Δt intervals. (A) M, CV, S and K of the Δt distributions measured in 14 different conditions. Error bars denote the standard error of the mean. (B) Pairwise differences (Δ) in M, CV, S and K between conditions (blue dots). The red diamond is the difference between LA(IPTG) and Mut1 conditions that illustrates how changes in S and K can be independent from changes in M and CV. (C and D) Percentage of Δt intervals (black dots) that are longer than a given threshold (from 2M to 6M) against (C) CV and S, and (D) CV and K. Also shown is the natural neighbor interpolation surface. Reproduced from **Publication IV**.

partially defines the values of skewness and kurtosis of a Δt distribution. These values affect threshold crossing propensities of RNA numbers and propagate to the values of skewness and kurtosis in the protein levels. Based on this, we suggest that changes in transcription initiation kinetics play a role in cellular decision-making processes.

5 Conclusions and Discussion

This thesis has focused on dissecting rate-limiting steps in transcription initiation and activation, studying the means to regulate these rate-limiting steps, and quantifying the influence of this regulation on the single-cell distributions of gene expression products. The study was mainly based on single-cell time-lapse microscopy data, heavily supported by novel applications of established molecular biology techniques. To interpret the empirical data, this study employed image and data processing strategies, stochastic modeling, and statistical analysis. Further, new methods for interpreting the data were developed and their usage was exemplified. Finally, this study resulted in novel insights into the regulation of rate-limiting steps in transcription initiation and its impact on RNA production kinetics and protein level distributions.

In order to understand the regulation of gene expression at the level of transcription, it is crucial to observe RNA production kinetics, and not only RNA and protein numbers, as these are affected by post-transcriptional and translational regulatory mechanisms (Picard et al., 2012; Van Assche et al., 2015). Thus, the distributions of the intervals between consecutive RNA production events (Δt distributions) are the backbone of the body of empirical data used in this work. To detect when a new RNA molecule is produced in a live *E. coil* cell, we used the MS2-GFP tagging method (see section 3.1.1). Although this tagging method significantly alters the properties of the RNA molecules, this is beneficial for detecting the intervals between RNA production events. Namely, the fact that MS2-GFP-coated RNA region degrades at a much slower rate than individual GFP or RNA molecules is utilized in the methodology for detection of RNA production events in live *E. coli* cells (Häkkinen and Ribeiro, 2015), including the method proposed in **Publication I**.

Publication I achieved the first aim of the study by presenting a new method for analyzing fluorescence time series data extracted from microscopy images. This method allowed to reduce the impact of transient nonzero-mean noise at previous time points on the estimation of an integer-valued RNA amount at the given moment by applying a stepwise approach to the time series data analysis. A search for the optimal values of the two parameters of this method shown that the optimal value of one of the parameters, v, is constant given realistic changes between the simulated conditions. Meanwhile, the optimal value of the other parameter, ω, depends on the sampling frequency of the time series and on the standard deviation of the constantly present zero-mean noise. While the sampling frequency of the data is always known because it is set up during the microscopy experiment, the standard deviation of the zero-mean noise was not estimated in this work. Instead, we used an average optimal value of ω that was obtained based on the analysis of the simulated data with realistic characteristics. Given this, augmenting our method with a step of estimating the standard deviation of the zero-mean noise and using this value to choose the optimal ω could result in higher accuracy of the method. However,

when the method was tested on the simulated data, using an optimal ω value provided only a slight increase in accuracy compared to using the average value.

The method developed in **Publication I** showed better performance than the existing method (Häkkinen and Ribeiro, 2015) when transient nonzero-mean noise was added to the simulated data, but also showed lower performance when no noise or only constantly present Gaussian noise with zero mean was introduced. Thus, the new method is not aimed to replace the already existing one, but instead the methods should be used in a complementary manner, in accordance with the characteristics of the data. When processing the data produced during the work on the following publications of the thesis, we visually evaluated performance of both the new and the reference methods and selected the one that performed better. In all cases, the performance of the new method was either worse or the same as that of the reference method. This is explained by an improvement in the microscopy setup that occurred after the completion of **Publication I**. Namely, the microscope laser alignment and coupling of the laser box were performed, which improved signal-to-noise ratio in the obtained images. The nonzero-mean transient noise posed a serious signal processing problem when using the MS2-GFP tagging system with the old microscopy setup but was not an issue anymore when using the new setup. Nevertheless, the method developed in this thesis could be of use in the future. For example, it is expected to perform well when the fluorescent time series are obtained from the molecules tagged by a small number (e.g. less than 40) of fluorescent proteins. This is based on the observation that as the size of the fluorescent tag reduces, the probability that the tagged molecule would not appear on the microscopy image in a given frame increases, adding tangible nonzero-mean noise to the fluorescence time series.

The second aim of the study was achieved in **Publication II**, which developed and applied a novel methodology for dissecting *in vivo*, single-cell transcription initiation kinetics. The prerequisite assumptions of the proposed approach were proved to hold true by a combination of molecular biology assays and statistical analysis techniques. The functionality of this new approach was demonstrated in a case study of *E. coli* promoter lac/ara-1. **Publication II** contributed to unraveling *in vivo* transcription initiation kinetics by approaching the question in two stages. First, it established that τ plots, previously used only *in vitro* (Bertrand-Burggraf et al., 1984; McClure, 1980), can also be used *in vivo* for estimating the average durations spent in transcription initiation prior and after closed complex formation. Next, it combined the model-fitting strategy recently developed to characterize the rate-limiting steps in transcription (Häkkinen and Ribeiro, 2016) with introducing the parameter constraints based on the experimental conditions into the model fitting procedure. This resulted in a novel methodology that allows studying *in vivo* transcription initiation at a higher level of detail than previously.

Although τ plots were known to be applicable to *in vitro* data for about 40 years, it was not clear whether the *in vivo* application of this method would be possible. In particular, varying RNAP concentration *in vivo* without significantly affecting other variables in transcription initiation was a non-trivial task, since RNAP concentration is known to affect cell growth rate, a parameter that is linked to changes in global regulation of gene expression (Klumpp and Hwa, 2008, 2014). Nevertheless, we were able to find a range of growth media composition where RNAP concentration changes while other variables stay approximately constant. This new methodology makes possible testing whether the *in vitro* measurements of transcription initiation kinetics are a good proxy for the *in vivo* dynamics. If this would be proven true, the large volume of measurements on *in vitro* transcription initiation kinetics could be used with more confidence than previously for

estimating the *in vivo* transcription kinetics.

The method developed in **Publication II** does not yet allow to estimate the number and durations of the multiple steps that compose the open complex formation and promoter escape. This could be possible if the kinetics of one of the sub-steps could also be altered without altering the other sub-steps. Interestingly, at about the same time as this work, the *in situ* studies of transcription initiation using single-molecule Förster resonance energy transfer were conducted (Duchi et al., 2016; Lerner et al., 2016). The approach used in these studies offer a great level of detail on certain steps in transcription initiation. However, the methodology used in this thesis is less invasive and allows for time-lapse observation of living cells during several hours, which is not possible with the *in situ* approach.

The third aim of the study was achieved in **Publication III**, which offered previously intangible insights into the temperature dependence of inducer intake times by combining the established molecular biology and statistical methods. The results indicated higher mean intake times and lower coefficient of variation in these times at lower temperatures, which suggests the emergence of additional rate-limiting steps. The results were consistent in that we found two exponential rate-limiting steps at 37 °C, three at 30 °C, and four or more at 24 °C. The fact that we cannot conclude on the number of the rate-limiting steps in inducer intake kinetics at 24 °C suggests that, at this temperature, the inducer intake kinetics have changed so drastically that a sequence of a small number of exponential steps is not a sufficient model anymore. One possible explanation for this phenomenon is an increase in the viscosity of the cytoplasm and the periplasm at lower temperatures (Oliveira et al., 2016b).

Publication III was based on the recent work that studied how the rate-limiting steps in transcription initiation are affected by temperature and induction schemes (Oliveira et al., 2016a). This study demonstrated that sub-optimal temperatures affect different rate-limiting steps to an unequal degree, and that these effects differ between promoters. However, it did not investigate whether inducer intake times are also significantly affected by the temperature shifts. Answering this question is crucial for understanding how bacterial cells function in constantly changing environmental conditions. Our results strongly suggest significant changes in the inducer intake kinetics at lower temperatures. It is yet to be understood how bacteria adapt to these changes, since they are known to survive in a wide range of temperature conditions. From previous studies, it is known that response to temperature shifts is gene-dependent (Chandraseelan et al., 2013; Touhami et al., 2006), which may play a role in how bacteria handle intake of molecules through cell walls at lower temperatures.

The fourth aim of the study was achieved in **Publication IV** by investigating the relationship between rate-limiting steps in transcription initiation, asymmetry and tailedness in RNA and protein expression levels, and threshold-crossing in RNA numbers in various conditions. It has been suggested that cellular decision-making mechanisms depend on threshold-crossing by RNAs or proteins (Alon, 2007; Arkin et al., 1998; McAdams and Arkin, 1997). In **Publication IV**, increasing skewness and kurtosis can serve as a means to cross higher thresholds in Δt intervals while not changing the mean and coefficient of variation. It also was shown that skewness and kurtosis in RNA production kinetics are negatively correlated with the skewness and kurtosis in protein level distributions, demonstrating that the impact of the rate-limiting steps is not necessarily lost in post-transcriptional noise. In addition, these skewdness and kurtosis were found to vary significantly between conditions differing only in promoter sequence or in RNAP or inducer

concentrations, indicating that the skewness and kurtosis are both sequence-dependent and subject to regulation. Thus, the work conducted in **Publication IV** suggests that skewness and kurtosis play a significant role in cellular decision-making.

Previously, it has been shown that the rate-limiting steps in transcription initiation can tune the mean and coefficient of variation of a Δt distribution (Häkkinen and Ribeiro, 2016; Mäkelä et al., 2017; Oliveira et al., 2016a). However, the regulation of the skewness and kurtosis of the Δt distribution has not been considered before. Given that changing the rate-limiting steps in transcription initiation allows tuning the skewness and kurtosis of this distribution independently from its mean and coefficient of variation, considering these skewness and kurtosis when, e.g., constructing genetic circuits with desirable properties could allow higher precision and flexibility in tuning the dynamics of these circuits. Previous studies have demonstrated that the mean and noise of a distribution of protein numbers affect different genetic circuits to various extent depending on the topology of the circuit (Cameron and Collins, 2014; Morelli and Jülicher, 2007; Purcell et al., 2010). In particular, different shapes of noise in protein numbers can either facilitate or affect detrimentally the functionality of a genetic oscillator, depending not only on the shape of noise but also on the topology of the oscillator (Purcell et al., 2010). Thus, it would be of interest to study how the skewness and kurtosis in RNA and protein numbers affect the functionality of various genetic circuits and how these effects are related to the topology of the circuits. For example, it would be of interest to test the hypothesis that lower absolute values of skewness and kurtosis in protein numbers increase robustness of a repressilator (a three-gene repression oscillator) by reducing the occurrence of short-term spikes in protein numbers that may cause skipped oscillations.

To summarize, the results of this study could be of use for investigating how regulation in transcription initiation affects functionality of network motifs. The presented work can serve as a starting point for future studies based both on stochastic models and on single-cell microscopy measurements of *in vivo* dynamics of genetic constructs. This thesis contributes to an emerging field of studying the rate-limiting steps in prokaryotic transcription initiation *in vivo*, at the single-cell level, and should facilitate further research in the area.

Bibliography

Acar, M., Mettetal, J. T., and van Oudenaarden, A., "Stochastic switching as a survival strategy in fluctuating environments," *Nature Genetics*, vol. 40, no. 4, pp. 471–475, 2008.

Adelman, K., La Porta, A., Santangelo, T. J., Lis, J. T., Roberts, J. W., and Wang, M. D., "Single molecule analysis of RNA polymerase elongation reveals uniform kinetic behavior," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 21, pp. 13 538–13 543, 2002.

Aghaeepour, N., Finak, G., The FlowCAP Consortium, The DREAM Consortium, Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., and Scheuermann, R. H., "Critical assessment of automated flow cytometry data analysis techniques," *Nature Methods*, vol. 10, no. 3, pp. 228–238, 2013.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., *Molecular Biology of the Cell*, 5th ed., Anderson, M. and Granum, S., Eds. Garland Science, 2008.

Alon, U., "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, 2007.

Arkin, A., Ross, J., and McAdams, H. H., "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells," *Genetics*, vol. 149, pp. 1633–1648, 1998.

Babu, M. M. and Teichmann, S. A., "Evolution of transcription factors and the gene regulatory network in *Escherichia coli*," *Nucleic Acids Research*, vol. 31, no. 4, pp. 1234–1244, 2003.

Beck, C. F. and Warren, R. A. J., "Divergent promoters, a common form of gene organization," *Microbiological Reviews*, vol. 52, no. 3, pp. 318–326, 1988.

Belogurov, G. A. and Artsimovitch, I., "Regulation of transcript elongation," *Annual Review of Microbiology*, vol. 69, pp. 49–69, 2015.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N., "Global analysis of mRNA decay and abundance in *Escherichia coli* cells at single-gene resolution using two-color fluorescent DNA microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9697–9702, 2002.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., and Long, R. M., "Localization of *ASH1* mRNA particles in living yeast," *Molecular Cell*, vol. 2, no. 4, pp. 437–445, 1998.

Bertrand-Burggraf, E., Lefèvre, J., and Daune, M., "A new experimental approach for studying the association between RNA polymerase and the tet promoter of pBR322," *Nucleic Acids Research*, vol. 12, no. 3, pp. 1697–1706, 1984.

Beveridge, T. J., "Structures of gram-negative cell walls and their derived membrane vesicles," *Journal of Bacteriology*, vol. 181, no. 16, pp. 4725–4733, 1999.

Bevington, P. R. and Robinson, D. K., "Least-squares fit to a straight line," in *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed. McGraw-Hill, 2003, pp. 98–115.

Blattner, F. R., Plunkett, III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y., "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

Blount, Z. D., "The unexhausted potential of *E. coli*," *eLife*, vol. 4, p. e05826, 2015.

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*, vol. 102, no. 41, pp. 14 593–14 598, 2005.

Bratton, B. P., Mooney, R. A., and Weisshaar, J. C., "Spatial distribution and diffusive motion of RNA polymerase in live *Escherichia coli*," *Journal of Bacteriology*, vol. 193, no. 19, pp. 5138–5146, 2011.

Brewster, R. C., Jones, D. L., and Phillips, R., "Tuning promoter strength through RNA polymerase binding site desgin in *Escherichia coli*," *PLoS Computational Biology*, vol. 8, no. 12, p. e1002811, 2012.

Browning, D. F. and Busby, S. J. W., "The regulation of bacterial transcription initiation," *Nature Reviews Microbiology*, vol. 2, no. 1, pp. 57–65, 2004.

Browning, D. F. and Busby, S. J. W., "Local and global regulation of transcription initiation in bacteria," *Nature Reviews Microbiology*, vol. 14, no. 10, pp. 638–650, 2016.

Bryant, J. A., Sellars, L. E., Busby, S. J. W., and Lee, D. J., "Chromosome position effects on gene expression in *Escherichia coli* K-12," *Nucleic Acids Research*, vol. 42, no. 18, pp. 11 383–11 392, 2014.

Buc, H. and McClure, W. R., "Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac uv5 promoter. evidence for a sequential mechanism involving three steps," *Biochemistry*, vol. 24, no. 11, pp. 2712–2723, 1985.

Burenina, O. Y., Elkina, D. A., Hartmann, R. K., Oretskaya, T. S., and Kubareva, E. A., "Small noncoding 6S RNAs of bacteria," *Biochemistry (Moscow)*, vol. 80, no. 11, pp. 1429–1446, 2015.

Burkhardt, D. H., Rouskin, S., Zhang, Y., Li, G.-W., Weissman, J. S., and Gross, C. A., "Operon mRNAs are organized into ORF-centric structures that predict translation efficiency," *eLife*, vol. 6, p. e22037, 2017.

Burnette, W. N., ""Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A," *Analytical Biochemistry*, vol. 112, no. 2, pp. 195 – 203, 1981.

Burnham, K. P. and Anderson, D. R., "Multimodel inference," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261 – 304, 2004.

Byrgazov, K., Vesper, O., and Moll, I., "Ribosome heterogeneity: another level of complexity in bacterial translation regulation," *Current Opinion in Microbiology*, vol. 16, no. 2, pp. 133 – 139, 2013.

Cabrera, J. E. and Jin, J. E., "The distribution of RNA polymerase in *Escherichia coli* is dynamic and sensitive to environmental cues," *Molecular Microbiology*, vol. 50, no. 5, pp. 1493 – 1505, 2003.

Cameron, E. D. and Collins, J. J., "Tunable protein degradation in bacteria," *Nature Biotechnology*, vol. 32, no. 12, pp. 1276 – 1281, 2014.

Cao, Y., Terebus, A., and Liang, J., "Accurate chemical master equation solution using multi-finite buffers read more: https://epubs.siam.org/doi/abs/10.1137/15m1034180," *Multiscale Modeling & Simulation*, vol. 14, no. 2, pp. 923 – 963, 2016.

Carpenter, J. and Bithell, J., "Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians," *Statistics in Medicine*, vol. 19, no. 9, pp. 1141 – 1164, 2000.

Casadesús, J. and Low, D., "Epigenetic gene regulation in the bacterial world," *Microbiology and Molecular Biology Reviews*, vol. 70, no. 3, pp. 830 – 856, 2006.

Casadesús, J. and Low, D. A., "Programmed heterogeneity: Epigenetic mechanisms in bacteria," *Journal of Biological Chemistry*, vol. 288, no. 20, pp. 13 929 – 13 935, 2013.

Casella, G. and Berger, R. L., "The Delta Method," in *Statistical Inference*, 2nd ed. Duxbury Press, 2001, pp. 240 – 245.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C., "Green fluorescent protein as a marker for gene expression," *Science*, vol. 263, no. 5148, pp. 802 – 805, 1994.

Chandraseelan, J. G., Oliveira, S. M. D., Häkkinen, A., Tran, H., Potapov, I., Sala, A., Kandhavelu, M., and Ribeiro, A. S., "Effects of temperature on the dynamics of the LacI-TetR-CI repressilator," *Molecular BioSystems*, vol. 9, no. 12, p. 3117, 2013.

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*, vol. 11, no. 1, p. 781, 2015.

Chen, W. W., Niepel, M., and Sorger, P. K., "Classic and contemporary approaches to modeling biochemical reactions," *Genes & Development*, vol. 24, no. 17, pp. 1861 – 1875, 2010.

Chen, Y., Pai, A. A., Herudek, J., Lubas, M., Meola, N., Järvelin, A. I., Andersson, R., Pelechano, V., Steinmetz, L. M., Jensen, T. H., and Sandelin, A., "Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters," *Nature Genetics*, vol. 48, no. 9, pp. 984 – 994, 2016.

Chong, S., Chen, C., Ge, H., and Xie, X. S., "Mechanism of transcriptional bursting in bacteria," *Cell*, vol. 158, no. 2, pp. 314–326, 2014.

Chowdhury, S., Kandhavelu, M., Yli-Harja, O., and Ribeiro, A. S., "An interacting multiple model filter-based autofocus strategy for confocal time-lapse microscopy," *Journal of Microscopy*, vol. 245, no. 3, pp. 265–275, 2012.

Chowdhury, S., Kandhavelu, M., Yli-Harja, O., and Ribeiro, A. S., "Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE)," *BMC Bioinformatics*, vol. 14, no. Suppl 10, p. S8, 2013.

Cooper, G. M., *The Cell: A Molecular Approach*, 2nd ed. Sinauer Associates, 2000.

Crick, F., "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–563, 1970.

Darst, S. A., "Bacterial RNA polymerase," *Current Opinion in Structural Biology*, vol. 11, no. 2, pp. 155–162, 2001.

Day, R. N. and Davidson, M. W., "The fluorescent protein palette: tools for cellular imaging," *Chemical Society Reviews*, vol. 38, no. 10, pp. 2887–2921, 2009.

De Lay, N. and Gottesman, S., "A complex network of small non-coding RNAs regulate motility in *Escherichia coli*," *Molecular Microbiology*, vol. 86, no. 3, pp. 524–538, 2012.

Degras, D., "Simultaneous confidence bands for the mean of functional data," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 9, no. 3, p. e1397, 2017.

deHaseth, P. L., Zupancic, M. L., and Record, Jr., T. M., "RNA polymerase-promoter interactions: the comings and goings of RNA polymerase," *Journal of Bacteriology*, vol. 180, no. 12, pp. 3019–3025, 1998.

Deuschle, U., Kammerer, W., Gentz, R., and Bujard, H., "Promoters of *Escherichia coli*: a hierarchy of *in vivo* strength indicates alternate structures," *The EMBO Journal*, vol. 5, no. 11, pp. 2987–2994, 1986.

Deutscher, M. P., "Degradation of RNA in bacteria: comparison of mRNA and stable RNA," *Nucleic Acids Research*, vol. 34, no. 2, pp. 659–666, 2006.

DiCiccio, T. J. and Efron, B., "Bootstrap confidence intervals," *Statistical Science*, vol. 11, no. 3, pp. 189–228, 1996.

Dobrzyński, 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*, vol. 106, no. 8, pp. 2583–2588, 2009.

Dominguez, A. A., Lim, W. A., and Qi, L. S., "Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 1, pp. 5–15, 2015.

Dougan, D. A., Mogk, A., and Bukau, B., "Protein folding and degradation in bacteria: to degrade or not to degrade? That is the question," *Cellular and Molecular Life Sciences*, vol. 59, no. 10, pp. 1607–1616, 2002.

Dove, S. L., Darst, S. A., and Hochschild, A., "Region 4 of σ as a target for transcription regulation," *Molecular Microbiology*, vol. 48, no. 4, pp. 863–874, 2003.

Duchi, D., Bauer, D. L. V., Fernandez, L., Evans, G., Robb, N., Hwang, L. C., Gryte, K., Tomescu, A., Zawadzki, P., Morichaud, Z., Brodolin, K., and Kapanidis, A. N., "RNA polymerase pausing during initial transcription," *Molecular Cell*, vol. 63, no. 6, pp. 939 – 950, 2016.

Ebright, R. H., "Transcription activation at Class I CAP-dependent promoters," *Molecular Microbiology*, vol. 8, no. 5, pp. 797 – 802, 1993.

Errington, J., "Regulation of endospore formation in *Bacillus subtilis*," *Nature Reviews Microbiology*, vol. 1, no. 2, pp. 117 – 126, 2003.

Feklístov, A., Sharon, B. D., Darst, S. A., and Gross, C. A., "Bacterial sigma factors: A historical, structural, and genomic perspective," *Annual Review of Microbiology*, vol. 68, pp. 357 – 376, 2014.

Finney, A. H., Blick, R. J., Murakami, K., Ishihama, A., and Stevens, A. M., "Role of the C-terminal domain of the alpha subunit of RNA polymerase in LuxR-dependent transcriptional activation of the *lux* operon during quorum sensing," *Journal of Bacteriology*, vol. 184, no. 16, pp. 4520 – 4528, 2002.

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*, vol. 13, no. 2, pp. 161 – 167, 2003.

Garcia, H. G. and Phillips, R., "Quantitative dissection of the simple repression input – output function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, pp. 12 173 – 12 178, 2011.

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*, vol. 2, no. 1, pp. 150 – 161, 2012.

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*, vol. 104, no. 9, pp. 1876 – 1889, 2000.

Gil, R. and Latorre, A., "Factors behind junk DNA in bacteria," *Genes*, vol. 3, no. 4, pp. 634 – 650, 2012.

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

Gillespie, D. T., "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340 – 2361, 1977.

Gillespie, D. T., "A rigorous derivation of the chemical master equation," *Physica A: Statistical Mechanics and its Applications*, vol. 188, no. 1-3, pp. 404 – 425, 1992.

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

Goldberg, A. L., "Protein degradation and protection against misfolded or damaged proteins," *Nature*, vol. 426, no. 6968, pp. 895 – 899, 2003.

Golding, I., "Decision making in living cells: Lessons from a simple system," *Annual Review of Biophysics*, vol. 40, pp. 63–80, 2011.

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*, vol. 101, no. 31, pp. 11 310–11 315, 2004.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C., "Real-time kinetics of gene activity in individual bacteria," *Cell*, vol. 123, no. 6, pp. 1025–1036, 2005.

Gourse, R. L., Ross, W., and Gaal, T., "UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition," *Molecular Microbiology*, vol. 37, no. 4, pp. 687–695, 2000.

Goutsias, J., "Classical versus stochastic kinetics modeling of biochemical reaction systems," *Biophysical Journal*, vol. 92, no. 7, pp. 2350–2365, 2007.

Grigorova, I. L., Phleger, N. J., Mutalik, V. K., and Gross, C. A., "Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 14, pp. 5332–5337, 2006.

Grohmann, D. and Werner, F., "Recent advances in the understanding of archaeal transcription," *Current Opinion in Microbiology*, vol. 14, no. 3, pp. 328–334, 2011.

Gross, C., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J., and Young, B., "The functional and regulatory roles of sigma factors in transcription," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 63, pp. 141–156, 1998.

Gunnelius, L., Hakkila, K., Kurkela, J., Wada, H., Tyystjärvi, E., and Tyystjärvi, T., "The omega subunit of the RNA polymerase core directs transcription efficiency in cyanobacteria," *Nucleic Acids Research*, vol. 42, no. 7, pp. 4606–4614, 2014.

Gupta, A., Lloyd-Price, J., Oliveira, S. M. D., Yli-Harja, O., Muthukrishnan, A.-B., and Ribeiro, A. S., "Robustness of the division symmetry in *Escherichia coli* and functional consequences of symmetry breaking," *Physical Biology*, vol. 11, no. 6, p. 066005, 2014.

Häkkinen, A. and Ribeiro, A. S., "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data," *Bioinformatics*, vol. 31, no. 1, pp. 69–75, 2015.

Häkkinen, A. and Ribeiro, A. S., "Characterizing rate limiting steps in transcription from RNA production times in live cells," *Bioinformatics*, vol. 32, no. 9, pp. 1346–1352, 2016.

Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., and Ribeiro, A. S., "CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*," *Bioinformatics*, vol. 29, no. 13, pp. 1708–1709, 2013.

Häkkinen, A., Kandhavelu, M., Garasto, S., and Ribeiro, A. S., "Estimation of fluorescence-tagged rna numbers from spot intensities," *Bioinformatics*, vol. 30, no. 8, pp. 1146–1153, 2014.

Haugen, S. P., Ross, W., and Gourse, R. L., "Advances in bacterial promoter recognition and its control by factors that do not bind DNA," *Nature Reviews Microbiology*, vol. 6, no. 7, pp. 507–519, 2008.

Hayashi-Takanaka, Y., Stasevich, T. J., Kurumizaka, H., Nozaki, N., and Kimura, H., "Evaluation of chemical fluorescent dyes as a protein conjugation partner for live cell imaging," *PLoS One*, vol. 9, no. 9, p. e106271, 2014.

Healy, T. M. and Schulte, P. M., "Phenotypic plasticity and divergence in gene expression," *Molecular Ecology*, vol. 24, no. 13, pp. 3220 – 3222, 2015.

Heltzel, A., Lee, I. W., Totis, P. A., and Summers, A. O., "Activator-dependent preinduction binding of σ-70 RNA polymerase at the metal-regulated *mer* promoter," *Biochemistry*, vol. 29, no. 41, pp. 9572 – 9584, 1990.

Henderson, K. L., Felth, L. C., Molzahn, C. M., Shkel, I., Wang, S., Chhabra, M., Ruff, E. F., Bieter, L., Kraft, J. E., and Record, Jr., M. T., "Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 15, pp. E3032 – E3040, 2017.

Hendrickson, W. and Schleif, R. F., "Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay," *Journal of Molecular Biology*, vol. 178, no. 3, pp. 611 – 628, 1984.

Hochschild, A., "Gene-specific regulation by a transcript cleavage factor: Facilitating promoter escape," *Journal of Bacteriology*, vol. 189, no. 24, pp. 8769 – 8771, 2007.

Hochschild, A. and Dove, S. L., "Protein – protein contacts that activate and repress prokaryotic transcription," *Cell*, vol. 92, no. 5, pp. 597 – 600, 1998.

Hsu, L. M., "Promoter escape by *Escherichia coli* RNA polymerase," *EcoSal Plus*, vol. 3, no. 1, pp. 1 – 16, 2008.

Huang, Y. and Agrawal, A. F., "Experimental evolution of gene expression and plasticity in alternative selective regimes," *PLoS Genetics*, vol. 12, no. 9, p. e1006336, 2016.

Ishihama, A., "Functional modulation of *Escherichia coli* RNA polymerase," *Annual Review of Microbiology*, vol. 54, no. 1, pp. 499 – 518, 2000.

Ishihama, A. and Nagata, K., "Viral RNA polymerases," *Critical Reviews in Biochemistry*, vol. 23, no. 1, pp. 27 – 76, 1988.

Jagger, J., "Effects of near-ultraviolet radiation on microorganisms," *Photochemistry and Photobiology*, vol. 23, no. 6, pp. 451 – 454, 1976.

Kandavalli, V. K., Tran, H., and Ribeiro, A. S., "Effects of σ factor competition are promoter initiation kinetics dependent," *Biochimica et Biophysica Acta*, vol. 1859, no. 10, pp. 1281 – 1288, 2016.

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*, vol. 314, no. 5802, pp. 1144 – 1147, 2006.

Kazeev, V., Khammash, M., Nip, M., and Schwab, C., "Direct solution of the chemical master equation using quantized tensor trains," *PLoS Computational Biology*, vol. 10, no. 3, p. e1003359, 2014.

Kepler, T. B. and Elston, T. C., "Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations," *Biophysical Journal*, vol. 81, no. 6, pp. 3116–3136, 2001.

Klumpp, S. and Hwa, T., "Growth-rate-dependent partitioning of RNA polymerases in bacteria," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 51, pp. 20 245–20 250, 2008.

Klumpp, S. and Hwa, T., "Bacterial growth: global effects on gene expression, growth feedback and proteome partition," *Current Opinion in Biotechnology*, vol. 28, pp. 96–102, 2014.

Kramer, G. F. and Ames, B. N., "Oxidative mechanisms of toxicity of low-intensity near-UV light in *Salmonella typhimurium*," *Journal of Bacteriology*, vol. 169, no. 5, pp. 2259–2266, 1987.

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

Krystek, M. and Anton, M., "A weighted total least-squares algorithm for fitting a straight line," *Measurement Science and Technology*, vol. 19, no. 11, p. 079801, 2008.

Kubori, T. and Shimamoto, N., "A branched pathway in the early stage of transcription by *Escherichia coli* RNA polymerase," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 449–457, 1996.

Kumar, P. and Libchaber, A., "Pressure and temperature dependence of growth and morphology of *Escherichia coli*: Experiments and stochastic model," *Biophysical Journal*, vol. 105, no. 3, pp. 783–793, 2013.

Kussell, E. and Leibler, S., "Phenotypic diversity, population growth, and information in fluctuating environments," *Science*, vol. 309, no. 5743, pp. 2075–2078, 2005.

Lane, W. J. and Darst, S. A., "Molecular evolution of multisubunit RNA polymerases: Structural analysis," *Journal of Molecular Biology*, vol. 395, no. 4, pp. 686–704, 2010.

Lee, C. H. and Kim, P., "An analytical approach to solutions of master equations for stochastic nonlinear reactions," *Journal of Mathematical Chemistry*, vol. 50, no. 6, pp. 1550–1569, 2012.

Lee, D. J., Minchin, S. D., and Busby, S. J. W., "Activating transcription in bacteria," *Annual Review of Microbiology*, vol. 66, no. 1, pp. 125–152, 2012.

Lee, S. K., Chou, H. H., Pfleger, B. F., Newman, J. D., Yoshikuni, Y., and Keasling, J. D., "Directed evolution of AraC for improved compatibility of arabinose- and lactose-inducible promoters," *Applied and Environmental Microbiology*, vol. 73, no. 18, pp. 5711–5715, 2007.

Lemon, B. and Tjian, R., "Orchestrated response: a symphony of transcription factors for gene control," *Genes & Development*, vol. 14, no. 20, pp. 2551–2569, 2000.

Lerner, E., Chung, S., Allen, B. L., Wang, S., Lee, J., Lu, S. W., Grimaud, L. W., Ingargiola, A., Michalet, X., Alhadid, Y., Borukhov, S., Strick, T. R., Taatjes, D. J., and Weiss, S., "Backtracked and paused transcription initiation intermediate of *Escherichia coli* RNA polymerase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 43, pp. E6562–E6571, 2016.

Lesnik, T., Solomovici, J., Deana, A., Ehrlich, R., and Reiss, C., "Ribosome traffic in *E. coli* and regulation of gene expression," *Journal of Theoretical Biology*, vol. 202, no. 2, pp. 175–185, 2000.

Li, J. J., Bickel, P. J., and Biggin, M. D., "System wide analyses have underestimated protein abundances and the importance of transcription in mammals," *PeerJ*, vol. 2, p. e270, 2014.

Lind, K., Ståhlberg, A., Zoric, N., and Kubista, M., "Combining sequence-specific probes and DNA binding dyes in real-time PCR for specific nucleic acid quantification and melting curve analysis," *BioTechniques*, vol. 40, no. 3, pp. 315–319, 2006.

Lineweaver, H. and Burk, D., "The determination of enzyme dissociation constants," *Journal of the American Chemical Society*, vol. 56, no. 3, pp. 658–666, 1934.

Liu, Y., Beyer, A., and Aebersold, R., "On the dependency of cellular protein levels on mRNA abundance," *Cell*, vol. 165, no. 3, pp. 535–550, 2016.

Livak, K. J. and Schmittgen, T. D., "Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

Lodish, H. F., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A., and Martin, K. C., *Molecular Cell Biology*, 8th ed. W. H. Freeman, 2016.

Lutz, R. and Bujard, H., "Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I_1-I_2 regulatory elements," *Nucleic Acids Research*, vol. 25, no. 6, pp. 1203–1210, 1997.

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*, vol. 29, no. 18, pp. 3873–3881, 2001.

Maeda, H., Fujita, N., and Ishihama, A., "Competition among seven *Escherichia coli* σ subunits: relative binding affinities to the core RNA polymerase," *Nucleic Acids Research*, vol. 28, no. 18, pp. 3497–3503, 2000.

Mahmood, T. and Yang, P.-C., "Western blot: Technique, theory, and trouble shooting," *North American Journal of Medical Sciences*, vol. 4, no. 9, pp. 429–434, 2012.

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

Mäkelä, J., Kandavalli, V., and Ribeiro, A. S., "Rate-limiting steps in transcription dictate sensitivity to variability in cellular components," *Scientific Reports*, vol. 7, no. 1, p. 10588, 2017.

Martínez-Antonio, A. and Collado-Vides, J., "Identifying global regulators in transcriptional regulatory networks in bacteria," *Current Opinion in Microbiology*, vol. 6, no. 5, pp. 482–489, 2003.

Mathew, R. and Chatterji, D., "The evolving story of the omega subunit of bacterial RNA polymerase," *Trends in Microbiology*, vol. 14, no. 10, pp. 450–455, 2006.

Mauri, M. and Klumpp, S., "A model for sigma factor competition in bacterial cells," *PLoS Computational Biology*, vol. 10, no. 10, p. e1003845, 2014.

McAdams, H. H. and Arkin, A., "Stochastic mechanisms in gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 3, pp. 814–819, 1997.

McAdams, H. H. and Arkin, A., "It's a noisy business! Genetic regulation at the nanomolar scale," *Trends in Genetics*, vol. 15, no. 2, pp. 65–69, 1999.

McClure, W. R., "Rate-limiting steps in RNA chain initiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 10, pp. 5634–5638, 1980.

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

McClure, W. R. and Cech, C. L., "On the mechanism of rifampicin inhibition on RNA synthesis," *Journal of Biological Chemistry*, vol. 253, no. 24, pp. 8949–8956, 1978.

McClure, W. R., Cech, C. L., and Johnston, D. E., "A steady state assay for the RNA polymerase initiation reaction," *Journal of Biological Chemistry*, vol. 253, no. 24, pp. 8941–8948, 1978.

McLeod, S. M. and Johnson, R. C., "Control of transcription by nucleoid proteins," *Current Opinion in Microbiology*, vol. 4, no. 2, pp. 152–159, 2001.

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*, vol. 95, no. 4, pp. 2103–2115, 2008.

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., Kapanidis, A. N., Niu, W., Ebright, Y. W., Levy, R., and Ebright, R. H., "Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex," *Cell*, vol. 108, no. 5, pp. 599–614, 2002.

Mileyko, Y., Joh, R. I., and Weitz, J. S., "Small-scale copy number variation and large-scale changes in gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 43, pp. 16 659–16 664, 2008.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

Minakhin, L., Bhagat, S., Brunning, A., Campbell, E. A., Darst, S. A., Ebright, R. H., and Severinov, K., "Bacterial RNA polymerase subunit ω and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 892–897, 2001.

Minsky, M., "Memoir on inventing the confocal scanning microscope," *Scanning*, vol. 10, pp. 128 – 138, 1988.

Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A., and Nudler, E., "Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria," *Cell*, vol. 111, no. 5, pp. 747 – 756, 2002.

Mitchell, A., Romano, G. H., Groisman, B., Yona, A., Dekel, E., Kupiec, M., Dahan, O., and Pilpel, Y., "Adaptive prediction of environmental changes by microorganisms," *Nature*, vol. 460, no. 7252, pp. 220 – 224, 2009.

Monsalve, M., Mencía, M., Salas, M., and Rojo, F., "Protein p4 represses phage ϕ29 A2c promoter by interacting with the α subunit of *Bacillus subtilis* RNA polymerase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 17, pp. 8913 – 8918, 1996.

Mooney, R. A., Darst, S. A., and Landick, R., "Sigma and RNA polymerase: An on-again, off-again relationship?" *Molecular Cell*, vol. 20, no. 3, pp. 335 – 345, 2005.

Morelli, L. G. and Jülicher, F., "Precision of genetic oscillators and clocks," *Physical Review Letters*, vol. 98, no. 22, p. 228101, 2007.

Morise, H., Shimomura, O., Johnson, F. H., and Winant, J., "Intermolecular energy transfer in the bioluminescent system of *Aequorea*," *Biochemistry*, vol. 13, no. 12, pp. 2656 – 2662, 1974.

Müller, J., Oehler, S., and Müller-Hill, B., "Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator," *Journal of Molecular Biology*, vol. 257, no. 1, pp. 21 – 29, 1996.

Munsky, B. and Khammash, M., "The finite state projection algorithm for the solution of the chemical master equation," *The Journal of Chemical Physics*, vol. 124, no. 4, p. 044104, 2006.

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

Murakami, K. S., "Structural biology of bacterial RNA polymerase," *Biomolecules*, vol. 5, no. 2, pp. 848 – 864, 2015.

Murakami, K. S. and Darst, S. A., "Bacterial RNA polymerases: the wholo story," *Current Opinion in Structural Biology*, vol. 13, no. 1, pp. 31 – 39, 2003.

Murakami, K. S., Masuda, S., and Darst, S. A., "Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution," *Science*, vol. 296, no. 5571, pp. 1280 – 1284, 2002.

Muthukrishnan, A.-B., Martikainen, A., Neeli-Venkata, R., and Ribeiro, A. S., "*In Vivo* transcription kinetics of a synthetic gene uninvolved in stress-response pathways in stressed *Escherichia coli* cells," *PLoS One*, vol. 9, no. 9, p. e109005, 2014.

Naryshkin, N., Revyakin, A., Kim, Y., Mekler, V., and Ebright, R. H., "Structural organization of the RNA polymerase-promoter open complex," *Cell*, vol. 101, no. 6, pp. 601 – 611, 2000.

Nielsen, O. and Løbner-Olesen, A., "Once in a lifetime: strategies for preventing re-replication in prokaryotic and eukaryotic cells," *EMBO Reports*, vol. 9, no. 2, pp. 151–156, 2008.

Oehler, S., Eismann, E. R., Krämer, H., and Müller-Hill, B., "The three operators of the *lac* operon cooperate in repression," *The EMBO Journal*, vol. 9, no. 4, pp. 973–979, 1990.

Oliveira, S. M. D., Häkkinen, A., Lloyd-Price, J., Tran, H., Kandavalli, V., and Ribeiro, A. S., "Temperature-dependent model of multi-step transcription initiation in *Escherichia coli* based on live single-cell measurements," *PLoS Computational Biology*, vol. 12, no. 10, p. e1005174, 2016.

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*, vol. 99, no. 4, pp. 686–699, 2016.

Osbourn, A. E. and Field, B., "Operons," *Cellular and Molecular Life Sciences*, vol. 66, no. 23, pp. 3755–3775, 2009.

Panovska-Griffiths, J., Page, K. M., and Briscoe, J., "A gene regulatory motif that generates oscillatory or multiway switch outputs," *Journal of The Royal Society Interface*, vol. 10, no. 79, p. 20120826, 2013.

Peabody, D. S., "The RNA binding site of bacteriophage MS2 coat protein," *The EMBO Journal*, vol. 12, no. 2, pp. 595–600, 1993.

Pérez-Rueda, E. and Collado-Vides, J., "The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12," *Nucleic Acids Research*, vol. 28, no. 8, pp. 1838–1847, 2000.

Philips, S. J., Canalizo-Hernandez, M., Yildirim, I., Schatz, G. C., Mondragón, A., and O'Halloran, T. V., "Allosteric transcriptional regulation *via* changes in the overall topology of the core promoter," *Science*, vol. 349, no. 6250, pp. 877–881, 2015.

Picard, F., Milhem, H., Loubière, P., Laurent, B., Cocaign-Bousquet, M., and Girbal, L., "Bacterial translational regulations: high diversity between all mRNAs and major role in gene expression," *BMC Genomics*, vol. 13, no. 1, p. 528, 2012.

Porter, J. R., Andrews, B. W., and Iglesias, P. A., "A framework for designing and analyzing binary decision-making strategies in cellular systems," *Integrative Biology*, vol. 4, no. 3, pp. 310–317, 2012.

Ptashne, M., "Regulation of transcription: from lambda to eukaryotes," *Trends in Biochemical Sciences*, vol. 30, no. 6, pp. 275–279, 2005.

Purcell, O., Savery, N. J., Grierson, C. S., and di Bernardo, M., "A comparative analysis of synthetic genetic oscillators," *Journal of The Royal Society Interface*, vol. 7, no. 52, pp. 1503–1524, 2010.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S., "Stochastic mRNA synthesis in mammalian cells," *PLoS Biology*, vol. 4, no. 10, p. e309, 2006.

Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O., and Ribeiro, A. S., "Effects of transcriptional pausing on gene expression dynamics," *PLoS Computational Biology*, vol. 6, no. 3, p. e1000704, 2010.

Rao, C. V., Wolf, D. M., and Arkin, A. P., "Control, exploitation and tolerance of intracellular noise," *Nature*, vol. 420, no. 6912, pp. 231–237, 2002.

Raser, J. M. and O'Shea, E. K., "Noise in gene expression: origins, consequences, and control," *Science*, vol. 309, no. 5743, pp. 2010–2013, 2005.

Ray-Soni, Ananya and. Bellecourt, M. J. and Landick, R., "Mechanisms of bacterial transcription termination: All good things must end," *Annual Review of Biochemistry*, vol. 85, pp. 319–347, 2016.

Razo-Mejia, M., Barnes, S. L., Belliveau, N. M., Chure, G., Einav, T., Lewis, M., and Phillips, R., "Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction," *Cell Systems*, vol. 6, no. 4, pp. 456–469.e10, 2018.

Record, Jr., T. M., Reznikoff, W. S., Craig, M. L., McQuade, K. L., and Schlax, P. J., "*Escherichia coli* RNA polymerase (Eσ70), promoters, and the kinetics of the steps of transcription initiation," in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, 2nd ed., Neidhardt, F. C., Curtiss, R., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schneider, D., and Umbarger, H. E., Eds. American Society for Microbiology, 1996, vol. 2, pp. 792–821.

Revyakin, A., Liu, C., Ebright, R. H., and Strick, T. R., "Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching," *Science*, vol. 314, no. 5802, pp. 1139–1143, 2006.

Ribeiro, A. S., "Dynamics and evolution of stochastic bistable gene networks with sensing in fluctuating environments," *Physical Review E*, vol. 78, no. 6, p. 061902, 2008.

Ribeiro, A. S., "Stochastic and delayed stochastic models of gene expression and regulation," *Mathematical Biosciences*, vol. 223, no. 1, pp. 1–11, 2010.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D., and Gold, L., "Translation initiation in *Escherichia coli*: sequences within the ribosome-binding site," *Molecular Microbiology*, vol. 6, no. 9, pp. 1219–1229, 1992.

Roeder, R. G. and Rutter, W. J., "Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms," *Nature*, vol. 224, no. 5216, pp. 234–237, 1969.

Rojo, F., "Repression of transcription initiation in bacteria," *Journal of Bacteriology*, vol. 181, no. 10, pp. 2987–2991, 1999.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., and Gourse, R. L., "A third recognition element in bacterial promoters: DNA binding by the α subunit of RNA polymerase," *Science*, vol. 262, no. 5138, pp. 1407–1413, 1993.

Ross, W., Vrentas, C. E., Sanchez-Vazquez, P., Gaal, T., and Gourse, R. L., "The magic spot: a ppGpp binding site on *E. coli* RNA polymerase responsible for regulation of transcription initiation," *Molecular Cell*, vol. 50, no. 3, pp. 420–429, 2013.

Roussel, M. R. and Zhu, R., "Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression," *Physical Biology*, vol. 3, no. 4, pp. 274–284, 2006.

Roy, S., Garges, S., and Adhya, S., "Activation and repression of transcription by differential contact: Two sides of a coin," *Journal of Biological Chemistry*, vol. 273, no. 23, pp. 14 059–14 062, 1998.

Ruff, E. F., Drennan, A. C., Capp, M. W., Poulos, M. A., Artsimovitch, I., and Record, Jr., T. M., "*E. coli* RNA polymerase determinants of open complex lifetime and structure," *Journal of Molecular Biology*, vol. 247, no. 15, pp. 2435–2450, 2015.

Ruff, E. F., Record, Jr., T. M., and Artsimovitch, I., "Initial events in bacterial transcription initiation," *Biomolecules*, vol. 5, no. 2, pp. 1035–1062, 2015.

Saecker, R. M., Record, Jr., T. M., 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*, vol. 412, no. 5, pp. 754–771, 2011.

Sanchez, A., Garcia, H. G., Jones, D., Phillips, R., and Kondev, J., "Effect of promoter architecture on the cell-to-cell variability in gene expression," *PLoS Computational Biology*, vol. 7, no. 3, p. e1001100, 2011.

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*, vol. 30, no. 19, pp. 3940–3946, 2011.

Sánchez-Romero, M. A., Cota, I., and Casadesús, J., "DNA methylation in bacteria: from the methyl group to the methylome," *Current Opinion in Microbiology*, vol. 25, pp. 9–16, 2015.

Schleif, R., "AraC protein, regulation of the L-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action," *FEMS Microbiology Reviews*, vol. 34, no. 5, pp. 779–796, 2010.

Schmittgen, T. D. and Livak, K. J., "Analyzing real-time PCR data by the comparative C_T method," *Nature Protocols*, vol. 3, no. 6, pp. 1101–1108, 2008.

Schneider, A. F. L. and Hackenberger, C. P. R., "Fluorescent labelling in living cells," *Current Opinion in Biotechnology*, vol. 48, pp. 61–68, 2017.

Selinummi, J., Ruusuvuori, P., Podolsky, I., Ozinsky, A., Gold, E., Yli-Harja, O., Aderem, A., and Shmulevich, I., "Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images," *PLoS ONE*, vol. 4, no. 10, p. e7497, 2009.

Sen, R., Nagai, H., and Shimamoto, N., "Conformational switching of *Escherichia coli* RNA polymerase-promoter binary complex is facilitated by elongation factor GreA and GreB," *Genes to Cells*, vol. 6, no. 5, pp. 389–401, 2001.

Shaner, N. C., Steinbach, P. A., and Tsien, R. Y., "A guide to choosing fluorescent proteins," *Nature Methods*, vol. 2, no. 12, pp. 905–909, 2005.

Shapiro, H. M., *Practical Flow Cytometry*, 4th ed. Wiley, 2005.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U., "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

Sheu, C.-F. and Ratcliff, R., "The application of Fourier deconvolution to reaction time data: A cautionary note." *Psychological Bulletin*, vol. 118, no. 2, pp. 285–299, 1995.

Shimomura, O., Johnson, F. H., and Saiga, Y., "Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*," *Journal of Cellular and Comparative Physiology*, vol. 59, no. 3, pp. 223–239, 1962.

Shultzaberger, R. K., Bucheimer, R. E., Rudd, K. E., and Schneider, T. D., "Anatomy of *Escherichia coli* ribosome binding sites," *Journal of Molecular Biology*, vol. 313, pp. 215–228, 2001.

Smith, P. L., "Obtaining meaningful results from Fourier deconvolution of reaction time data," *Psychological Bulletin*, vol. 108, no. 3, pp. 533–550, 1990.

Smits, W. K., Kuipers, O. P., and Veening, J.-W., "Phenotypic variation in bacteria: the role of feedback regulation," *Nature Reviews Microbiology*, vol. 4, no. 4, pp. 259–271, 2006.

Stephens, D. J. and Allan, V. J., "Light microscopy techniques for live cell imaging," *Science*, vol. 300, no. 5616, pp. 82–86, 2003.

Stricker, J., Cookson, S., Bennett, M. R., Mather, W. H., Tsimring, L. S., and Hasty, J., "A fast, robust and tunable synthetic gene oscillator," *Nature*, vol. 456, no. 7221, pp. 516–519, 2008.

Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M., and Elowitz, M. B., "An excitable gene regulatory circuit induces transient cellular differentiation," *Nature*, vol. 440, no. 7083, pp. 545–550, 2006.

Suh, W. C., Leirmo, S., and Record, Jr., M. T., "Roles of Mg2+ in the mechanism of formation and dissociation of open complexes between escherichia coli RNA polymerase and the λ P_R promoter: kinetic evidence for a second open complex requiring Mg2+," *Biochemistry*, vol. 31, no. 34, pp. 7815–7825, 1992.

Susa, M., Sen, R., and Shimamoto, N., "Generality of the branched pathway in transcription initiation by *Escherichia coli* RNA polymerase," *Journal of Biological Chemistry*, vol. 277, no. 18, pp. 15 407–15 412, 2002.

Susa, M., Kubori, T., and Shimamoto, N., "A pathway branching in transcription initiation in *Escherichia coli*," *Molecular Microbiology*, vol. 59, no. 6, pp. 1807–1817, 2006.

Swint-Kruse, L. and Matthews, K. S., "Allostery in the LacI/GalR family: Variations on a theme," *Current Opinion in Microbiology*, vol. 12, no. 2, pp. 129–137, 2009.

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*, vol. 329, no. 5991, pp. 533–538, 2010.

Thattai, M. and van Oudenaarden, A., "Intrinsic noise in gene regulatory networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 15, pp. 8614–8619, 2001.

Tokunaga, M., Imamoto, N., and Sakata-Sogawa, K., "Highly inclined thin illumination enables clear single-molecule imaging in cells," *Nature Methods*, vol. 5, no. 2, pp. 159–161, 2008.

Touhami, A., Jericho, M., and Rutenberg, A. D., "Temperature dependence of MinD oscillation in *Escherichia coli*: Running hot and fast," *Journal of Bacteriology*, vol. 188, no. 21, pp. 7661–7667, 2006.

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*, vol. 11, no. 9, pp. 2579–2587, 2015.

Travers, A. and Muskhelishvili, G., "A common topology for bacterial and eukaryotic transcription initiation?" *EMBO Reports*, vol. 8, no. 2, pp. 147–151, 2007.

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

Valentin-Hansen, P., Søgaard-Andersen, L., and Pedersen, H., "A flexible partnership: the CytR anti-activator and the cAMP–CRP activator protein, comrades in transcription control," *Molecular Microbiology*, vol. 20, no. 3, pp. 461–466, 1996.

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., and Steenackers, H. P., "RNA-binding proteins involved in post-transcriptional regulation in bacteria," *Frontiers in Microbiology*, vol. 6, p. 141, 2015.

Veening, J.-W., Smits, W. K., and Kuipers, O. P., "Bistability, epigenetics, and bet-hedging in bacteria," *Annual Review of Microbiology*, vol. 62, pp. 193–210, 2008.

Vogel, C. and Marcotte, E. M., "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 227–232, 2012.

von Hippel, P. H. and Pasman, Z., "Reaction pathways in transcript elongation," *Biophysical Chemistry*, vol. 101–102, pp. 401–423, 2002.

von Hippel, P. H., Bear, D. G., Morgan, W. D., and McSwiggen, J. A., "Protein-nucleic acid interactions in transcription: A molecular analysis," *Annual Review of Biochemistry*, vol. 53, pp. 389–446, 1984.

Ward, W. W., Cody, C. W., Hart, R. C., and Cormier, M. J., "Spectrophotometric identity of the energy-transfer chromophores in *Renilla* and *Aequorea* green-fluorescent protein," *Photochemistry and Photobiology*, vol. 31, no. 6, pp. 611–615, 1980.

Wassarman, K. M. and Storz, G., "6S RNA regulates *E. coli* RNA polymerase activity," *Cell*, vol. 101, no. 6, pp. 613–623, 2000.

Waters, J. C., "Accuracy and precision in quantitative fluorescence microscopy," *The Journal of Cell Biology*, vol. 185, no. 7, pp. 1135–1148, 2009.

Werner, F. and Weinzierl, R. O. J., "A recombinant RNA polymerase II-like enzyme capable of promoter-specific transcription," *Molecular Cell*, vol. 10, no. 3, pp. 635–646, 2002.

Wilks, S. S., "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60 – 62, 1938.

Wilson, D. N., Arenz, S., and Beckmann, R., "Translation regulation via nascent polypeptide-mediated ribosome stalling," *Current Opinion in Structural Biology*, vol. 37, pp. 123 – 133, 2016.

Wisniewski-Dyé, F. and Vial, L., "Phase and antigenic variation mediated by genome modifications," *Antonie van Leeuwenhoek*, vol. 94, no. 4, pp. 493 – 515, 2008.

Wolf, D. M. and Arkin, A. P., "Fifteen minutes of *fim*: Control of type 1 pili expression in *E. coli*," *OMICS: A Journal of Integrative Biology*, vol. 6, no. 1, pp. 91 – 114, 2002.

Wolf, D. M. and Arkin, A. P., "Motifs, modules and games in bacteria," *Current Opinion in Microbiology*, vol. 6, no. 2, pp. 125 – 134, 2003.

Wolf, V., Goel, R., Mateescu, M., and Henzinger, T. A., "Solving the chemical master equation using sliding windows," *BMC Systems Biology*, vol. 4, p. 42, 2010.

Yanofsky, C., "The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis*," *Trends in Genetics*, vol. 20, no. 8, pp. 367 – 374, 2004.

Young, B. A., Gruber, T. M., and Gross, C. A., "Views of transcription initiation," *Cell*, vol. 109, no. 4, pp. 417 – 420, 2002.

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*, vol. 311, no. 5767, pp. 1600 – 1603, 2006.

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 of the United States of America*, vol. 111, no. 4, pp. 1557 – 1561, 2014.

Zernike, F., "Phase contrast, a new method for the microscopic observation of transparent objects part II," *Physica*, vol. 9, no. 10, pp. 974 – 986, 1942.

Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K., and Darst, S. A., "Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution," *Cell*, vol. 98, no. 6, pp. 811 – 824, 1999.

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*, vol. 246, no. 4, pp. 725 – 745, 2007.

Publications

Errata for Publications

Publication III: In Methods, subsection "Estimation of intake times by deconvolution from empirical data on activation times and active transcription interval duration", the text "0.05 and 0.95" should be replaced with "0.025 and 0.975".

(This error is only in the text; the calculations were done using the correct numbers.)

Publication I

S. Startceva, J.G. Chandraseelan, A. Visa, and A.S. Ribeiro "Quantitative Estimation of Long-living Fluorescent Molecules from Temporal Fluorescence Intensity Data Corrupted by Nonzero-mean Noise", *In Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016)*

Quantitative Estimation of Long-living Fluorescent Molecules from Temporal Fluorescence Intensity Data Corrupted by Nonzero-mean Noise

Sofia Startceva, Jerome G. Chandraseelan, Ari Visa and Andre S. Ribeiro
Department of Signal Processing, Tampere University of Technology, Tampere, Finland
{sofiia.startceva, jerome.chandraseelan, ari.visa, andre.ribeiro}@tut.fi

Keywords: Fluorescence-tagged RNA Quantification, Single-molecule Time-lapse Microscopy, Biosignal Processing.

Abstract: We present a new quantitative method of estimation of fluorescent molecule numbers from time-lapse, single-cell, fluorescence microscopy data. Its main aim is to eradicate backward propagation of noise, which is present in previous methods. The method is first validated using Monte Carlo simulations. These tests show that when the time-lapse data are corrupted with negative noise, the method obtains significantly more precise results than current techniques. The applicability of the method is demonstrated on novel time-lapse, single-cell measurements of fluorescently tagged ribonucleic acid (RNA) molecules. Interestingly, we find that the intervals inferred by the new method have the same mean but reduced variability when compared to the previously existing method, which, in accordance to human observers, is a more accurate estimation.

1 INTRODUCTION

Gene expression is a complex, multi-step process (McClure, 1985; Lutz and Bujard, 1997; deHaseth et al., 1998; Yarchuk et al., 1992; Wen et al., 2008; Zhang et al., 2014). In addition, the underlying steps of this process are stochastic in nature, generating a variability in RNA and protein numbers that mostly explains the phenotypic diversity of monoclonal cell populations (McAdams and Arkin, 1997; Elowitz et al., 2002; Rao et al., 2002; Raser and O'Shea, 2005). To study this process, specialised techniques in molecular biology (Golding and Cox, 2004; Yu et al., 2006), microscopy (Rutter et al., 1998; Chowdhury et al., 2012), image analysis (Chowdhury et al. 2013; Häkkinen et al., 2013), computational biology (Zhu et al., 2007) and signal processing (Häkkinen and Ribeiro, 2014) were developed.

Methods of signal processing should consider the characteristics of the underlying processes. For example, in the RNA tracking technique based on MS2-GFP tagging, the MS2-GFP proteins (composed of the bacteriophage MS2 coat protein fused to the GFPmut3 protein (Golding et al., 2005)) bind to multiple MS2 binding sites of the target RNA soon after its production, and once formed, those RNA-MS2-GFP complexes remain in a cell for the duration of the experiment (Golding and Cox, 2004; Muthukrishnan et al., 2012). Thus, in this case, when estimating the numbers of target RNAs, any signal reduction can be classified as noise.

Since complexes can co-localize, the number of target RNAs in each cell is estimated from the total fluorescence of the complexes at a given moment (Golding and Cox, 2004; Kandhavelu et al., 2012; Häkkinen and Ribeiro, 2014). However, the signal can be disrupted (i.e. subject to nonzero-mean noise), which hampers an exact determination of fluorescent molecules' numbers. That is, though the number of RNA-MS2-GFP complexes in a cell is considered as a monotonic non-decreasing function during the experiment (Muthukrishnan et al., 2012), the total fluorescence intensity of the tagged RNA molecules can decrease, transiently or permanently, in the course of an experiment. These decreases are usually caused by the RNA complexes moving away from the focal plane, or as a result of photobleaching. While the latter corrupts the data permanently, the former are isolated events in single cell time series and usually cause a steep, transient decrease in the fluorescence intensity of tagged RNA molecules.

Here, we present a new quantitative method of estimation of fluorescent molecule numbers from single-cell fluorescent intensity data obtained by

Startceva, S., Chandraseelan, J., Visa, A. and Ribeiro, A.
Quantitative Estimation of Long-living Fluorescent Molecules from Temporal Fluorescence Intensity Data Corrupted by Nonzero-mean Noise.
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 4: BIOSIGNALS*, pages 17-24
ISBN: 978-989-758-170-0

time-lapse microscopy. The method aims to eliminate backward noise propagation, caused by molecules 'moving out of focus', which currently is one of the main sources of noise in the estimation of the numbers of fluorescent molecules from time-lapse, live cell images.

2 METHODS

The technique of RNA detection by MS2-GFP tagging allows observing individual RNA molecules in live cells, soon after they are transcribed (Golding et al., 2005). In order to extract information from the images in an automated fashion, it is necessary to detect the tagged RNA molecules, which appear as bright spots in the image. Then, the intensity of the spots is extracted and summed, so as to obtain the "total RNA intensity signal" in a cell, at a given point in time.

This RNA intensity signal from non-degradable fluorescent tagged RNA molecules contains noise accumulated through each step of signal registration (microscope settings, image registration and image processing). From observation of the data (Muthukrishnan et al., 2012; Kandhavelu et al., 2012; Häkkinen et al., 2014), we assume that the signal behaves as a monotonic non-decreasing function corrupted with three types of noise:

1. Consistent, normally distributed independent noise (probability of occurrence $p_1 = 1$), with zero mean and given standard deviation, which is introduced by imprecisions of the microscope and detector (Chowdhury et al., 2012; Waters, 2009).

2. Negative noise, which in our measurements corresponds to fluorescent molecules moving out of focus and remaining there for a certain amount of time. Probabilities $p_{2\,out}$ of going out of focus and $p_{2\,in}$ of returning to focus depends, e.g., on the type of fluorescent molecule, temperature, etc.

3. Inconsistent positive noise (low probability of occurrence, $p_3 < 0.01$), caused, for instance, by false-positive detection of fluorescent molecules. These events are independent from each other, so the probability of occurring n times is p_3^n, which is negligible for $n \geq 3$. Note that, the limit value of p_3 is set by empirical observations that these events are rare.

2.1 Previous Computational Methods

In (Häkkinen and Ribeiro, 2014), a method was proposed for estimating RNA numbers and production intervals from temporal data of tagged RNAs fluorescence intensity in individual cells. This method, here denominated as a 'reference method', has three steps. First, a monotonically increasing curve is fitted to the time series, and temporal information on related samples is extracted. Second, the intensity of a single fluorescent molecule, or a 'jump size', is estimated from the information obtained at the first step. In the third and final step, a quantized curve is fit to the time series, given the parameters, enforcing the quantization to the fit. From this, the RNA numbers are extracted.

The third step in (Häkkinen and Ribeiro, 2014) goes as follows. Given the jump size, time series are fitted quantitatively, and the fit obtained is an estimation of the number of fluorescent molecules.

For the fits performed throughout the method, one can use least squares (LSQ) or least absolute deviations (LD) fitting. The LD was found to be more robust to signal disruptions.

In order to exploit the characteristics of the empirical data, this method assumes that all fluorescent molecules have the same intensity and that, once formed, they do not degrade before the end of the measurements (experimental evidence for this assumption is provided in (Muthukrishnan et al., 2012)). The first assumption is equivalent to assuming that the jump size is a constant. The second assumption corresponds to forbidding non-monotonic behaviour of quantitative estimation of the molecules over time.

This method fits full time series to a curve in one step, which aids in eliminating a consistent zero-mean noise, but also allows a backward propagation of any inconsistent disruption of the signal. Hence, although this method fully addresses the problems of the first and the third types of noise described in the Methods section, the problem of the second type of noise is addressed only to a limited extent (a fluorescent molecule is detectable only if it is in focus for at least more than 50% of the time series length).

2.2 Experimental Methods

2.2.1 Cells, Plasmids, Chemicals and Media

For live, single cell, time-lapse measurements of the RNA production times, the MS2d-GFP tagging system was used. Fluorescent RNA-MS2d-GFP complexes were observed in *Escherichia coli* DH5α-PRO strain (generously provided by Ido Golding, University of Illinois, IL). The strain contains a single copy plasmid (coding for the RNA with 96 MS2d binding sites under the control of the

promoter P_{lac}) and a high-copy reporter plasmid coding for MS2d-GFP under the control of the promoter $P_{LtetO-1}$ (Golding and Cox, 2004).

For growth media, we used the following composition per 100 ml: 1.5 g tryptone, 0.75 g yeast extract and 1 g NaCl (pH of 7.0). Media components were purchased from LabM (UK), while antibiotics, Isopropyl b-D-1-thiogalactopyranoside (IPTG), arabinose, and anhydrotetracycline (aTc) are from Sigma-Aldrich (USA).

2.2.2 Cell Growth and Microscopy

Cells from the DH5α-PRO strain, containing the target and the reporter plasmids, were grown overnight, diluted into fresh media to an OD_{600} of 0.1 (measured with an Ultraspec 10 cell density meter), and allowed to grow to an OD_{600} of ~0.3. For the reporter plasmid induction, aTc (100 ng/ml) was added 1 h before the start of the measurements. For the target plasmid, IPTG (1mM) was added 10 min before the start of the measurements. Cells were pelleted and resuspended into fresh medium. A few µl of the cells were placed between a coverslip and an agarose gel pad (2%), containing the respective inducers, in a thermal imaging chamber (FCS2, Bioptechs), heated to 37 ºC. The cells were visualized using a Nikon Eclipse (Ti-E, Nikon, Japan) inverted microscope with a C2+ confocal laser-scanning system using a 100x Apo TIRF objective. Images were acquired using the Nikon Nis-Elements software. GFP fluorescence was measured using a 488 nm argon ion laser (Melles-Griot) and 514/30 nm emission filter. Phase contrast images were acquired with an external phase contrast system and a Nikon DS-Fi2 camera. Fluorescence images were acquired every 1 min for 2 hours. Phase-contrast images were acquired every 5 min.

2.2.3 Image Analysis

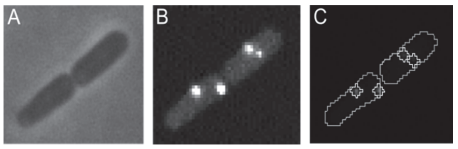


Figure 1. Panel A and B exemplify phase contrast and confocal images, correspondingly, of the same cells. Panel C shows masks of those cells and their fluorescent spots.

Cells were detected from phase contrast images as in (Gupta et al., 2014). First, the images were temporally aligned using cross-correlation. Next, an automatic segmentation of the cells was obtained with MAMLE (Chowdhury et al., 2013). The results were corrected manually. Cell lineages were constructed by CellAging (Häkkinen et al., 2013). Alignment of the phase contrast images with the confocal images was done by manually selecting 5-7 landmarks in both images, and using thin-plate spline interpolation for the registration transform. After the registration, the cell masks were adjusted to the borders of corresponding cells from the confocal images based on the fluorescent intensity. Finally, fluorescent spots and their intensities were detected from confocal images using a Gaussian surface-fitting algorithm from (Häkkinen et al., 2014). Examples of original images and obtained masks are shown in Figure 1.

3 RESULTS

3.1 Algorithm

Our algorithm for the quantitative estimation of fluorescent molecules from the data is described in Figure 2.

3.1.1 Initial Parameters

To obtain the intensity of one fluorescent molecule, μ, we combine the first two steps of the 'reference method' in their original form with visual inspection of the time series of fluorescence intensity. Other methodologies could be used instead.

To account for positive noise (type 3 noise), the 'trusted interval', w, is introduced. If an increase in intensity persists for w frames, then we assume that this increase is not due to noise. Otherwise, the assumption that it is positive noise cannot be rejected.

The choice of the value of w is based on the standard deviation σ of a consistent noise (type 1). The optimum value of w rises with the increase of σ (Figure 4). Also, we found by inspection that, to be resistant to the type 3 noise, w should not be smaller than 5 data points.

The parameter v is introduced to account for deviations in the mean of type 1 noise. The exploration of the parameter space of the fit (Figure 4) shows that, for a signal without a consistent non-zero mean noise, $v \approx 0.25$ is an optimal value. However, the optimal v increases up to 0.4 in the case of fitting a signal with $\sigma = 2$.

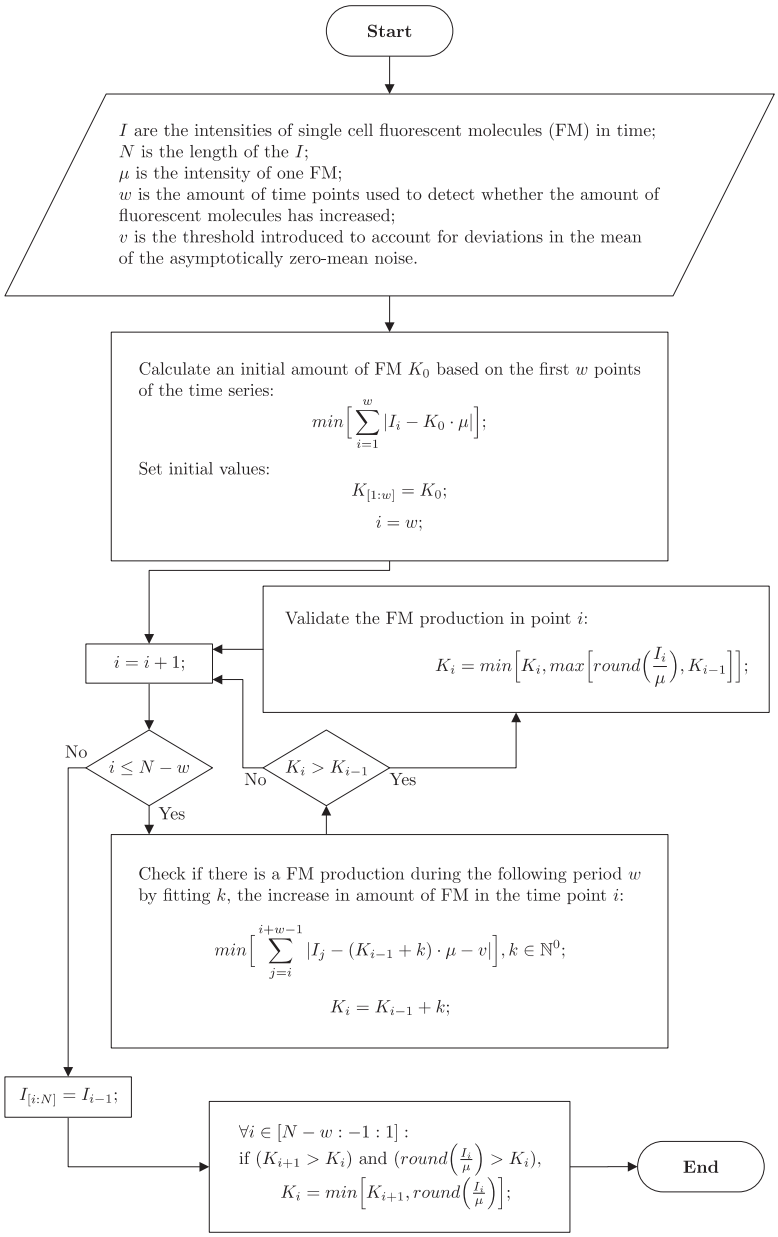


Figure 2: Algorithm used for the quantization of fluorescent molecules.

3.1.2 Computational Procedure

The procedure of the algorithm can be represented as a set of interval-fitting events. Each interval l has length w, the values of each fit at each time point are a constant proportional to μ, the fit is performed using least absolute deviations and, the coefficient of proportion K of the best fit is an initial estimate

ofthe amount of fluorescent molecules. Given this, first, we estimate the amount of fluorescent molecules in the first w time points. For each following data point I_i, where $w < i \leq N - w$, the fit is performed. If $K_i > K_{i-1}$, then the estimated amount of fluorescent molecules at time point t_i is the maximum value of the estimated amount $round(I_i/\mu)$ at t_i, and the estimated amount K_{i-1} at t_{i-1}.

Since it is not possible to determine whether any increase in the signal at the time points $[N - w + 1{:}N]$ is caused by noise or by the production of fluorescent molecules, no estimation is performed on this interval.

Finally, the obtained time series of estimated amounts of fluorescent molecules K are checked at each time point i (from N to 1). If $(K_{i+1} > K_i)$ and $(round(I_i/\mu) > K_i)$ are true, K_i is set to $min(K_{i+1}, round(I_i/\mu))$. We note that the production events at these moments were not detected during the fitting procedure because of the local disruptions of the signal in subsequent moments.

3.2 Analysis of *in Silico* Data

Monte Carlo simulations were performed using a model of transcription that assumes that RNA molecules are produced in exponentially distributed intervals (with mean interval of 15 min (Muthukrishnan et al., 2012)). The sampling frequency f used is 10 sec^{-1} and 1 min^{-1}, for comparison.

The obtained time series are then corrupted by adding zero-mean independent and normally distributed noise. To introduce significant, transient disruptions of the signal (i.e. to model RNA-MS2d-GFP complexes going out of focus), we set the RNA signal to zero at random moments, for a randomly selected duration. For that, we set the probability that an RNA goes out of focus to $p_{2\,out} = 60$ min^{-1} and the probability of the zeroed RNA to be fully recovered to $p_{2\,in} = 20$ min^{-1}.

In Figure 3 we exemplify the outcome of simulating the model for 120 min.

We use this model's ground truth data to test the accuracy of the RNA numbers estimation by our method. To quantify the accuracy, we define it to be the proportion of time moments where the RNA numbers in a cell were correctly detected (Häkkinen and Ribeiro 2014).

First, the parameter space of the proposed model was investigated in order to detect a combination of values of w and v that maximize the accuracy.

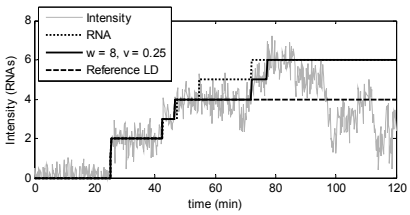


Figure 3: Simulated data. $f = 10$sec^{-1}. $\sigma = 0.5$. $p_{2\,out} = 60$min^{-1} and $p_{2\,in} = 20$min^{-1}, $w = 8$, $v = 0.25$.

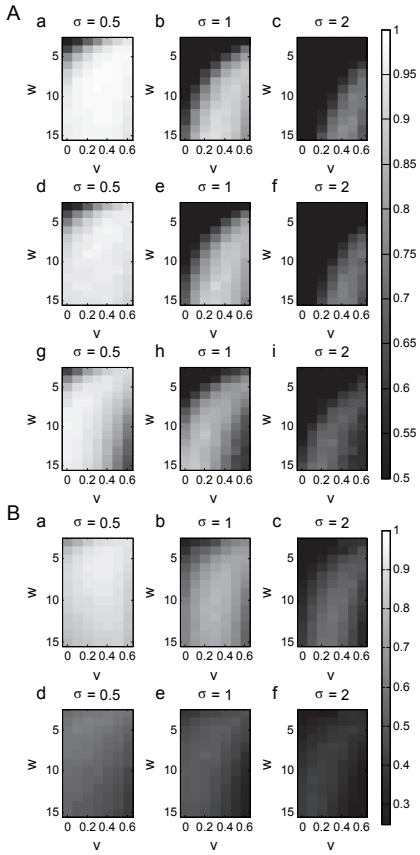


Figure 4: Mean accuracy along the parameter space of $w$and $v$for $\sigma = 0.5$, $\sigma = 1$, and $\sigma = 2$. In panel A, $f = 10$sec^{-1} and in panel B, $f = 1$min^{-1}. In both panels, from a-c: $p_{2\,out} = 0$ min^{-1} and $p_{2\,in} = 0$min^{-1}; from d-f: $p_{2\,out} = 60$min^{-1} and $p_{2\,in} = 20$min^{-1}; from g-i: 25% time series points were randomly selected and set to zero. In all sub-panels of panel A and in sub-panels a-c of panel B, each accuracy value is a mean of 1000 simulations. In sub panels d-f of panel B, each accuracy value is a mean of 2500 simulations.

For that, we performed a set of at least 1000 simulations for each combination of values of v, in the range $[0, 0.6]$, and w, in the range $[3, 15]$ for $\sigma = 0.5, 1, 2$ for each of the following sets of parameter values: a) $p_{2\,out} = 0$ min^{-1}, $p_{2\,in} = 0$ min^{-1} ($f = 10$ sec^{-1} and $f = 1$ min^{-1}); b) $p_{2\,out} = 60$ min^{-1}, $p_{2\,in} = 20$ min^{-1} ($f = 10$ sec^{-1} and $f = 1$ min^{-1}); and c) 25% time series points randomly selected and set to zero ($f = 10$ sec^{-1}). Results are shown in Figure 4.

From Figure 4, w depends on the variation of σ of the consistent noise (namely, as it increases monotonically with increasing σ), whereas v depends on the mean consistent noise (which becomes negative due to zeroing 25% of the time moments). Also, the optimal trusted interval w suffered only a slight reduction with a sixfold decrease of the sampling frequency, f.

In addition, we found that for $\sigma = 0,\ 0.5,\ 1,\ 1.5,\ 2$ and $f = 10$ sec^{-1}, we obtain $w_{opt} = 5, 7, 13, 13, 13$, respectively. Meanwhile, for $f = 1$ min^{-1}, we obtain $w_{opt} = 5, 5, 10, 10, 12$, respectively. Finally, we found that the optimal $v \approx 0.25$.

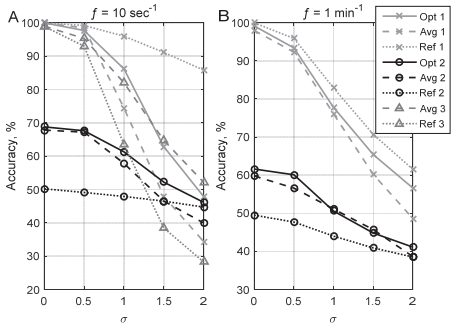


Figure 5: Mean accuracy of the counting of fluorescent molecules using a given method (Opt, Avg, or Ref) with a given noise model (1, 2, or 3) from STD $\sigma$of zero-mean noise. Panel A: $f = 10$sec^{-1}; panel B: $f = 1$min^{-1}. Opt is the proposed method with $w = w_{opt}$; Avg is the proposed method with $w = 10$(panel A) and $w = 8$(panel B); Ref is the reference method. In case 1, $p_{2\,out} = 0$min^{-1} and $p_{2\,in} = 0$min^{-1}. In case 2, $p_{2\,out} = 60$min^{-1} and $p_{2\,in} = 20$min^{-1}. In case 3, 25% of the data points are randomly selected and set to zero. Each accuracy value is a mean of 10000 simulations (using $v = 0.2$ in case 3 and $v = 0.25$ otherwise).

Next, we analysed the simulated data with and without going-out-of-focus events using the proposed method and the LD version of the reference method, and compared their accuracies.

In particular, we measured the accuracy of our method for $\sigma = 0,\ 0.5,\ 1,\ 1.5,\ 2$, along with an optimal w ('Opt' method) as well as with a mean w ('Avg' method), in order to study the impact of this parameter as a function of σ. An estimated optimal v was chosen separately for data with zero-mean noise and for data with negative-mean noise. Also, we measured the accuracy of the reference method ('Ref') on the same data, for comparison.

From Figure 5, in general, the proposed method has higher precision when analysing data with out-of-focus events (i.e. is more robust to type 2 noise). For $\sigma = 0.5$, its accuracy is improved from 49.1% to 67.6% for $f = 10$ sec^{-1}, and from 47.7% to 60.1% for $f = 1$ min^{-1}. However, our method is less robust to type 1 noise, which is expected because the data is processed piecewise.

Also from Figure 5, note how the precision is lowered for mean w versus optimal w. This difference in precision increases with increasing σ.

Finally, we made use of the *in silico* data to assess the timing of the proposed algorithm. For this, we measured the time required to analyse 10000 simulated time series with $f = 1$ min^{-1}, $\sigma = 1$, $p_{2\,out} = 60$ min^{-1}, $p_{2\,in} = 20$ min^{-1}, and length of 120 min. For w=4, 8, 16 the duration was 16 s, 12 s, and 10 s respectively (processor Intel Core i5-2400, 3.10GHz), while v does not have a noticeable impact on the time length of this process.

3.3 Analysis of Empirical Data

We next applied our method to empirical data, obtained as described in the methods section. This data was processed using our method and the reference method, for comparison (Table 1). The fluorescent RNA complexes have a non-negligible tendency to go out of focus, which makes it possible to demonstrate the usefulness of the proposed method.

Table 1: Comparative analysis of the mean and variability of the intervals between consecutive RNA production events obtained by our method ($w = 8$, $v = 0.25$) and the reference method. The data was collected from 178 cells.

Method	No. intervals	Mean interval	Interval CV^2
Our method	158	1047	1.15
Ref. method	153	1018	1.43

From the Table 1, the two methods differ in performance. Namely, while the two methods infer similar mean intervals between transcription events (the new method detected 3% more intervals), the

CV^2 of those intervals duration is significantly smaller when using the new method (19.6% smaller). Inspection of the data by two expert human observers indicated that the new method's detection process was the more accurate one (see example Figure 6).

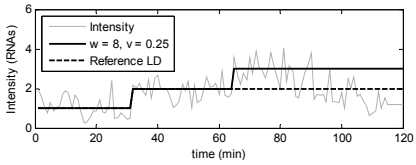


Figure 6: Example intensity series and estimated RNA numbers with the proposed method ($w = 8,\ v = 0.25$), and with the reference method (LD version).

4 CONCLUSIONS

Here we proposed a new method for the quantitative estimation of fluorescent molecules from temporal intensity microscopy data. This method was developed to handle transient, nonzero-mean noise in the data, i.e. it aims to cope with temporary absences of fluorescent molecules from the focal plane in time-lapse microscopy measurements. This is particularly important in studies requiring a consistent tracking of tagged molecules, such as studies of, e.g., chemotaxis mechanisms which rely on chemoreceptor clusters (Sourjik and Berg, 2004; Wadhams and Armitage, 2004; Parkinson et al., 2005; Kentner and Sourjik, 2006) and protein aggregates' accumulation, which is associated with cellular aging processes (Maisonneuve et al., 2008; Tyedmers et al., 2010; Winkler et al., 2010; Lindner et al., 2008; Gupta et al., 2014; Lloyd-Price et al., 2012).

We validated our method by tests on *in silico* data. Next, we applied it to empirical data to show that its results can differ from those of the previous method. By inspection, we found, as expected, that the reason why the results of the two methods differ is the enhanced robustness of our method to 'negative', inconsistent noise. Another reason is its weaker robustness to consistent, type 1 noise.

The causes of the two main differences are that, in the new method: i) previous values of a tagged RNA intensity confine the next ones into boundaries defined by the known properties of the signal. The main benefit of this is that it restricts backward propagation of inconsistent noise, which results in more precise results when $p_{2\,out} > 0$; ii) the stepwise analysis of the signal hampers the removal of consistent zero-mean noise.

We expect our method to be of use to a broad range of time-lapse microscopy measurements making use of fluorescence molecules in live cells, particular when the phenomenon of moving out of the focus plane is common for those molecules.

ACKNOWLEDGEMENTS

Work supported by TUT's Graduate School (SS) and Academy of Finland (257603, ASR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Chowdhury, S. et al., 2012. An interacting multiple model filter-based autofocus strategy for confocal time-lapse microscopy. *J.Microscopy*, 245, pp.265–75.

Chowdhury, S. et al., 2013. Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE). *BMC Bioinformatics*, 14 (Suppl 10), p.S8.

deHaseth, P.L., Zupancic, M.L. and Record, M.T., 1998. RNA polymerase-promoter interactions: The comings and goings of RNA polymerase. *J Bacteriology*, 180(12), pp.3019–25.

Elowitz, M.B. et al., 2002. Stochastic gene expression in a single cell. *Science*, 297(5584), pp.1183–6.

Golding, I. et al., 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6), pp.1025–36.

Golding, I. and Cox, E.C., 2004. RNA dynamics in live *Escherichia coli* cells. *Proc. of the National Academy of Sciences of the USA*, 101(31), pp.11310–5.

Gupta, A., Lloyd-Price, J., Neeli-Venkata, R., et al., 2014. In Vivo Kinetics of Segregation and Polar Retention of MS2-GFP-RNA Complexes in *Escherichia coli*. *Biophysical J.*, 106(9), pp.1928–37.

Gupta, A., Lloyd-Price, J., Oliveira, S.M.D., et al., 2014. Robustness of the division symmetry in *Escherichia coli* and functional consequences of symmetry breaking. *Physical Biology*, 11(6), p.066005.

Häkkinen, A. and Ribeiro, A.S., 2014. Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics*, 31(1), pp.69–75.

Häkkinen, A. et al., 2013. CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, 29(13), pp.1708–9.

Häkkinen, A. et al., 2014. Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics*, 30(8), pp.1146–53.

Kandhavelu, M. et al., 2012. Single-molecule dynamics of transcription of the *lar* promoter. *Physical Biology*, 9(2), p.026004.

Kentner, D. and Sourjik, V., 2006. Spatial organization of the bacterial chemotaxis system. *Current Opinion in Microbiology*, 9(6), pp.619–24.

Lindner, A.B. et al., 2008. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proc. of the National Academy of Sciences of the USA*, 105(8), pp.3076–81.

Lloyd-Price, J. et al., 2012. Asymmetric disposal of individual protein aggregates in *Escherichia coli*, one aggregate at a time. *J.Bacteriology*, 194(7), pp.1747–52.

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 Research*, 25(6), pp.1203–10.

Maisonneuve, E., Ezraty, B. and Dukan, S., 2008. Protein aggregates: An aging factor involved in cell death. *J.Bacteriology*, 190(18), pp.6070–5.

McAdams, H.H. and Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proc. of the National Academy of Sciences of the USA*, 94(3), pp.814–9.

McClure, W.R., 1985. Mechanism and control of transcription initiation in prokaryotes. *Ann. Rev. of Biochemistry*, 54, pp.171–204.

Muthukrishnan, A.-B. et al., 2012. Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucleic Acids Research*, 40(17), pp.8472–83.

Parkinson, J.S., Ames, P. and Studdert, C.A., 2005. Collaborative signaling by bacterial chemoreceptors. *Current opinion in microbiology*, 8(2), pp.116–21.

Rao, C. V, Wolf, D.M. and Arkin, A.P., 2002. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912), pp.231–7.

Raser, J.M. and O'Shea, E.K., 2005. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743), pp.2010–3.

Rutter, G.A. et al., 1998. Real-time imaging of gene expression in single living cells. *Chemistry and biology*, 5(11), pp.R285–90.

Sourjik, V. and Berg, H.C., 2004. Functional interactions between receptors in bacterial chemotaxis. *Nature*, 428(March), pp.1–4.

Tyedmers, J., Mogk, A. and Bukau, B., 2010. Cellular strategies for controlling protein aggregation. *Nature rev. Mol. cell biology*, 11(11), pp.777–88.

Wadhams, G.H. and Armitage, J.P., 2004. Making sense of it all: bacterial chemotaxis. *Nature rev. Mol. Cell Biology.*, 5(12), pp.1024–37.

Waters, J.C., 2009. Accuracy and precision in quantitative fluorescence microscopy. *J.Cell Biology*, 185(7), pp.1135–48.

Wen, J.-D. et al., 2008. Following translation by single ribosomes one codon at a time. *Nature*, 452(7187), pp.598–603.

Winkler, J. et al., 2010. Quantitative and spatio-temporal features of protein aggregation in *Escherichia coli* and consequences on protein quality control and cellular ageing. *The EMBO J.*, 29(5), pp.910–23.

Yarchuk, O., Guillerez, J. and Dreyfus, M., 1992. Interdependence of Translation, Transcription Degradation in the ZacZ Gene and mRNA. *J.Molecular Biology*, pp.581–96.

Yu, J. et al., 2006. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767), pp.1600–03.

Zhang, Z. et al., 2014. Single-molecule tracking of the transcription cycle by sub-second RNA detection. *eLife*, 2014(3), pp.1–20.

Zhu, R. et al., 2007. Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *J. Theoretical Biology*, 246(4), pp.725–45.

Publication II

J. Lloyd-Price, S. Startceva, V. Kandavalli, J.G. Chandraseelan, N. Goncalves, S.M.D. Oliveira, A. Häkkinen, and A.S. Ribeiro "Dissecting the stochastic transcription initiation process in live *Escherichia coli*", *DNA Research*

DNA Research, 2016, 23(3), 203–214
doi: 10.1093/dnares/dsw009
Advance Access Publication Date: 28 March 2016
Full Paper

OXFORD

Full Paper

Dissecting the stochastic transcription initiation process in live *Escherichia coli*

Jason Lloyd-Price, Sofia Startceva, Vinodh Kandavalli, Jerome G. Chandraseelan, Nadia Goncalves, Samuel M. D. Oliveira, Antti Häkkinen, and Andre S. Ribeiro*

Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, PO Box 553, Office TC336, 33101 Tampere, Finland

*To whom correspondence should be addressed. Tel. +358 408490736. Fax. +358 331154989. Email: andre.ribeiro@tut.fi

Edited by Prof. Kenta Nakai

Received 1 December 2015; Accepted 11 February 2016

Abstract

We investigate the hypothesis that, in *Escherichia coli*, while the concentration of RNA polymerases differs in different growth conditions, the fraction of RNA polymerases free for transcription remains approximately constant within a certain range of these conditions. After establishing this, we apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the $P_{lac/ara-1}$ promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations. We find that, under full induction, the closed complex lasts ~788 s while subsequent steps last ~193 s, on average. We then establish that the closed complex formation usually occurs multiple times prior to each successful initiation event. Furthermore, the promoter intermittently switches to an inactive state that, on average, lasts ~87 s. This is shown to arise from the intermittent repression of the promoter by LacI. The methods employed here should be of use to resolve the rate-limiting steps governing the *in vivo* dynamics of initiation of prokaryotic promoters, similar to established steady-state assays to resolve the *in vitro* dynamics.

Key words: free RNA polymerase, *in vivo* transcription dynamics, rate-limiting steps, reversible closed complex formation, repressor binding dynamics

1. Introduction

Gene expression has been intensively studied with the relatively new tools provided by fluorescent proteins and microscopy techniques with single-molecule resolution, in both prokaryotic[1–5] and eukaryotic[6,7] systems. These studies have established that this process cannot be fully characterized by the mean protein production rate,[8–12] since cells exhibit fluctuations (i.e. noise) over time and diversity in numbers across populations,[13] which, among other things, generates phenotypic diversity.[8] The noise has generally been investigated through indirect means, such as by observing the diversity in RNA and protein numbers in cell populations.[2,3,10,11,14] Other, more direct means consist of observing the distribution of intervals between RNA productions[2,4,5] and between protein bursts in individual cells.[3,15]

From these observations, a wide range of gene expression behaviours have been reported and, therefore, significantly different probabilistic models of transcription have been proposed.[2,4,16–18] In general, higher-than-Poissonian variability in RNA numbers has been explained by models in which the promoter intermittently switched into an inactive state, resulting in bursty RNA production dynamics.[2,16,19] Meanwhile, lower-than-Poissonian variability appears to be more consistent with models assuming multiple rate-limiting steps.[4,5,16,20,21]

There is direct experimental evidence for the existence of both mechanisms. Recently, Chong et al.[19] showed that bursts of RNA production can emerge due to positive supercoiling build-up on a DNA segment, which eventually stops transcription initiation for a short period until the release of the supercoiling by gyrase. On the other hand, the existence of rate-limiting steps was established by studies using steady-state assays.[22–24] Also, more recently, by fitting a monotone piecewise-constant function to the fluorescence signal from MS2-GFP tagged RNAs in individual cells, it was shown that *in vivo* RNA production can be a sub-Poissonian process.[4,5,20,21]

Recent studies have considered the possibility that both mechanisms can be present in a single promoter.[16,25] In ref. 25, a model including both mechanisms was proposed, and statistical methods were developed to select the relevant components and estimate the kinetics of the intermediate steps in initiation based on empirical data. However, this method cannot distinguish the order of the steps which occur after the start of transcription initiation, nor can it determine their reversibility, which recent evidence suggests may play a significant role in the dynamics of RNA production.[26]

A complete model for transcription in prokaryotes must account, apart from the genome-wide variability in noise levels,[17,27,28] for the well-established genome-wide variability in mean transcription rate[2,3,8] and in fold change (ratio of production rate between zero and full induction)[29] in response to induction found, e.g. in *Escherichia coli* promoters. For example, *in vitro* measurements on fully induced variants of the *lar* promoter showed that the mean interval between transcription events of these variants differs by hundreds of seconds.[29] Promoters also differ widely in range of induction, even when differing only by a couple of nucleotides.[29,30] For example, while P_{larS17} has an induction range of 500 fold, $P_{larconS17}$ has an induction range of 4.5-fold, even though it only differs by 3 point mutations.[29] This wide behavioural diversity is likely made possible by the sequence dependence of each step in transcription initiation.[29]

Thus far, the strategies used *in vitro* to characterize the kinetics of the steps involved in transcription initiation[22,26] have not been applied *in vivo* since they rely on measuring transcription for different RNA polymerase (RNAp) concentrations. Such a change in cells is expected to have a multitude of unforeseen effects[31] (in addition to the side effects of the means used to alter RNAp concentrations), which hampers the assessment of its consequences to the duration of the closed complex formation of a specific promoter. However, it is reasonable to hypothesize that, for certain small ranges of RNAp concentrations, these side effects will be negligible and thus, in such ranges, the inverse of the rate of transcription will be linear with respect to the inverse of the free RNAp concentration.

Importantly, in *E. coli*, RNAp concentrations have been shown to vary widely with differing growth conditions.[32] As such, here we make use of different media richness to achieve different RNAp concentrations and test whether within this range of conditions, the RNA production rate changes hyperbolically with the RNAp concentrations (i.e. if the inverse of this rate changes linearly with the inverse of the RNAp concentration). Having established this relationship, we make use of it to study the *in vivo* kinetics of transcription initiation of $P_{lac/ara\text{-}1}$. In particular, we perform measurements of the time intervals between RNA productions at the single molecule level in different intracellular RNAp and inducer concentration conditions, which we use to derive a more detailed model of transcription initiation of $P_{lac/ara\text{-}1}$. For this, we first extrapolate the mean interval between production events to the limit of infinite RNAp concentration, so as to estimate the *in vivo* durations of the open and closed complex formations of this promoter. Next, we examine the significance of an intermittent inactive promoter state, and the role of LacI in the emergence of this state. Finally, for the first time *in vivo*, we determine the reversibility of the closed complex formation.

2. Materials and methods

2.1. Cells and plasmids

For single-cell RNAp fluorescence measurements, we used *E. coli* W3110 and RL1314,[33] generously provided by Robert Landick, University of Wisconsin-Madison. For single-cell transcription interval measurements, we used *E. coli* DH5α-PRO (generously provided by Ido Golding, Baylor College of Medicine, Houston). The strain information is: deoR, endA1, gyrA96, hsdR17(rK- mK+), recA1, relA1, supE44, thi-1, Δ(lacZYA-argF)U169, Φ80δlacZΔM15, F-, λ-, PN25/tetR, PlacIq/lacI and SpR. This strain contains two constructs: a high-copy reporter plasmid vector PROTET-K133 (carrying MS2d-GFP under the control of $P_{LtetO\text{-}1}$) and a single-copy plasmid vector pIG-BAC carrying the target transcript (mRFP1 followed by 96 MS2-binding sites) under the control of $P_{lac/ara\text{-}1}$.[2] This promoter is located approximately 2 and 9 kb from the origin of replication (Ori2) and the plasmid size is 11.5 kb.[2] This system has been used to measure the distribution of time intervals between RNA production events due to its ability to detect individual target RNA molecules consisting of numerous MS2 coat protein binding sites, which are rapidly bound by fluorescently tagged MS2 coat proteins. These can be seen as they are produced under a fluorescence microscope as fluorescent foci.[2,4,5,20,21] Finally, we used the plasmid pAB332 carrying *hupA-mCherry* to visualize nucleoids (generously provided by Nancy Kleckner, Harvard University, Cambridge, MA, USA). For our measurements, we inserted this plasmid into DH5α-PRO cells so as to detect nucleoids in individual cells during the live cell microscopy sessions. HupA is a major nucleoid associated protein (NAP) that participates in its structural organization.[34]

2.2. Chemicals

The components of Lysogeny Broth (LB) were purchased from LabM (UK), and antibiotics from Sigma-Aldrich (USA). For RT-PCR, cells were fixed with RNAprotect bacteria reagent (Qiagen, USA). Tris and EDTA for lysis buffer were purchased from Sigma-Aldrich and lysozyme from Fermentas (USA). The total RNA extraction was done with RNeasy RNA purification kit (Qiagen). DNase I, RNase-free for RNA purification, was purchased from Promega (USA). iScript Reverse Transcription Supermix for cDNA synthesis and iQ SYBR Green supermix for RT-PCR were purchased from Biorad (USA). Agarose, isopropyl β-D-1-thiogalactopyranoside (IPTG), arabinose, and anhydrotetracycline (aTc) are from Sigma-Aldrich.

2.3. Growth media

To achieve different RNAp concentrations in cells, we altered their growth conditions as in.[35] For this, we used modified LB media which differed in the concentrations of some of their components. The media used are denoted as *m*×, where the composition per 100 ml are: *m* grams of tryptone, *m*/2 gram of yeast extract and 1 g of NaCl (pH = 7.0). For example, 0.25× media has 0.25 g of tryptone and 0.125 g of yeast extract per 100 ml.

2.4. Relative RNAp quantification

We measured relative RNAp concentrations in cells using four different methods. First, relative RNAp concentrations in the strains W3110 and DH5α-PRO were measured from the relative *rpoC* transcript

levels obtained using RT-PCR. Cells containing the target plasmid with $P_{lac/ara\text{-}1}$-mRFP1-96BS and the reporter plasmids were grown overnight in respective media. Cells were diluted into fresh media to an OD_{600} of 0.05. After 110 min, cells were re-diluted to an OD_{600} of 0.05 into respective media containing IPTG (1 mM) and arabinose (1%). After 70 min, RNA protect reagent was added to fix the cells, followed by enzymatic lysis with Tris–EDTA lysozyme buffer (pH 8.3). RNA was isolated from cells using RNeasy mini-kit (Qiagen). One microgram of RNA was used as the starting material. The RNA samples were treated with DNase free of RNase to remove residual DNA. Next, RNA was reverse transcribed into cDNA using iSCRIPT reverse transcription super mix (Biorad). RT-PCR was performed using Power SYBR-green master mix (Life Technologies) with primers for the amplification of the target gene at a concentration of 200 nM. Reactions were carried out in triplicate with 500 nM per primer with a total reaction volume 20 µl. The following primers were used for quantification: RpoC-F: CGTCAGATGCTGCGTAAAGC, RpoC-R: GCGATCTTGACGCGAGAGTA, mRFP1-F: TACGACG CCGAGGTCAAG, mRFP1-R: TTGTGGGAGGTGATGTCCA. Estimated relative RNAp concentrations $\hat{R}_m$ in each condition m, and their standard uncertainties $\sigma(\hat{R}_m)$, were calculated according to the $\Delta C'_T$ method.[36]

Second, *E. coli* RL1314 cells with fluorescently tagged β′ subunits were grown overnight in respective media. A pre-culture was prepared by diluting cells to an OD_{600} of 0.1 with fresh specific medium, and grown to an OD_{600} of 0.5 at 37°C at 250 rpm. Cells were pelleted by centrifugation and re-suspended in saline. Fluorescence from the cell population was measured using a fluorescent plate-reader (Thermo Scientific Fluoroskan Ascent Microplate Fluorometer).

Third, relative RNAp concentrations were also estimated based on the growth rates of DH5α-PRO cells in Supplementary Fig. S1. First, we fit a power law function to the 'RNA polymerase molecules per cell' row of Table 3 from ref. 32, which we found to be $R = 10^6 \mu^{-1.426}$, where μ is the cell doubling time. Relative RNAp concentrations were then estimated from the measured cell doubling times.

Lastly, we measured the relative RNAp concentrations in RL1314 cells under the microscope using fluorescently tagged RpoC (described in the next section).

2.5. Microscopy

DH5α-PRO cells containing the target and the reporter plasmids were grown as described previously. Briefly, cells were grown overnight in respective media, diluted into fresh media to an OD_{600} of 0.1, and allowed to grow to an OD_{600} of ~0.3. For the reporter plasmid induction, aTc (100 ng/ml) was added 1 h before the start of the measurements. For the target plasmid, arabinose (1%) was added at the same time as aTc (following the protocol in ref. 2), and IPTG (1 mM) was added 10 min before the start of the measurements. Cells were pelleted and resuspended to fresh medium. A few microliters of cells were placed between a coverslip and an agarose gel pad (2%), which contains the respective inducers, in a thermal imaging chamber (FCS2, Bioptechs), heated to 37°C. The cells were visualized using a Nikon Eclipse (Ti-E, Nikon, Japan) inverted microscope with a C2+ confocal laser-scanning system using a 100× Apo TIRF objective. Images were acquired using the Nikon Nis-Elements software. GFP fluorescence was measured using a 488 nm argon ion laser (Melles-Griot) and 514/30 nm emission filter. Phase-contrast images were acquired with the external phase contrast system and a Nikon DS-Fi2 camera. Fluorescence images were acquired every 1 min for a total duration of 2 h. Phase-contrast images were acquired simultaneously every 5 min during the measurements.

We tested for phototoxicity due to the fluorescence and the phase-contrast imaging in these measurements. Supplementary Results suggest that there is no significant phototoxicity. Additionally, we verified that the relative RNAp concentrations under the microscope are similar to those measured in the previous section by repeating the above procedure with RL1314 cells and imaging RpoC::GFP fluorescence, 1 h after being placed in the thermal imaging chamber (see Supplementary Fig. S4). The relative RNAp concentration was estimated from the mean fluorescence concentrations of cells growing in each media.

2.6. Image analysis

Cells were detected from the phase contrast images as described in ref. 37. First, the images were temporally aligned using cross-correlation. Next, an automatic segmentation of the cells was performed by MAMLE,[38] which was checked and corrected manually. Next, cell lineages were constructed by CellAging.[39] Alignment of the phase-contrast images with the confocal images was done by manually selecting 5–7 landmarks in both images, and using thin-plate spline interpolation for the registration transform. Fluorescent spots and their intensities were detected from the confocal images using the Gaussian surface-fitting algorithm from.[40]

Jumps were detected in each cell's spot intensity timeseries using a least-deviation jump-detection method.[41] Given the level of noise in the timeseries, jump sizes, i.e. the intensity of 'one RNA', were selected by manual inspection of the timeseries of total foreground spot intensities within cells of a given timeseries, and cross-referencing these values with the observed numbers of spots in the cells. After performing the jump detection process making use of the complete timeseries, jumps occurring within 5 min of the beginning or end of a cell's lifetime were disregarded due to our observation that the jump detection method tends to produce spurious jumps in these regions due to insufficient data. The remaining jumps were interpreted as RNA production times, from which intervals between transcription events were calculated. Finally, censored intervals were calculated as the time from the last RNA production in a cell until the last time at which a jump could have been observed (i.e. until 5 min prior to cell division or the end of the timeseries). This removes the possibility of false positives while not affecting the distribution of intervals.

This method, when first proposed, made two assumptions on the fluorescence of MS2-GFP tagged RNAs (named 'spots'). Importantly, both assumptions were recently shown to be valid.[42] First, an individual spot is bound sufficiently rapidly by MS2-GFPs such that its fluorescence intensity, when first detected, is already within the range of fluorescence of fully formed MS2-GFP-RNA spots (when taking one image per minute). In other words, the spot intensity of a newly transcribed RNA jumps from 0 to 'full' in <1 min, rather than slowly ramping up. Namely, since the transcription elongation rate of mRNA in *E. coli* is ~50 nt/s[32] and the target gene is ~3,200 bp long,[1] the time to elongate the MS2-binding site region of the target RNA is ~60 s. Provided that MS2-GFP binding to its RNA-binding sites is fast, there will therefore be a maximum of one timepoint at which the fully transcribed target RNA may have reduced fluorescence. Since MS2-GFP is produced in excess in the cell and its binding affinity is strong (dissociation constant of ~0.04 nM[43]), most binding sites will be saturated very shortly after being produced. In agreement with ref. 42, no gradual increase in spot fluorescence was observed around the time of the first appearance of a spot.

Second, once formed, MS2-GFP-RNA spots, as well as their fluorescence, are resistant to degradation for the duration of our

measurements (2 h). This was shown by measurements of the dissociation rate of MS2 coat proteins from their RNA binding sites (on the order of several hours[43]), and by measurements of the lifetimes of the fluorescence of MS2-GFP tagged RNAs kept under observation for more than 2 h.[1,2,5,42,44] Relevantly, no detectable decrease in fluorescence was observed during this time.[42]

2.7. Model of transcription initiation

We first consider a model that allows for RNA production dynamics to range from sub-Poissonian to super-Poissonian, given the results from genome-wide studies of the variability in RNA numbers[27,45] and from studies of the transcription dynamics of individual genes.[2,4,5,17,20] The features of the model that allow it to reproduce these numbers are based on processes known to occur during transcription initiation in *E. coli* (e.g. the open complex formation[16,22,23] and an ON/OFF mechanism[16,19]). Then, based on our novel empirical data and methodology, we aim to obtain the most parsimonious version of the model that fits the data for a given promoter. We expect this procedure to be applicable to any promoter, and to result in slightly different models due to their differing dynamics and regulatory mechanisms.

The full model of transcription initiation considered here consists of the following set of reactions:

$$\mathrm{R} + \mathrm{P_{ON}} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP_c} \xrightarrow{k_2} \mathrm{RP_o} \xrightarrow{k_3} \mathrm{R} + \mathrm{P_{ON}} + \mathrm{RNA} \qquad (1)$$

$$\mathrm{P_{ON}} \underset{k_{ON}}{\overset{k_{OFF}}{\rightleftarrows}} \mathrm{P_{OFF}} \qquad (2)$$

Reaction (1) represents the multi-step process of transcription initiation of an active promoter in prokaryotes.[23,24,46,47] It begins with the formation of the closed complex ($\mathrm{RP_c}$), i.e. the binding of the RNA polymerase (R) to a free promoter ($\mathrm{P_{ON}}$). Once at the start site, the polymerase must open the DNA double helix, a process that includes several long-lived intermediate states,[23,26,46,48] resulting in the open complex ($\mathrm{RP_o}$). Finally, the polymerase begins RNA elongation, though before clearing the promoter, it may engage in abortive RNA synthesis in which short RNA transcripts (<10 nt) are produced.[47,49] The reactions in (1) should not be interpreted as elementary transitions. Rather, they represent the effective rates of the rate-limiting steps in the process, thus defining the promoter strength, and have been shown to be sequence-dependent.[50]

Specifically, k_1 represents the rate at which polymerases find and bind to the promoter region, which is the overall result of the promoter search process which includes non-specific binding of the polymerases to the DNA, followed by a 1D diffusive search,[51,52] collectively referred to here as the closed complex formation. Subsequently, several rapid, possibly reversible isomerization reactions occur until the polymerase melts the DNA and forms the transcription 'bubble'.[51] In Reaction (1), the $\mathrm{RP_c}$ state represents all substates until the first irreversible reaction in this chain. Consequently, k_2 and k_{-1} should be interpreted as the product of the rates of the elementary reactions which exit from this group of substates, and the steady-state probability of being in the appropriate substates for these reactions to occur.

Similarly, the $\mathrm{RP_o}$ state may represent numerous substates between the first state after which the complex is committed to initiation, and successful initiation. However, after this point, we cannot distinguish the reversibility of any of the following steps, since the time-interval distribution of a sequence of elementary reversible reactions of arbitrary rates is observationally equivalent to a sequence of irreversible reactions.[25] The remaining steps (here, only k_3) therefore represent the rates of the slowest of these irreversible reactions. Such steps may include additional isomerization reactions, abortive RNA synthesis and promoter escape and clearance.[35]

Reaction (2) represents the promoter intermittently transitioning to a transcriptionally inactive state (P_{OFF}). Experimentally verified mechanisms by which this can occur are the binding and unbinding of repressors and activators,[29] the accumulation of positive supercoiling in the DNA.[19] Additional mechanisms have also been hypothesized, such as transcriptional pausing[53,54] and others.[55]

For a given concentration of R, the interval distribution between transcription events described by Reactions (1) and (2) (i.e. the first-passage time distribution to reach the final state, starting in the P_{ON} state) is observationally equivalent to the interval distribution described by a model of the form:

$$S_{\mathrm{OFF}} \underset{\lambda_{\mathrm{OFF}}}{\overset{\lambda_{\mathrm{ON}}}{\rightleftarrows}} S_0 \xrightarrow{\lambda_1} S_1 \xrightarrow{\lambda_2} S_2 \xrightarrow{\lambda_3} S_0 + \mathrm{RNA}, \qquad (3)$$

where the system starts in state S_0. The relationship between the parameters of these two models is described in Supplementary Table S1. Note that the states S_i do not correspond to the promoter states in Reactions (1) and (2). For details on how to derive and evaluate the distribution function for this model, see Supplementary Material and.[25]

It is noted that this model assumes that only one copy of the promoter is present in each cell at any given time. In the experiments performed here, in all conditions tested, the bacteria divided sufficiently slowly such that they spent most lifetime with only one chromosome. Specifically, cells spent no more than 11.4 ± 1.0% of their lifetime with two copies of the target promoter (Supplementary Material).

Finally, it is noted that the present model does not consider the influence of σ factors' numbers on the dynamics of transcription initiation, focussing instead solely on the concentration of RNA polymerases (in particular, on the concentration of holoenzymes containing a σ^{70}, i.e. $\mathrm{E}\sigma^{70}$, since our promoter of interest can only be transcribed by $\mathrm{E}\sigma^{70}$). This is based on the fact that, in all conditions tested, most RNA polymerases are occupied by σ factors.[56,57] Further, this occupation is made largely by σ^{70} since, first, when altering media richness, only σ^{32}'s concentration is significantly altered[56] and, second, the binding affinity of σ^{70} to E is much higher than that of any other σ factor (e.g. it is approximately 9 times higher than that of σ^{32}).[57]

2.8. Parameter estimation

Parameter estimates in Tables 1–3 were obtained by a maximum likelihood fit using the samples of the distribution of time intervals between production events obtained above (the intervals and censored intervals), as in.[25] The complete model-fitting procedure is detailed in the Supplementary Material. The uncertainty of the fit of the model parameters was estimated using the negative of the Hessian of the log-likelihood surface, evaluated at the maximum likelihood estimate.

The mean of the time interval distribution between transcription initiation events, $I(R)$, predicted by Reactions (1) and (2) is, for a given RNAp concentration R:

$$I(R) = \frac{(k_{\mathrm{ON}} + k_{\mathrm{OFF}})(k_{-1} + k_2)}{R k_1 k_2 k_{\mathrm{ON}}} + \frac{1}{k_2} + \frac{1}{k_3} = \tau_{\mathrm{CC}}(R) + \tau_{\overline{\mathrm{CC}}} \qquad (4)$$

where $\tau_{\mathrm{CC}}(R) = k_{\mathrm{CC}}^{-1} R^{-1}$ is the mean time taken by the initial binding of RNAp for a given RNAp concentration, and $\tau_{\overline{\mathrm{CC}}}$ is the mean time taken by the steps occurring after the polymerase has committed to transcription until the clearance of the promoter region (due to the

initiation of elongation). As such, we expect the majority of the duration of $\tau_{\overline{CC}}$ to consist of the open complex formation as defined in.[46] The remaining of its duration we attribute to failures in promoter escape.[59]

Estimates of $\tau_{\overline{CC}}$ and $k_{\overline{CC}}^{-1}$, denoted $\hat{\tau}_{\overline{CC}}$ and $\hat{k}_{\overline{CC}}^{-1}$, were obtained from the best-fit parameters of the most parsimonious model, as given in Table 3. The standard uncertainties of the estimators $\hat{\tau}_{\overline{CC}}$ and $\hat{k}_{\overline{CC}}^{-1}$, denoted $\sigma(\hat{\tau}_{\overline{CC}})$ and $\sigma(\hat{k}_{\overline{CC}}^{-1})$, were obtained using the Delta Method[60] from the uncertainties of the model parameters.

Finally, mean durations of intervals between transcription events for each media condition $\hat{I}_m$, were estimated by fitting the model in Reaction (3) to the data from only that condition, and taking the mean of the distribution. This procedure was followed to include the censored intervals in the estimate of $\hat{I}_m$ to avoid underestimating the mean interval duration due to the limited observation times. The standard uncertainty $\sigma(\hat{I}_m)$ was estimated using the Delta Method.[60]

2.9. Validation of the τ-plot slope

We verified the slope of the τ-plot in Fig. 4 using the RT-PCR measurements from Fig. 3. These measurements are both linear with respect to $\hat{R}_m^{-1}$, but differ by an unknown scaling factor. We denote the estimated production rate as measured by RT-PCR in media condition m as $\hat{S}_m$, with standard uncertainty $\sigma(\hat{S}_m)$. We found this scaling factor by fitting the parameter c in $\hat{I}_m = c\hat{S}_m^{-1}$ by weighted total least squares[61] (WTLS), with the measurements weighted by the inverse of their uncertainty (i.e. $\sigma^{-2}(\hat{S}_m^{-1})$ and $\sigma^{-2}(\hat{I}_m)$). This method was chosen since it accounts for the uncertainty in both of the measurements. It results in the estimate $\hat{c}$. The dashed line in Fig. 4 was obtained by fitting the scaled points $\hat{c}\hat{S}_m^{-1}$ against $\hat{R}_m^{-1}$ by WTLS. The uncertainty shown includes both the uncertainty in the WTLS fit of this line, as well as the uncertainty in $\hat{c}$.

2.10. Method to infer the duration of the closed complex of a promoter

The method to infer the kinetics of transcription initiation *in vivo* is illustrated in Fig. 1. First, conditions are selected that differ widely in free intracellular RNAp concentrations (step A in Fig. 1). Next, an *in vivo* single-molecule detection technique is used to sample the time interval distribution between consecutive transcription events in individual cells in each of the conditions (step C in Fig. 1). To obtain these intervals, here we used the MS2d-GFP single RNA detection system[4] (step B in Fig. 1). Then, we fit a general model of transcription initiation to the empirical data (see above), which includes both the multi-step nature of transcription initiation as well as the possibility of an intermittently inactive promoter state[25] (Reactions (1) and (2)). From this fit, we obtain an estimate of the *in vivo* mean duration of the open complex formation by extrapolating the duration of intervals between transcription events to infinite RNAp concentrations, similar to the *in vitro* extrapolation presented in ref. 22 (step D in Fig. 1). The model fit will also assess the importance of an intermittent inactive promoter state and the reversibility and kinetics of the closed complex formation.

3. Results

3.1. Changing free RNA polymerase concentrations

We first verified that it is possible to change intracellular RNAp concentration by a wide range by changing the growth conditions of the cells.[32,35,62] As such, we grew cells in four media (described in the Materials and methods), labelled 1×, 0.75×, 0.5×, and 0.25×, which solely differ in richness of two components (tryptone and yeast extract). We then measured the relative RNAp concentrations in cells grown in these four media using RT-PCR of the *rpoC* gene, i.e. the gene coding for the β′ subunit, which is the limiting factor in the assembly of the RNAp holoenzyme.[48,57,62] Results in Fig. 2 (dark grey bars) show that, in the range tested, the RNAp concentration in the cells increases significantly with increasing media richness.

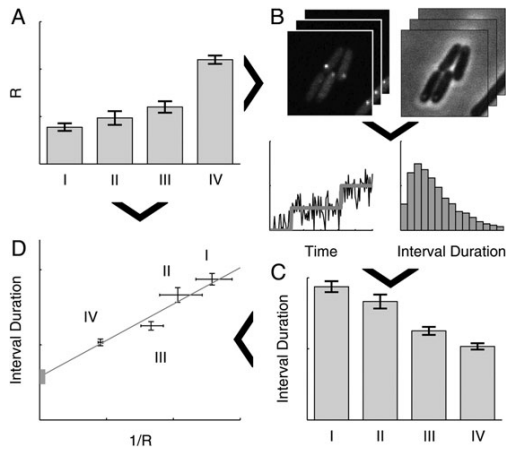


Figure 1. Schematic representation of the *in vivo* measurement of the initiation kinetics, using simulated data. (A) First, several conditions are selected, labelled I–IV, differing in intracellular RNAp concentration, *R*. (B) Next, we obtain timeseries of fluorescence and phase contrast (for cell segmentation purposes) images of cells expressing MS2d-GFP and target RNA under the control of the promoter of interest in each condition, from which time intervals between individual transcription events are determined. This is done by jump detection in the total RNA spot intensity of each cell (lower-left in B), from which the interval distribution is obtained (lower-right in B). (C) Mean interval durations are then estimated from these interval distributions for each condition. (D) Finally, the mean interval durations and measurements of *R* are combined into a τ-plot,[22] from which estimates of the mean times taken by the closed complex and open complex formation are obtained for each condition. Arrows depict the flow of information in the measurement procedure.

To validate this result, we measured the relative RNAp concentrations by plate reader in cells expressing fluorescently tagged RpoC in the strain RL1314 (derived from W3110),[33] in the same four media. In addition, we also measured the levels of the *rpoC* transcripts in the strain W3110 by RT-PCR in the 0.5× and 1× conditions. Results (Fig. 2) show that the relative changes in the protein and mRNA levels of *rpoC* match the measurements by RT-PCR of the *rpoC* gene in DH5α-PRO.

Note that, even though the experimental procedures and strains differ, our measurements are in agreement with the relative changes in RNAp concentrations reported in ref. 32, for the difference in growth rates observed here between the 0.25× and 1× conditions (Supplementary Fig. S1), which we estimate to be ~0.48 (Materials and methods). In this regard, given that the same result applies to (at least) three different strains, we expect it to be significantly strain-independent.

Finally, to verify that the relative RNAp concentrations measured in Fig. 2 are maintained under the microscope, we measured the relative RNAp concentration in the RL1314 cells expressing fluorescently

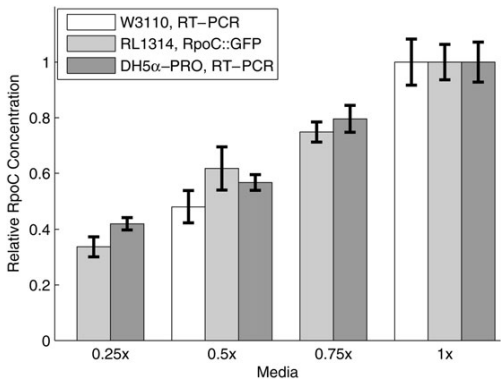


Figure 2. Measurements of the relative intracellular RNAp concentrations ($\hat{R}_m$) for cells growing in the four different media. Bars show the standard uncertainties ($\sigma(\hat{R}_m)$) of the measurements. Data is from two replicates with 3 technical replicates each (DH5α-PRO, RT-PCR, and W3110, RT-PCR), and three replicates with three technical replicates each (RL1314, RpoC::GFP). All data are presented relative to the RNAp concentration at 1×. The media used are denoted as *m*×, where the composition per 100 ml is: *m* grams of tryptone, *m*/2 grams of yeast extract and 1 g of NaCl (pH = 7.0).For example, 0.25× media has 0.25 g of tryptone and 0.125 g of yeast extract per 100 ml.

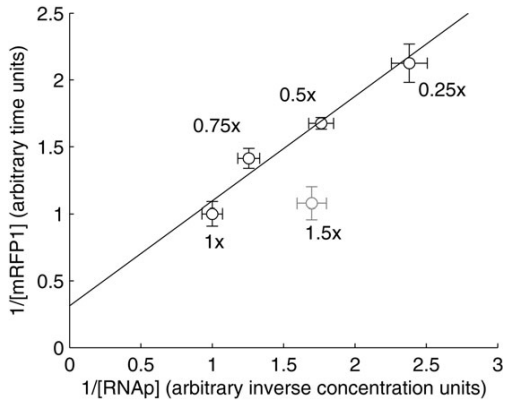


Figure 3. Lineweaver–Burk plot of the inverse of the production rate of mRFP1 from the $P_{lac/ara\text{-}1}$ promoter against the inverse of the total RNAp concentrations for the same growth conditions as in Fig. 2 (black points), and for 1.50× media (grey point). Standard uncertainties are shown for both quantities (horizontal and vertical error bars). Relative production rates were measured by RT-PCR with two biological replicates with three technical replicates each.

tagged RpoC under the microscope between the two extreme conditions (0.25× and 1×), after 1 h in the thermal imaging chamber (Materials and methods). The relative RNAp concentration between the conditions was measured to be 0.367 ± 0.012, which is consistent with the measurements in Fig. 2. Lastly, from these images, we did not observe significant cell-to-cell variability in the RNAp concentrations (Supplementary Fig. S4), indicating that the mean concentrations reported in Fig. 2 are representative of the populations.

These measurements show that the relative RNAp concentration changes widely between the selected growth conditions. However, the variable affecting transcription kinetics is the relative free RNAp concentration. As such, we must verify whether the relative total RNAp concentration can be used as a proxy for the relative free RNAp concentrations. If this holds true and there are no other factors affecting the production rate of the promoter of interest in these conditions, then the RNA production rate should be hyperbolic with respect to the RNAp concentration. That is, the reciprocal of the RNA production rate from this promoter should be linear when plotted against the reciprocal of the measured relative RNAp concentrations, and one should obtain a line on a Lineweaver–Burk plot.

There are several reasons why this plot may not be linear. If, for example, the ratio of free RNAp to total RNAp is not constant in this range of growth conditions, with a higher fraction of free RNAp in the poorer growth conditions due to increased ppGpp,[31] then we expect a curve with positive curvature on this plot. Meanwhile, a negative curvature would be obtained if the promoter of interest could be induced by increased cAMP in the poorer growth conditions, or if the cells spent, on average, a significantly increased amount of time with multiple copies of the plasmid in the richer growth conditions, among other possibilities. In these cases, to dissect the transcription initiation kinetics of such promoters, another method of modifying the free RNAp concentration will be required.

Given the above, we interpret a straight line on the Lineweaver–Burk plot as evidence that, for the conditions tested, (i) the relative free RNAp concentrations can be assessed from the total RNAp concentrations, and (ii) no factors other than the changes in the free RNAp concentration affect the target promoter.

Here, we tested this by measuring the RNA production rate from $P_{lac/ara\text{-}1}$ in *E. coli* DH5α-PRO by RT-PCR in the same four media conditions as in Fig. 2. We selected this promoter, since its dynamics has been extensively characterized[2,21,29,63–67] and because it has the same logical structure as the *lac* promoter, with an activator and a repressor.[63] The resulting Lineweaver–Burk plot is shown in Fig. 3 where a linear relationship is clearly observed between these points (black points). To determine whether the small deviations from linearity are statistically significant, we performed a likelihood ratio test between a linear fit by WTLS[61] (shown as a line in Fig. 3), and fits with higher order polynomials (also by WTLS by minimizing χ^2 as in[61]). No test rejected the linear model (all $P > 0.25$). As noted earlier, this relationship is only expected to occur in a limited range of growth conditions. To illustrate this, we repeated the same measurements in 1.5× media (grey point in Fig. 3). The result shows that this hyperbolic relationship is lost in very rich media (including this point causes the likelihood ratio test to reject the linear model, $P = 0.0014$). We conclude that, for the growth conditions in Fig. 2, the relative free RNAp concentrations are well-approximated by the total RNAp concentrations, and there are no significant other factors affecting the initiation dynamics of $P_{lac/ara\text{-}1}$.

3.2. Interval distributions between consecutive RNA productions

Given this, it is possible to apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the $P_{lac/ara\text{-}1}$ promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations.

We measured the distribution of time intervals between transcription events (hereafter referred to as 'intervals') for $P_{lac/ara\text{-}1}$ in each cell growth condition using the MS2d-GFP single-RNA detection system,[1] with a least-deviation jump-detection procedure[41] (Materials and

Table 1. Statistics of the measured distributions of intervals between transcription events from *lac/ara-1* promoters

Condition	Number of cells	Number of intervals	Number of censored intervals	Inferred interval mean and uncertainty (s)	Inferred CV^2
0.25×	196	371	323	1,899 ± 105	1.08
0.5×	302	1,027	605	1,553 ± 50	1.06
0.75×	146	620	345	1,205 ± 51	1.09
1×	206	1,202	573	1,005 ± 112	1.21

Shown are the condition, the number of cells (which is the cell count at the start of the measurements), the numbers of whole and censored intervals extracted, and finally the inferred mean (and its standard uncertainty) and CV^2 of the interval distribution.

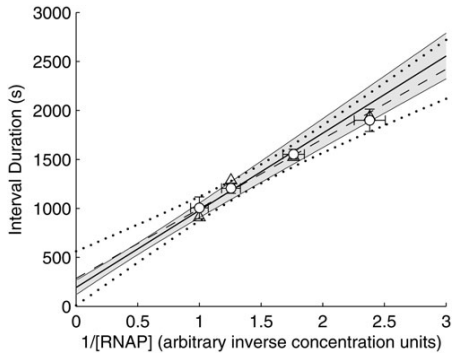


Figure 4. τ-plot for $P_{lac/ara\text{-}1}$, showing the mean interval between transcription events in individual cells for each media condition (white circles), with their standard uncertainties (vertical error bars) and the standard uncertainties of the relative RNAp concentrations (horizontal error bars). Also shown is the best-fit line (solid line), as determined by the intercept and slope obtained from the best-fitting model (Table 3), with one standard uncertainty estimated by Scheffé's method[69] combined with the Delta Method[60] (grey area). In addition, the figure shows the data from Fig. 3 (triangles), and the best-fitting line (dashed line, see Materials and methods) with one standard uncertainty estimated by Scheffé's method[69] (dotted black curves).

methods). This measurement results in samples from the interval distribution as well as 'censored' intervals, i.e. intervals for which we only observe the beginning due to cell division or the end of the time series. Both censored and uncensored intervals were accounted for in all parameter estimates to avoid biasing the estimates. For example, note that taking the mean of the uncensored intervals alone would underestimate the mean of the true interval distribution since long unobservable intervals would be absent from the estimate. Including the censored intervals balances this by considering long intervals that are at least as long as the censored interval length.[25]

From these distributions, we estimated the true mean and the squared coefficient of variation (CV^2, defined as the variance over the squared mean) of the interval distributions (Materials and methods). We chose CV^2 for quantifying the noise in the interval distribution since, to a good approximation, this quantity reflects the level of noise in the protein levels regardless of the actual shape of the transcription interval distribution.[68] Further, this variable equals 1 for the interval distribution of a Poisson process (i.e. an exponential distribution), regardless of the mean rate. These results, along with the amount of empirical data used, are shown in Table 1.

From Table 1, the mean interval decreases significantly with increasing media richness, as expected from the increased RNAp concentrations. Meanwhile, the CV^2 does not exhibit the same dependence on the media richness, and remains slightly >1 in all conditions tested.

3.3. Decomposition of the *in vivo* kinetics

From the data in Table 1, we next recreate the Lineweaver–Burk plot in Fig. 3 (white circles in Fig. 4), using the mean interval durations between RNA productions, as this quantity is an absolute measure of the inverse rate of RNA production (this plot is called a τ-plot).

Previously, using *in vitro* techniques, it has only been possible to extract from a τ-plot the mean duration of the open complex formation (the y-intercept of the plot, here denoted $\tau_{\overline{CC}}$), because the plot is based on the steady-state assay which only measures the mean rate of abortive transcription initiations. However, the distributions of time intervals between RNA productions contain information about the stochasticity of the process (i.e. the variability between intervals). As such, it is possible to extract a more complete model of the process of transcription. Namely, aside from the open complex formation, as mentioned in Materials and methods, it is possible to extract information on the closed complex and on an intermittent state prior to the closed complex formation.

In particular, we consider the detailed model of transcription initiation presented in Materials and methods (Reactions (1) and (2)), along with simplified models that can be considered if certain steps of the more detailed model do not influence the distribution of intervals. This model assumes that only one copy of the promoter is present in each cell at any given time, since in all conditions, the bacteria divided slowly, which suggests that they spent most lifetime with only one chromosome. We then consider three simplified models. First, if the time spent in the OFF state is very small, or if the system switches between OFF and ON very rapidly when compared with the forward reaction, then Reaction (2) will not affect the RNA production dynamics. A sufficient condition for both of these situations is that $k_{ON} >> k_1$. The other two simplifications are two limits of the closed complex formation, first considered in[22]: (i) $k_{-1} >> k_2$, i.e. it is reversible (Limiting Mechanism I), and (ii) $k_2 >> k_{-1}$, i.e. irreversible (Limiting Mechanism II). Limiting Mechanism I was found to be more likely in several *in vitro* measurements of various promoters.[22,23,26]

While all three simplifications are consistent with a line on a τ-plot, they produce significantly different distributions of intervals between RNA production events. For example, a significant ON/OFF mechanism will result in a more noisy distribution (a higher CV^2).[25] Similarly, Limiting Mechanism I effectively eliminates one limiting step, which also results in higher noise when compared with Limiting Mechanism II (Supplementary Fig. S2).

We fit the full and simplified models of transcription initiation to the observed dynamics of $P_{lac/ara\text{-}1}$ from all media conditions (Materials and methods). We used the Bayesian Information Criterion[70] (BIC) to compare the fits. The BIC is a model selection criterion which balances goodness-of-fit with the number of parameters to determine which model is most likely the 'truth'. The difference between BIC values (ΔBIC) can be interpreted as evidence *against* the model with *higher* BIC, with a ΔBIC > 5 being interpreted as strong evidence.[58] Results are shown in Table 2. Since, for several of the models, the optimal fit was for $k_3^{-1} = 0$, we also considered models that do not include another rate-limiting step after the open complex formation.

From Table 2, the initiation kinetics of $P_{lac/ara\text{-}1}$ is best-fit by Limiting Mechanism I (i.e. a reversible closed complex), with very high

Table 2. Fit parameters of the transcription initiation model in Reactions (1) and (2), and the models derived by applying the listed simplifying assumptions

Limiting mechanisms	Simplifications	k_{ON}^{-1} (s)	k_{OFF}^{-1} (s)	$k_1 k_{OFF}^{-1}$ (R^{-1})	k_1^{-1} (Rs)	$k_1 k_{-1}^{-1}$ (R^{-1})	k_{-1}^{-1} (s)	k_2^{-1} (s)	k_3^{-1} (s)	ΔBIC	ΔBIC_C
Full model		87	Fast	8,313	Fast	2,247	Fast	177	Fast	14.8	15.7
I	$k_{-1} \gg k_2, k_1 \gg k_{OFF}$	87	Fast[a]			7,446		192	Fast	8.1	8.5
I, $k_3 = \infty$	$k_{-1} \gg k_2, k_1 \gg k_{OFF}, k_3 = \infty$	87	Fast[a]			6,469		192		0.0	0.0
II	$k_2 \gg k_{-1}$	90	Fast	0.10	Fast			7	7	18.3	18.8
II, $k_3 = \infty$	$k_2 \gg k_{-1}, k_3 = \infty$	86	Fast	0.09	Fast			10		10.7	10.7
No ON/OFF	$k_{ON} \gg k_1$				Fast	0.49	Fast	326	Fast	188.1	188.1
No ON/OFF, I	$k_{ON} \gg k_1, k_{-1} \gg k_2$					0.50		328	Fast	180.1	179.6
No ON/OFF, I, $k_3 = \infty$	$k_{ON} \gg k_1, k_{-1} \gg k_2, k_3 = \infty$					0.50		328		172.0	171.1
No ON/OFF, II	$k_{ON} \gg k_1, k_2 \gg k_{-1}$				910			Fast	Fast	201.0	200.6
No ON/OFF, II, $k_3 = \infty$	$k_{ON} \gg k_1, k_2 \gg k_{-1}, k_3 = \infty$				910			Fast		192.9	192.0

Parameters denoted 'fast' are too fast to present on the timescale of seconds. When competing fast reactions occur, relevant ratios are given. ΔBIC values are given as the difference of the model's BIC from the BIC of the best-fitting model (the one with ΔBIC = 0). Models with lower ΔBIC are favoured over models with higher ΔBIC.[58] Censored intervals were included in ΔBIC_C, but not in ΔBIC. The best-fitting model is shaded. Rates (and ratios) involving k_1^{-1} are given relative to the intracellular RNAp concentration in the 1× media.

[a] $k_1 k_2 k_{-1}^{-1} k_{OFF}^{-1} = 0.11$.

certainty (ΔBIC of all other models >8). We also find evidence for a significant ON/OFF mechanism. Though the time spent in each OFF state is short (~87 s), it will turn OFF, on average, ~9.1 times before committing to transcription in the 1× case (see Supplementary Material). This results in an interval distribution which is only slightly more noisy than what would be expected if the production process were Poissonian (i.e. a CV^2 of the interval distribution of 1; see the CV^2 values in Table 1). Interestingly, this implies that the noise in transcription of this promoter is representative of the behaviour of the majority of promoters in *E. coli*.[27] Finally, the steps after the commitment to transcription are fast, indicating that abortive initiation events do not play a significant role in the dynamics of RNA production by $P_{lac/ara\text{-}1}$. This model is depicted graphically in Fig. 5.

In addition, from Table 2, we find that $\tau_{\overline{CC}}$ is 193 ± 49 s. Meanwhile, the slope of the line on the τ-plot, here denoted k_{CC}^{-1}, is 788 ± 59 R·s (R is the polymerase concentration such that $R = 1$ is the polymerase concentration in 1× media). The line given by these values is shown in Fig. 4 (solid line). As a side note, the uncertainties of these estimates exaggerate the uncertainty of the inference, since the estimates are highly correlated (correlation coefficient of −0.6). This correlation is responsible for the hyperbolic shape of the confidence bounds (grey region in Fig. 4).

We verified the slope of the solid line in Fig. 4 using the RT-PCR measurements presented in Fig. 3, scaled to match the timescale of the intervals (Materials and methods). The resulting line is shown in Fig. 4 (dashed line), and is in good agreement with both the line given by our estimates of $\tau_{\overline{CC}}$ and k_{CC}^{-1} (solid line), and the inferred interval means (white circles).

Lastly, we note that the BIC depends on the number of samples used to calculate the likelihood. Thus, BIC values calculated assuming that each censored interval is 'one sample' will over-penalize models with more parameters, while removing them will under-penalize them. Both sets of ΔBIC values are presented in Table 2 and, in our case, both result in the same conclusion, and thus the distinction does not affect the results for $P_{lac/ara\text{-}1}$. If, for another promoter, this turns out to be the case, additional measurements will be required to distinguish between the models.

Our results are in agreement with previous measurements of the kinetics of this and similar promoters. For example, a previous study reported that, under full induction in LB media (1× media here), $P_{lac/ara\text{-}1}$ expresses ~4 RNA/h[2] (i.e. 1 RNA every ~900 s), while we inferred the time between transcription events to be ~980 s. Using the steady-state assay, $\tau_{\overline{CC}}$ was measured to be ~330 s for P_{lac}[71] (with or without CRP-cAMP), while we obtained ~193 s.

3.4. Determining the source of the intermittent inactive state for $P_{lac/ara\text{-}1}$

We identified the presence of an ON/OFF mechanism in the dynamics of $P_{lac/ara\text{-}1}$. It is worth noting that this ON/OFF phenomenon differs from the one reported in refs 2 and 19 since, first, we only observe OFF periods on the order of ~87 s, while in ref. 2 the OFF periods reported for $P_{lac/ara\text{-}1}$were on the order of 37 min. In addition, both here and in ref. 2, the promoter of interest is integrated in a single-copy plasmid, and thus the OFF periods cannot be explained by the buildup of positive supercoiling, since the plasmid is not topologically constrained.[19] We therefore hypothesized that the OFF periods observed here more likely result from the intermittent formation of a DNA loop, due to the transient binding of LacI, which exists in high concentration in DH5α-PRO (~3,000 copies vs. ~20 in wild type[63]).

If LacI is responsible for the ON/OFF behaviour, then reducing the concentration of IPTG should affect the ON/OFF dynamics, and not

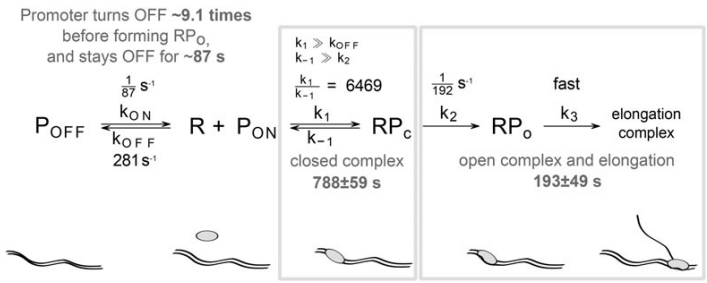


Figure 5. Best fitting model of transcription initiation (with ON/OFF mechanism and reversible close complex formation). The model parameters are specified in black and estimated durations of the transcription initiation steps for 1× LB media are shown in grey.

change the dynamics following the closed complex formation.[29] To test this prediction, and demonstrate the utility of the model-fitting approach, besides considering the interval measurements in 1× in Table 1, we also measured the interval distribution of $P_{lac/ara\text{-}1}$ using MS2d-GFP in the 1× media without induction by IPTG. From 130 cells, we extracted 57 intervals and 117 censored intervals between transcription events. From these, we inferred a mean interval of 3,374 ± 462 s, and a CV^2 of 1.03. This mean is significantly greater than the mean measured in the fully induced condition (1,005 ± 112 s), consistent with the much stronger repression of the promoter by LacI in this condition.

Given the wide difference in dynamics of RNA production between the induced and non-induced cases, we used the model fitting procedure to determine which steps are significantly affected by LacI. For this, we performed independent fits of a reduced model of initiation to the induced and the non-induced conditions. This model is observationally equivalent to the full model of initiation (Reactions (1) and (2)) for a single value of R, and is presented in Reaction (3). This reduced model is necessary since we do not have measurements of the uninduced case at multiple values of *R* with which to fit all parameters of the full model. The reduced model's parameters are denoted by λ_x, which are related to, but are not equal to the values of k_x. Their relationship is presented in Supplementary Table S1. The fitting results are shown in Table 3 (labelled 'Independent'). We also considered joint models where parameters were fixed between conditions, and used the BIC to select the most likely model.

The first three models with joint parameters test for whether or not the parameters controlling the ON/OFF mechanism change with induction strength. Consistent with this hypothesis, the models with joint λ_{OFF}^{-1} are strongly rejected (ΔBIC much higher than that of the Independent model). Surprisingly, the model with only joint λ_{ON}^{-1} was also rejected, implying that the mean OFF times might also vary with induction strength. Additional studies are needed to elucidate why such OFF times depend on the induction strength.

Having established that λ_{ON}^{-1} and λ_{OFF}^{-1} differ between conditions, we next assessed whether only these parameters differ. For that, we fixed λ_1^{-1} and λ_2^{-1}, and verified that this model is the most parsimonious model (ΔBIC relative to the Independent model of −14.3). We conclude that only λ_{ON}^{-1} and λ_{OFF}^{-1} differ between conditions, confirming the prediction that LacI is responsible for the ON/OFF mechanism affecting the RNA production dynamics.

Finally, other models were considered, e.g. the hypothesis that λ_1^{-1}, λ_2^{-1}, and/or λ_{ON}^{-1} do not differ between conditions. These models were also strongly rejected in favour of the parsimonious model, and are not shown for brevity.

Table 3. Fit parameters of the transcription initiation model in Reaction (3) to the measured intervals in the 1× media with and without induction by IPTG

Joint parameters	Condition	λ_{ON}^{-1} (s)	λ_{OFF}^{-1} (s)	$\lambda_1\lambda_{OFF}^{-1}$	λ_1^{-1} (s)	λ_2^{-1} (s)	ΔBIC
Independent	IPTG+	110	Fast	0.11	Fast	5	14.3
	IPTG−	48	Fast	0.01	Fast	Fast	
λ_{ON}^{-1}	IPTG+	4,444	Fast	11.50	Fast	964	120.3
	IPTG−		Fast	∞	Fast	2,919	
λ_{OFF}^{-1}	IPTG+	7	Fast	∞	Fast	964	152.9
	IPTG−	320		1.86	Fast	2,919	
λ_{ON}^{-1}, λ_{OFF}^{-1}	IPTG+	326	Fast	∞	Fast	964	145.7
	IPTG−			1.94	Fast	2,918	
λ_1^{-1}, λ_2^{-1}	IPTG+	106	Fast	0.11	Fast	Fast	0.0
	IPTG−	48	Fast	0.01			

The relationship between these parameters and the parameters in Table 2 are discussed in the Materials and methods and Supplementary Material. Five models are considered, differing in which parameters are assumed to be the same between the two induction conditions. Parameters denoted 'fast' are too fast to present on the timescale of seconds. As λ_{OFF}^{-1} and λ_1^{-1} were found to be fast in all models, the $\lambda_1\lambda_{OFF}^{-1}$ ratio is also shown. ΔBIC values are given as the difference of the model's BIC from the BIC of the best-fitting model (the one with ΔBIC = 0). Models with lower ΔBIC are favoured over models with higher ΔBIC.[58]

3.5. Precision of the estimates

We define the precision of the estimates of $\tau_{\overline{CC}}$ and k_{CC}^{-1} as the ratio between the timescale of the intervals (i.e. the mean interval in the condition with greatest *R*) and the standard uncertainties of $\hat{\tau}_{\overline{CC}}$ and $\hat{k}_{CC}^{-1}$, respectively. Specifically, the precision of $\hat{\tau}_{\overline{CC}}$'s estimate is $P_{\overline{CC}} = \hat{I}_1/\sigma(\hat{\tau}_{\overline{CC}})$, and the precision of $\hat{k}_{CC}^{-1}$'s estimate is $P_{CC} = \hat{I}_1/\sigma(\hat{k}_{CC}^{-1})$. Given this, here, with the volume of data in Table 1, we achieved $P_{\overline{CC}} = 20.7$ and $P_{CC} = 17.0$, corresponding to errors of ∼5 and ∼6%, respectively.

In addition, we found that this precision is highly dependent on the dynamic range of RNAp concentrations. For example, for a small dynamic range of 1.5 (our measurements in Fig. 2 have a range of ∼2.4), the precisions $P_{\overline{CC}}$ (in $\hat{\tau}_{\overline{CC}}$) and P_{CC}(in $\hat{k}_{CC}^{-1}$) would have been reduced to ∼11.2 and ∼6.7, respectively. Losses in precision due to reduced dynamic ranges can, however, to some extent, be offset by collecting more samples for the interval distributions (see estimation of precision in Supplementary Material).

4. Discussion

We established that, in *E. coli*, the concentration of free RNA polymerases differs significantly within a certain range of growth conditions, and that the inverse of the target RNA production rate under the control of $P_{lac/ara\text{-}1}$ varies linearly with the inverse of the free RNAp concentration (which are the conditions imposed in the *in vitro* measurements the open complex formation by steady state assays[22,24,72]). Thus, we were able to apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the $P_{lac/ara\text{-}1}$ promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations. This revealed that this promoter has two rate-limiting steps: a reversible closed complex formation and a significant open complex formation. Further, it also intermittently switches to a short-lived inactive state. Based on the inferred timescale of this inactive state, we predicted that this state is the result of the intermittent binding of the repressor LacI, which we verified by measuring the interval distribution when the promoter is not induced by IPTG. We believe that the complexity of this process is the reason why it has not been reported before. Namely, previous studies only considered either multiple rate-limiting steps,[4,5,22,23,66] or an ON/OFF process,[2,17,19,73,74] while this promoter exhibits both.

We note that, provided that the promoter has a reversible closed complex formation, the model fitting procedure proposed here allows the duration and order of two steps following the closed complex to be obtained (specifically, the ratio between k_2 and k_3 can be determined from how the CV^2 of the interval distribution changes with R; see Supplementary Fig. S2). Here, this additional step was not found. However, we expect that, for other promoters, or in different conditions (e.g. low temperatures[72]), this step may be significant. Meanwhile, if Limiting Mechanism II is found to be the best-fitting model, the order of the last two steps will remain ambiguous due to the lack of reversibility.

Finally, it is worth noting that in previous works, we have not found evidence for an ON/OFF mechanism for $P_{lac/ara\text{-}1}$, due to the low levels of noise detected in the time intervals between transcription events.[4,21,66] This can be explained by, first, we did not consider censored intervals, which contribute significantly to the increase of the tail of the distribution of intervals.[25] Second, the OFF period is quite short, and thus its detection requires a large volume of data and a sensitive inference methodology.[25] Our results show that, by solving these two issues (by applying the methods in refs 41 and 25), our methodology can identify and characterize many relevant steps in transcription initiation, including those with lesser influence.

In the future, it would be of interest to extend the model to consider what occurs when more than one copy of a promoter is present in the cell. We expect that variations in the promoter copy numbers would, in that case, explain some of the variance of the data, instead of this variance being solely determined by the ON/OFF mechanism and the sequential steps.

We expect the methodology employed here to be applicable to promoters, native or synthetic, whose changes in the inverse of the transcription rate are linear with the inverse of the free RNAp concentrations. Also, it should be applicable to promoters evolved to interact with multiple transcription factors (TF), provided their fast binding and unbinding (compared with competing events), as they could be accounted for by tuning the rate constants of some of the reactions of the model. Further, multiple slow TFs, including activators, can be accounted for by adding appropriate TF-bound states, with differing production rates, in a similar manner to the ON/OFF model. As such, the methodology should be applicable at a genome wide scale. It should also be applicable to eukaryotes, provided suitable means to alter polymerase concentrations. Lastly, it should be useful in detecting differences in transcription initiation kinetics of a promoter subject to different intra- or extra-cellular conditions.

Acknowledgements

We thank Axel Oikari, Abhishekh Gupta, and Antti Martikainen for valuable advice.

Supplementary Data

Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the Academy of Finland (257603 to A.S.R.); Centre of International Mobility (13.1.2014/TM-14-91361/CIMO to A.S.R.); Jenny and Antti Wihuri Foundation (to A.H.); and the Tampere University of Technology President's Graduate Programme (to J.L.-P. and S.S.). Funding to pay the Open Access publication charges for this article was provided by Academy of Finland (257603 to A.S.R.).

References

1. Golding, I. and Cox, E.C. 2004, RNA dynamics in live *Escherichia coli* cells, *Proc. Natl Acad. Sci. USA*, **101**, 11310–5.
2. Golding, I., Paulsson, J., Zawilski, S.M. and Cox, E.C. 2005, Real-time kinetics of gene activity in individual bacteria, *Cell*, **123**, 1025–36.
3. 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**, 1600–3.
4. Kandhavelu, M., Mannerström, H., Gupta, A., et al. 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.
5. Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., et al. 2012, Dynamics of transcription driven by the *tetA* promoter, one event at a time, in live *Escherichia coli* cells, *Nucleic Acids Res.*, **40**, 8472–83.
6. Fusco, D., Accornero, N., Lavoie, B., et al. 2003, Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells, *Curr. Biol.*, **13**, 161–7.
7. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S. 2006, Stochastic mRNA synthesis in mammalian cells, *PLoS Biol.*, **4**, 1707–19.
8. Kaern, M., Elston, T.C., Blake, W.J. and Collins, J.J. 2005, Stochasticity in gene expression: from theories to phenotypes, *Nat. Rev. Genet.*, **6**, 451–64.
9. Arkin, A.P., Ross, J. and McAdams, H.H. 1998, Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells, *Genetics*, **149**, 1633–48.
10. Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. 2002, Stochastic gene expression in a single cell, *Science*, **297**, 1183–6.
11. Raser, J.M. and O'Shea, E.K. 2005, Noise in gene expression: origins, consequences, and control, *Science*, **309**, 2010–3.
12. Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D. and van Oudenaarden, A. 2002, Regulation of noise in the expression of a single gene, *Nat. Genet.*, **31**, 69–73.
13. McAdams, H.H. and Arkin, A.P. 1999, It's a noisy business! Genetic regulation at the nanomolar scale, *Trends Genet.*, **15**, 65–9.
14. Süel, G.M., Garcia-Ojalvo, J., Liberman, L.M. and Elowitz, M.B. 2006, An excitable gene regulatory circuit induces transient cellular differentiation, *Nature*, **440**, 545–50.
15. Cai, L., Friedman, N. and Xie, X.S. 2006, Stochastic protein expression in individual cells at the single molecule level, *Nature*, **440**, 358–62.
16. Mitarai, N., Dodd, I.B., Crooks, M.T. and Sneppen, K. 2008, The generation of promoter-mediated transcriptional noise in bacteria, *PLoS Comput. Biol.*, **4**, e1000109.

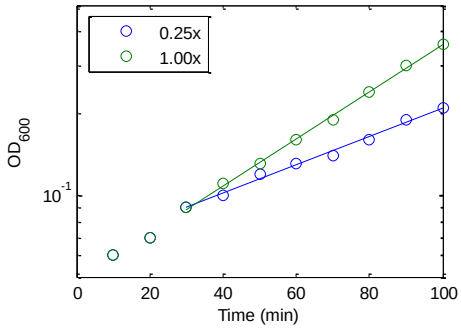
17. So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L.A., Segev, R. and Golding, I. 2011, General properties of transcriptional time series in *Escherichia coli*, *Nat. Genet.*, **43**, 554–60.
18. Zhdanov, V.P. 2011, Kinetic models of gene expression including noncoding RNAs, *Phys. Rep.*, **500**, 1–42.
19. Chong, S., Chen, C., Ge, H. and Xie, X.S. 2014, Mechanism of transcriptional bursting in bacteria, *Cell*, **158**, 314–26.
20. Kandhavelu, M., Häkkinen, A., Yli-Harja, O. and Ribeiro, A.S. 2012, Single-molecule dynamics of transcription of the *lar* promoter, *Phys. Biol.*, **9**, 026004.
21. 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**, 3870–5.
22. McClure, W.R. 1980, Rate-limiting steps in RNA chain initiation, *Proc. Natl Acad. Sci. USA*, **77**, 5634–8.
23. McClure, W.R. 1985, Mechanism and control of transcription initiation in prokaryotes, *Annu. Rev. Biochem.*, **54**, 171–204.
24. Bertrand-Burggraf, E., Lefèvre, J.F. and Daune, M. 1984, A new experimental approach for studying the association between RNA polymeras and the *tet* promoter of pBR322, *Nucleic Acids Res.*, **12**, 1697–706.
25. Häkkinen, A. and Ribeiro, A.S. 2016, Characterizing rate limiting steps in transcription from RNA production times in live cells, *Bioinformatics*, http://bioinformatics.oxfordjournals.org/content/early/2016/01/28/bioinformatics.btv744.abstract.
26. Friedman, L.J. and Gelles, J. 2012, Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation, *Cell*, **148**, 679–89.
27. Taniguchi, Y., Choi, P.J., Li, G.-W., et al. 2010, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells, *Science*, **329**, 533–8.
28. Sanchez, A., Garcia, H.G., Jones, D., Phillips, R. and Kondev, J. 2011, Effect of promoter architecture on the cell-to-cell variability in gene expression, *PLoS Comput. Biol.*, **7**, e1001100.
29. 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**, 3873–81.
30. Garcia, H.G., Sanchez, A., Boedicker, J.Q., et al. 2012, Operator sequence alters gene expression independently of transcription factor occupancy in bacteria, *Cell Rep.*, **2**, 150–61.
31. Gummesson, B., Magnusson, L.U., Lovmar, M., et al. 2009, Increased RNA polymerase availability directs resources towards growth at the expense of maintenance, *EMBO J.*, **28**, 2209–19.
32. Bremer, H. and Dennis, P.P. 1996, Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In: Neidhardt, F.C., (ed.), *Escherichia Coli and Salmonella*, 2nd ed. ASM Press, Washington, DC, pp. 1553–69.
33. Bratton, B.P., Mooney, R.A. and Weisshaar, J.C. 2011, Spatial distribution and diffusive motion of RNA polymerase in live *Escherichia coli*., *J. Bacteriol.*, **193**, 5138–46.
34. Dillon, S.C. and Dorman, C.J. 2010, Bacterial nucleoid-associated proteins, nucleoid structure and gene expression, *Nat. Rev. Microbiol.*, **8**, 185–95.
35. Liang, S.-T., Bipatnath, M., Xu, Y.-C., et al. 1999, Activities of constitutive promoters in *Escherichia coli*., *J. Mol. Biol.*, **292**, 19–37.
36. Livak, K.J. and Schmittgen, T.D. 2001, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ Method, *Methods*, **25**, 402–8.
37. Gupta, A., Lloyd-Price, J., Oliveira, S.M.D., Yli-Harja, O., Muthukrishnan, A.-B. and Ribeiro, A.S. 2014, Robustness of the division symmetry in *Escherichia coli* and functional consequences of symmetry breaking, *Phys. Biol.*, **11**, 066005.
38. Chowdhury, S., Kandhavelu, M., Yli-Harja, O. and Ribeiro, A.S. 2013, Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE), *BMC Bioinformatics*, **14**, S8.
39. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J.M. and Ribeiro, A.S. 2013, CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*, *Bioinformatics*, **29**, 1708–9.
40. Häkkinen, A., Kandhavelu, M., Garasto, S. and Ribeiro, A.S. 2014, Estimation of fluorescence-tagged RNA numbers from spot intensities, *Bioinformatics*, **30**, 1146–53.
41. Häkkinen, A. and Ribeiro, A.S. 2015, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, *Bioinformatics*, **31**, 69–75.
42. Tran, H., Oliveira, S.M.D., Goncalves, N. and Ribeiro, A.S. 2015, Kinetics of the cellular intake of a gene expression inducer at high concentrations, *Mol. Biosyst.*, **11**, 2579–87.
43. Johansson, H.E., Dertinger, D., LeCuyer, K.A., Behlen, L.S., Greef, C.H. and Uhlenbeck, O.C. 1998, A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein, *Proc. Natl Acad. Sci. USA*, **95**, 9244–9.
44. Golding, I. and Cox, E.C. 2006, Physical Nature of Bacterial Cytoplasm, *Phys. Rev. Lett.*, **96**, 98102.
45. Bernstein, J.A., Khodursky, A.B., Pei-Hsun, L., 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**, 9697–702.
46. Saecker, R.M., Record, M.T. and 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**, 754–71.
47. 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**, 3019–25.
48. Chamberlin, M.J. 1974, The selectivity of transcription, *Annu. Rev. Biochem.*, **43**, 721–75.
49. Hsu, L.M. 2009, Monitoring abortive initiation, *Methods*, **47**, 25–36.
50. Mulligan, M.E., Hawley, D.K., Entriken, R. and McClure, W.R. 1984, *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity, *Nucleic Acids Res.*, **12**, 789–800.
51. Bai, L., Santangelo, T.J. and Wang, M.D. 2006, Single-molecule analysis of RNA polymerase transcription, *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 343–60.
52. Wang, F. and Greene, E.C. 2011, Single-molecule studies of transcription: From one RNA polymerase at a time to the gene expression profile of a cell, *J. Mol. Biol.*, **412**, 814–31.
53. Artsimovitch, I. and Landick, R. 2000, Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals, *Proc. Natl Acad. Sci. USA*, **97**, 7090–5.
54. Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O. and Ribeiro, A.S. 2010, Effects of transcriptional pausing on gene expression dynamics, *PLoS Comput. Biol.*, **6**, e1000704.
55. Bar-Nahum, G. and Nudler, E. 2001, Isolation and characterization of σ^{70}-retaining transcription elongation complexes from *Escherichia coli*, *Cell*, **106**, 443–51.
56. Grigorova, I.L., Phleger, N.J., Mutalik, V.K. and Gross, C.a. 2006, Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA, *Proc. Natl Acad. Sci. USA*, **103**, 5332–7.
57. Maeda, H., Fujita, N. and Ishihama, A. 2000, Competition among seven *Escherichia coli* σ subunits: relative binding affinities to the core RNA polymerase, *Nucleic Acids Res.*, **28**, 3497–503.
58. Kass, R.E. and Raftery, A.E. 1995, Bayes Factors, *J. Am. Stat. Assoc.*, **90**, 773–95.
59. Hsu, L.M. 2002, Promoter clearance and escape in prokaryotes, *Biochim. Biophys. Acta*, **1577**, 191–207.
60. Casella, G. and Berger, R.L. 2001, *The Delta Method. Statistical Inference*, 2nd ed. Duxbury Press, Pacific Grove, CA, pp. 240–5.
61. Krystek, M. and Anton, M. 2008, A weighted total least-squares algorithm for fitting a straight line, *Meas. Sci. Technol.*, **19**, 79801.
62. Klumpp, S. and Hwa, T. 2008, Growth-rate-dependent partitioning of RNA polymerases in bacteria, *Proc. Natl Acad. Sci. USA*, **105**, 20245–50.

63. 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/I_1-I_2 regulatory elements, *Nucleic Acids Res.*, **25**, 1203–10.
64. Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S. and Hasty, J. 2008, A fast, robust and tunable synthetic gene oscillator, *Nature*, **456**, 516–9.
65. Martins, L., Mäkelä, J., Häkkinen, A., et al. 2012, Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time, *J. Theor. Biol.*, **301**, 83–94.
66. Mäkelä, J., Kandhavelu, M., Oliveira, S.M.D., et al. 2013, *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter, *Nucleic Acids Res.*, **41**, 6544–52.
67. Kandhavelu, M., Lihavainen, E., Muthukrishnan, A.B., Yli-Harja, O. and Ribeiro, A.S. 2012, Effects of Mg^{2+} on *in vivo* transcriptional dynamics of the *lar* promoter, *BioSystems*, **107**, 129–34.
68. Pedraza, J.M. and Paulsson, J. 2008, Effects of molecular memory and bursting on fluctuations in gene expression, *Science*, **319**, 339–43.
69. Casella, G. and Berger, R.L. 2001, *Simultaneous Estimation and Confidence Bands. Statistical Inference*, 2nd ed. Duxbury Press, Pacific Grove, CA, USA, pp. 559–63.
70. Schwarz, G. 1978, Estimating the dimension of a model, *Ann. Stat.*, **6**, 461–4.
71. Malan, T.P., Kolb, A., Buc, H. and McClure, W.R. 1984, Mechanism of CRP-cAMP activation of *lac* operon transcription initiation activation of the *P1* promoter, *J. Mol. Biol.*, **180**, 881–909.
72. 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**, 2712–23.
73. Sanchez, A., Choubey, S. and Kondev, J. 2013, Stochastic models of transcription: From single molecules to single cells, *Methods*, **62**, 13–25.
74. Schwabe, A., Rybakova, K.N. and Bruggeman, F.J. 2012, Transcription stochasticity of complex gene regulation models, *Biophys. J.*, **103**, 1152–61.

Supplement to "Dissecting the stochastic transcription initiation process in live *Escherichia coli*"

Jason Lloyd-Price, Sofia Startceva, Vinodh Kandavalli, Jerome G. Chandraseelan, Nadia Goncalves, Samuel M. D. Oliveira, Antti Häkkinen and Andre S. Ribeiro

I. Growth Curves



Supplementary Figure S1: Growth curves (OD_{600}, measured with an Ultraspec 10 cell density meter) of cells in 1x and 0.25x media (circles) at 37 °C. DH5α-PRO cells were grown overnight in 1x media at 30 °C with aeration of 250 rpm, and diluted into fresh 1x media to an initial OD_{600} of 0.05. Cells were incubated at 37 °C at 250 rpm until reaching the mid-log phase (~2 h), and re-diluted into the appropriate medium to an OD_{600} of 0.05. Their OD_{600} was measured every 10 minutes thereafter. At ~30 min, the cells in 0.25x media adjusted their growth rate (before this, the measurements overlap). Thus, growth rates were measured by least-squares fits (lines) from the data from 30 min onward. The slopes of the fits correspond to doubling times of 34.4 min (1.00x) and 57.9 min (0.25x).

II. Models of transcription initiation

To evaluate the cumulative distribution function (CDF) of the distribution of time intervals between production events from the full model of transcription initiation for a given value of R, we first translate this model into an observationally equivalent model of the form in equation 3. For the full model, this translation is given in the first row of Supplementary Table S1. The translated model's CDF can be evaluated using [1]. This CDF, when there are n steps after S_0, is referred to here as $F_{\mathrm{ON/OFF}+n}$. This distribution has a mean and variance of:

$$\mu_{\mathrm{ON/OFF}+n} = \frac{\lambda_{\mathrm{OFF}}}{\lambda_1 \lambda_{\mathrm{ON}}} + \sum_{i=1}^{n} \lambda_i^{-1} \tag{S1}$$

$$\sigma^2_{\mathrm{ON/OFF}+n} = \frac{\lambda_{\mathrm{OFF}}}{\lambda_1^2 \lambda_{\mathrm{ON}}} \left(2 + \frac{2\lambda_1 + \lambda_{\mathrm{OFF}}}{\lambda_{\mathrm{ON}}} \right) + \sum_{i=1}^{n} \lambda_i^{-2} \tag{S2}$$

Assumptions	CDF	λ_{ON}	λ_{OFF}	λ_1	λ_2	λ_3
	$F_{ON/OFF+3}$	k_{ON}	$\frac{-k_{ON}}{(Q_0-k_{ON})(Q_2-k_{ON})}$	$\frac{Q_1}{k_1 k_2}$	Q_1^{-1}	k_3
$k_{-1} \gg k_2$, $k_1 \gg k_{OFF}$	$F_{ON/OFF+2}$	k_{ON}	$\frac{k_{OFF}}{RK_a+1}$	$\frac{k_2 RK_a}{RK_a+1}$	k_3	
$k_2 \gg k_{-1}$	$F_{ON/OFF+3}$	k_{ON}	k_{OFF}	Rk_1	k_2	k_3
$k_{ON} \gg k_1$	$F_{Hypo(3)}$			$\frac{u+v}{2}$	$\frac{u-v}{2}$	k_3
$k_{ON} \gg k_1$, $k_{-1} \gg k_2$	$F_{Hypo(2)}$			$\frac{k_2 RK_a}{RK_a+1}$	k_3	
$k_{ON} \gg k_1$, $k_2 \gg k_{-1}$	$F_{Hypo(3)}$			Rk_1	k_2	k_3

Supplementary Table S1: Relation between kinetic parameters from equations (1) and (2) of the main manuscript with the parameters of the model from equation (3), for a given value of R. Here, $K_a = k_1 k_{-1}^{-1}$, $u = Rk_1 + k_{-1} + k_2$, $v = \sqrt{(k_{-1} + k_2 - Rk_1)^2 + 4Rk_1 k_{-1}}$, and Q_n are the roots of $-x^3 + bx^2 - cx + d$ *, where $b = u + k_{ON} + k_{OFF}$, $c = uk_{ON} + k_{OFF}(k_{-1} + k_2) + Rk_1 k_2$ $d = Rk_1 k_2 k_{ON}$, ordered such that $\lambda_{OFF} \geq 0$.

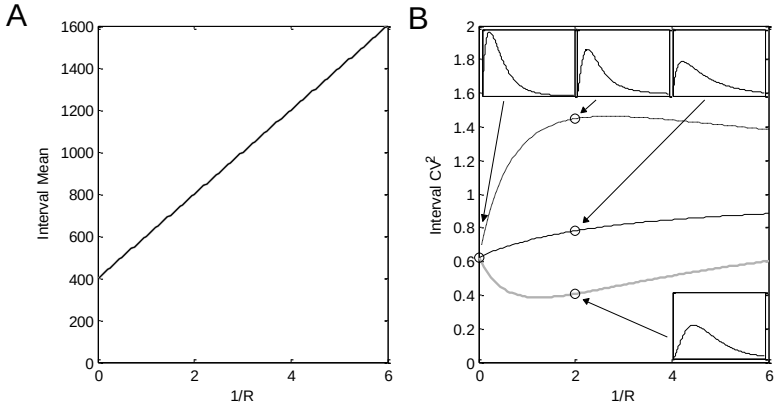
In the manuscript, several limiting cases of this model are considered. The first is that the ON/OFF mechanism is fast relative to initiation, i.e. $k_{ON} \gg k_1$. In this case, the model's CDF simplifies to that of a hypoexponential distribution with three exponentials with rates λ_1, λ_2 and λ_3, which relate to the parameters of 0 as shown in the fourth row of Supplementary Table S1. The hypoexponential CDF with n exponentials is referred to here as $F_{Hypo(n)}$.

Two further simplifications are considered, referred to in the manuscript as Limiting Mechanisms I and II. Both of these result in models with CDFs that are equivalent to either $F_{ON/OFF+n}$ or $F_{Hypo(n)}$. The parameters of the CDFs of the models derived from these three simplifying assumptions are presented in Supplementary Table S1. The final model simplification considered in the manuscript is when $k_3 = \infty$, i.e. when there is no rate-limiting third step in initiation, which removes the step parameterized by k_3 from the model.

The model of transcription initiation predicts the same linear change in the mean interval duration with $1/R$, regardless of the model simplifications (Figure S2A). However, the different simplifications result in different distributions of intervals as a function of $1/R$, which will differ in, e.g., noise (Figure S2B).

---

* Q_n can be evaluated with $Q_n = -2\sqrt{p}\cos\left[\frac{1}{3}\left(\cos^{-1}\left(\frac{-q}{2p^{3/2}}\right) - 2\pi n\right)\right] + \frac{b}{3}$, where $p = \frac{b^2 - 3c}{9}$ and $q = \frac{b}{3}\left(\frac{9b^2}{2} - c\right) + d$.



Supplementary Figure S2: Model prediction for (A) mean and (B) CV^2 of intervals as a function of 1/R with assumptions $k_2 \gg k_{-1}$ (dashed black line, $k_{\mathrm{ON}}^{-1} = 1000$, $k_{\mathrm{OFF}}^{-1} = 200$, $k_1^{-1} = 200 k_{\mathrm{ON}} \left(k_{\mathrm{ON}} + k_{\mathrm{OFF}}\right)^{-1}$, $k_2^{-1} = 300$, $k_3^{-1} = 100$), $k_{\mathrm{ON}} \gg k_1$, $k_{-1} \gg k_2$ (black lines, $K_a = 1.5$, $k_2^{-1} = 300$, $k_3^{-1} = 100$), and $k_{\mathrm{ON}} \gg k_1$, $k_2 \gg k_{-1}$ (grey lines, $k_1^{-1} = 200$, $k_2^{-1} = 300$, $k_3^{-1} = 100$). Note that in (A), all three lines overlap. Interval distributions for several parameter sets are shown in the insets of (B) (the axes of the insets are the same).

III. Parameter Estimation

Model parameter estimation was performed using a censored log-likelihood objective function as in [1], which accounts for uncertainty in the measurement of R, and for the uncertainty in the interval durations that arises from the limited framerate of the measurements and from the limited observation time:

$$\log L(\boldsymbol{\theta}) = \sum_m \mathbb{E} \log L_m\left(\boldsymbol{\theta}; R^{-1}\right) \tag{S3}$$

where $\mathbb{E}$ is the expectation over R^{-1}, and the conditional log-likelihood for condition m at relative RNAp concentration R is:

$$\begin{aligned} \log L_m(\boldsymbol{\theta}; R^{-1}) = &\sum_i \log\left[F_{\mathcal{M}}(t_{m,i} + T_M; \boldsymbol{\theta}, R^{-1}) - F_{\mathcal{M}}(\max(0, t_{m,i} - T_M); \boldsymbol{\theta}, R^{-1}) \right] \\ &+ \sum_i \log\left[1 - F_{\mathcal{M}}(c_{m,i}; \boldsymbol{\theta}, R^{-1}) \right] \end{aligned} \tag{S4}$$

where $F_{\mathcal{M}}\left(x; \boldsymbol{\theta}, R^{-1}\right)$ is the CDF of the model being fit (either $F_{\mathrm{ON/OFF}+n}$ or $F_{\mathrm{Hypo}(n)}$) with parameters translated as appropriate using Supplementary Table S1, $\boldsymbol{\theta}$ is the parameter vector, $t_{m,i}$ are measured intervals in condition m, T_M is the time between frames, and $c_{m,i}$ are the right-censored intervals.

The expectation of $\log L_m\left(\boldsymbol{\theta}; R^{-1}\right)$ over R in equation (S3) accounts for the uncertainty in the measurement of R. This was performed with $R^{-1} \sim \mathcal{N}\left(\hat{R}_m^{-1}, \sigma^2\left(\hat{R}_m^{-1}\right)\right)$, which was approximated by

evaluating the conditional log likelihood at 21 equally-spaced points in the interval $\left[\hat{R}_m^{-1}-3\sigma\left(\hat{R}_m^{-1}\right),\hat{R}_m^{-1}+3\sigma\left(\hat{R}_m^{-1}\right)\right]$.

Fitting was performed using the 'fminsearch' function in Matlab, with multiple restarts, to ensure that a local minimum was not selected. Each restart was started randomly in the parameter subspace where the model's mean interval at $R=1$ matched the corresponding measured mean interval.

The Bayesian Information Criterion (BIC) was used to compare models. We selected it over other candidates, such as the Akaike Information Criterion (AIC), due to its consistency. That is, as the number of samples $n\to\infty$, the probability that the BIC will select the true model (assuming it is among the candidate models) approaches 1, while the AIC will tend to over-fit the data[2]. We note, however, that in the case of all model comparisons in the manuscript, none of the conclusions are altered by utilizing the AIC over the BIC.

The BIC is calculated as follows:

$$\mathrm{BIC}=-2\log L(\boldsymbol{\theta}_{\max})+\log n \tag{S5}$$

where $\boldsymbol{\theta}_{\max}$ is the parameter set which maximizes $\log L(\boldsymbol{\theta})$.

IV. Number of transitions into the OFF state per RNA production event

In this section, we estimate the number of times that, on average, a promoter will transit into the OFF state for each time it commits to transcription. This estimation is made for the best fitting model (see Table 2 in the main manuscript).

For the best-fitting model (Limiting Mechanism I), the back-and-forward transitions between P_{ON} and RP_c states can be considered to be fast (since $k_{-1}>>k_2$ and $k_1>>k_{\mathrm{OFF}}$). We can therefore apply the slow-scale SSA to merge these two states[3]. In this limit, the probabilities $\mathrm{P}(P_{ON})$ and $\mathrm{P}(P_c)$ of being in P_{ON} and P_c states, respectively, are:

$$\mathrm{P}\left(P_{ON}\right)=\frac{1}{K_a+1}\ \text{and}\ \mathrm{P}\left(P_c\right)=\frac{K_a}{K_a+1},\ \text{with}\ K_a=k_1k_{-1}^{-1} \tag{S6}$$

The propensity of changing from the merged state to RP_o is then $\left[\mathrm{P}\left(P_c\right)k_2\right]$, while the propensity to move from the merged state to P_{OFF} equals $\left[\mathrm{P}\left(P_{ON}\right)k_{OFF}\right]$. The probability of moving into P_c instead of P_{OFF} is therefore given by:

$$\mathrm{P}_{c/OFF}=\frac{\mathrm{P}\left(P_c\right)k_2}{\mathrm{P}\left(P_{ON}\right)k_{OFF}+\mathrm{P}\left(P_c\right)k_2} \tag{S7}$$

Since each attempt at transcription is independent in the model, and has a constant probability of committing at each attempt, the number of times that the systems changes into the OFF state prior to committing to transcription follows a geometric distribution with a probability of success of $\mathrm{P}_{c/OFF}$. The mean of this distribution is:

$$\mu = \frac{1 - \mathrm{P}_{c/OFF}}{\mathrm{P}_{c/OFF}} \tag{S8}$$

Converting this in terms of model parameters (and given $k_1 k_{-1}^{-1} k_{\mathrm{OFF}}^{-1} = 0.11$ from Table 2) one obtains:

$$\mu = \frac{k_{OFF}}{k_2 K_a} = \frac{1}{0.11} \tag{S9}$$

V. Minimum samples required for a given precision

To estimate the number of samples required to obtain a given precision in the estimates of $\tau_{\overline{\mathrm{CC}}}$ and τ_{CC}, consider the following alternate method of measuring these values if we could sample the uncensored interval distribution between transcription events.

Let these measurements be at two RNAp concentrations $\hat{R}_m$, where $m = \{1,2\}$ such that $D = \hat{R}_1 / \hat{R}_2 > 1$. Let I_m be the population mean of the inter-transcription intervals in medium m, with corresponding standard deviation σ_m, and that we have n_m samples of this distribution (we assume, without significant loss of generality, that $n_1 = n_2 = n$). For sufficient n, estimates of the population means $\hat{I}_m$ will follow Normal distributions with $\sigma^2\left(\hat{I}_m\right) = \sigma_m^2 / n$. The least-squares fit of a line to these points will thus result in:

$$\hat{\tau}_{\overline{\mathrm{CC}}} = \frac{\hat{I}_1 D - \hat{I}_2}{D-1},\ \sigma\left(\hat{\tau}_{\overline{\mathrm{CC}}}\right) = \frac{D^2\sigma_1^2 + \sigma_2^2}{n(D-1)^2} \tag{S10}$$

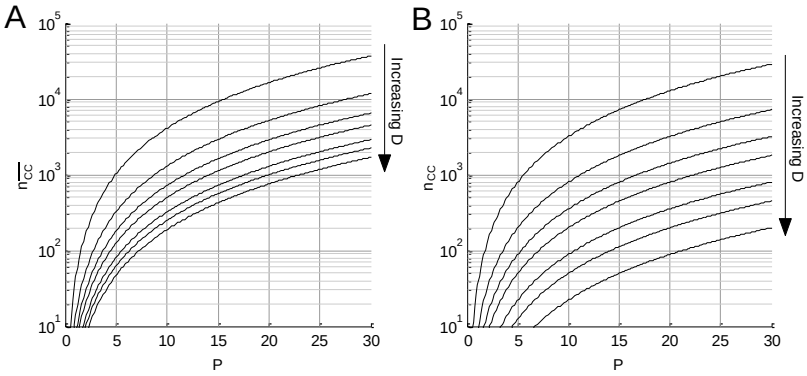
$$\hat{k}_{\mathrm{CC}}^{-1} = \frac{\hat{I}_2 - \hat{I}_1}{D-1},\ \sigma\left(\hat{k}_{\mathrm{CC}}^{-1}\right) = \frac{\sigma_1^2 + \sigma_2^2}{n(D-1)^2} \tag{S11}$$

Note that this method will overestimate the uncertainty in $\hat{\tau}_{\overline{\mathrm{CC}}}$ and $\hat{k}_{\mathrm{CC}}^{-1}$ since these estimates are highly anti-correlated. We define the precision of the measurement as $P = I_1 / \sigma(\hat{\tau}_x)$, where $\hat{\tau}_x$ is $\hat{\tau}_{\overline{\mathrm{CC}}}$ or $\hat{k}_{\mathrm{CC}}^{-1}$. Intuitively, this definition relates the uncertainty in the estimate with the mean timescale of the intervals. For example, if the intervals are on a timescale of ~500 s, to achieve a precision of 10 in $\hat{\tau}_{\overline{\mathrm{CC}}}$, we must know it to within 50 s. Assuming that $\sigma_1^2 I_1^{-2} \approx \sigma_2^2 I_2^{-2} = \eta^2$, i.e. that the CV^2 of the interval distribution is similar between the two RNAp concentrations, the number of samples required to achieve a given precisions in $\hat{\tau}_{\overline{\mathrm{CC}}}$ and $\hat{k}_{\mathrm{CC}}^{-1}$ is:

$$n_{\overline{\mathrm{CC}}} = \eta^2 P^2 \frac{D^2 + 1}{(D-1)^2} \tag{S12}$$

$$n_{\mathrm{CC}} = \eta^2 P^2 \frac{2}{(D-1)^2} \tag{S13}$$

Note that the above assumes that there is no variance in the estimate of the RNAp concentration, and that all n samples are uncensored. Equations (S12) and (S13) should therefore be considered as only a rough guide for the number of samples required. The number of samples required for a range of precisions and possible dynamic ranges in RNAp concentrations is shown in Supplementary Figure S3.



Supplementary Figure S3: Number of samples required in two conditions to achieve a given precision in (A) $\hat{\tau}_{\overline{CC}}$ and (B) $\hat{k}_{CC}^{-1}$, with production interval measurements at only two RNAp concentrations with ratio D and assuming $\eta^2 = 1$. Lines are shown for values of D of 1.25, 1.5, 1.75, 2, 2.5, 3, and 4 (from top to bottom).

VI. Photo-toxicity measurements

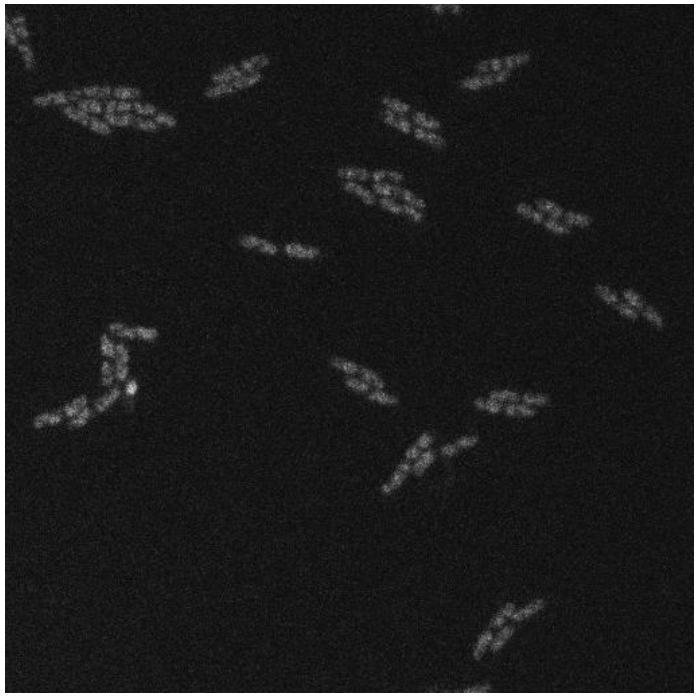
To assess the level of phototoxicity from the imaging procedure under the microscope, we took the measurements in the 1.00x case (Table 1, main manuscript), and estimated the cells' doubling time under the microscope by counting the number of cells at the start and end of the two hour measurement period (first row of Table S2). In this case, cells were imaged by phase contrast every 5 minutes, and confocal microscopy every minute for two hours. We then imaged two new populations of cells, but in the first, we only imaged the cells with phase contrast (i.e. no confocal, row 2 of Table S2), while in the second, only two images were taken in total, one at the start and one at the end (row 3 of Table S2).

Phase Contrast	Confocal	Cells at start	Cells at end	Doubling Time
5 min	1 min	206	468	52.8 min
5 min	Not used	399	962	49.8 min
2 h	Not used	480	1189	48.4 min

Supplementary Table S2: Phototoxicity under the microscope for different imaging intervals and channels. All measurements took 2 hours. The first two columns of the table show the intervals at which images were taken. The subsequent columns show the number of cells at the start and end of the measurements, obtained from single phase contrast images. Finally, it is shown the estimated doubling time of the cells, which was determined from the fold change.

From Supplementary Table S2, the estimated doubling time while taking images with both channels is only 4.4 minutes longer than in the case with minimal imaging. Thus, while there is an observable effect on the doubling time, it is not expected to cause significant differences in the transcription initiation dynamics. In any case, any changes would affect all conditions similarly, and will not affect *relative* RNAp concentrations. Finally, we note that the effect from phase contrast imaging appears to be negligible.

VII. Cell-to-cell variability in RNAp concentrations



Supplementary Figure S4: Confocal image of RL1314 cells expressing fluorescently-tagged RpoC in 1x media, one hour after being placed in the thermal imaging chamber at 37 ^{o}C. Contrast was enhanced for easier visualization.

VIII. Number of promoter copies during the cell lifetime

The model fitting procedure employed in the main text assumes that there is only one copy of the target promoter in a cell at all times. To determine to what extent this assumption is not true in our experimental system, we measured the fraction of time cells contain two chromosomes. Since the F-plasmid replicates at the same time[4] or shortly after[5] the chromosome, this provides an upper bound for the fraction of time the cells spend with more than one promoter of interest (it is worth noting that, in our measurements, we did not observe cells with more than 2 nucleoids at any given point).

For this, *E. coli* DH5α-PRO cells (see main text) were transformed with the pAB332 plasmid carrying the gene *hupA-mcherry* that encodes a fluorescent protein tag under the control of the *hupA* constitutive promoter[6]. This tagging protein, composed of a nucleoid-associated protein (HupA) fused with a red fluorescent protein (mCherry), can be used to assess the location and size of nucleoids in live cells[7] (see Methods).

Cells were diluted from overnight culture to an OD_{600} of 0.05 in fresh 1x media, supplemented with appropriate antibiotics, and kept at 37°C in a shaker at 250 rpm, until reaching an OD_{600} of 0.3. Cells were then placed in a thermal chamber (FCS2, Bioptechs, USA), set to 37°C, and imaged once every minute for 1 hour (the red signal was too weak to continue after 1 hour) using a Nikon Eclipse (Ti-E, Nikon) inverted microscope equipped with C2+ (Nikon) confocal laser-scanning system. To visualise HupA-mCherry-tagged nucleoids, we used a 543 nm HeNe laser (Melles-Griot) and an emission filter (HQ585/65, Nikon). Phase contrast images of cells were captured every 5 minutes by a CCD camera (DS-Fi2, Nikon).

Cells were segmented from phase contrast images using CellAging[8]. Fluorescent nucleoids were segmented and quantified from confocal images as in [7,9]. Of the cells that were born and divided during the time series (124 cells), we found that the mean fraction of time points in which cells had two nucleoids was 0.114 ± 0.010.

Thus, we estimate the fraction of time spent with multiple target promoters to be at most 11.4 ± 1.0% in 1x media. As this was the most nutrient-rich condition tested, other conditions should have even lower fractions[5].

References

1. Häkkinen, A., and Ribeiro, A. S. 2015, Characterizing rate limiting steps in transcription from RNA production times in live cells. *Bioinformatics*, in press. DOI: 10.1093/bioinformatics/btv744.

2. Burnham, K. P., and Anderson, D. R. 2004, Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.*, **33**, 261–304.

3. Cao, Y., Gillespie, D. T., and Petzold, L. R. 2005, The slow-scale stochastic simulation algorithm. *J. Chem. Phys.*, **122**, 14116.

4. Cooper, S., and Keasling, J. D. 1998, Cycle-specific replication of chromosomal and F plasmid origins. *FEMS Microbiol. Lett.*, **163**, 217–22.

5. Keasling, J. D., Palsson, B. Ø., and Cooper, S. 1991, Cell-cycle-specific F plasmid replication: Regulation by cell size control of initiation. *J. Bacteriol.*, **173**, 2673–80.

6. Fisher, J. K., Bourniquel, A., Witz, G., Weiner, B., Prentiss, M., and Kleckner, N. 2013, Four-dimensional imaging of *E. coli* nucleoid organization and dynamics in living cells. *Cell*, **153**, 882–95.

7. Oliveira, S. M. D., Neeli-Venkata, R., Goncalves, N. S. M., et al. 2016, Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli*. *Mol. Microbiol.*, **99**, 686–99.

8. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., and Ribeiro, A. S. 2013, CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, **29**, 1708–9.

9. Mora, A. D., Vieira, P. M., Manivannan, A., and Fonseca, J. M. 2011, Automated drusen detection in retinal images using analytical modelling algorithms. *Biomed. Eng. Online*, **10**, 59.

Publication III

N.S.M. Goncalves, S. Startceva, C.S.D. Palma, M.N.M. Bahrudeen, S.M.D. Oliveira and A.S. Ribeiro "Temperature-dependence of the single-cell variability in the kinetics of transcription activation in *Escherichia coli*", *Physical Biology*

Temperature-dependence of the single-cell variability in the kinetics of transcription activation in *Escherichia coli*

Nadia S.M. Goncalves[1], Sofia Startceva[1], Cristina S.D. Palma[1,2], Mohamed N.M. Bahrudeen[1], Samuel M.D. Oliveira[1] and Andre S. Ribeiro[1,2,3,4]

Running Head: Temperature-dependence of transcription activation in *E. coli*

Keywords: Single-cell intake kinetics; gene expression activation times; *Escherichia coli*; critically low temperatures;

Accepted for Publication: 28 November 2017

doi: 10.1088/1478-3975/aa9ddf/meta

Link for the Version of Record of this article:
http://iopscience.iop.org/article/10.1088/1478-3975/aa9ddf/meta

[1] Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101, Tampere, Finland.
[2] CA3 CTS/UNINOVA. Faculdade de Ciencias e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516, Caparica, Portugal.
[3] Multi-scaled Biodata Analysis and Modelling Research Community, Tampere University of Technology, 33101, Tampere, Finland.
[4] Corresponding author. E-mail: andre.ribeiro@tut.fi, Tel: +358408490736.

Abstract

From *in vivo* single-cell, single-RNA measurements of the activation times and subsequent steady-state active transcription kinetics of a single-copy Lac-ara-1 promoter in *Escherichia coli*, we characterize the intake kinetics of the inducer (IPTG) from the media, following temperature shifts. For this, for temperature shifts of various degrees, we obtain the distributions of transcription activation times as well as the distributions of intervals between consecutive RNA productions following activation in individual cells. We then propose a novel methodology that makes use of deconvolution techniques to extract the mean and the variability of the distribution of intake times. We find that cells, following shifts to low temperatures, have higher intake times, although, counter-intuitively, the cell-to-cell variability of these times is lower. We validate the results using a new methodology for direct estimation of mean intake times from measurements of activation times at various inducer concentrations. The results confirm that *E. coli*'s inducer intake times from the environment are significantly higher, following a shift to a sub-optimal temperature. Finally, we provide evidence that this is likely due to the emergence of additional rate-limiting steps in the intake process at low temperatures, explaining the reduced cell-to-cell variability in intake times.

Introduction

RNA and protein numbers differ between cells of monoclonal populations, due to the stochastic nature of the chemical reactions composing gene expression ('intrinsic' noise) [1,2] and the cell-to-cell variability in the numbers of the molecules involved ('extrinsic' noise) [3].

Besides these 'constant' sources of cell-to-cell variability, recent studies have shown that, following the appearance of an inducer of gene expression in the media, there is an additional transient cell-to-cell diversity in RNA and protein numbers of the target gene [4–6], which cannot be explained by the intrinsic and extrinsic noise of active gene expression. This additional source can be strong enough and the transient long enough to affect the phenotypic diversity of cell lineages for generations [4–12].

The origin of this transient phenotypic diversity has been shown to be the noise in the intake time of the inducers, which causes the time for transcription to be activated (following the introduction of the inducers in the media) to differ widely between cells [5]. At the RNA numbers level, this transient diversity can be higher than the diversity caused by the intrinsic and extrinsic noise in active transcription for long periods of time [5].

Similarly to noise in gene expression, noise in intake times has two sources. One is the stochasticity of the intake process, caused by the random nature of the chemical reactions and the membrane crossing processes [2,6]. The other is likely a non-negligible degree of cell-to-cell heterogeneity in the efficiency of the mechanisms involved in the intake of inducers [5]. This heterogeneity can be caused by, among other, cell-to-cell diversity in the number of transmembrane proteins involved in the active uptake of

inducer/repressor molecules [5]. One example is the lactose permease (LacY), which, while being produced by an all-or-nothing system that minimizes cellular heterogeneity, it nevertheless exhibits significant cell-to-cell diversity in numbers, following the appearance of the inducer (e.g. TMG) in the media [13].

As natural environmental conditions fluctuate and many genes in *E. coli* are only activated in specific conditions, cellular heterogeneity in gene expression activation times is expected to affect significantly the phenotypic diversity of cell populations.

One environmental parameter that we expect to have a tangible impact on both the mean and variability of intake times of external inducers and repressors of gene expression is temperature. This assumption originates from the fact that temperature affects not only proteins functionality and numbers in cells [14], but also the physical properties of cell walls, periplasm and cytoplasm (e.g. the cytoplasm's viscosity is temperature dependent [15]), and these variables are expected to affect the kinetics of intake of inducers from the environment.

However, there is yet no direct experimental validation and, as many variables are involved, model-based predictions of the quantitative degree of changes with temperature in inducers intake times and subsequent transcription initiation times are unreliable.

Here, we characterize quantitatively the changes in cell-to-cell variability in gene expression activation times of the Lac-ara-1 promoter and, more importantly, of the intake times of its inducer, Isopropyl β-D-1-thiogalactopyranoside (IPTG), caused by rapid physical changes following temperature shifts.

For this, we use time-lapse microscopy measurements of RNA production at the single-cell, single-RNA level at various temperatures, along with several recently developed techniques [6,14,16], including a new strategy here proposed to dissect the kinetics of the intake process. Our results provide novel information for the understanding of the effects of temperature shifts of bacterial populations at the single-cell level.

Methods

Bacterial strains and plasmids

We use *E. coli* strain DH5α-PRO, generously provided by I. Golding, University of Illinois, U.S.A. The genotype is deoR, endA1, gyrA96, hsdR17(rK^- mK^+), recA1, relA1, supE44, thi-1, Δ(lacZYA-argF)U169, Φ80δlacZΔM15, F^-, λ^-, P_{N25}/tetR, P_{lacIq} /lacI, SpR. The strain contains two genes, *lacI* and *tetR*, constitutively expressed under the control of $P_{laci}{}^q$ and P_{N25} promoters, respectively [17]. Relevantly, the native lac operon (*lacZYA*) is mutated, to prevent production of permease (*lacY*) and activation of the lactose metabolic system [18]. I.e., these cells lack the native positive feedback mechanism involving lactose [6,19].

In addition to this strain, we also use *E. coli* JW0334 strain. The genotype is F^- (Δ (araD-araB)567 ΔlacY784 ΔlacZ4787(::rrnB-3) λ^-rph-1 Δ(rhaD-rhaB)568 hsdR514) [18].

This strain also lacks the ability to produce lacY [18]). Here, we only make use of this strain to show that the changes in the target gene activation time with temperature are, qualitatively, only weakly strain dependent. Unless stated otherwise, measurements are made using DH5α-PRO cells.

Both strains lack the ability to express lacY permease [18], which is responsible for a feedback response to the intake of IPTG, which would result in more complex, time-dependent single-cell intake times, as they would not be solely determined by the induction level and temperature.

Two constructs were added to DH5α-PRO cells: pROTET-K133 with $P_{LtetO-1}$-MS2d-GFP and pIG-BAC, a single-copy plasmid with $P_{Lac-ara-1}$-mRFP1-MS2d-96bs [20] (Figure 1). In the case of JW0334 cells, another reporter is used (P_{RHAM}-MS2d-GFP), as these cells lack the ability to express TetR.

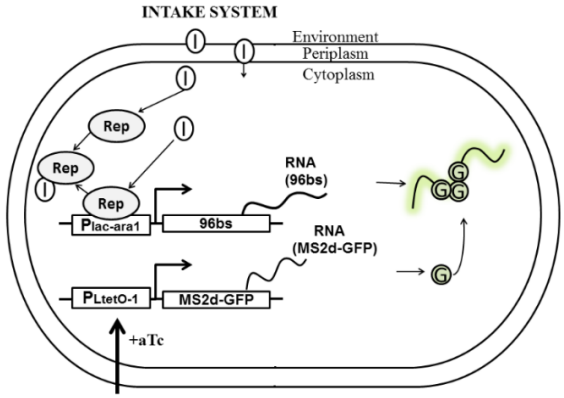


Figure 1. Diagram of the target gene and its RNA tagging system, along with the intake system of inducers of the target gene: IPTG molecules (I) are added to the media and enter the cytoplasm by passing through two membrane layers, with a periplasmic space in between. When in the cytoplasm, they neutralize lacI repressors (R) by forming inducer-repressor complexes (RI). This allows $P_{Lac-ara-1}$ to express RNAs that include an array of 96 MS2d-binding sites. Meanwhile, MS2d-GFP expression is controlled by the $P_{LtetO-1}$ promoter and anhydrotetracycline (aTc). Once produced, each target RNA is rapidly bound by multiple tagging MS2d-GFP proteins (G), and appears as a bright spot, significantly above background fluorescence, under the confocal microscope [6,20]. The tagging provides the RNA a long lifetime, with constant fluorescence, beyond our observation times [6].

Finally, it is noted that previous measurements [6] have shown that, provided full induction of the reporter gene (1 hour) prior to induction of the target gene, any newly produced target RNA molecule becomes ‘fully fluorescent’ (i.e. its RNA MS2-GFP binding

sites become fully occupied) in less than 1 minute. These measurements were conducted in the same strain and media employed here. Given this, and since our microscopy time-lapse images are separated by 1 minute intervals, it is reasonable to assume that, once a new RNA appears, the full occupation of its MS2-GFP binding sites will take less time than the time between two consecutive images. This is agreement with measurements in [21].

Growth Conditions, Microscopy, Data Extraction on Transcription Activation Times

Cells were grown overnight at 30 °C with aeration and shaking in lysogeny broth (LB) medium, supplemented with the appropriate antibiotics (35 μg/ml Kanamycin and 34 μg/ml Chloramphenicol). From the overnight cultures, cells were diluted into fresh LB medium, supplemented with antibiotics, to an optical density of $OD_{600} \approx 0.05$, and allowed to grow at 37 °C, 250 rpm, until reaching an $OD_{600} \approx 0.3$. Next, 100 $ng.ml^{-1}$ anhydrotetracycline (aTc) was added to induce $P_{LtetO-1}$ and produce MS2d-GFP, and 0.1% L-Arabinose to pre-activate the target gene, controlled by $P_{Lac-ara-1}$ [17,20]. Afterwards, cells were centrifuged (8000 rpm, for 1 minute), and re-suspended in the remaining LB medium. From this, a few microliters of cells were taken and placed between a 3% agarose gel pad and a glass coverslip, before assembling the FCS2 imaging chamber (Bioptechs, see Figure S1). Finally, the chamber was heated to the desired temperature (24 °C, 30 °C, 37 °C and 41 °C) and placed under the microscope.

We observed that, in the absence of IPTG, the cells produce the same (spurious) amount of RNA, with or without Arabinose (data not shown), in agreement with previous studies [20]. However, pre-induction by Arabinose much prior to induction by IPTG, enhances slightly the RNA production rate [16,18]. As such, we pre-induced cells with Arabinose [17,20] 45 minutes prior to introducing IPTG in the media. As such, we pre-induced cells with Arabinose [17]. This implies that, by the time IPTG is added, the cells already contain a constant amount of Arabinose. This is ensured by the presence of Arabinose in the original media and by the constant replenishment of this media during microscopy measurements (Methods and Figure S1). Thus, we do not expect any potential feedback mechanism associated to the Arabinose intake process to influence the transcription activation times measured here, following the introduction of IPTG in the media.

Cells were visualized by a 488 nm argon ion laser (Melles-Griot), and an emission filter (HQ514/30, Nikon) using a Nikon Eclipse (Ti-E, Nikon) inverted microscope with a 100x Apo TIRF (1.49 NA, oil) objective. Fluorescence images were acquired by C2+ (Nikon), a point scanning confocal microscope system, and Highly Inclined and Laminated Optical sheet (HILO) microscopy, using an EMCCD camera (iXon3 897, Andor Technology). The laser shutter was open only during exposure time to minimize photobleaching. All images were acquired with NIS-Elements software (Nikon). While imaging, cells were supplied with a constant flow of fresh LB medium (pre-warmed to the

same temperature as in the chamber), containing 1 mM of IPTG, 0.1% of L-Arabinose, and 100 ng.ml^{-1} of aTc, using a peristaltic pump (Bioptechs), at a rate of 0.1 mL min^{-1}. Images were taken once per minute for 2.5 hours. At each moment, we imaged 6 specific locations, to attain information on multiple lineages.

After performing a semi-automated cell segmentation and lineage construction [22], the moment of production of the first RNA by each cell lineage was obtained by selecting cells absent of RNA spots at the start of the imaging period (i.e., without leaky expression), and then detecting by visual inspection (from fluorescence images) when the first production occurs in each branch of each lineage (Figure 2B), after introducing the inducers.

Aside from visual inspection, fluorescent RNA spots and their intensities were also detected from the confocal images using the Gaussian surface-fitting algorithm proposed in [23] specifically for the purpose of detecting and quantifying MS2-GFP tagged RNAs. We found no significant difference between using this automatic algorithm and the visual inspection of the moment when the first RNA appears.

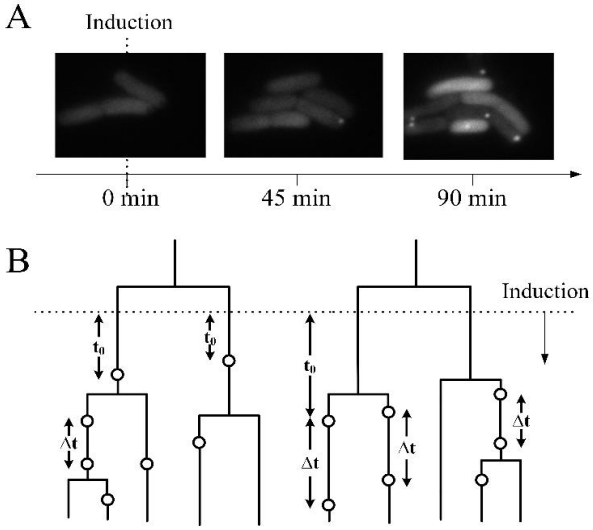


Figure 2. Data collection: (A) Cells are placed under the microscope at t=0 min and continuously supplemented with fresh medium. The reporter system (MS2d-GFP) is induced in liquid culture at t = -45 min. At t = 0 min, with the cells already having sufficient MS2d-GFP proteins for accurate RNA detection, transcription of the target RNA for MS2d-GFP is induced. (B) Illustration of RNA production events (circles) in cell lineages. Shown are the time for the first RNAs to appear (t_0) and the subsequent time intervals between consecutive RNA production events (Δt) in single cells. A dotted line indicates when the inducer of the target promoter is introduced.

As a side note, we found the rate of leaky expression to be very weak (less than 1 spot per ~20 cells prior to induction).

Finally, we note that the data on time intervals between consecutive RNA productions in individual cells used here was entirely obtained from [15]. There, time lapse microscopy was conducted on cells of the same strain, with the same constructs, and under the same induction and growth conditions as the ones used here.

Quantitative PCR for mean RNA quantification

Quantitative PCR (qPCR) was used to attain the induction curve of $P_{Lac\text{-}ara\text{-}1}$ as a function of IPTG concentration at 37 °C (for details, see Supplementary Material). This induction curve is shown in Figure S2. Visibly, for 0.5 mM IPTG and above, $P_{Lac\text{-}ara\text{-}1}$ is fully induced.

Estimation of intake times by deconvolution from empirical data on activation times and active transcription interval duration

The empirical method of MS2-GFP tagging of RNA allows for new RNAs containing multiple MS2-GFP binding sites to be detected shortly after they are produced [20]. From this data, one can directly extract the time intervals between consecutive RNA productions in individual cells following induction, as well as the time for the first RNA to be produced once inducers are added in the media. However, one cannot directly measure the time that inducers take to enter the cells and activate the target promoter. To obtain this information, we next propose a methodology based on deconvolution techniques for extracting this information from the data.

Given the model above, the mean time for the first RNA to appear in a cell following the addition of inducers in the media (here named t_0) depends on the time for inducers to enter the cell (reactions (1-2) in Supplementary Material) [4,5], here denoted as t_{int}. Also, it depends on the time for RNA production by an active promoter (which depends on the rate-limiting steps in transcription) [24,25], determined by reactions (3-6) in Supplementary Material, and here represented by Δt since, under full induction, this time should equal the time between consecutive RNA productions in active promoters [5]. In particular, we have:

$$t_0 = t_{int} + \Delta t \tag{1}$$

As the inducer intake and the production of the first RNA are independent, consecutive processes, one can use deconvolution to obtain a distribution of values of t_{int} (and, thus, mean and variance) from the data. Namely, for each temperature, one can deconvolve the probability density function (PDF) of the Δt distribution from the PDF of the t_0 distribution, provided that these two distributions are known [26].

For this, we estimate the PDFs of Δt and t_0 distributions as their best-fitted gamma distributions to the respective empirical distributions. We choose the gamma distribution as a model, since such distributions allow the mean and the variance to change independently, thus facilitating the fitting to the empirical distribution [14].

First, we use the gamma fits to the empirical Δt distributions reported in a previous work [14]. This fit used censored intervals between productions of consecutive transcripts extracted from live-cell measurements. The censoring accounts for the effects of finite sampling rate (60 s sampling interval), and thus improves the accuracy of the parameter estimation [27]. It also accounts for right-censored intervals, to compensate for the truncation of the right tail of the Δt distribution due to the finite cell division times. This fitting follows the maximum likelihood criteria [14].

Afterwards, to the measured t_0 distributions, we apply the same censored fitting procedure, but without right-censoring (as t_0 durations are not restricted by cell lifetime). Finally, we obtain the PDF of the t_{int} distribution using the Fast Fourier Transform (FFT) deconvolution method, as proposed by Sheu and Ratcliff [26], except that we do not apply frequency filtering, since our estimated t_0 and Δt PDFs do not contain high-frequency noise. As outlined by Sheu and Ratcliff [26], the result of the deconvolution may contain negative values, even though the PDF, by definition, cannot have values below zero. Those negative values should be interpreted as resulting from the uncertainty in the best-fit gamma distributions to t_0 and Δt empirical data, which, in turn, originates from uncertainty in the t_0 and Δt measurements. However, even if the selected models do not precisely depict the PDFs of the corresponding processes, the results of the deconvolution are still interpretable, even though the uncertainty in the deconvolution product is undefined [26]. Here, to allow such interpretation, we set the negative values of the t_{int} PDF to zero.

To estimate the uncertainty of our findings, we constructed bootstrap 95% confidence intervals (CIs) for mean and noise of the t_{int} distribution using non-parametric resampling of t_0 and Δt empirical data [28,29]. For this, for each temperature condition, we perform 2000 random resamples with replacement of the t_0 and Δt empirical distributions (using an original amount of samples), and obtain the t_{int} PDF for each resampled pair of t_0 and Δt distributions, which then allows obtaining the bootstrap distributions of the mean and CV^2 (squared coefficient of variation) of the t_{int} PDFs. We take 0.05 and 0.95 percentiles of those distributions as the 95% CIs of the estimated mean and CV^2 of the t_{int} distribution.

Estimation of intake times using Lineweaver-Burk equation

Aside from the method above, we make use of the Lineweaver-Burk equation [30] to estimate mean intake times. For this, from (1) and the model of gene expression (reactions 1-6 in Supplementary Material), note that as the amount of inducers in the media is increased, in a first stage, the inducers inside the cell will increase in number. As such, during this stage, both t_{int} and Δt will decrease with increasing inducer concentration. However, beyond a certain concentration of inducers in the media, further increases in this concentration will no longer lead to increases in the rate of RNA production (i.e. when the regime of full induction is reached), due to the rate-limiting steps in transcription and the

finite number of RNA polymerases inside the cell (reaction 6 in Supplementary Material). This well-known fact is also demonstrated here by Figure S2, which shows that, beyond a certain inducer concentration (both in the microscopy measurements and in the qPCR measurements) the rate of RNA production no longer increases with further increases in the IPTG concentration in the media.

Meanwhile, the time taken by the cell to intake inducers should continue to decrease with increases in inducer concentration in the media, even in the regime of full induction of transcription. Namely, in theory, for an infinite amount of inducers in the media, t_{int} should equal zero. In this regime, following the introduction of infinite number of inducers in the media, the total mean time taken to produce the first RNA will be equal to the duration of subsequent intervals between consecutive RNA productions, i.e.:

$$t_0\left([\mathrm{IPTG}]=\infty\right)=\Delta t \tag{2}$$

Thus, provided that the decrease in t_{int} with the decrease of the inverse of the inducer concentration is linear (as assumed in our model reactions (1) and (2) in Supplementary Material), we can derive t_{int} in the 'control condition' using the Lineweaver-Burk equation [30] as follows:

$$t_{int}=\frac{[\mathrm{IPTG}]_2(t_{0_2}-t_{0_1})}{([\mathrm{IPTG}]_1-[\mathrm{IPTG}]_2)} \tag{3}$$

In equation (3), t_{0_1} and $[\mathrm{IPTG}]_1$ are, respectively, the mean t_0 and the inducer concentration in the control condition. Meanwhile, t_{0_2} and $[\mathrm{IPTG}]_2$ are the corresponding values in a condition where the inducer concentration differs from the control, and is above the minimum concentration to achieve maximum RNA production rate.

Also, one can calculate 95% CIs for the obtained mean t_{int} value based on the method of propagation of errors [31].

As a side note, this methodology is similar to the usage of τ plots, from which, by fitting a line to the results of measurements of the transcription rate for increasing RNA polymerase concentrations one can extract the duration of the events following the initiation of the open complex formation [16,24,32].

Inference of the number and duration of the sequential steps in the intake process by fitting with a sum of exponential steps

Our model of intake (reactions (1) and (2) in Supplementary Material) assumes 2 steps, each with a duration following an exponential distribution, in agreement with measurements at optimal temperatures [5,6]. However, as noted, our modelling strategy allows considering the possibility that, at different temperatures, additional or less steps may be rate-limiting.

To determine the number of steps, one can perform fittings of *d*-steps models (each step following an exponential distribution) for increasing number of steps, until adding a step no longer improves the fitting. In such a model, as more steps are added and if the overall mean duration of the d-steps process is kept constant, the variance of the durations between events will decrease. The closer the d-exponential steps distribution is to a gamma distribution with a shape parameter set to d, the smaller will be its variance.

The d-exponential step model was chosen due to how we model transcription, namely, as a set of consecutive of chemical reactions, each of which having a distribution of intervals between consecutive occurrences that is expected to follow an exponential distribution. Also, there is significant accumulated evidence that, in *E. coli*, this model fits very well, in a statistical sense, the empirical distributions of many promoters [5,6,16,33,34].

Here we perform this fitting to a *d*-steps model for each temperature condition. For this, by deconvolution of the empirical data, we obtain a distribution of the duration of the intake process. From it, we determine the maximum likelihood fit of a model with *d* statistically independent steps, whose time lengths each follow an exponential distribution, with possibly different rates.

The likelihoods are compared using the likelihood ratio test, and the model with smallest *d* that cannot be rejected at the significance level 0.01 is selected in favor of a higher order model.

Note that this method does not allow determining the order of the steps, only their number and durations. Note also that, while changing temperature may not alter the number of rate-limiting steps, it may instead (or also) cause them to no longer be well modeled by elementary reactions as our model assumes. In that case, we expect the fitting to d exponential steps to require a higher number of steps than if the steps were elementary.

Results and Conclusions

$P_{Lac\text{-}ara\text{-}1}$ transcription activation kinetics is temperature dependent

We first studied, at the single cell level, the temperature dependence of the kinetics of transcription activation of $P_{Lac\text{-}ara\text{-}1}$ by IPTG. All empirical data were obtained from observing individual cells over time, using MS2d-GFP tagging of the target RNA, fluorescence microscopy, and image analysis techniques (Methods).

For this, we placed *E. coli* cells (DH5α-PRO) with a single-copy plasmid coding for the RNA target for MS2d-GFP under the control of $P_{Lac\text{-}ara\text{-}1}$, and fully activated its expression by adding IPTG (1 mM) to the media (Figure S2) while already under microscope observation (Figure 2). The MS2d-GFP reporters, expressed by a multi-copy plasmid, were induced prior to this, so that cells were flooded with MS2d-GFP by the time $P_{Lac\text{-}ara\text{-}1}$ was induced (Methods).

From the time series obtained (~2.5 hour long, with images taken every minute), for each temperature, we extracted t_0, the time taken by individual cells to produce the first RNA, following the addition of inducers in the media (Methods). Note that only one such event per lineage is considered and that cells already with one or more RNAs at the start of the observation period were discarded.

From these data, we calculated the mean, standard error, and CV^2 of t_0 values. Finally, we performed Kolmogorov-Smirnov (KS) tests to compare each distribution of t_0 values with the distribution at 37 °C (named 'control' condition). Results are shown in Table 1.

Table 1. Measurements of t_0 vs temperature. Shown are the number of measurements (N_{t_0}), mean (μ_{t_0}) standard error (SE) and CV^2 of the distribution of t_0 values ($CV^2_{t_0}$). The table also shows the *p*-value from the KS tests comparing the t_0 distributions at each temperature, with the distribution at 37 ˚C (control). For *p*-values smaller than 0.01, the null hypothesis that the two sets of data are from the same distribution can be rejected.

T (˚C)	N_{t_0}	$\mu_{t_0} \pm$ SE (s)	$CV^2_{t_0}$	KS-test for t_0 values vs 37 ˚C (*p*-value)
24	93	2743 ± 102	0.13	< 0.01
30	162	3020 ± 119	0.25	< 0.01
37	60	2109 ± 215	0.63	-
41	93	2379 ± 144	0.34	0.19

From the data in Table 1, we find that for temperatures lower than 37 °C, the activation time t_0 differs significantly from the control (in a statistically sense), with its mean (μ_{t_0}) being higher and its $CV^2_{t_0}$ (surprisingly) being lower for lower temperatures.

Qualitatively similar results were obtained (Table S1) using the *E. coli* JW0334 strain (see section 'Bacterial strains and plasmids').

Cell-to-cell variability of t_{int} decreases with decreasing temperature

Next, we investigate how the time for inducers to enter the cell, t_{int}, changes with temperature. For this, besides the data above, we make use of the data from [14], which consists of empirical distributions of intervals between consecutive RNA productions by active promoters in individual cells (Δt), under the same temperature conditions as above. These data therefore informs on the kinetics of active transcription (i.e. is not affected by intake times).

As mentioned in Methods, in accordance to our model (reactions 1-6 in Supplementary Material) and equation 1, the time for the production of the first RNA in each cell, following the introduction of inducers in the media (t_0), should consist of the time for the intake of the inducer by the cell (t_{int}) and the time taken by the active promoter to produce the first RNA (Δt). As these processes are consecutive and independent, it should

be possible to obtain the time-length for intake of the inducers (t_{int}) by deconvolving Δt from t_0.

For this, we performed model fitting with censoring to the data from live-cell measurements of t_0 (Table 1) and used the model fitting of empirical Δt values from [14]. In Figure 3, we show the empirical distribution and the best gamma fits of t_0.

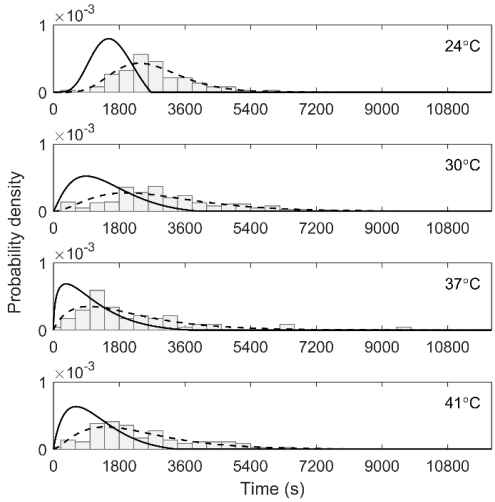


Figure 3. Empirical distribution of t_0 (histogram), along with the best gamma fit to t_0 (dashed line) and the deconvolved t_{int} (solid line), as function of temperature.

Next, we obtained the t_{int} distribution for each temperature condition from the deconvolution of Δt from t_0 (Methods). Results for the mean and CV^2 values of the distributions of t_{int} obtained from this deconvolution are shown in Table 2, along with the 95% CI. It is noted that the values at 37 °C are in agreement with previously reported measurements [5,6].

Meanwhile, the deconvolved distributions are shown in Figure 3. From these, we find a clear change in the shape of the t_{int} distribution as temperature is lowered.

Table 2. Mean and CV^2 of the deconvolved t_{int}, along with the 95% CI for each temperature condition.

T (˚C)	$\mu_{\widehat{t_{int}}}$ (s)	95% CI of $\mu_{\widehat{t_{int}}}$ (s)	$CV^2_{\widehat{t_{int}}}$	95% CI of $CV^2_{\widehat{t_{int}}}$
24	1548	[1316, 1799]	0.10	[0.06, 0.18]
30	1369	[1113, 1671]	0.32	[0.20, 0.48]
37	986	[726, 1329]	0.52	[0.28, 0.95]
41	1083	[807, 1402]	0.37	[0.23, 0.63]

From Table 2, we find that the mean duration of the intake process, $\mu_{\widehat{t_{int}}}$, is the lowest while the variability, $CV^2_{\widehat{t_{int}}}$, is the highest at 37 °C. Meanwhile, at the lowest temperature tested (24 °C) the opposite occurs ($\mu_{\widehat{t_{int}}}$ is the highest and $CV^2_{\widehat{t_{int}}}$ is the lowest).

Also, from the values of t_0 (Table 1) and t_{int} (Table 2), we find that the dynamics of intake plays a major role in the dynamics of transcription activation in all temperature conditions, both regarding the mean duration of activation and its cell-to-cell variability. Thus, it is not a surprise that t_{int} behaves similarly to t_0 with changes in temperature.

Finally, note that the fact that noise is reduced with decreasing temperature suggests that the process becomes more sub-Poissonian, which could occur, e.g., if the number of the rate-limiting steps in the intake process increases with decreasing temperature.

As a side note, we also conducted similar experiments in the absence of IPTG, so as to estimate the level of toxicity due to induction by 1 mM IPTG. We found no difference in cell growth rate between the two conditions, and thus conclude that the levels of toxicity are not significant.

Validation of the inferred mean t_{int} using the Lineweaver-Burk equation

It is possible to empirically validate the mean value of the deconvolved t_{int} using the Lineweaver-Burk equation (Methods). For this, from individual cells at 24 °C, 37 °C and 41 °C, we measured the time between the moment of induction and the moment when the first RNA is produced for IPTG concentrations of 1 mM and of 0.5 mM. Note that both of these concentrations suffice to reach maximum induction in cells under the microscope (as shown in Figure S2). Because of this, Δt does not differ between the two conditions, and only affects t_{int}. From the measurements of t_0 in these two induction levels at a given temperature, using the Lineweaver-Burk equation, one can extrapolate the value of t_0 for infinite inducer concentration, which allows estimating the mean intake time at that temperature (Table 3).

Table 3. Mean t_{int} ($\mu_{t_{int}}$) obtained from the Lineweaver-Burk equation and 95% CI of $\mu_{t_{int}}$ for various temperatures.

T (˚C)	$\mu_{t_{int}}$ (s)	95% CI of $\mu_{t_{int}}$ (s)
24	2434	[1949, 2918]
37	1322	[842, 1801]
41	1459	[1113, 1804]

From Table 3, we find that, in accordance with the results of deconvolution (Table 2), the mean t_{int} is highest at 24 °C, and is similar at 37 °C and 41 °C, being slightly smaller at 37 °C.

Quantitatively, we find that these values are ~ 35% larger (for 37 °C and 41 °C) and ~50% for 24 °C than those in Table 2. This is expected, as the deconvolution method is known to underestimate the peak value of the PDF [26].

Finally, we note that the value at 37 °C is also in clear agreement with a previous estimation of intake times at this temperature [6].

Number and duration of the rate-limiting steps of the intake process differs with temperature

To investigate the hypothesis that temperature affects the number and duration of the rate-limiting steps of the intake process, next, from the deconvolved t_{int} distributions of each temperature condition, we estimated the number and duration of these steps in maximum likelihood sense (Methods).

For this, we generalize the model of intake depicted by reactions (1) and (2) in Supplementary Material to a *d*-steps model, each exponentially distributed in duration, so that the number and duration of the rate-limiting steps are allowed to differ between the temperature conditions.

Results of this estimation are shown in Table 4, where we present the number and duration of the steps of the best fit model, along with the log-likelihood values. Meanwhile, in Table S2, we show the results for each condition when assuming specifically 1, 2, 3, and 4 steps, along with the *p*-values of the tests comparing pairs of models that are used to select the best model. Finally, in Figure 4 we show the best fit to the deconvolved t_{int} for each condition.

Table 4. Rate-limiting steps in the intake process determined by maximum likelihood estimation. Shown are the number of steps, the log-likelihood, the durations of the steps of the inferred models for each condition, and the CV^2 of the best fit. We fit the models to 10^5 random samples from the deconvolved t_{int} distribution. Note that there is no implied temporal order of the steps.

T (˚C)	No. Steps	Steps Durations	Log-likelihood	$CV^2_{\widetilde{t_{int}}}$
24	≥4	(387, 387, 387, 386)	-774461	0.25
30	3	(457, 457, 457)	-801201	0.33
37	2	(667, 319)	-783576	0.56
41	3	(532, 532, 20)	-783350	0.48

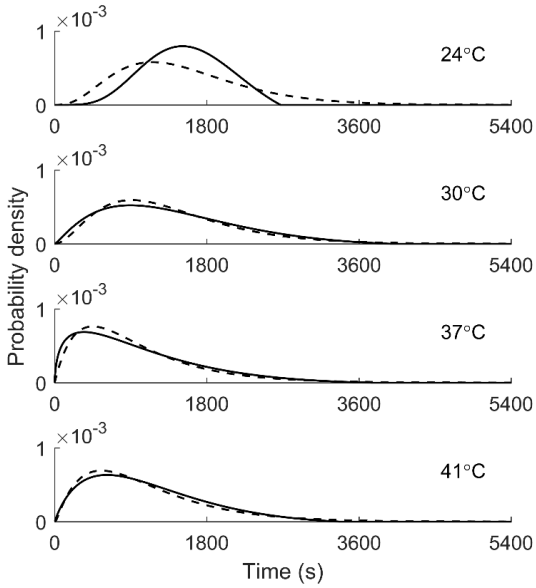


Figure 4. Deconvolved t_{int} distributions (solid line) and their best-fit d-steps model (dashed line).

Importantly, this result is in agreement with previous studies using data from cells at 37 °C [5].

From Tables 4 and S1, for all conditions, the test rejects the 1-step model in favor of a higher order model. This is expected, given the existence of the two membranes in the cell walls of *E. coli* cells and the time that inducers are expected to take to cross the periplasm in between [6].

Also, interestingly, the 2-steps model is the preferred one for cells at 37 °C and 41 °C (the step with a 20 s duration for the 41 °C condition can be disregarded, as the microscopy images are separated by 60 s intervals).

Meanwhile, at lower temperatures, higher order models (3 or more steps) are preferred, indicating that other steps become rate-limiting (in agreement with the deconvolution results), and/or that the steps duration may no longer follow an exponential distribution.

In this regard, we interpret the fact that a 4-steps model did not suffice to model the 24 °C condition (see Figure 4) as evidence for a significant change in the kinetics of intake with temperature, which renders the multi-step, exponentially distributed model incapable of fully capturing the dynamics. We hypothesize that this may be the consequence of increased viscosity of the cytoplasm and periplasm [14], along with changes in the physical properties and functionality of the intake 'machinery' in the cell walls.

Note that the CV^2 values of the best fits for 30 °C, 37 °C and 41 °C match the estimated values of the corresponding t_{int} distributions deconvolved from the fits to the empirical data. While the best fit in 24 °C condition has higher CV^2 than the deconvolved t_{int} (which is expected from the fact that the 4-steps model did not suffice to model the 24 °C condition), the trend in CV^2 of the deconvolved distributions and of their best fits is the same.

Finally, note that, in several cases, the time scales of the steps are identical. This may be due to an unknown artefact of the inference method or be representative of the real kinetics of intake of this inducer.

Discussion

In this work, we studied the single-cell dynamics of intake of IPTG, an inducer of the promoter $P_{Lac\text{-}ara\text{-}1}$, as a function of temperature. Rather than focusing on biological cellular adaptations, we focused solely on rapid physical changes due to temperature shifts in the process of inducer intake and consequent transcription kinetics.

For this, we first measured *in vivo* the time taken by individual cells to produce the first RNA, following the start of induction. From this, and previously collected data on the dynamics of RNA production by $P_{Lac\text{-}ara\text{-}1}$ [14], we applied two novel, independent methods to obtain the single-cell intake kinetics of the inducers, for each temperature condition. These methods' results were consistent with one another.

From this, first, we established that the response of the distribution of intake times of individual cells to temperature changes remains similar to that of the distribution of transcription activation times as temperature is changed, much due to the fact that most of the activation time is spent in the intake process in all conditions. Interestingly, the mean value of these distributions increases while their variability decreases for decreasing temperatures.

Since the intake process is bound to consist of multiple consecutive steps (in the case of IPTG, it was previously shown to be well modeled by a 2-step process for cells at 37 °C [5,6], we hypothesize that the decrease in variability could be the result of the emergence of additional rate-limiting steps in this process with decreasing temperature. The results of the maximum likelihood estimation tests support this view.

Further, they suggest that, at the lowest temperature condition tested here, the process is, from a dynamical point of view, 'too complex' to be well fitted by a sum of a small number (less than 5) of exponential steps. We hypothesize that this is clear evidence that the duration of one of the steps of the intake process becomes non-exponential-like at low temperatures. There are several potential causes for this (and perhaps multiple causes), and they are likely not accounted by our model (else, the increase in number of exponential steps would have allowed to fit the data well). We expect these potentials causes to range from malfunctioning of the porins in the membrane responsible for the diffusive intake of

the inducers, increased viscosity of the cytoplasm and periplasm, alteration of the physical properties of the outer and inner membranes, etc.

It is worth noting that the application of the Lineweaver-Burk equation to extract the mean value of the intake times is a methodology that has not been previously used, but we expect it to be of use in future works as well. It requires measuring transcription activation times for various inducer concentrations (at least 2) above the minimum concentration required for maximum induction. It is limited by the fact that the speed of intake is assumed to change linearly with the inverse of the inducer concentration, which may not always be the case. However, we expect this to be the case within certain ranges of inducer concentrations for simpler intake (mostly diffusion-based), mechanisms. Thus, it should be applicable to the study of a wide range of cellular intake mechanisms.

Overall, we conclude that different environmental conditions cause significant changes in the single-cell distributions of intake times of transcription inducers, which is expected to have a significant effect on the degree of heterogeneity in cell populations and cell lineages, due to the longevity of the transients during which this phenomenon has a strong effect in RNA numbers.

In the future, one important aspect that requires further research is the cause for the reduced cell-to-cell diversity in response times with decreasing temperatures, which we believe to be due to the emergence of rate-limiting steps in the intake process. Which steps and how they emerge are open questions, whose answers will help better understanding the robustness of the intake systems of *E. coli*.

Acknowledgements

Work supported by Academy of Finland (295027 and 305342 to ASR), Jane & Aatos Erkko Foundation (610536 to ASR), Tampere University of Technology President's Graduate Program (SS), Finnish Academy of Science and Letters (SO), Doctoral Programme of Computing and Electrical Engineering of TUT (NG) and Erasmus+ program 2919(713)2915/2016/SMS (CP). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.

References

[1] Arkin A, Ross J and McAdams H 1998 Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells *Genetics* **149** 1633–48

[2] van Kampen NG, Reinhardt WP 1983 Stochastic Processes in Physics and Chemistry Physics Today **36(2)** 78–80

[3] Elowitz B, Siggia D, Levine AJ, Swain PS, Siggia E, and Swain P 2002 Stochastic gene expression in a single cell *Science* **297** 1183–6

[4] Megerle J, Fritz G, Gerland U, Jung K, and Rädler J 2008 Timing and dynamics of single cell gene expression in the arabinose utilization system *Biophys J* **95** 2103–15

[5] Mäkelä J, Kandavalli V, and Ribeiro AS 2017 Rate-limiting steps in transcription dictate sensitivity to variability in cellular components *Sci Rep* **7** 10588

[6] Tran H, Oliveira SMD, Goncalves N, and Ribeiro AS 2015 Kinetics of the cellular intake of a gene expression inducer at high concentrations *Mol Biosyst* **11** 2579–87

[7] Hensel Z, Feng H, Han B, Hatem C, Wang J, and Xiao J 2012 Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis *Nat Struct Mol Biol* **19** 797–802

[8] Rosenfeld N, Young J, Alon U, Swain P, and Elowitz M 2005 Gene regulation at the single-cell level *Science* **307** 1962–5

[9] Robert L, Paul G, Chen Y, Taddei F, Baigl D, and Lindner AB 2010 Pre-dispositions and epigenetic inheritance in the *Escherichia coli* lactose operon bistable switch *Mol Syst Biol* **6** 357

[10] Kiviet D, Nghe P, Walker N, Boulineau S, Sunderlikova V, and Tans SJ 2014 Stochasticity of metabolism and growth at the single-cell level *Nature* **514** 376–9

[11] Yun HS, Hong J, and Lim HC 1996 Regulation of Ribosome Synthesis in *Escherichia coli*: Effects of Temperature and Dilution Rate Changes *Biotechnol Bioeng* **52** 615–24

[12] Gupta A, Lloyd-Price J, Neeli-Venkata R, Oliveira SMD, and Ribeiro AS 2014 *In vivo* kinetics of segregation and polar retention of MS2-GFP-RNA complexes in *Escherichia coli Biophysical Journal* **106(9)** 1928–37

[13] Choi PJ, Cai L, Frieda K, and Xie XS 2008 A stochastic single-molecule event triggers phenotype switching of a bacterial cell *Science* **322** 442–6

[14] Oliveira SMD, Häkkinen A, Lloyd-Price J, Tran H, Kandavalli V, and Ribeiro AS 2016 Temperature-Dependent Model of Multi-step Transcription Initiation in *Escherichia coli* Based on Live Single-Cell Measurements *PLoS Comput Biol* **12** e1005174

[15] Oliveira SMD, Neeli-Venkata R, Goncalves N, Santinha JA, Martins L, Tran H, Mäkelä J, Gupta A, Barandas M, Häkkinen A, Lloyd-Price J, Fonseca JM and Ribeiro AS 2016 Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli Mol Microbiol* **99** 686–99

[16] Lloyd-Price J, Startceva S, Chandraseelan JG, Kandavalli V, Goncalves N, Häkkinen A, and Ribeiro AS 2016 Dissecting the stochastic transcription initiation process in live *Escherichia coli DNA Res* **23** 203–14

[17] Lutz R, and Bujard H 1997 Independent and tight regulation of trasncriptional units in *Escherichia coli* via the LacR/O, the TetR/O and Arac/I1-I2 regulatory elements *Nucleic Acids Res* **25** 1203–10

[18] Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, and Mori H 2006 Construction of *Escherichia coli* K-12 in-frame, single-

gene knockout mutants: the Keio collection *Mol Syst Biol* **2** 1234–44

[19] Marbach A, and Bettenbrock K 2012 Lac operon induction in *Escherichia coli*: Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA *J Biotechnol* **157** 82–8

[20] Golding I, Paulsson J, Zawilski S, and Cox EC 2005 Real-time kinetics of gene activity in individual bacteria *Cell* **123** 1025–36

[21] Golding I, and Cox EC 2004 RNA dynamics in live *Escherichia coli* cells *Proc Natl Acad Sci U S A* **101** 11310–5

[22] Häkkinen A, Muthukrishnan A, Mora A, Fonseca JM, and Ribeiro AS 2013 CellAging: a tool to study segregation and partitioning in division in cell lineages of *Escherichia coli* *Bioinformatics* **29** 1708–9

[23] Häkkinen A, and Ribeiro AS 2015 Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data *Bioinformatics* **31** 69–75

[24] McClure WR 1985 Mechanism and control of transcription initiation in prokaryotes *Annu Rev Biochem* **54** 171–204

[25] 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** 3873–81

[26] Sheu CF, and Ratcliff R 1995 The application of fourier deconvoution to reaction time data: a cautionary note *Psychol Bull* **118** 285–99

[27] Häkkinen A and Ribeiro AS 2016 Characterizing rate limiting steps in transcription from RNA production times in live cells *Bioinformatics* **32** 1346–52

[28] DiCiccio T and Efron B 1996 Bootstrap Confidence Intervals *Stat Sci* **11** 189–228

[29] Carpenter J and Bithell J 2000 Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians *Stat Med* **19** 1141–64

[30] Lineweaver H and Burk D 1934 The Determination of Enzyme Dissociation Constants *J Am Chem Soc* **56** 658–66

[31] Press W, Teukolsky S, Vetterling W, and Flannery B 1992 *Numerical Recipes in C* (Cambrigde: Cambrigde University Press) 661–5

[32] Bertrand-Burggraf E, Lefèvre J, and Daune M 1984 A new experimental approach for studying the association between RNA polymerase and the tet promoter of pBR322 *Nucleic Acids Res* **12** 1697–706

[33] Kandavalli VK, Tran H, and Ribeiro AS 2016 Effects of σ factor competition are promoter initiation kinetics dependent *BBA - Gene Regul Mech* **1859** 1281–8

[34] Muthukrishnan A, Martikainen A, Neeli-Venkata R, and Ribeiro AS 2014 In vivo transcription kinetics of a synthetic gene uninvolved in stress-response pathways in stressed *Escherichia coli* cells *PLoS One* **9** e109005

Supplementary Material for "Temperature-dependence of the single-cell variability in the kinetics of transcription activation in *Escherichia coli*"

Nadia S.M. Goncalves[†], Sofia Startceva[†], Cristina S.D. Palma[†,‡], Mohamed N.M. Bahrudeen[†], Samuel M.D. Oliveira[†] and Andre S. Ribeiro[†,§,*]

[†] Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101, Tampere, Finland.

[‡] CA3 CTS/UNINOVA. Faculdade de Ciencias e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516, Caparica, Portugal.

[§] Multi-scaled Biodata Analysis and Modelling Research Community, Tampere University of Technology, 33101, Tampere, Finland.

[*] Corresponding author. E-mail: andre.ribeiro@tut.fi, Tel: +358408490736.

1. Supplementary Methods

Quantification of target gene activity by qPCR and microscopy

Cells with the plasmid carrying the target gene (pIG-BAC-$P_{lac\text{-}ara\text{-}1}$-mRFP1-96xMS2) were grown overnight at 30 °C with aeration and shaking in lysogeny broth (LB) medium, supplemented with the appropriate antibiotics (35 µg/ml Kanamycin and 34 µg/ml Chloramphenicol). From the overnight cultures, cells were diluted into fresh LB medium, supplemented with antibiotics, to an optical density of $OD_{600} \approx 0.05$, and allowed to grow at 37 °C, 250 rpm, until reaching an $OD_{600} \approx 0.3$.

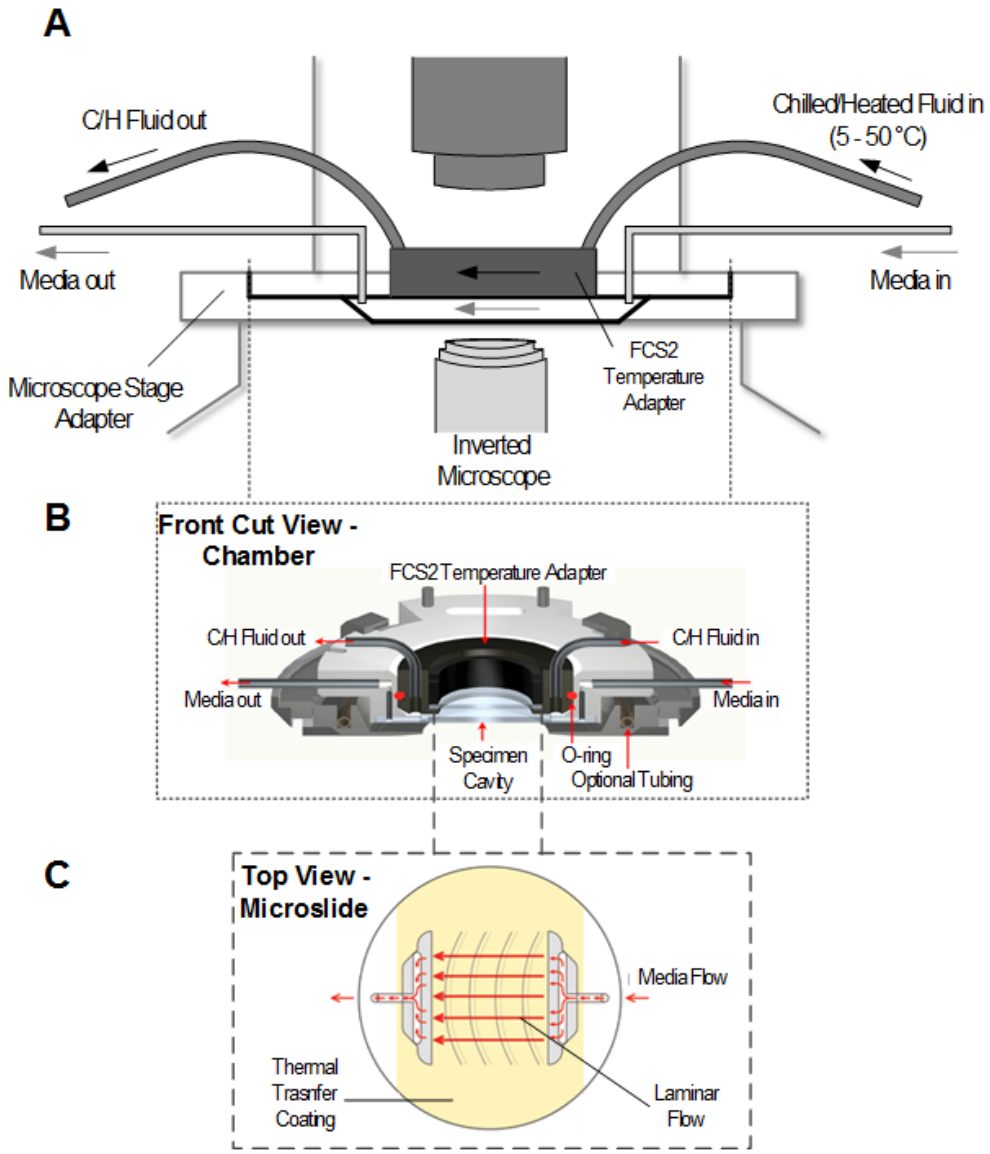
Next, qPCR was performed to analyze the fold change in mRNA production with induction of the target gene. From the culture described above, cells were then grown in LB media, at 37 ºC. To obtain the induction curve of this promoter, we first pre-induced with L-Arabinose (0.1%). Next, we induced with different concentrations of IPTG (0, 0.05, 0.1, 0.25, 0.5 and 1 mM). After collecting the cells by centrifugation at 8000 rpm for 5 minutes, twice the cell culture volume of RNA protect reagent (Qiagen) was added to the reaction tube, following the addition of Tris EDTA Lysozyme (15mg/ml) buffer (pH 8.0) for enzymatic lysis. Total RNA was isolated using RNeasy kit (Qiagen), according to the manufacturer instructions. Samples with total isolated RNA were treated with DNase for residual DNA removal. The A260/A280nm ratio of the isolated RNA samples was assessed using a Nanovue plus spectrophotometer (GE Healthcare) with the value obtained (2.0-2.1) indicating a highly purified RNA. Additionally, the resulting RNA yield was used to normalize the RNA concentration in the samples with varying IPTG concentrations. Following that, iSCRIPT reverse transcription super mix (Biorad) was added for cDNA synthesis. Next, the cDNA samples were mixed with the qPCR master mix, containing iQ SYBR Green supermix (Biorad), with specific primers for the target and reference genes. The

qPCR reaction was carried out in technical triplicates with a total reaction volume of 20 µL. To amplify the target gene mRPF1 and reference gene 16SrRNA, we used the following primers, respectively: i) forward 5' TACGACGCC GAGGTCAAG 3' and reverse 5' TTGTGGGAGGTGATGTCCA 3', and ii) forward: 5' CGTCAG CTCGTGTTGTGAA 3' and reverse: 5' GGACCGCTGGCAA CAAAG 3'. The qPCR experiments were performed using a MiniOpticon Real time PCR system (Biorad). The following thermal cycling protocol was used: 40 cycles at 95 ºC for 10 s, 52 ºC for 30 s, and 72 ºC for 30 s. No-RT and No-Template controls were used to crosscheck non-specific signals and contamination, and the efficiency of the PCR reactions were found to be greater than 95%. The data from the CFX Manager™ software was then used to calculate the fold change in mRNA production and its standard error [1], which are presented in Figure S2.

Meanwhile, microscopy measurements were conducted as described in Methods (section "Growth Conditions, Microscopy, Data Extraction on Transcription Activation Times").

We note two main differences between the qPCR and the microscopy measurements. First, in the qPCR measurements, the report system is not activated, since it is not required to obtain the measurements and since this does not cause significant differences in target RNA production rates (data not shown). Second, in the qPCR measurements, the activation of the target gene is performed in liquid. In general, this results in higher absolute expression levels than if the induction is performed under the microscope. However, this does not constitute a problem as it does not alter the inducer concentration at which maximum induction is reached (Figure S2).

FCS2 imaging system



Supplementary Figure S1. (A) Schematic illustration of the CFCS2 microfluidics and the temperature control system for cell cultures while under microscope observation. The CFCS2 chamber is mounted on the stage of an inverted microscope. This device is comprised of two independent fluidic systems. One is a thermo-chiller device (not shown), which is connected to two inlets and two outlets of the CFCS2 chamber. This device controls the temperature of the system (i.e. of the metal chamber and the optical cavity, where cells are placed) through the flow of heat/chilled fluidics, whose temperature can range from 5 ºC to 50 ºC ± 0.2 ºC. The second

device, a micro-perfusion device (not shown), is connected to one inlet and one outlet of the CFCS2. It constantly provides cells with fresh media and chemicals required for cell growth. (B) Illustrative front cut view of the optical cavity of the cooled FCS2 adapter (CFCS2). The CFCS2 is a modified version of the FCS2 system, in that it has an additional, independent tubing system to facilitate the circulation of a heat/chilled fluid, that increases/reduces temperature of the metal base and of the optical cavity of the chamber. (C) Schematic top view of the micro-aqueduct slide, which is placed inside the optical cavity. The slide allows laminar flow of fluids, when a uniform and rapid exchange of media is required across the cell population. Images shown in (B) and (C) are adapted from Bioptechs Inc. (http://www.bioptechs.com).

Model of Inducer Intake and Active Transcription

We assume the following models of transcription activation and active transcription [2]. First, regarding activation, when an inducer is added to the media, the gene is only activated after a multi-step process that includes events such as the entry and diffusion of inducers in the periplasm and then cytoplasm, binding of an inducer to a transcription factor repressing gene activity, etc. These events and their kinetics differ with the induction and repression systems of the gene [3, 4].

Relevantly, as mentioned, the strain used here lacks the ability to produce LacY [5]. As such, we expect the intake process of inducers (IPTG) to be purely diffusive-like. Also, as *E. coli* is gram-negative, the cell walls have an outer and inner membranes, with a periplasmic space in between. Thus, the activation process is expected to have at least two, consecutive rate-limiting steps: entering of inducers into the periplasm, followed by entering into the cytoplasm. In agreement, previous studies of the intake of IPTG at optimal temperatures (37 ºC) [2] have shown that the activation process of our target promoter, $P_{Lac\text{-}ara\text{-}1}$, in cells lacking LacY as those used here [5], is well modelled by a 2-step stochastic process of the form [29]:

$$I_{env} \xrightarrow{Int_1} I_{peri} \quad (1)$$

$$I_{peri} \xrightarrow{Int_2} I \quad (2)$$

Reaction (1) represents the entrance of an inducer molecule (I) into the periplasm, while reaction (2) models the passage of that inducer from the periplasm into the cytoplasm, where it can activate the target gene, e.g. by interacting with repressor molecules. In general, additional rate-limiting steps could, in theory, be modelled by a sequence of d-steps with exponential duration: $I_1 \rightarrow \ldots \rightarrow I_d$. This is particularly important when selecting a strategy to decompose the rate-limiting steps from the empirical data.

Meanwhile, transcription activation is modeled as follows:

$$Rep + I \underset{k_{inact}}{\overset{k_{act}}{\rightleftharpoons}} Rep.I \quad (3)$$

$$P_{OFF} + I \xrightarrow{k_{act}} P_{ON} + Rep.I \quad (4)$$

$$P_{OFF} \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} P_{ON} + \text{Rep} \quad (5)$$

In (3), an inducer (I) binds to the repressor (Rep), creating a complex (Rep.I) that cannot repress the promoter (see reactions (5)). In our case, the repressor are LacI tetramers, and IPTG, the inducer, acts by binding to these tetramers, greatly reducing their binding affinity to the promoter [6]. We assume that this binding reaction is very weakly reversible. Also, the reactions necessary to form LacI tetramers are not explicitly considered since most LacI molecules in the cell are present in the form of tetramers.

In reactions (4), again an inducer binds to a repressor, but the repressor is bound to the promoter, which frees the promoter. We note that, for such to occur, the LacI tetramer must unbind from both DNA binding sites [6].

Reaction (5) allows for the repression of the promoter by free repressors and for the possibility of a repressor unbinding the promoter, without direct intervention of inducers.

Finally, active transcription by a free promoter is modeled as a multi-step process [7-11]:

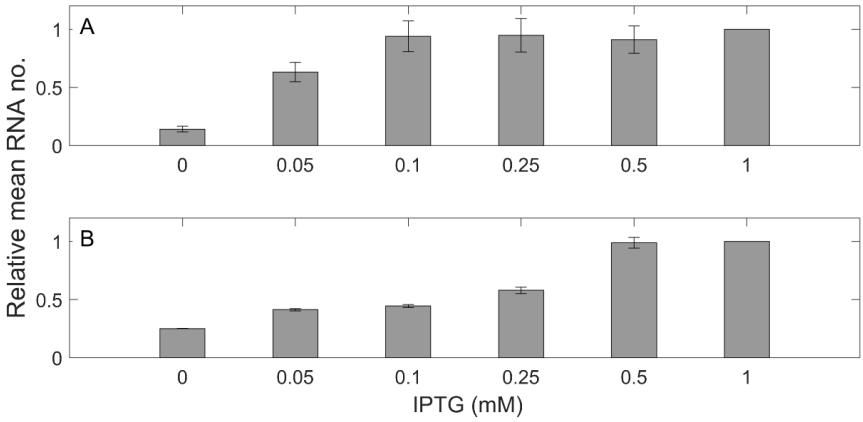
$$R+P_{ON} \xrightarrow{k_{cc}} RP_{cc} \xrightarrow{k_{oc}} RP_{oc} \xrightarrow{\infty} P_{ON} + R+RNA \quad (6)$$

In (6), R is the RNA polymerase. Once bound to the promoter, it forms a closed complex (RP_{cc}), which is followed by the open complex (RP_{oc}) formation, elongation (not rate-limiting, and thus not represented), and, finally, RNA production and RNA polymerase release (also not rate-limiting, and thus having an 'infinite' rate,∞).

It is of importance to note that, while not represented, the steps in (6) are all considered to be reversible (except the open complex formation, which, once initiated, is nearly irreversible [9]. I.e., the reactions in (6) are not to be interpreted as elementary transitions. Instead, they represent effective rates of the rate-limiting steps in transcription, thus defining the promoter strength [11].

2. Supplementary Results

2.1 Induction curve at 37 ºC



Supplementary Figure S2. Induction curve of $P_{Lac/ara-1}$ in cells at 37 ºC as obtained by microscopy (A) and qPCR (B). In the case of microscopy, from the single cell measurements, we calculated the mean and standard uncertainty from the distribution of RNA numbers produced in each cell during 1 hour following induction. In the case of qPCR, the mean RNA fold change and its standard uncertainty in each condition were extracted from 3 technical replicates. In both (A) and (B), values relative to the 1 mM induction condition were calculated in each condition using the Delta Method [12]. Also in both, measurements were conducted after induction of the target gene (IPTG added 1 hour prior to the measurements and 0.1% of L-Arabinose added 1 hour and 45 minutes prior to the measurements, see Methods).

2.2 Measurements of t_0 vs temperature for *E. coli* JW0334 strain

T (˚C)	N_{t_0}	$\mu_{t_0} \pm$ SE (s)	$CV^2_{t_0}$	KS-test for t_0 values vs 37 ˚C (p-value)
24	76	1939 ± 99	0.20	< 0.01
30	64	1725 ± 108	0.25	< 0.01
37	97	1089 ± 67	0.37	-
41	106	1729 ± 86	0.26	< 0.01

Supplementary Table S1. Measurements of t_0 vs temperature for *E. coli* JW0334 strain. Shown are the number of measurements (N_{t_0}), mean (μ_{t_0}) standard error (SE) and CV^2 of the distribution of t_0 values ($CV^2_{t_0}$). The table also shows the p-value from the KS tests comparing the t_0 distributions at each temperature, with the distribution at 37 ˚C (control). For p-values smaller than 0.01, the null hypothesis that the two sets of data are from the same distribution can be rejected.

2.3 Maximum log-likelihood fit to the deconvolved distributions of intake times

24 ˚C			
d	Log-likelihood	Durations	*p*-values ($d_1 = d_0 + 1$)
1	-834550	(1549)	0.00
2	-801008	(775, 775)	0.00
3	-784454	(517 , 517, 516)	0.00
4	-774461	(388 , 388, 388, 387)	-
30 ˚C			
d	Log-likelihood	Durations	*p*-values ($d_1 = d_0 + 1$)
1	-822278	(1370)	0.00
2	-803246	(685, 685)	0.00
3	-801201	(457, 457, 457)	1.00
4	-801204	(458, 456, 456, 0)	-
37 ˚C			
d	Log-likelihood	Durations	*p*-values ($d_1 = d_0 + 1$)
1	-789385	(986)	0.00
2	-783576	(667, 319)	1.00
3	-783576	(667, 319, 0)	1.00
4	-783576	(667, 319, 0, 0)	-
41 ˚C			
d	Log-likelihood	Durations	*p*-values ($d_1 = d_0 + 1$)
1	-798760	(1083)	0.00
2	-783444	(542, 542)	0.00
3	-783350	(532, 532, 20)	1.00
4	-783350	(532, 532, 20, 0)	-

Supplementary Table S2. Log-likelihood and durations of the steps of the inferred models with d-steps, for each temperature condition. The table shows, first, the number of steps (d) assumed, followed by the log-likehood, and the duration of the steps (the order of these steps cannot be determined by this method). The last column shows the *p*-values of the likelihood-ratio tests between pairs of models for each condition. The null model is the $d_0 = [1:3]$ step model, while the alternative model is the $d_1 = d_0 + 1$ step model.

References

[1] Livak K J and Schmittgen T D 2001 Analysis of relative gene expression data using real-time quantitative PCR and the 2-DDCT method. *Methods* **25** 402–8

[2] Tran H, Oliveira S M D, Goncalves N and Ribeiro A S 2015 Kinetics of the cellular intake of a gene expression inducer at high concentrations *Mol. Biosyst.* **11** 2579–87

[3] 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** 2103–15

[4] Schleif R 2000 Regulation of the L-arabinose operon of Escherichia coli *Trends Genet.* **16** 559–65

[5] Lutz R and Bujard H 1997 Independent and tight regulation of trasncriptional units in Escherichia coli via the LacR/O, the TetR/O and Arac/I1-I2 regulatory elements. *Nucleic Acids Res.* **25** 1203–10

[6] Lewis M 2005 The lac repressor *Comptes Rendus - Biol.* **328** 521–48

[7] Ribeiro A S, Zhu R and Kauffman S A 2006 A General modeling strategy for gene regulatory networks with stochastic dynamics *J. Comput. Biol.* **13** 1630–9

[8] 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. Theor. Biol.* **246** 725–45

[9] McClure W R 1985 Mechanism and control of transcription initiation in prokaryotes *Annu. Rev. Biochem.* **54** 171–204

[10] 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** 3019–25

[11] Mulligan M E, Hawley D K, Entriken R and McClure W R 1984 Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity *Nucleic Acids Res.* **12** 789–800

[12] Casella G and Berger R L 2001 The Delta Method. Statistical Inference, 2nd ed. Duxbury Press, Pacific Grove, CA, 240–5

Publication IV

S. Startceva, V.K. Kandavalli, A. Visa, and A.S. Ribeiro "Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression", *BBA - Gene Regulatory Mechanisms*

BBA - Gene Regulatory Mechanisms 1862 (2019) 119–128

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression

Sofia Startceva[a], Vinodh K. Kandavalli[a], Ari Visa[b], Andre S. Ribeiro[a,*]

[a] Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101 Tampere, Finland

[b] Faculty of Computing and Electrical Engineering, Tampere University of Technology, Tampere 33101, Finland

ARTICLE INFO

Keywords:
Single-cell time-lapse microscopy
Transcription initiation
RNA and protein numbers
Asymmetry and tailedness
Threshold crossing

ABSTRACT

Genetic circuits change the *status quo* of cellular processes when their protein numbers cross thresholds. We investigate the regulation of RNA and protein threshold crossing propensities in *Escherichia coli*. From in vivo single RNA time-lapse microscopy data from multiple promoters, mutants, induction schemes and media, we study the asymmetry and tailedness (quantified by the skewness and kurtosis, respectively) of the distributions of time intervals between transcription events. We find that higher thresholds can be reached by increasing the skewness and kurtosis, which is shown to be achievable without affecting mean and coefficient of variation, by regulating the rate-limiting steps in transcription initiation. Also, they propagate to the skewness and kurtosis of the distributions of protein expression levels in cell populations. The results suggest that the asymmetry and tailedness of RNA and protein numbers in cell populations, by controlling the propensity for threshold crossing, and due to being sequence dependent and subject to regulation, may be key regulatory variables of decision-making processes in *E. coli*.

1. Introduction

The gene regulatory networks of bacteria, such as *Escherichia coli*, include network motifs [1,2]. Some of these are responsible for decision-making processes that assist cells in adapting to environmental changes [3,4]. Significant behavioural changes in these motifs usually occur when the numbers of one or more of the component proteins cross thresholds [3]. The underlying mechanisms that define the propensity for the protein numbers of a given gene to cross a specific threshold are not yet fully understood.

In *E. coli*, it is common for the protein numbers to follow the corresponding RNA numbers [5,6]. These are determined by the rates of RNA production and degradation. Interestingly, RNA degradation in *E. coli* appears to be largely independent from the RNA sequence, abundance and metabolic function [7–9], suggesting that little regulation occurs at this stage. Meanwhile, various regulatory mechanisms of transcription have been identified, which usually act at the stage of initiation, suggesting that control over the RNA numbers is exerted at this stage [10–12].

From the dynamics point of view, the regulation of transcription initiation kinetics occurs via the tuning of the time-length of the rate-limiting steps of initiation, respectively, the events prior and after committing to open complex formation [13–17]. In particular, recent studies [14,16–18] have shown that, under full induction, the in vivo kinetics of these rate-limiting steps, along with supercoiling buildups [19], define, to a great extent, the distribution of time intervals between consecutive RNA production events (here referred to as 'Δt distribution'). Further, it was shown that not only the first moment (mean), but also the second moment of this distribution (variance) can be tuned by the kinetics of these steps [16,18].

Given this, we hypothesise that, by tuning the kinetics of these rate-limiting steps, one can also tune the third and fourth moments of the Δt distribution (respectively, the skewness and kurtosis). Further, we hypothesise that these two moments can be tuned independently from the mean and coefficient of variation. To test these hypotheses, we perform in vivo time-lapse microscopy employing single-RNA detection by MS2-GFP tagging [20–22], from which we extract the Δt distributions for various promoters, media, induction schemes, growth phases, mutants and a stress condition. Next, for each condition, we estimate their mean, coefficient of variation, skewness and kurtosis. Subsequently, we estimate the kinetics of the rate-limiting steps in each condition and assess their influence on the skewness and kurtosis. Finally, to test whether changing the skewness and kurtosis of the Δt distribution has functional consequences, we measure the corresponding values of the

* Corresponding author.
E-mail address: andre.ribeiro@tut.fi (A.S. Ribeiro).

https://doi.org/10.1016/j.bbagrm.2018.12.005
Received 29 October 2018; Received in revised form 4 December 2018; Accepted 5 December 2018
Available online 14 December 2018

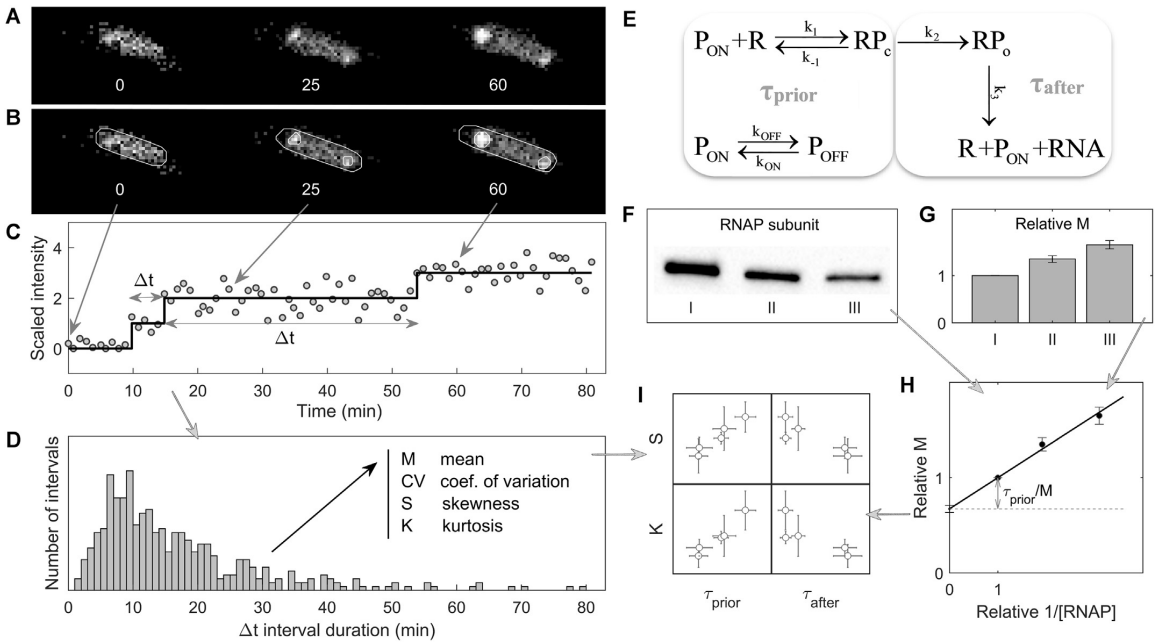


Fig. 1. Schematic representation of the steps for the analysis of the dynamics of RNA production in individual cells, from in vivo single-RNA, single-cell measurements. (A) Example confocal microscopy images over time of a cell expressing MS2-GFP and the target RNAs. (B) Segmentation of a cell and the MS2-GFP tagged RNA spots within (white lines). (C) Scaled RNA spots intensity over time (grey circles) of the example cell, along with the best-fitting monotonic piecewise-constant curve (black line) from which Δt intervals are estimated. (D) The distribution of time intervals between consecutive RNA production events in individual cells (Δt) from which mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) are extracted. (E) Model of transcription initiation. The first box contains the reactions occurring before commitment to open complex formation, with their mean time-length denoted as τ_{prior}. The second box contains the reactions occurring after commitment to open complex formation, with their mean time-length equals τ_{after}. For a detailed description of these reactions and parameters see Supplementary materials and methods, Section 1.6. (F) Western blot image of the RNA polymerase (RNAP) subunit in different media richness. (G) Relative inverse transcription rate of the target gene, measured by qPCR. (H) Relative τ plot (Lineweaver–Burk plot [25] of the inverse of the RNA production rate versus the inverse of the RNAP concentration, [RNAP]) for estimating τ_{prior} relative to M. (I) S and K versus τ_{prior} and τ_{after} in different conditions.

skewness and kurtosis of the distributions of single-cell protein expression levels.

2. Materials and methods

Fig. 1 informs on the models and methods used. In short, the main empirical data (Δt distributions) are obtained by measuring when each RNA appears in each cell. Also, we measure the average intracellular RNAP concentration. From these concentrations and the corresponding mean of the Δt distribution in each condition, we estimate the time spent in transcription initiation prior and after commitment to open complex formation (τ_{prior} and τ_{after}, respectively, with their sum equalling Δt) (model in Fig. 1E).

In summary, we first estimate τ_{prior}/M from τ plots [23]. For this, the inverse of the RNA production rate relative to the control (as measured by qPCR) is plotted against the inverse of the RNAP concentration relative to the control (as measured by Western blot, Supplementary materials and methods, Section 1.4). Next, a line is fitted to the data. The point where this line intersects the Y axis equals the extrapolated value of the inverse of the transcription rate for an 'infinite' RNAP concentration. As such it should equal τ_{after}/M, according to the model in Fig. 1E. From this and the value of M, one can calculate τ_{after} and τ_{prior} (Supplementary materials and methods, Section 1.5). Next, from the same Δt distributions, we extract the coefficient of variation, skewness and kurtosis in each condition.

Note that, although genes replicate during the cells lifetime by a process that is not absent of noise and many variables control when each specific gene is replicated [24], we assume that the rate constants controlling the kinetics of RNA production of our gene of interest (Fig. 1E), which is on a single-copy F-plasmid, do not change significantly during the lifetime of the cells. To validate this assumption we compared the distributions of time intervals (between consecutive RNA production events) that started and ended in the first half of the lifetime with intervals that started and ended in the second half (Supplementary results, Section 2.1). From the comparisons of these distributions in each condition (Table 1) we conclude that the assumption is sufficiently accurate.

2.1. Bacterial strains, plasmids, growth conditions, MS2-GFP tagging system, induction of the reporter and target genes, and measurement conditions

The *E. coli* strain used was DH5α-PRO (identical to DH5αZ1) [26] whose genotype is: deoR, endA1, gyrA96, hsdR17(rK− mK+), recA1, relA1, supE44, thi-1, Δ(lacZYA-argF)U169, Φ80δlacZΔM15, F-, λ-, PN25/tetR, PlacIq/lacI and SpR. This strain produces, from the chromosome and in abundance, the necessary regulatory proteins for their constructs, namely, LacI, AraC and TetR [26]. E.g. LacI, the main repressor of the control promoter ($P_{lac/ara-1}$), exists in a concentration much higher than the wild type (~3000 copies vs ~20 in wild type [26]). These characteristics allow tight regulation of both target and reporter genes, ensuring that the observed RNAs are due to active transcription and not the result of transcription leakiness (i.e. in the absence of activation). In particular, we measured leaky expression of

Table 1
Description of conditions. Shown are the name by which the condition is identified, the target plasmid and corresponding inducer, the reporter plasmid and corresponding inducer, and the media.

Conditions	Target promoter	Target inducers	Reporter promoter	Reporter inducer	Growth media
LA	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1×
LA(75)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	0.75×
LA(50)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	0.5×
LA(ara)	$P_{lac/ara-1}$	1% ara	$P_{LtetO-1}$	100 ng aTc	1×
LA(IPTG)	$P_{lac/ara-1}$	1 mM IPTG	$P_{LtetO-1}$	100 ng aTc	1×
LA(oxi)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1× + 0.6 mM H_2O_2
Mut1	$P_{lac/ara-1}$ (Mut-1)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1×
Mut2	$P_{lac/ara-1}$ (Mut-2)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1×
Mut3	$P_{lac/ara-1}$ (Mut-3)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1×
Mut4	$P_{lac/ara-1}$ (Mut-4)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1×
tetA	P_{tetA}	–	P_{lac}	1 mM IPTG	1×
tetA(st)	P_{tetA}	–	P_{lac}	1 mM IPTG	Stationary phase
BAD	P_{BAD}	0.1% ara	P_{lac}	1 mM IPTG	1×
BAD(st)	P_{BAD}	0.1% ara	P_{lac}	1 mM IPTG	Stationary phase

$P_{lac/ara-1}$, in the absence of IPTG and arabinose, and found only ~5% or less cells with an MS2-GFP tagged RNA, 2 h after inducing the reporter expressing MS2-GFP.

We also use BW25113, whose genotype is F-, DE(araD-araB)567, lacZ4787(del)::rrnB-3, LAM-, rph-1, DE(rhaD-rhaB)568, hsdR514, which expresses LacI and AraC from the genotype. The absence of TetR allows the Tet promoter to express constitutively.

All cells carry two plasmids: a multi-copy reporter plasmid coding for MS2-GFP under the control of an inducible promoter and a single-copy F-based target plasmid coding for the transcript with multiple MS2-GFP binding sites under the control of another promoter (Table 1). Also, in all target plasmids, we inserted a sequence coding for a red fluorescent protein, between the target promoter and MS2 binding sites. Promoter sequences are specified in Supplementary Fig. S1. Tagged RNAs can be visualized as fluorescent spots [14,20–23] (Fig. 1A).

In general, to observe RNAs tagged by MS2-GFP proteins, cells were grown overnight in LB media with the respective antibiotics at 30 °C in an orbital shaker with aeration of 250 rpm. From the overnight culture, cells were diluted using fresh LB media (unless stated otherwise in Table 1) to an initial OD_{600} of 0.05 (measured with a spectrophotometer, Ultrospec 10; GE Healthcare) and incubated at 37 °C at 250 rpm to allow growth until reaching an OD_{600} of 0.25. In general, the reporter gene was induced 1 h prior to the target gene, to allow for sufficient MS2-GFP proteins to be produced prior to the appearance of the target RNAs. For a detailed description, see Supplementary materials and methods, Section 1.1. Inducers of target and reporter genes are described in Table 1.

The MS2-GFP RNA tagging technique, proposed in [27], is at present the only direct method to measure time intervals between RNA production events in live, individual cells [14,16,21,22]. This is possible because, first, once appearing, each tagged RNA spot exhibits 'full' fluorescence (assuming 1 min interval between microscopy images) [22]. This removes uncertainty in the process of RNA counting as it reduces the possibility for 'partially fluorescent RNAs'. This uncertainty is further reduced in that, once tagged, the fluorescence of the spots remains near constant for longer than our measurement time (2 h or more) [22]. This provides significant reliability to the quantification of the time-length of intervals between consecutive RNA production events [21].

MS2-GFP tagging affects the spatial organization of the RNAs inside the cell [28]. However, this does not affect the precision of quantification of the intervals between consecutive RNA production events, which are based solely on the total intensity of the MS2-GFP tagged RNAs in a cell, not on their location.

To assess whether this technique has a negative impact on cell physiology, we compared cell growth rates and morphology with and without activating the expression of the MS2-GFP reporter. Supplementary results in Section 2.2 show that growth rates and cell morphology are not significantly affected by expression of MS2-GFP, in agreement with previous studies [14,23].

Finally, it is also reasonable to assume that MS2-GFP tagging could affect the protein expression levels of the target gene, due to partially interfering with the target RNA (albeit in a different region from the one coding for the red fluorescent protein). We tested this by comparing protein expression levels when and when not activating the expression on MS2-GFP (Supplementary results, Section 2.3). The results confirm that the expression levels of the red fluorescent protein are not perturbed significantly by MS2-GFP tagging (Fig. S9).

Meanwhile, to measure the single cell distributions of RNAP concentration, we used *E. coli* RL1314 strain with fluorescently tagged β' subunits (a kind gift from Robert Landick, University of Wisconsin-Madison) [29]. From the overnight culture, we diluted the cells to an OD_{600} of 0.1 in various media richness (Materials and methods) and allowed them to grow to an OD_{600} of 0.5 at 37 °C at 250 rpm. Cells were then pelleted by centrifugation and visualized under the microscope.

The plasmids (Table 1) construction and transformation were performed using standard molecular cloning techniques [30]. To construct $P_{lac/ara-1}$-mCherry-48 binding sites (bs) mutants, we used a plasmid carrying mCherry followed by a 48bs array in the pBELO vector backbone, originally constructed in [31]. To obtain the mutant promoters (Supplementary Fig. S1), we synthesized new promoter sequences of $P_{lac/ara-1}$ with specific point mutants with support from Gene Script, USA. Next, we inserted them into the pBELO vector backbone by Gibson Assembly [32], to obtain a single copy F-based plasmid carrying the target region $P_{lac/ara-1}$-mCherry-48bs mutants. This product was transferred into competent *E. coli* host cells. The recombinants were selected by antibiotic screening and confirmed with sequence analysis. It is noted that the mutant promoters were selected solely based on that their Δt distributions differed from the one of $P_{lac/ara-1}$.

2.2. Chemicals

The chemical components of LB media are Tryptone, Yeast extract and NaCl, purchased from LabM (Topley House, Bury, Lancashire, UK). The antibiotics used are Kanamycin 34 μg/ml, Ampicillin 50 μg/ml and Chloramphenicol 35 μg/ml, purchased from Sigma-Aldrich (St. Louis, MO). The inducers used are isopropyl β-D-1-thiogalactopyranoside (IPTG), anhydrotetracycline (aTc) and arabinose (ara), purchased from Sigma-Aldrich. Agarose (Sigma-Aldrich) was used for preparing the microscope gel pads. For PCR, Phusion high-fidelity polymerase and other PCR reagents were purchased from Finnzymes (Finland). Qiagen kits (USA) were used for plasmid isolation. For qPCR, cells were treated with RNA protect bacteria reagent (Qiagen, USA). iScript Reverse Transcription Supermix for cDNA synthesis and iQ SYBR green

supermix for qPCR were purchased from Biorad (USA).

2.3. Growth media

In all experiments, we used the LB media and its altered versions, first described in [14]. Namely, we used the following media compositions per 100 ml: 1 g tryptone, 0.5 g yeast extract and 1 g NaCl (pH 7.0), referred to as '1 ×' (Table 1); 0.75 g tryptone, 0.375 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.75 ×'; 0.5 g tryptone, 0.25 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.5 ×'; 0.25 g tryptone, 0.125 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.25 ×'. These four media are used to attain various mean intracellular RNA polymerase concentrations ([RNAP]) in cell populations, while not affecting normal cell physiology and morphology [14,16,23] (Supplementary Fig. S2A). Additionally, in two conditions, as in [23], we used the stationary phase media obtained by centrifuging the overnight culture of LB media at 10000 rpm for 10 min followed by filtration [23] (growth rates shown in Supplementary Fig. S2B).

2.4. qPCR measurements

Cells with target plasmids were harvested by centrifuging them at 8000 × *g* for 5 min. To the pelleted cells, twice the amount of RNA protect reagent (Qiagen) was added, followed by the enzymatic lysis with Tris EDTA lysozyme buffer (pH 8.0). Total RNA was isolated using RNeasy kit (Qiagen) according to the kit instructions. The concentration of RNA was quantified using the Nanovue plus spectrophotometer (GE Healthcare). The RNA samples were treated with DNase to remove the residual DNA, followed by cDNA synthesis, using the iSCRIPT reverse transcription super mix. The cDNA samples were mixed with the qPCR master mix containing iQ SYBR Green Supermix (Biorad) with primers for the target and reference genes. The reaction was carried out in triplicates with the total reaction volume of 20 µl. For quantifying the target gene we used following primers: for mRFP1 (Forward: 5′ TACG ACGCCGAGGTCAAG 3′ and Reverse: 5′ TTGTGGGAGGTGATGTCCA 3′), for mCherry (Forward: 5′ CACCTACAAGGCCAAGAAGC 3′ Reverse: 5′ TGGTGTAGTCCTCGTTGTGG 3′). For the reference gene, 16S RNA primers (Forward: 5′ CGTCAGCTCGTGTTGTGAA 3′ and Reverse: 5′ GGACCGCTGGCAACAAAG 3′) were used. The qPCR experiments were performed by a MiniOpticon Real- time PCR system (Biorad). The following conditions were used during the reaction: 40 cycles of 95 °C for 10 s, 52 °C for 30 s and 72 °C for 30 s for each cDNA replicate. We used no-RT controls and no-template controls to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were > 95%. The data from CFX Manager TM Software was used to calculate the relative gene expression and its standard error [33].

2.5. Microscopy

Measurements of integer-valued numbers of RNAs or of the moments when a new RNA appears in individual cells were conducted using microscopy. For this, a few µl of cells carrying the induced reporter and target plasmids were placed between a coverslip and agarose gel pad (2.5%), with the respective inducers and antibiotics. Next, an FCS2 chamber (Bioptechs) was heated to 37 °C and placed under the microscope. Cells were visualized using a Nikon Eclipse (Ti-E, Nikon) inverted microscope, equipped with a 100× Apo TIRF (1.49 NA, oil) objective. Confocal images were obtained by a C2+ (Nikon) confocal laser-scanning system. For measuring GFP fluorescence (to visualize MS2-GFP 'spots' or RNAP-GFP), we used a 488 nm laser (Melles-Griot) and an emission filter (HQ514/30, Nikon). For time series, confocal images were taken every 1 min for 2 h. Previous studies [14] have shown that these microscopy settings do not cause significant phototoxicity in this strain. Finally, phase-contrast images were obtained simultaneously, with an external phase-contrast system and CCD camera (DS-Fi2, Nikon), every 5 min. Images were extracted using Nikon Nis-Elements software.

2.6. Image and data analysis

Microscopy images were analysed using the software 'CellAging' [34]. For details see Supplementary materials and methods, Section 1.2. From these analysed time-lapse images, we extracted intervals between consecutive RNA production events in individual cells, from which empirical distributions of these intervals (Δt distributions) were obtained (Fig. 1A–D). Data analysis was conducted using tailored algorithms implemented in MATLAB R2017b (MathWorks).

2.7. Flow cytometry

Measurements of protein expression levels were conducted using flow cytometry (FC). For this, cells from 5 ml of bacterial culture were diluted 1:10,000 into 1 ml PBS vortexed for 10 s. We performed measurements under various conditions. In each condition, a total of 50,000 cells were observed. Measurements were performed using an ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego, USA) with a yellow laser (561 nm) for excitation and the PE-Texas Red (mCherry) fluorescence detection channel (615/20 nm filter) for emission, at a flow rate of 14 µl/min and a core diameter of 7.7 µM. The PMT voltage of 584 was used for mCherry. To avoid background signal from particles smaller than bacteria, the detection threshold was set to 5000 in FSC-H analyses.

We applied unsupervised gating [35] (implemented in Python 3.6) to the flow cytometry data. We set the fraction of the cells whose data is used in the analysis (α) to 0.9, as it was sufficient to remove data points produced by debris, cell doublets and other undesired events. Reducing α further did not change the results qualitatively.

3. Results

3.1. Mean, coefficient of variation, skewness and kurtosis of the distributions of time intervals between consecutive RNA productions in individual cells differ with promoter sequence, regulatory factors and growth conditions

First, we obtained empirical data on the Δt distributions in 14 conditions (see Table 1 for details). These conditions were selected so as to test if the promoter sequence (conditions LA, Mut1, Mut2, Mut3, and Mut4, see Supplementary Fig. S1), regulatory factors such as RNAP and inducer concentrations (conditions LA, LA(75), LA(50), LA(ara), LA(IPTG)), and variables associated to the environment (e.g. media and stress) affect the skewness and kurtosis of the Δt distribution.

Results are shown in Supplementary Fig. S3. From these distributions, we estimated their mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) (Supplementary materials and methods, Section 1.3). The data was produced from at least 3 repeats per condition. Since no significant differences were found between repeats, the data for each condition were merged. Noteworthy, all target genes used have identical sequences upstream and downstream of the promoter region (Materials and methods). Also, as noted above, as they are integrated into single-copy F-plasmids, not anchored to the membrane, they are not expected to be significantly influenced by transcription halting due to positive supercoiling buildup [19,36].

From Fig. 2A, M and CV differ between conditions. S and K also differ between conditions, but do so following a similar trend to one another. Importantly, changes in S and K seem uncorrelated with the values of M and CV. These results suggest that altering the promoter sequence and/or the active regulation allows altering M, CV and S independently.

Observing only subsets of this data, we find it to be in accordance with the model considered (Fig. 1E). E.g., consider the conditions LA, LA(75) and LA(50), which differ only in [RNAP] [14]. In these, as

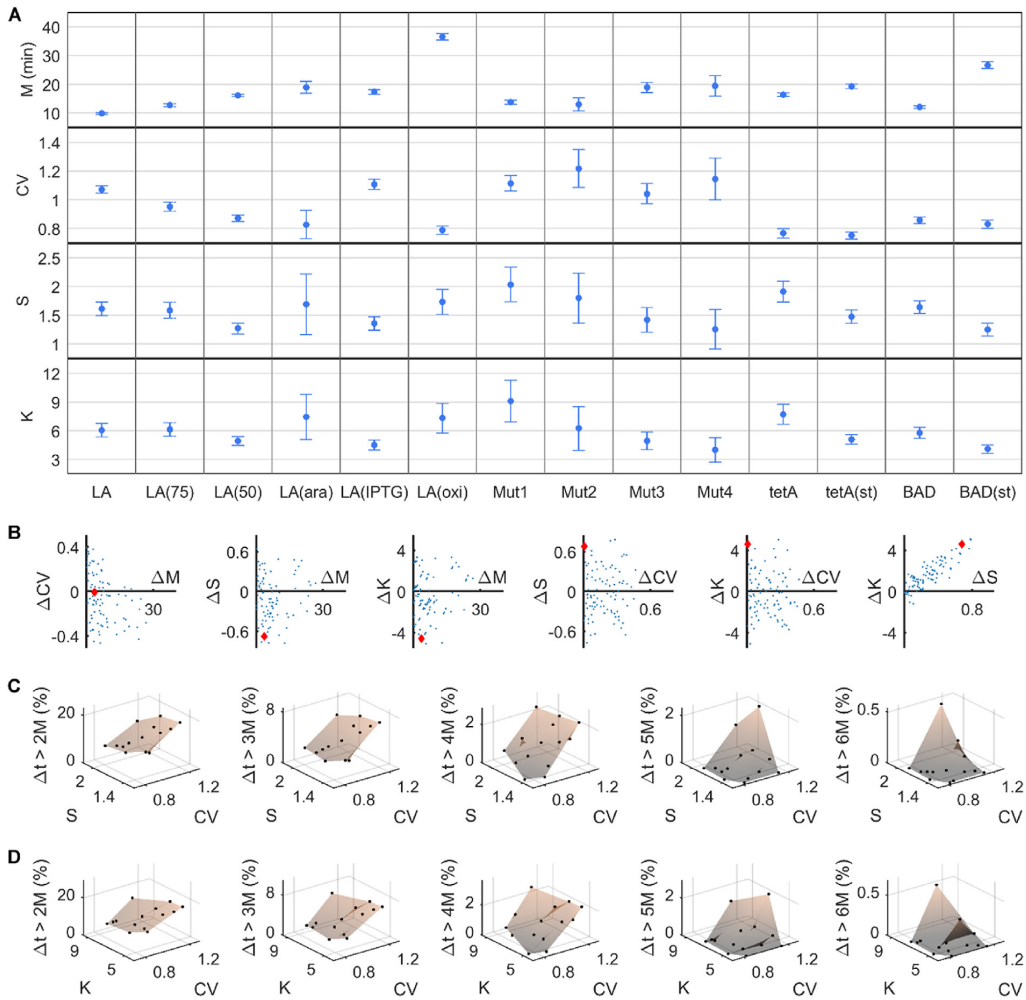


Fig. 2. Skewness (S) and kurtosis (K) affect the probability of crossing upper-bound thresholds in the time length of the intervals between consecutive RNA production events in individual cells (Δt). (A) Mean (M), coefficient of variation (CV), S and K of the distribution of Δt intervals (~600 cells per condition). S and K vary independently from M and CV. Error bars denote SEM. (B) Pairwise differences (Δ) in M, CV, S and K between conditions (blue dots). The red diamond is the difference between LA(IPTG) and Mut1 conditions that illustrates how changes in S and K can be independent from changes in M and CV. (C and D) Percentage of Δt intervals (black dots) that are longer than a given threshold (from 2 M to 6 M) against (C) CV and S, and (D) CV and K. Also shown is the natural neighbour interpolation surface.

[RNAP] decreases, M increases and CV decreases. Meanwhile, S and K decrease (weakly) as [RNAP] decreases. This change is weak enough so that, as shown in the next section, the only significant difference in S is between the two extreme conditions, LA and LA(50), and differences in K are not statistically significant (Supplementary Table S1).

Mutations in $P_{lac/ara\text{-}1}$ (Supplementary Fig. S1) also cause significant behavioural changes. Namely, M, CV and S differ between the mutants independently from each other, and only changes in S and K appear to be correlated. The same is observed when considering only the induction schemes of $P_{lac/ara\text{-}1}$ (LA, LA(ara) and LA(IPTG) conditions). Oxidative stress also affects M, CV, S and K significantly, when compared to the control. Further, comparing the three promoters tested here ($P_{lac/ara\text{-}1}$, P_{tetA} and P_{BAD}), again M, CV and S differ in an independent way, and only the differences between conditions in S and K exhibit a similar trend.

Finally, comparing P_{tetA} and P_{BAD} in the exponential and stationary growth phases (Supplementary Fig. S2A,B), we find that both differ significantly in M, S and K with the growth phase. This agrees with the findings in [23], which reported that the kinetics of rate-limiting steps in transcription changes with σ^{38} numbers (even in σ^{70}-dependent promoters). Interestingly, the differences in M, CV, S and K between growth phases are, qualitatively, the same in both promoters, supporting that they have the same cause.

We also tested whether the differences in M, CV, S and K between conditions could be explained by differences between the distributions of cell lifetimes or between the distributions of intracellular RNAP concentrations. The results of this test indicate that the features of the Δt distribution cannot be explained by the features of either these distributions (Supplementary results, Section 2.4; Supplementary Figs. S4A and S5).

Table 2
Pearson's correlation coefficient r (with the corresponding two-tailed *p*-value) for all conditions, for the subset 'Mutants', where only the promoter sequence differs between conditions, and for the subset 'Regulatory factors', where only the inducers or RNA polymerase concentrations differ between conditions. For *p*-values ≤ 0.05, the null hypothesis that there is no correlation is rejected.

	M vs CV	M vs S	M vs K	CV vs S	CV vs K	S vs K
All conditions	−0.44 (0.12)	−0.19 (0.52)	−0.08 (0.80)	0.01 (0.98)	−0.10 (0.73)	0.94 (< 0.001)
Mutants	−0.12 (0.85)	−0.64 (0.24)	−0.56 (0.32)	0.27 (0.66)	0.07 (0.91)	0.96 (< 0.01)
Regulatory factors	−0.47 (0.43)	−0.24 (0.70)	0.02 (0.98)	−0.17 (0.79)	−0.54 (0.34)	0.91 (0.03)

3.2. Promoter sequence and regulatory factors suffice to alter skewness and kurtosis of RNA production kinetics independently from its mean and coefficient of variation

To determine whether changes in M, CV, S and K between conditions are uncorrelated in a statistical sense, we first calculated linear correlations between each pair of these features when considering all 14 conditions (Fig. 2A). Results in Table 2 show no significant correlation between all pairs, except between S and K. The result holds also when applying the Bonferroni-Holm correction for multiple comparisons (the corrected *p*-value in the case of S and K is < 0.001). Tests for non-linear correlations (Kendall's and Spearman's rank correlation coefficients) give the same qualitative results. While this could be due to the lack of significant changes in M and CV, results in Fig. 2A reject this hypothesis. We thus conclude that all features can differ between conditions in an uncorrelated way, aside from S and K.

We also performed pairwise comparisons of M, CV, S and K between each pair of the 14 conditions. The results (Supplementary Table S1) show statistically significant differences between many pairs of conditions, indicating that all features differ widely between conditions. In detail, one observes that it is possible to alter S and K significantly, while CV is kept unchanged (e.g. between LA(IPTG) and Mut1). Similarly, the same is possible keeping M unchanged (e.g. between LA(50) and tetA).

Next, we quantified the degree with which each feature can differ between conditions while another feature is kept constant. In Fig. 2B we show all pairwise differences in M, CV, S and K between conditions. In all cases, we find that a feature can differ widely while the others remain mostly unchanged, except between S and K.

Finally, we investigated how S and K change as a function of the promoter sequence and the regulatory factors. For this, we considered two subsets of the data above. The first subset ('Mutants') includes the original $P_{lac/ara\text{-}1}$ promoter (LA) and the 4 mutants, specifically 1 single-point mutant (Mut1) and 3 three-point mutants (Mut2, Mut3 and Mut4) (Supplementary Fig. S1). The second subset ('Regulatory factors') includes the control (LA), two conditions with different [RNAP] (LA(75) and LA(50)) and two induction schemes (LA(IPTG) and LA(ara)). From Table 2, we conclude that changes in S (and K), due to point mutations and/or due to altering the concentrations of the regulatory factors, are not correlated to the changes in CV and M.

As before, for both subsets, we tested whether the differences in M, CV, S and K between conditions could be explained by differences between the distributions of cell lifetimes. Again, the results showed that the features of the cell lifetimes distributions cannot explain the features of the Δt distribution (Supplementary results, Section 2.4; Supplementary Fig. S4B,C).

3.3. Increasing the skewness and kurtosis of RNA production kinetics enhances the probability of crossing upper bound thresholds in intervals between consecutive RNA production events

Stochastic models of gene expression assuming transcription initiation as a two-step process predict that changing these steps' kinetics can alter the noise in RNA production without changing the mean rate of RNA production [37]. If the intrinsic noise in transcription changes, so will the probability of crossing thresholds based on RNA numbers. Here we quantify this noise by the CV of the Δt distribution [17,18], because this distribution is not affected by noise in RNA degradation.

If this noise was symmetric around the mean of the Δt distribution, the CV would suffice to estimate the probability of threshold crossing. However, recent results [16,17] suggest that it can be significantly asymmetric. As such, a more accurate estimation of threshold crossing probabilities in RNA numbers requires calculating S and K of the Δt distribution.

To test whether S and K differ significantly between the conditions (Supplementary materials and methods, Section 1.3), we first obtained, for each condition, the fraction of individual Δt intervals that are longer than a given threshold. We considered the thresholds 2 M, 3 M, 4 M, 5 M and 6 M, to eliminate influences by the value of M. Results in Supplementary Table S2 indicate that the fraction of intervals that cross a specific threshold differ between conditions, particularly for higher thresholds.

Next, to determine whether it is CV or S (and K) that is responsible for the differences in threshold crossing probabilities between conditions, we plotted the percentage of intervals in each condition that crossed each threshold against CV and S. We also calculated the natural neighbour interpolation surfaces (using MATLAB R2017b function scatteredInterpolant [38]).

Results in Fig. 2C show that for the lower thresholds (2 M and 3 M), varying S does not alter significantly the chance of threshold crossing, while changing CV does. For higher thresholds (4 M and 5 M), both S and CV are relevant. For the highest threshold (6 M), the relevance of S further increases. Equivalent conclusions are reached when considering K instead of S (Fig. 2D).

Overall, tuning S and K of the Δt distribution allows altering significantly the probability of crossing upper-bound thresholds in Δt values and, thus, of crossing lower-bound thresholds of RNA numbers in individual cells.

3.4. Skewness and kurtosis of RNA production kinetics can be tuned by the rate-limiting steps in transcription initiation

Previous studies have established that CV can be tuned by changing the kinetics of the rate-limiting steps in transcription initiation [14,16,17]. In particular, for example, changing the average time spent in the events prior (τ_{prior}) and after (τ_{after}) commitment to open complex formation without changing M, allows tuning noise in RNA production without affecting the rate of this production [16]. We hypothesised that S and K could be similarly regulated.

To test this, for each condition, we first estimated the mean fraction of time spent in the events prior to commitment to open complex formation (τ_{prior}/M) from τ plots (Materials and methods, paragraphs 1–2). Namely, we plotted the inverse of the relative RNA production rate, as measured by qPCR, against the inverse of the relative RNAP concentration, as measured by Western blot (Supplementary materials and methods, Section 1.4). Then, we fitted a line to the data from which we obtain τ_{prior}/M (Fig. 3A and Supplementary Table S3). Finally, from this and the value of M (Fig. 2A), we obtained the absolute values of τ_{prior} and τ_{after} for each condition (Supplementary Table S3).

Cells in the stationary phase (conditions tetA(st) and BAD(st)) are

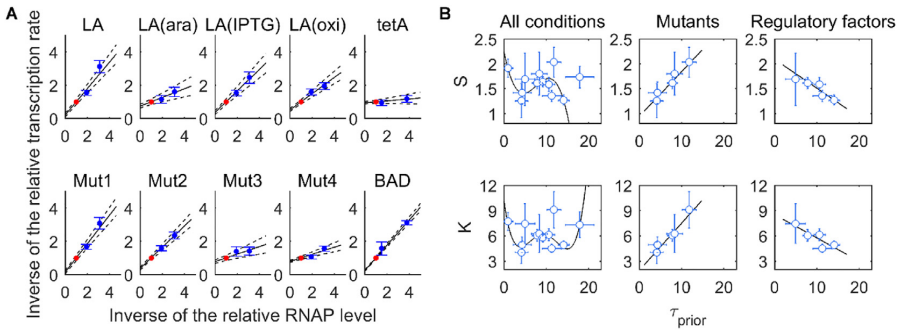


Fig. 3. Skewness (S) and kurtosis (K) of the distribution of intervals between consecutive RNA production events in individual cells change linearly with the fraction of time spent in events prior to commitment to the open complex formation (τ_{prior}). (A) Relative τ plots. Transcription rates are measured by qPCR, and RNA polymerase (RNAP) levels are measured by Western blot (Supplementary Fig. S2C). Values are shown relative to the control condition (red dot). Error bars denote the standard error. The solid line is the best-fitting line, and the dashed lines denote the standard error of the fit. (B) S and K plotted against τ_{prior}. Values plotted for all conditions and for subsets ('Mutants' and 'Regulatory factors'). Error bars denote SEM. The black line is the best-fitting model. The linear relationships are statistically significant when the set of variables allowed to change between conditions is restricted to either the sequence of the promoter or the regulatory factors. When all variables are allowed to differ simultaneously, the best-fitting model is a polynomial of the third or fourth degree.

not considered since, in these conditions, σ^{38} numbers are sufficiently high for the amount of core RNAP enzymes to become a less accurate proxy of the RNAP-σ^{70} holoenzymes levels [23]. Additional factors that may differ include potential sRNA regulation [39,40], ppGpp [41], cAMP (see e.g. [42]) contribute to these differences.

We assessed whether S and K change with τ_{prior}. For this, we plotted S and K against τ_{prior} in each condition (Fig. 3B) and performed likelihood ratio tests (at significance level of 0.05) between the best-fit polynomial models (using weighted total least squares approach [14,43]) with degrees ranging from 0 to N-1, with N being the number of conditions (*p*-values are shown in Supplementary Table S4). We also tested whether the data can be better explained by a model where τ_{prior} does not differ between conditions, by performing a likelihood ratio test between this model and the selected best-fitting polynomial (Supplementary Table S4). For both S and K, the zero-degree and the first-degree polynomial models, as well as the models where τ_{prior} is constant, are rejected in favour of higher-degree polynomials.

The fact that S and K are best fit by, respectively, third and fourth degree polynomials (that still do not explain all data points) illustrates the level of complexity of the data. This is likely due to the conditions differing in several factors (promoter, induction scheme, etc.). We thus next consider, as above, the subsets 'Mutants' and 'Regulatory factors'. For each, we perform, also as above, likelihood ratio tests to determine the best fitting models (Supplementary Table S4). In both subsets, a 1st degree model is preferred.

Meanwhile, from the Pearson's correlation coefficient (with the corresponding two-tailed *p*-value) between τ_{prior} and skewness (S) and kurtosis (K), for the subsets 'Mutants' and 'Regulatory factors', we find a significant correlation in all cases (absolute correlation values above 0.85 and *p*-values $\leq$ 0.05), except for K in 'Regulatory factors', where the *p*-value equals 0.06. Overall, the results suggest that, similarly to M and CV, tuning τ_{prior} can regulate S and K. This implies that the lower bound threshold crossing probability of RNA numbers over time can be tuned.

Next, we performed the same analysis for changing τ_{after} and τ_{prior}/M. Contrary to when considering τ_{prior}, the results (Supplementary Fig. S6 and Supplementary Tables S5-S6) do not allow establishing statistically significant relationships (also the *p*-values from the Pearson's correlation were larger than 0.05).

Interestingly, the linear relationships of S and K with τ_{prior} are positive in the subset 'Mutants' and negative in the subset 'Regulatory factors'. This strongly indicates that τ_{prior} is not the only parameter defining these features. Namely, we hypothesise that these relationships may depend on what causes τ_{prior} to differ between the conditions. For instance, in one subset, the difference may be due to differences in the mean time required by the RNAP to complete a closed complex formation, while in the other subset the differences may be in the number of times that the RNAP fails to commit to the open complex formation. These potential differences could be accounted for in the model by tuning k_1, k_{-1} and k_2 (Supplementary materials and methods, Section 1.6), but cannot be detected by the measurements conducted here. Future work is needed to test this hypothesis.

3.5. Skewness and kurtosis of the RNA production kinetics and of the distribution of protein expression levels in individual cells are negatively correlated

To assess if changes in S and K of the Δt distribution could affect the phenotypic distribution of cell populations, we next investigate whether these changes result in significant changes in the distribution of protein expression levels of a cell population. This is expected given the known coupling between transcription and translation in prokaryotes [44–46]. Nevertheless, it is reasonable to assume that noise in the stochastic process of translation (e.g. on the time to be completed once initiated) would render changes in S and K ineffectual on protein expression levels. A model of gene expression in prokaryotes accounting for the coupling between the two processes is shown in Supplementary materials and methods, Section 1.6.

We first tested whether the mean protein expression levels of the cell populations follow their mean RNA numbers. For that, we measured RNA numbers (by microscopy) and protein mean expression levels (by flow cytometry) produced under the control of $P_{lac/ara-1}$ for various induction conditions. We expect the same relationship in all other constructs used here, as they have identical sequences following the promoter sequence. Results in Supplementary Fig. S7 show that the average number of proteins in a cell population follows the average RNA numbers.

Given this, since M of the Δt distribution is negatively correlated with the mean RNA numbers of the cell population, one can expect it to also be negatively correlated to the mean number of proteins. Using the same promoter as a case-study, we tested whether the skewness and kurtosis of the distribution of protein expression levels of a cell population are sensitive to the induction strength. For this, we measured the total fluorescence intensity level of the proteins expressed by $P_{lac/ara-1}$ in individual cells for various induction levels using flow cytometry (Materials and methods). From these, for each induction level, we

obtained the distribution of fluorescence of individual cells (in arbitrary units). For each of these distributions, we estimated the mean (M_P), skewness (S_P) and kurtosis (K_P) as previously (Supplementary materials and methods, Section 1.3). From Supplementary Fig. S8, we find that S_P and K_P can differ with induction strength. Also, it is possible to have, for similar values of M_P, significantly different values of S_P and K_P (e.g. conditions 0 to 25 μM). Further, conditions differing in M_P can have similar values of S_P and K_P (beyond 100 μM). Overall, we find that, as for the Δt distributions, S_P and K_P can change independently from M_P and vice versa.

Next, we investigate whether changes in S and K of the Δt distribution due to changing the promoter sequence or its regulation reflect on the distribution of protein expression levels, as expected from the model. For this, we consider, respectively, the subsets 'Mutations' and 'Induction schemes'. We note that, within these subsets, the cells are grown under identical culture conditions and do not differ in their fundamental physiology, and are therefore not expected to differ in, e.g., ribosome population and/or in any other global gene expression regulators, such as [RNAP] or σ factors. For these reasons, here we do not consider the other conditions in Table 1, as the translation rate or protein maturation time may differ significantly from the control.

For each condition considered, we measured the fluorescence intensity from the target proteins by flow cytometry (Materials and methods) and obtained the single-cell distributions of protein fluorescence intensity. Next, we estimated its M_P (in arbitrary units), S_P and K_P, as previously. We also measured M_P for cells with an uninduced $P_{lac/ara\text{-}1}$ to obtain a reference point for the values of M_P. In this regard, the LA(ara) condition was not included in the subsequent analysis since, for unknown reasons, its protein expression levels were not significantly above those of the uninduced $P_{lac/ara\text{-}1}$ (Fig. 4).

In Fig. 4, we show M_P, S_P and K_P plotted against M, S and K, respectively, along with the best-fitting models obtained by likelihood ratio tests (Supplementary Table S7). In all cases, the linear model is preferred. We also calculated the Pearson's correlation coefficient for each case. The results agree with the likelihood ratio tests. Namely, there are strong, statistically significant (p-values ≤ 0.05), negative correlations between M and M_P (−0.82) and between S and S_P (−0.86). Between K vs K_P the negative correlation is also strong (−0.70), but the p-value is 0.12, likely due to higher uncertainty. From the statistically significant linear relationships, we conclude that the differences in skewness and kurtosis of the Δt distribution between conditions result in statistically significant differences between the skewness and kurtosis of the corresponding protein distributions, in a manner that is consistent with the model. As a side note, our data does not allow investigating whether a similar (expected) correlation exist in the case of CV and CV_P, since LA, LA(IPTG), and the mutant promoters have CV values that cannot be distinguished in a statistical sense (Supplementary Table S1).

Finally, to assess if the values of M could explain the values of S_P and K_P, we performed likelihood ratio tests (as above) between M and S_P and between M and K_P. A polynomial model of the 1st order was rejected in both cases (p-values equal 0.04 and 0.02, respectively). Also, we failed to find linear correlations (p-values equal 0.06 and 0.25, respectively). We conclude that M is not correlated with either S_P or K_P, as expected from the lack of the correlation between M and S or K.

4. Discussion and conclusions

Previous research have established that bacterial transcription is mostly regulated at the stage of initiation [10–12,47]. This regulation, e.g. by transcription factors and σ factors, affects the mean and variance in RNA and protein numbers [10–12,19,48]. From the dynamics point of view, these and similar regulatory molecules were shown to have direct effect on the kinetics of the rate-limiting steps in transcription initiation of a gene (assessed here by τ_{prior} and τ_{after}), resulting in changes in the mean and variance of its distribution of intervals between consecutive RNA production events in individual cells (Δt distribution) [14,23].

Here we provided evidence that the fraction of cells that reach high thresholds in RNA and protein numbers of an externally regulated gene can be tuned by altering the skewness and kurtosis of its Δt distribution. Also, we showed that this can be achieved without significantly altering the mean and CV of this distribution. Further, this regulation is possible by tuning τ_{prior} and τ_{after} alone which can be altered by changing the promoter sequence, the induction scheme, or the intracellular RNAP concentration.

On the other hand, we did not find significant evidence that the skewness and kurtosis could be altered independently from one another. Instead, they exhibit a strong positive correlation (Fig. 2B, ΔK vs ΔS, and Table 2). We suggest that this may be due to the variability of the time length between transcription events along with the existence of mechanical constraints imposed by the transcription machinery. This variability is visible in Fig. S3, which shows that the distributions of intervals between transcriptions are broad, with several intervals having a short time-length. This limits how much the kurtosis of this distribution can increase by increasing the tail on the left side. This limit does not exist on the right side. Thus, increasing the kurtosis of one of these distributions by increasing the size of the right tail cannot be easily compensated on the left side so that the skewness remains unaltered.

Regulation of asymmetry and tailedness of gene expression, so far, has only been considered in the context of small genetic circuits or complex regulatory pathways (e.g. [3]). Given the above, our findings suggest that regulatory mechanisms of individual genes suffice for this regulation as well. In particular, based on the data from the conditions in Table 1, we found statistically significant linear relationships between τ_{prior} and the skewness and kurtosis of the Δt distribution, provided that either only the promoter sequence or the regulatory factors (i.e. inducers and RNAP concentrations) differ between the conditions. We hypothesise that relationships more complex than linear are also possible, if more than one parameter is allowed to change. E.g. in the future it would be of interest to investigate whether the data in Fig. 3B

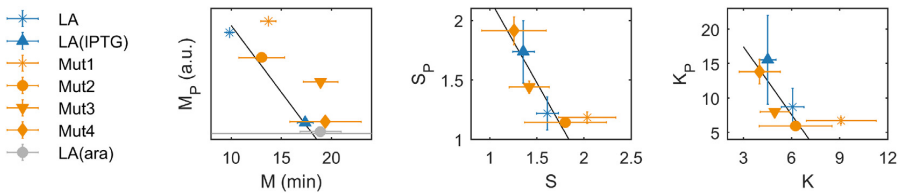


Fig. 4. Mean (M), skewness (S), and kurtosis (K) of the distribution of protein expression levels in individual cells change linearly with the corresponding features of the distribution of time intervals between consecutive RNA production events (Δt distribution). (From left to right) M_P, S_P and K_P of the single-cell distributions of protein levels against the corresponding feature of the Δt distributions (extracted from Fig. 2A). Error bars denote SEM (in some cases, the SEM is too small to produce visible error bars). The solid line is the best fitting model. On the left plot, the horizontal grey line corresponds to M_P for an uninduced $P_{lac/ara\text{-}1}$ which is used as a reference point (SEM is too small to be represented). M_P of LA(ara) is not considered in model fitting.

could be better explained by consider both τ_{prior} and τ_{after} simultaneously. Nevertheless, the linear relationships found here are evidence that the skewness and kurtosis are evolvable (i.e. sequence dependent) and adaptable (i.e. subject to regulation). Meanwhile, the strong correlation between RNA production kinetics and single-cell distribution of protein levels suggests that tuning these skewness and kurtosis can have a significant impact on the phenotypic distribution of the cell population.

It is well known that the two rate-limiting steps of transcription initiation here considered (i.e. the events prior and after commitment to open complex formation) are composed of specific 'sub-steps', such as promoter escape [49–51], reversibility of the closed complex formation and isomerization [13,52,53]. Further developments in the dissection techniques of the in vivo kinetics of these sub-steps during transcription initiation should allow characterising, in greater detail, their contributions to the regulation of the skewness and kurtosis of the distributions of RNA production kinetics and corresponding protein numbers. This should also allow establishing precise methods for tuning the skewness and kurtosis of these distributions.

It is worth noting that the findings here reported do not discard the importance of other mechanisms of regulation of protein numbers in *E. coli*, such as regulation by sRNAs [39,40,54]. Here we did not consider this mechanism since all target genes studied shared the same elongation region. It will be of interest to study whether this post-transcription regulation process also allows tuning the skewness and kurtosis of single-cell distributions of protein numbers, particularly given its known effects on the cell-to-cell variability in protein numbers [55,56] and protein numbers' threshold-crossing propensities [39,57].

Finally, while a strict relationship between the skewness and kurtosis in the RNA and protein numbers was established here, the implications of these findings in the context of the qualitative behaviour of genetic circuits remain to be demonstrated. We expect the amplitude of these effects to differ with the circuit topology, as in the case of mean and variance [58–60]. If the effects are significant, direct regulation of these features in genetic circuits (by tuning the rate limiting steps of the component genes) should allow a more precise control of their kinetics, towards enhancing their robustness to fluctuations in molecular numbers or environmental changes, and sensitivity to external signals.

Funding

Work supported by Tampere University of Technology Graduate School Grant (Finland) [to S.S.]; Pirkanmaa Regional Fund [to V.K.K.]; Academy of Finland [295027, 305342 to A.S.R.]; and Jane and Aatos Erkko Foundation [610536 to A.S.R.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding for open access charge: Academy of Finland [295027].

Conflict of interest

The authors declare that they have no conflict of interest.

Transparency document

The Transparency document associated with this article can be found, in online version.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2018.12.005.

References

[1] S.S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, Nat. Genet. 31 (2002) 64–68, https://doi.org/10.1038/ng881.
[2] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (2002) 824–827, https://doi.org/10.1126/science.298.5594.824.
[3] U. Alon, Network motifs: theory and experimental approaches, Nat. Rev. Genet. 8 (2007) 450–461, https://doi.org/10.1038/nrg2102.
[4] E. Kussell, S. Leibler, Phenotypic diversity, population growth, and information in fluctuating environments, Science 309 (2005) 2075–2078, https://doi.org/10.1126/science.1114383.
[5] Y. Liu, A. Beyer, R. Aebersold, On the dependency of cellular protein levels on mRNA abundance, Cell 165 (2016) 535–550, https://doi.org/10.1016/j.cell.2016.03.014.
[6] C. Vogel, E.M. Marcotte, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses, Nat. Rev. Genet. 13 (2012) 227–232, https://doi.org/10.1038/nrg3185.
[7] J.A. Bernstein, A.B. Khodursky, P.-H. Lin, S. Lin-Chao, S.N. Cohen, Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 9697–9702, https://doi.org/10.1073/pnas.112318199.
[8] H. Chen, K. Shiroguchi, H. Ge, X.S. Xie, Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*, Mol. Syst. Biol. 11 (2015) 781, https://doi.org/10.15252/msb.20145794.
[9] M.P. Deutscher, Degradation of RNA in bacteria: comparison of mRNA and stable RNA, Nucleic Acids Res. 34 (2006) 659–666, https://doi.org/10.1093/nar/gkj472.
[10] S.M. McLeod, R.C. Johnson, Control of transcription by nucleoid proteins, Curr. Opin. Microbiol. 4 (2001) 152–159, https://doi.org/10.1016/S1369-5274(00)00181-8.
[11] E.F. Ruff, A.C. Drennan, M.W. Capp, M.A. Poulos, I. Artsimovitch, T.M. Record Jr., *E. coli* RNA polymerase determinants of open complex lifetime and structure, J. Mol. Biol. 247 (2015) 2435–2450, https://doi.org/10.1016/j.jmb.2015.05.024.
[12] D.F. Browning, S.J.W. Busby, Local and global regulation of transcription initiation in bacteria, Nat. Rev. Microbiol. 14 (2016) 638–650, https://doi.org/10.1038/nrmicro.2016.103.
[13] P.L. deHaseth, M.L. Zupancic, T.M. Record Jr., RNA polymerase-promoter interactions: the comings and goings of RNA polymerase, J. Bacteriol. 180 (1998) 3019–3025 (PMID: 9620948).
[14] J. Lloyd-Price, S. Startceva, V. Kandavalli, J.G. Chandraseelan, N. Goncalves, S.M.D. Oliveira, A. Häkkinen, A.S. Ribeiro, Dissecting the stochastic transcription initiation process in live *Escherichia coli*, DNA Res. 23 (2016) 203–214, https://doi.org/10.1093/dnares/dsw009.
[15] W.R. McClure, Rate-limiting steps in RNA chain initiation, Proc. Natl. Acad. Sci. U. S. A. 77 (1980) 5634–5638, https://doi.org/10.1073/pnas.77.10.5634.
[16] J. Mäkelä, V. Kandavalli, A.S. Ribeiro, Rate-limiting steps in transcription dictate sensitivity to variability in cellular components, Sci. Rep. 7 (2017) 10588, https://doi.org/10.1038/s41598-017-11257-2.
[17] S.M.D. Oliveira, A. Häkkinen, J. Lloyd-Price, H. Tran, V. Kandavalli, A.S. Ribeiro, Temperature-dependent model of multi-step transcription initiation in *Escherichia coli* based on live single-cell measurements, PLoS Comput. Biol. 12 (2016) e1005174, , https://doi.org/10.1371/journal.pcbi.1005174.
[18] A. Häkkinen, A.S. Ribeiro, Characterizing rate limiting steps in transcription from RNA production times in live cells, Bioinformatics 32 (2016) 1346–1352, https://doi.org/10.1093/bioinformatics/btv744.
[19] S. Chong, C. Chen, H. Ge, X.S. Xie, Mechanism of transcriptional bursting in bacteria, Cell 158 (2014) 314–326, https://doi.org/10.1016/j.cell.2014.05.038.
[20] I. Golding, J. Paulsson, S.M. Zawilski, E.C. Cox, Real-time kinetics of gene activity in individual bacteria, Cell 123 (2005) 1025–1036, https://doi.org/10.1016/j.cell.2005.09.031.
[21] A. Häkkinen, A.S. Ribeiro, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, Bioinformatics 31 (2015) 69–75, https://doi.org/10.1093/bioinformatics/btu592.
[22] H. Tran, S.M.D. Oliveira, N. Goncalves, A.S. Ribeiro, Kinetics of the cellular intake of a gene expression inducer at high concentrations, Mol. BioSyst. 11 (2015) 2579–2587, https://doi.org/10.1039/C5MB00244C.
[23] V.K. Kandavalli, H. Tran, A.S. Ribeiro, Effects of σ factor competition are promoter initiation kinetics dependent, Biochim. Biophys. Acta 1859 (2016) 1281–1288, https://doi.org/10.1016/j.bbagrm.2016.07.011.
[24] J.R. Peterson, J.A. Cole, J. Fei, T. Ha, Z.A. Luthey-Schulten, Effects of DNA replication on mRNA noise, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 15886–15891, https://doi.org/10.1073/pnas.1516246112.
[25] H. Lineweaver, D. Burk, The determination of enzyme dissociation constants, J. Am. Chem. Soc. 56 (1934) 658–666, https://doi.org/10.1021/ja01318a036.
[26] R. Lutz, H. Bujard, Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I_1-I_2 regulatory elements, Nucleic Acids Res. 25 (1997) 1203–1210, https://doi.org/10.1093/nar/25.6.1203.
[27] I. Golding, E.C. Cox, RNA dynamics in live *Escherichia coli* cells, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 11310–11315, https://doi.org/10.1073/pnas.0404443101.
[28] A. Gupta, J. Lloyd-Price, R. Neeli-Venkata, S.M.D. Oliveira, A.S. Ribeiro, In vivo kinetics of segregation and polar retention of MS2-GFP-RNA complexes in *Escherichia coli*, Biophys. J. 106 (2014) 1928–1937, https://doi.org/10.1016/j.bpj.2014.03.035.
[29] B.P. Bratton, R.A. Mooney, J.C. Weisshaar, Spatial distribution and diffusive motion of RNA polymerase in live *Escherichia coli*, J. Bacteriol. 193 (2011) 5138–5146, https://doi.org/10.1128/JB.00198-11.
[30] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York,

0879693096, 1989.
[31] N.S.M. Goncalves, L. Martins, H. Tran, S.M.D. Oliveira, R. Neeli-Venkata, J.M. Fonseca, A.S. Ribeiro, *In vivo* single-molecule dynamics of transcription of the viral T7 Phi 10 promoter in *Escherichia coli*, The 8th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2016), 2016 978-1-61208-488-6, pp. 9–15.
[32] D.G. Gibson, L. Young, R.-Y. Chuang, J.C. Venter, C.A. Hutchison III, H.O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases, Nat. Methods 6 (2009) 343–345, https://doi.org/10.1038/nmeth.1318.
[33] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method, Methods 25 (2001) 402–408, https://doi.org/10.1006/meth.2001.1262.
[34] A. Häkkinen, A.-B. Muthukrishnan, A. Mora, J.M. Fonseca, A.S. Ribeiro, CellAging: a tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*, Bioinformatics 29 (2013) 1708–1709, https://doi.org/10.1093/bioinformatics/btt194.
[35] M. Razo-Mejia, S.L. Barnes, N.M. Belliveau, G. Chure, T. Einav, M. Lewis, R. Phillips, Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction, Cell Syst. 6 (2018) 456–469.e10, https://doi.org/10.1016/j.cels.2018.02.004.
[36] J.D. Boeke, P. Model, A prokaryotic membrane anchor sequence: carboxyl terminus of bacteriophage f1 gene *III* protein retains it in the membrane, Proc. Natl. Acad. Sci. U. S. A. 79 (1982) 5200–5204, https://doi.org/10.1073/pnas.79.17.5200.
[37] S.M.D. Oliveira, M.N.M. Bahrudeen, S. Startceva, A.S. Ribeiro, Estimating effects of extrinsic noise on model genes and circuits with empirically validated kinetics, in: M. Pelillo, I. Poli, A. Roli, R. Serra, D. Slanzi, M. Villani (Eds.), Artificial Life and Evolutionary Computation. WIVACE 2017, Springer, 2018, pp. 181–193, , https://doi.org/10.1007/978-3-319-78658-2_14.
[38] I. Amidror, Scattered data interpolation methods for electronic imaging systems: a survey, J. Electron. Imaging 11 (2002) 157–176, https://doi.org/10.1117/1.1455013.
[39] E. Levine, Z. Zhang, T. Kuhlman, T. Hwa, Quantitative characteristics of gene regulation by small RNA, PLoS Biol. 5 (2007) e229, , https://doi.org/10.1371/journal.pbio.0050229.
[40] E.G.H. Wagner, P. Romby, Small RNAs in bacteria and archaea: who they are, what they do, and how they do it, Adv. Genet. 90 (2015) 133–208, https://doi.org/10.1016/bs.adgen.2015.05.001.
[41] C. Condon, C. Squires, C.L. Squires, Control of rRNA transcription in *Escherichia coli*, Microbiol. Rev. 59 (1995) 623–645 (PMID: 8531889).
[42] C.M. Johnson, R.F. Schleif, In vivo induction kinetics of the arabinose promoters in *Escherichia coli*, J. Bacteriol. 177 (1995) 3438–3442 (PMID: 7768852).
[43] M. Krystek, M. Anton, A weighted total least-squares algorithm for fitting a straight line, Meas. Sci. Technol. 18 (2007) 3438–3442, https://doi.org/10.1088/0957-0233/18/11/025.
[44] O. Dahan, H. Gingold, Y. Pilpel, Regulatory mechanisms and networks couple the different phases of gene expression, Trends Genet. 27 (2011) 316–322, https://doi.org/10.1016/j.tig.2011.05.008.
[45] C. Yanofsky, Attenuation in the control of expression of bacterial operons, Nature 289 (1981) 751–758, https://doi.org/10.1038/289751a0.
[46] S. Proshkin, A.R. Rahmouni, A. Mironov, E. Nudler, Cooperation between translating ribosomes and RNA polymerase in transcription elongation, Science 328 (2010) 504–508, https://doi.org/10.1126/science.1184939.
[47] D.F. Browning, S.J.W. Busby, The regulation of bacterial transcription initiation, Nat. Rev. Microbiol. 2 (2004) 57–65, https://doi.org/10.1038/nrmicro787.
[48] W.R. McClure, Mechanism and control of transcription initiation in prokaryotes, Annu. Rev. Biochem. 54 (1985) 171–204, https://doi.org/10.1146/annurev.bi.54.070185.001131.
[49] D. Duchi, D.L.V. Bauer, L. Fernandez, G. Evans, N. Robb, L.C. Hwang, K. Gryte, A. Tomescu, P. Zawadzki, Z. Morichaud, et al., RNA polymerase pausing during initial transcription, Mol. Cell 63 (2016) 939–950, https://doi.org/10.1016/j.molcel.2016.08.011.
[50] L.M. Hsu, Promoter escape by *Escherichia coli* RNA polymerase, EcoSal Plus 3 (2008) 1–16, https://doi.org/10.1128/ecosalplus.4.5.2.2.
[51] A.N. Kapanidis, E. Margeat, S.O. Ho, E. Kortkhonjia, S. Weiss, R.H. Ebright, Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism, Science 314 (2006) 1144–1147, https://doi.org/10.1126/science.1131399.
[52] L.M. Hsu, Promoter clearance and escape in prokaryotes, Biochim. Biophys. Acta 1577 (2002) 191–207, https://doi.org/10.1016/S0167-4781(02)00452-9.
[53] T.M. Record Jr., W.S. Reznikoff, M.L. Craig, K.L. McQuade, P.J. Schlax, *Escherichia coli* RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription initiation, in: F.C. Neidhardt, R. Curtiss, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, D. Schneider, H.E. Umbarger (Eds.), *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology, 2nd ed., ASM press, Washington, DC, 1555810845, 1996, pp. 792–821.
[54] G. Storz, J. Vogel, K.M. Wassarman, Regulation by small RNAs in bacteria: expanding frontiers, Mol. Cell 43 (2011) 880–891, https://doi.org/10.1016/j.molcel.2011.08.022.
[55] R. Arbel-Goren, A. Tal, T. Friedlander, S. Meshner, N. Costantino, D.L. Court, J. Stavans, Effects of post-transcriptional regulation on phenotypic noise in *Escherichia coli*, Nucleic Acids Res. 41 (2013) 4825–4834, https://doi.org/10.1093/nar/gkt184.
[56] R. Arbel-Goren, A. Tal, B. Parasar, A. Dym, N. Costantino, J. Muñoz-García, D.L. Court, J. Stavans, Transcript degradation and noise of small RNA-controlled genes in a switch activated network in *Escherichia coli*, Nucleic Acids Res. 44 (2016) 6707–6720, https://doi.org/10.1093/nar/gkw273.
[57] P. Mehta, S. Goyal, N.S. Wingreen, A quantitative comparison of sRNA-based and protein-based gene regulation, Mol. Syst. Biol. 4 (2016) 221, https://doi.org/10.1038/msb.2008.58.
[58] E.D. Cameron, J.J. Collins, Tunable protein degradation in bacteria, Nat. Biotechnol. 32 (2014) 1276–1281, https://doi.org/10.1038/nbt.3053.
[59] L.G. Morelli, F. Jülicher, Precision of genetic oscillators and clocks, Phys. Rev. Lett. 98 (2007) 228101, https://doi.org/10.1103/PhysRevLett.98.228101.
[60] R. Zhu, A.S. Ribeiro, D. Salahub, S.A. Kauffman, Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models, J. Theor. Biol. 246 (2007) 725–745, https://doi.org/10.1016/j.jtbi.2007.01.021.

Supplementary Material for:

Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression

Sofia Startceva, Vinodh K. Kandavalli, Ari Visa and Andre S. Ribeiro

1. Supplementary Materials and Methods

1.1. Measuring times of RNA and proteins following induction

When measuring the integer-valued number of RNAs or the moment when a new RNA appears in a cell, for $P_{lac/ara-1}$ and its variants, following the procedure above, we induce the reporter gene with aTc and the target gene with arabinose (when appropriate, as in [1]). Next, 50 min later, we induce the target gene with a given amount of IPTG (Table 1). Images of cells are taken 1 h after that, from which RNA numbers are quantified. In time series measurements, imaging starts 10 min after induction with IPTG (for details, see Materials and Methods, section 2.5). For other promoters (P_{tetA} and P_{BAD}), the reporter gene, under the control of a P_{lac}, is induced with IPTG. Next, 50 min later, we induce the target gene (using the inducer specified in Table 1).

When measuring protein expression levels, we followed the same protocols as for measuring RNA numbers (aside from inducing MS2-GFP production), but we waited 90 min after induction of the target before performing the flow cytometry measurements. The additional 30 min compared to the RNA measurements are to account for the time for protein translation and maturation, in agreement with [2]. We also tested other waiting times (15, 45 and 60 min), but 30 min was the time interval that generated more consistent results between RNA and protein numbers in all conditions.

1.2. Image analysis of microscopy data

We used the software 'CellAging' [3]. It performs automated segmentation of phase-contrast images, followed by a manual correction. Next, confocal images are semi-automatically aligned with the phase-contrast images using thin-plate spline interpolation for the registration transform (for that, we manually select 5-8 landmarks that adjust the cell masks to the borders of the corresponding cells from the confocal images). After alignment, cell lineages are constructed (when applicable), by establishing the relationships between cell masks in sequential frames. Next, from each segmented cell, at each time point, fluorescent spots are detected automatically by the Gaussian surface-fitting algorithm [4]. From these data, time-series of fluorescent spots intensity were obtained for each cell, and the time points when novel RNA molecules appear in each cells were estimated [4]. This allows obtaining the time between consecutive RNA production events in individual cells (see Materials and Methods, section 2.6).

1.3. Analysing the mean, coefficient of variation, skewness and kurtosis of the Δt distribution

From the Δt distributions, we calculated M, CV, S and K in accordance with the definitions below, where $\langle \Delta t \rangle$, $\sigma_{\Delta t}$ and n denote the average, SD and sample size of the Δt distribution, respectively. In the case of S and K, we also applied the sample size correction [5]).

Feature	M	CV	S	K
Definition	$\langle \Delta t \rangle$	$\frac{\sigma_{\Delta t}}{\langle \Delta t \rangle}$	$\frac{\left\langle \left(\Delta t - \langle \Delta t \rangle\right)^3 \right\rangle}{\sigma_{\Delta t}^3}$	$\frac{\left\langle \left(\Delta t - \langle \Delta t \rangle\right)^4 \right\rangle}{\sigma_{\Delta t}^4}$
Corrected value	-	-	$\frac{\sqrt{n(n-1)}}{n-2} S$	$\frac{(n-1)}{(n-2)(n-3)}\left((n+1)K - 3(n-1)\right) + 3$

Next, we estimated the standard error of the mean (SEM) of these features using a non-parametric bootstrap method [6,7]. Namely, for each Δt distribution, we performed 10^5 random resamples with replacement and obtained the bootstrapped distributions of M, CV, S and K values. Since a bootstrapped distribution is expected to converge to Gaussian according to the central limit theorem, the standard deviation (SD) of each bootstrapped distribution is equivalent to the SEM of the corresponding feature. This allows using a 2-sample z-test to compare the estimated features between conditions.

The same methodology was also applied when extracting mean, coefficient of variation, skewness and kurtosis from other distributions, such as the distribution of protein expression levels in single cells.

1.4. Western blot measurements

Mean RNA production rates differ with the free RNAP concentration in the cells [8,9]. The RNAP concentrations in each condition, relative to the control, were assessed by measuring the level of the RpoC protein by Western blot. The results confirmed that the relative RNAP levels change linearly with media richness as first reported in [1] and then confirmed in [10–12]. To attain different concentrations of intracellular RNAP without altering significantly the growth rates of the cells, we grow the cultures in media of different richness (1x, 0.5x and 0.25x), as described above. Results are shown in Supplementary Figure S2C.

Pelleted cells were lysed with B-PER bacterial protein extraction reagent supplemented with a protease inhibitor for 10 min, at room temperature. Afterwards, the lysed cells were centrifuged at 15000xg for 10 min, and the supernatant was collected and diluted in the 4X laemmli sample loading buffer containing β-mercaptoethanol, after which it was boiled for 5 min at 95 °C. Each sample containing ~30 µg of total soluble proteins, were resolved by 4% to 20% TGX stain free precast gels (Biorad). Proteins were separated by electrophoresis and then electro-transferred to the PVDF membrane. Membranes were blocked with 5% non-fat milk for 1 h at room temperature and incubated with respective primary RpoC antibodies of 1:2000 dilutions (Biolegend) overnight at 4 °C, followed by the appropriate HRP-secondary antibodies 1:5000 dilutions (Sigma Aldrich) for 1 h at room temperature. For detection, chemiluminescence reagent (Biorad) was

used. Images were generated by the Chemidoc XRS system (Biorad). Quantification of the band intensity was done using Image lab software (version 5.2.1).

1.5. Estimating the time spent in transcription initiation prior and after commitment to open complex formation

To estimate τ_{prior} and τ_{after}, we use a methodology based on measuring RNA production rates at different intracellular RNAP concentrations in live cells. The method follows a similar protocol, established using *in vitro* techniques [13], and was adapted for *in vivo*, single-cell, single-RNA detection measurement techniques [1].

From the *in vitro* measurements one can directly measure the time-length of the closed and open complex formations since one can limit which components are in the reaction vessels and which reactions can take place during transcription initiation [13]. This is not possible in live cells. Also, one can only measure (by microscopy and single-RNA detection by MS2-GFP tagging) the time intervals between consecutive RNA production events in individual cells (Δt) at different intracellular RNAP concentrations [1]. As such, all normal events during transcription initiation can occur, unlike when using *in vitro* techniques. Consequently, τ_{prior} (Figure 1) is not the mean time-length of the closed complex formation since, among other, it also is affected by transient promoter locking events. Similarly, τ_{after} is not the mean time-length of the open complex formation since it is affected by other events, such as promoter escape. Rather, τ_{prior} is the mean time-length of all events preceding commitment to open complex formation, while τ_{after} is the mean time-length of all events subsequent to this commitment.

According to the model (Figure 1E), τ_{prior} depends on the intracellular concentration of RNAP while τ_{after} does not. Thus, provided knowledge on M (mean of the Δt distribution, which equals the inverse of the mean RNA production rate), τ_{prior} and τ_{after} can be estimated from measurements of the rates of RNA production at different RNAP concentrations [1,10,13–15] (Materials and Methods, section 2.3). For that, one can use a Lineweaver–Burk plot [16] of the inverse of the RNA production rate versus the inverse of the RNAP concentration ([RNAP]) (also named 'τ plot'). From this, one can estimate τ_{after} (which equals the inverse of the rate of RNA production for infinite [RNAP]). Next, τ_{prior} at a given [RNAP] can be obtained by subtracting τ_{after} from M at that [RNAP].

Here, we measure [RNAP] by Western blot [10,11] and RNA production rates by qPCR [10], relative to the control condition (1x LB media) (Figure 1F-G). From these, we estimate τ_{prior}/M (Figure 1H), where the line is obtained by a maximum likelihood fit [17]. We also calculate the standard error of the estimate using the Delta Method [18]. Next, given M for each condition, we calculate the absolute values of τ_{prior} and τ_{after} for that condition (Figure 1I).

1.6. Stochastic model of transcription

In vitro studies have shown that, in normal conditions, the kinetics of active transcription initiation in *E. coli* can be well described as a stochastic, two rate-limiting steps process [1,14,15,19–21]. The kinetics of these steps can be regulated separately from one another [1,10,13,15,19,21–24]. The first rate-limiting step is the set of events that take place from the freeing of a promoter from a preceding RNAP until the successful binding of

the ‘next’ RNAP to the promoter and commitment to open complex formation (including, among other, the sporadic repression states and the finding of the transcription start site by the RNAP). The average time-length of these events is here denoted as τ_{prior}. We note that this time includes the fractions of time that the promoter may be under the influence of a repressor molecule.

The second step is the set of events (e.g. isomerization) that occur from the commitment to open complex formation up to its completion and promoter escape [22,25–30]. The average time-length of these events is here denoted as τ_{after}. The sum of these two average time-lengths (τ_{prior} and τ_{after}) is denoted as M, which corresponds to the mean time-length between two consecutive transcription events.

Given this, the empirical data is analysed assuming that transcription is well modelled by a two rate-limiting steps stochastic process [1] (depicted in Figure 1E). In detail, in this model, an active promoter (P_{ON}) can participate in either of two competing processes. The first is a transition with the rate k_{OFF} of P_{ON} to an intermittent inactive state (P_{OFF}), e.g. due to repression. This step is reversible (e.g. due to the unbinding of the repressor) with the rate k_{ON}.

The other competing step is P_{ON} being bound by an RNAP (R) at the rate k_1 and forming a closed complex (RP_c). This step is also reversible [1,13,15] at the rate k_{-1} and competes with the formation of an open complex (RP_o) whose rate constant is k_2. Once committed to the open complex formation, it is assumed that, in normal conditions, transcription is no longer reversible [13]. The subsequent steps are accounted for by a single-step reaction with the rate k_3 [14,31,32] (also see [33] and references within). These steps include, among other, promoter escape (freeing the promoter for new events), transcription elongation, and termination, at which point the RNA and RNAP are also released.

This stochastic model does not consider positive supercoiling buildups, as we do not model genes exhibiting particularly high expression levels [34], in accordance with the empirical data (Figure 2A).

1.7. Stochastic model of coupled transcription and translation

To model the dynamic coupling between transcription and translation, one needs a more complex stochastic model of transcription than the one considered in Figure 1E. For this, we model explicitly the ribosome binding site (RBS) region of the RNA, while still also modelling the complete RNA molecule. This is because the production of the RBS occurs soon after promoter escape (following the completion of transcription initiation) and, once this occurs, translation can begin (but not be completed before the transcript is complete). For a detailed description of this modelling strategy see e.g. [35,36] and references within. The multi-delayed stochastic model on RNA production considered here is:

$$P_{ON} + R \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} RP_c \xrightarrow{k_2} RP_o \xrightarrow{k_e} P_{ON}(\tau_1) + RBS(\tau_1) + R(\tau_2) + RNA(\tau_2) \quad \text{(S1)}$$

$$P_{ON} \underset{k_{ON}}{\overset{k_{OFF}}{\rightleftarrows}} P_{OFF} \quad \text{(S2)}$$

In reactions S1 and S2, aside from the rate constants defined in the previous section, k_e is the rate of promoter escape. Meanwhile τ_1 is the time for the RNAP to move 30 to 60 base pairs (bp) downstream of the transcription start site. This allows a new RNAP to bind [37]. At approximately the same time, an RBS is produced (since this region of the RNA is up to ~40 nucleotides long [38]). As such, and given the much longer time-length of the intervals between consecutive RNA production events, we assume that this process time-length also equals τ_1. Finally, τ_2 is the time-length of completion of transcription elongation along with RNAP and RNA release.

As a side note, this model can also account for elongation along with alternative pathways, such as pausing, arrests, editing, pyrophosphorolysis and RNA polymerase traffic. Namely, the effects of such events can be accounted for by the distribution from which the values of τ_1 and τ_2 are randomly extracted [39].

Next, translation is modelled by reaction S3, using the RBS above as a reactant (thus allowing it to initiate prior to the complete production of the corresponding RNA). The other reactant is a ribosome (Rib) [35]:

$$\mathrm{RBS} + \mathrm{Rib} \xrightarrow{k_{tr}} \mathrm{RBS}(\tau_3) + \mathrm{Rib}(\tau_4) + \mathrm{Protein}(\tau_5) \qquad \text{(S3)}$$

In (S3), k_{tr} is the binding rate of a ribosome to the RBS of the target RNA. Meanwhile, τ_3 is the time for the RBS to be available for a new ribosome to bind and τ_4 is the time for a polypeptide to be produced and the ribosome to be released. Finally, τ_5 includes the time for the previous events plus the time for protein folding and maturation. As above, one can consider in the distribution from which τ_4 and τ_5 are extracted, events such as variable codon translation rates, ribosome traffic, back-translocation and trans-translation.

Known events not accounted in this model are premature termination during transcription and drop-off in translation, whose occurrence is rare in normal growth conditions [39].

Based on this model, while affected by noise, we expect a positive correlation between the mean number of proteins and the RNA numbers. This correlation should be maximal if the moments when RNA and proteins numbers are counted are distanced by the mean time taken to produce a functional protein from the RNA.

2. Supplementary Results

2.1. RNA production kinetics during the lifetime of the cells

It is reasonable to hypothesize that the kinetics of RNA production of the target gene may differ following gene replication. Meanwhile, we interpret our measurements of RNA production intervals assuming that in each cell there is only one gene active coding for this target RNA. For this to be valid, on average, there should not exist a significant difference in the kinetics of RNA production (e.g. mean rate) between the first and second half of the cells lifetime.

To test this, we compared distributions of Δt intervals extracted from cells during the first half of their lifetime and during the second half of their lifetime (during which the DNA replicates). In particular, we compared the

distribution of intervals that started and ended in the first half of the lifetime with the distribution of intervals that started and ended in the second half of the lifetime. For this, we performed 2-sample Kolmogorov-Smirnov tests for each condition (see Table 1 for the list of conditions), and applied a Bonferroni-Holm correction for multiple comparisons to the *p*-values obtained. We found that, at the significance level of 0.05, the two distributions cannot be distinguished (*p*-values > 0.31) except for the LA(75) condition (*p*-value = 0.04). As it is unlikely that DNA replication would affect this condition differently when compared to the other conditions, we conclude that there are no significant differences in the kinetics of RNA production of the target gene during the cell lifetime. This suggests that, in our measurements, DNA replication does not disturb significantly the RNA production kinetics of our target genes.

2.2. Cell growth rates and morphology

We tested whether the expression of MS2-GFP proteins, at the induction levels employed in this study, affects cell growth rates and/or cell morphology. For this, first, we measured mean cell division times. Their mean and standard error were found to equal 44.3 ± 1.4 min, when expressing, and 43.2 ± 1.3 min, when not expressing MS2-GFP, from which we conclude that they do not differ significantly. Next, using phase-contrast microscopy and image analysis [3], we compared the morphology of the cells with and without the expression of the MS2-GFP proteins, and found no significant differences.

2.3 Distribution of protein expression levels in individual cells is not affected by MS2-GFP tagging

To test if the MS2-GFP tagging system could affect the protein expression levels of the target gene, we measured the distribution of single-cell protein expression levels (by flow cytometry) under the control of $P_{lac/ara-1}$ (LA condition, Table 1 in main manuscript) when and when not activating the expression of MS2-GFP. From the distributions, we extract M, CV, S, and K, as these are the features of interest.

To quantify the degree to which two distributions differ (i.e. the distance D between them), we obtained the distance between the values of M, CV, S and K of these distributions, and normalized them by dividing by the mean value of that feature in the conditions considered. Assuming that Δ is the difference between two features, this distance between two distributions equals:

$$D = \frac{|\Delta M_P|}{\langle M_P \rangle} + \frac{|\Delta CV_P|}{\langle CV_P \rangle} + \frac{|\Delta S_P|}{\langle S_P \rangle} + \frac{|\Delta K_P|}{\langle K_P \rangle} \quad \text{(S4)}$$

In order to determine whether this distance is significant, we also considered the distances between pairs of distributions obtained in different conditions. Shortly, if the distance D between the LA conditions expressing and not expressing MS2-GFP is smaller than the distances between different conditions, we can conclude that the expression of MS2-GFP followed by tagging of the target RNA does not perturb significantly the relationship between RNA and protein numbers of the target gene.

For this, we make use of the single-cell distributions of protein expression levels of the control condition (LA) along with the subset 'Mutants' (Mut1, Mut2, Mut3, Mut4) and the LA(IPTG) condition, since those are the conditions used in Figure 4. Since we make use of more than two conditions, the normalization in equation

(S4) is performed by dividing the difference in each feature between a pair of conditions by the mean of all conditions considered.

In Figure S9, it is visible that, in general, the pair of conditions LA, differing in whether MS2-GFP is expressed, exhibits one of the smallest differences in each of the features considered. More importantly, when considering the four features together (using distance D as defined above), they are the pair of conditions whose distributions of single-cell protein expression levels are most similar. We thus conclude that the expression of MS2-GFP does not affect significantly the observed protein expression levels.

This result can be explained by the location of the coding regions of the RNA target for MS2-GFP in the plasmid, relative to the transcription start site. Namely, it starts with a ribosome binding site (RBS), followed by the region coding for the red fluorescent protein. Only afterwards is the region coding for the MS2-GFP binding sites, thus minimizing interference with the RBS activity and with the degradation rate of the region coding for the red fluorescent protein.

2.4. Skewness and kurtosis of RNA production kinetics are not correlated to the distributions of cell lifetimes or to the distributions of intracellular RNAP concentrations

It is reasonable to assume that differences in the shapes of the distributions of cell lifetimes between the conditions considered above could also affect the Δt distributions. To test this, we measured cell lifetimes in the conditions where cells are in the exponential growth phase (Table 1). Next, we calculated the mean, coefficient of variation, skewness and kurtosis of each of these distributions of cell lifetimes (here named M_L, CV_L, S_L, and K_L, respectively) and plotted them against the corresponding M, CV, S, and K of the Δt distribution (Figure S4A). In each case, we calculated the Pearson's correlation coefficient (with the corresponding two-tailed *p*-value), and found no significant correlation (all *p*-values > 0.05).

We conclude that, in our data, S and K of the Δt distribution are not correlated with any feature of the distribution of cell lifetimes. This is in agreement with our observation that the cell morphology (Materials and Methods) and physiology do not differ significantly between the conditions considered.

Further, as in the main manuscript, section 3.2, we applied the same calculations when considering the subsets 'Mutations' and 'Regulatory factors' separately (Figure S4B,C). Again, when applying the Bonferroni-Holm correction for multiple comparisons, the only potential correlation (K vs. K_L in the subset 'Regulatory factors') is not statistically significant. These results show that even when reducing the number of variables differing between conditions, there is no visible significant correlation between the features of the distributions of cell lifetimes and the features of the Δt distribution.

Finally, we obtained the single-cell distributions of RNAP concentrations using a cell strain where RNAPs are fluorescently tagged with GFP (Materials and Methods) in the media richness conditions 1x, 0.75x and 0.5x (Supplementary Figure S5). We found no relationship between the skewness of these distributions and S of the corresponding Δt distributions.

The results on the various conditions differing in target promoter or regulatory factors are expected since the cells are from the same strain and in the same media conditions. Similarly, the results on the conditions

differing in medium are expected given that these media (1x, 0.75x and 0.5x) were specially tuned for having cells with differing RNAP levels but similar average growth rates [1]. In this regard, it is worth mentioning that when observing the lifetimes of a small number of cells (values of M_L in Figure S4) there are visible differences between the conditions. However, the growth curves (Supplementary Figure S2A) indicate that this variability is due to the small number of cells that are observed during their entire lifetime by microscopy (when compared to the growth curves).

Supplementary References

[1] Lloyd-Price J., Startceva S., Kandavalli V., Chandraseelan J.G., Goncalves N., Oliveira S.M.D., Häkkinen A., and Ribeiro A.S. (2016). Dissecting the stochastic transcription initiation process in live *Escherichia coli*. *DNA Res.*, 23: 203–214. DOI: https://doi.org/10.1093/dnares/dsw009

[2] Hebisch E., Knebel J., Landsberg J., Frey E., and Leisner M. (2013). High variation of fluorescence protein maturation times in closely related *Escherichia coli* strains. *PLoS One*, 8: e75991. DOI: http://dx.doi.org/10.1371/journal.pone.0075991

[3] Häkkinen A., Muthukrishnan A.-B., Mora A., Fonseca J.M., and Ribeiro A.S. (2013). CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, 29: 1708–1709. DOI: https://doi.org/10.1093/bioinformatics/btt194

[4] Häkkinen A., and Ribeiro A.S. (2015). Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics*, 31: 69–75. DOI: https://doi.org/10.1093/bioinformatics/btu592

[5] Joanes D.N., and Gill C.A. (1998). Comparing Measures of Sample Skewness and Kurtosis. *J. R. Stat. Soc., Ser. D (Stat.)*, 47: 183–189. DOI: http://dx.doi.org/10.1111/1467-9884.00122

[6] Carpenter J., and Bithell J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, 19: 1141–1164. DOI: http://dx.doi.org/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F

[7] DiCiccio T.J., and Efron B. (1996). Bootstrap Confidence Intervals. *Stat. Sci.*, 11: 189–228. DOI: http://dx.doi.org/10.1214/ss/1032280214

[8] Liang S.-T., Bipatnath M., Xu Y.-C., Chen S.-L., Dennis P., Ehrenberg M., and Bremer H. (1999). Activities of constitutive promoters in *Escherichia coli*. *J. Mol. Biol.*, 292: 19–37. DOI: https://doi.org/10.1006/jmbi.1999.3056

[9] Ehrenberg M., Bremer H., and Dennis P.P. (2013). Medium-dependent control of the bacterial growth rate. *Biochimie*, 95: 643–58. DOI: http://dx.doi.org/10.1016/j.biochi.2012.11.012

[10] Kandavalli V.K., Tran H., and Ribeiro A.S. (2016). Effects of σ factor competition are promoter initiation kinetics dependent. *Biochim Biophys. Acta*, 1859: 1281–1288. DOI: http://dx.doi.org/10.1016/j.bbagrm.2016.07.011

[11] Mäkelä J., Kandavalli V., and Ribeiro A.S. (2017). Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. *Sci. Rep.*, 7: 10588. DOI: https://doi.org/10.1038/s41598-017-11257-2

[12] Oliveira S.M.D., Häkkinen A., Lloyd-Price J., Tran H., Kandavalli V., and Ribeiro A.S. (2016). Temperature-dependent model of multi-step transcription initiation in *Escherichia coli* based on live single-cell measurements. *PLoS Comput. Biol.*, 12: e1005174. DOI: http://dx.doi.org/10.1371/journal.pcbi.1005174

[13] McClure W.R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, 54: 171–204. DOI: http://dx.doi.org/10.1146/annurev.bi.54.070185.001131

[14] 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. DOI: http://dx.doi.org/10.1093/nar/29.18.3873

[15] McClure W.R. (1980). Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. U.S.A.*, 77: 5634–5638. DOI: http://dx.doi.org/10.1073/pnas.77.10.5634

[16] Lineweaver H., and Burk D. (1934). The Determination of Enzyme Dissociation Constants. *J. Am. Chem. Soc.*, 56: 658–666. DOI: http://dx.doi.org/10.1021/ja01318a036

[17] Bevington P.R., and Robinson D.K. (2003). Least-Squares Fit to a Straight Line. *In* Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill), pp. 98–115. ISBN: 0-07-247227-8

[18] Casella G., and Berger R.L. (2001). The Delta Method. *In* Statistical Inference, 2nd ed (Pacific Grove, CA: Duxbury Press), pp. 240–245. ISBN: 0-534-24312-6

[19] 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. DOI: http://dx.doi.org/10.1021/bi00332a018

[20] Chamberlin M.J. (1974). The selectivity of transcription. *Annu. Rev. Biochem.*, 43: 721–775. DOI: http://dx.doi.org/10.1146/annurev.bi.43.070174.003445

[21] Browning D.F., and Busby S.J.W. (2016). Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.*, 14: 638–650. DOI: http://dx.doi.org/10.1038/nrmicro.2016.103

[22] deHaseth P.L., Zupancic M.L., and Record T.M. Jr. (1998). RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.*, 180: 3019–3025. PMID: 9620948

[23] Jones D.L., Brewster R.C., and Phillips R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346: 1533–1536. DOI: http://dx.doi.org/10.1126/science.1255301

[24] Feklístov A., Sharon B.D., Darst S.A., and Gross C.A. (2014) Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.*, 68: 357–376. DOI: http://dx.doi.org/10.1146/annurev-micro-092412-155737

[25] Duchi D., Bauer D.L.V., Fernandez L., Evans G., Robb N., Hwang L.C., Gryte K., Tomescu A., Zawadzki P., Morichaud Z., *et al*. (2016). RNA polymerase pausing during initial transcription. *Mol. Cell*, 63: 939–950. DOI: http://dx.doi.org/10.1016/j.molcel.2016.08.011

[26] Duchi D., Gryte K., Robb N.C., Morichaud Z., Sheppard C., Brodolin K., Wigneshweraraj S., and Kapanidis A.N. (2018). Conformational heterogeneity and bubble dynamics in single bacterial transcription initiation complexes. *Nucleic Acids Res.*, 46: 677–688. DOI: http://dx.doi.org/10.1093/nar/gkx1146

[27] Hsu L.M. (2002). Promoter clearance and escape in prokaryotes. *Biochim Biophys. Acta*, 1577: 191–207. DOI: http://dx.doi.org/10.1016/S0167-4781(02)00452-9

[28] Hsu L.M. (2008). Promoter escape by *Escherichia coli* RNA polymerase. *EcoSal Plus*, 3: 1–16. DOI: http://dx.doi.org/10.1128/ecosalplus.4.5.2.2

[29] Kapanidis A.N., Margeat E., Ho S.O., Kortkhonjia E., Weiss S., and Ebright R.H. (2006). Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science*, 314: 1144–1147. DOI: http://dx.doi.org/10.1126/science.1131399

[30] Lerner E., Chung S., Allen B.L., Wang S., Lee J., Lu S.W., Grimaud L.W., Ingargiola A., Michalet X., Alhadid Y., *et al*. (2016). Backtracked and paused transcription initiation intermediate of *Escherichia coli* RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, 113: E6562–E6571. DOI: https://doi.org/10.1073/pnas.1605038113

[31] Greive S.J., and von Hippel P.H. (2005). Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.*, 6: 221–232. DOI: http://dx.doi.org/10.1038/nrm1588

[32] Herbert K.M., La Porta A., Wong B.J., Mooney R.A., Neuman K.C., Landick R., and Block S.M. (2006). Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, 125: 1083–1094. DOI: http://dx.doi.org/10.1016/j.cell.2006.04.032

[33] Rajala T., Häkkinen A., Healy S., Yli-Harja O., and Ribeiro A.S. (2010). Effects of Transcriptional Pausing on Gene Expression Dynamics. *PLoS Comput. Biol.*, 6: e1000704. DOI: https://doi.org/10.1371/journal.pcbi.1000704

[34] Chong S., Chen C., Ge H., and Xie X.S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, 158: 314–326. DOI: http://dx.doi.org/10.1016/j.cell.2014.05.038

[35] Ribeiro A.S. (2010). Stochastic and delayed stochastic models of gene expression and regulation. *Math. Biosci.*, 223: 1–11. DOI: http://dx.doi.org/10.1016/j.mbs.2009.10.007

[36] Zhu R., Ribeiro A.S., Salahub D., Kauffman S.A. (2007). Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *J. Theor. Biol.*, 246: 725–745. DOI: http://dx.doi.org/10.1016/j.jtbi.2007.01.021

[37] Record T.M. Jr., Reznikoff W.S., Craig M.L., McQuade K.L., Schlax P.J. (1996). *Escherichia coli* RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription initiation. *In Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology, 2nd ed, F.C. Neidhardt, R. Curtiss, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, D. Schneider, and H.E. Umbarger, eds, (Washington, DC: ASM press), pp. 792–821. ISBN: 1555810845

[38] Shultzaberger R.K., Bucheimer R.E., Rudd K.E., and Schneider T.D. (2001). Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.*, 313: 215–228. DOI: https://doi.org/10.1006/jmbi.2001.5040

[39] Mäkelä J., Lloyd-Price J., Yli-Harja O., Ribeiro A.S. (2011). Stochastic sequence-level model of coupled transcription and translation in prokaryotes. *BMC Bioinformatics*, 12: 121. DOI: https://doi.org/10.1186/1471-2105-12-121

Supplementary Figures

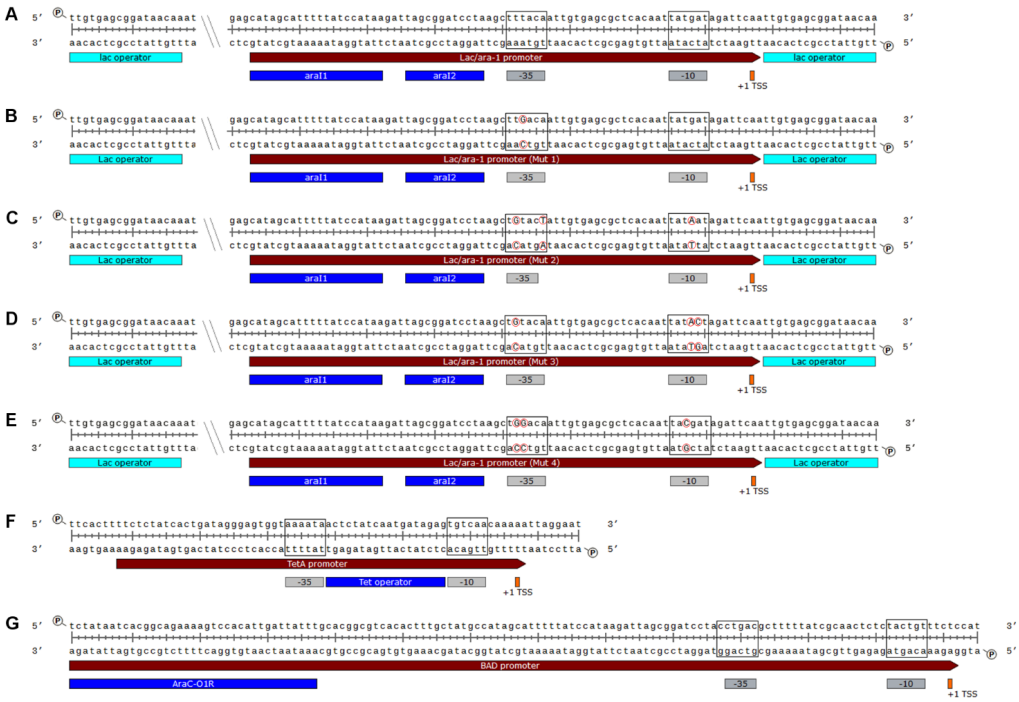


Figure S1. Related to Figure 2A and Tables 1 and 2. Schematic representation of the target promoter's sequences. The -35 and -10 promoter elements are shown in black boxes. The transcription start sites (+1 TSS) are marked in orange. Operator sites are marked in cyan and blue. In the mutants, specific nucleotide changes in the -35 and -10 region are marked by red circles. These promoters were used in the studied conditions (Table 1) as follows: (**A**) LA, LA(75), LA(50), LA(ara), LA(IPTG) and LA(oxi); (**B**) Mut1; (**C**) Mut2; (**D**) Mut3; (**E**) Mut4; (**F**) tetA and tetA(st); (**G**) BAD and BAD(st).

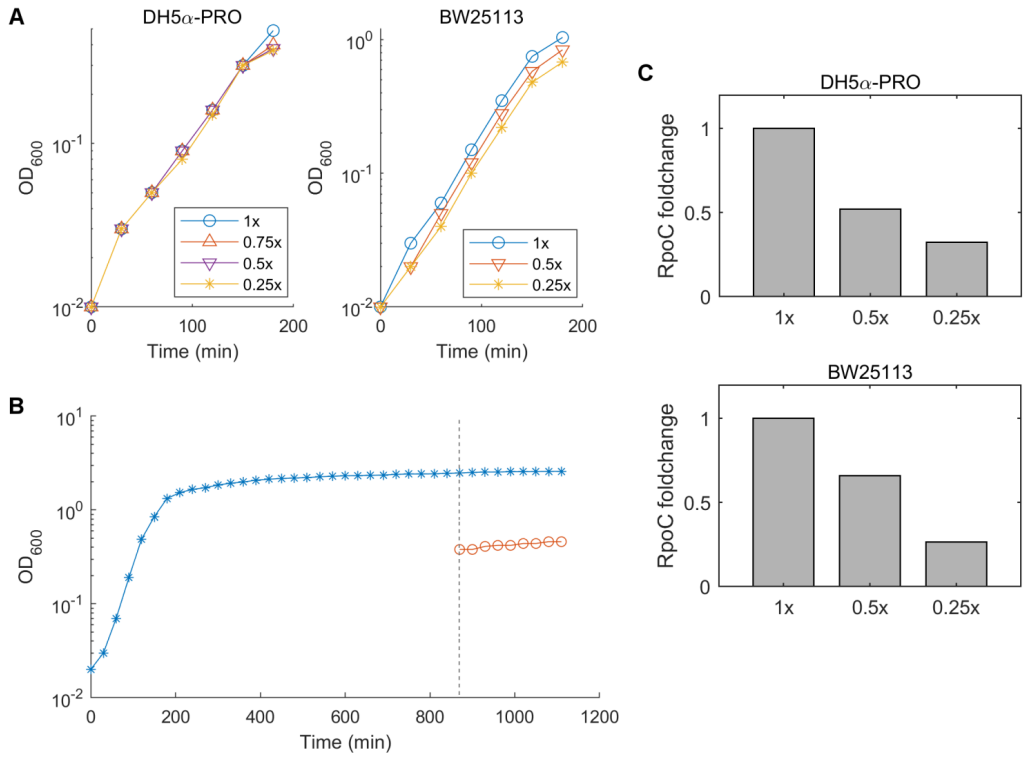


Figure S2. Related to Figures 2A and 3A. Bacterial growth curves and RNAP levels as a function of media richness, relative to the control condition. (**A**) Bacterial growth curves of the DH5α-PRO and BW25113 *E. coli* strains when grown in LB media with different richness (1x, 0.75x, 0.5x and 0.25x, see Materials and Methods for a detailed description). The optical density at the wavelength of 600 nm (OD_{600}) was measured every 30 min for 3 h. (**B**) Bacterial growth curve of the BW25113 strain reaching the stationary phase. Cells were grown in 1x LB media (see Materials and Methods for a detailed description) at 37 °C with shaking at 250 rpm, and the OD_{600} values were monitored every 30 min (blue stars). After the cells reached the stationary phase, we diluted them in stationary phase media (Materials and Methods) and monitored the OD_{600} every 30 min (red circles) for 4 h. The vertical dashed line shows the time of the dilution. (**C**) RNAP levels (relative to the 1x condition) of the DH5α-PRO and BW25113 *E. coli* strains grown in LB media with different richness (1x, 0.5x, and 0.25x) as assessed by Western blot measurements of the RpoC protein (Supplementary Materials and Methods, section 1.4).

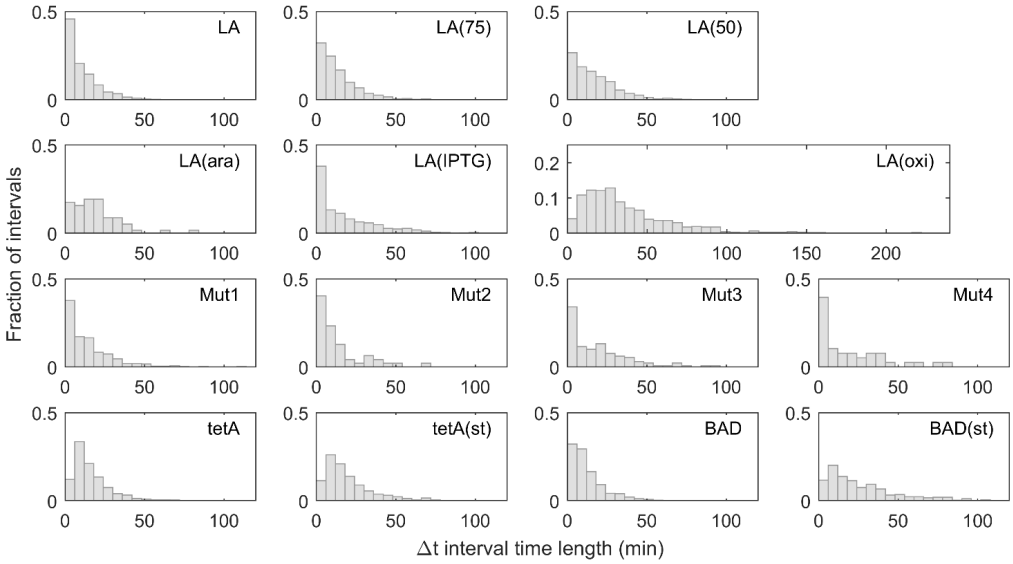


Figure S3. Related to Figure 2A. The shape of the distribution of time intervals between consecutive RNA production events in individual cells (Δt distribution) changes significantly between mutants as well as with the promoter, induction scheme and media. See Table 1 for a detailed description of each condition and Supplementary Table S8 for the statistical tests to assess whether the distributions normalized by the mean differ significantly. Data were collected from approximately 600 cells per condition.

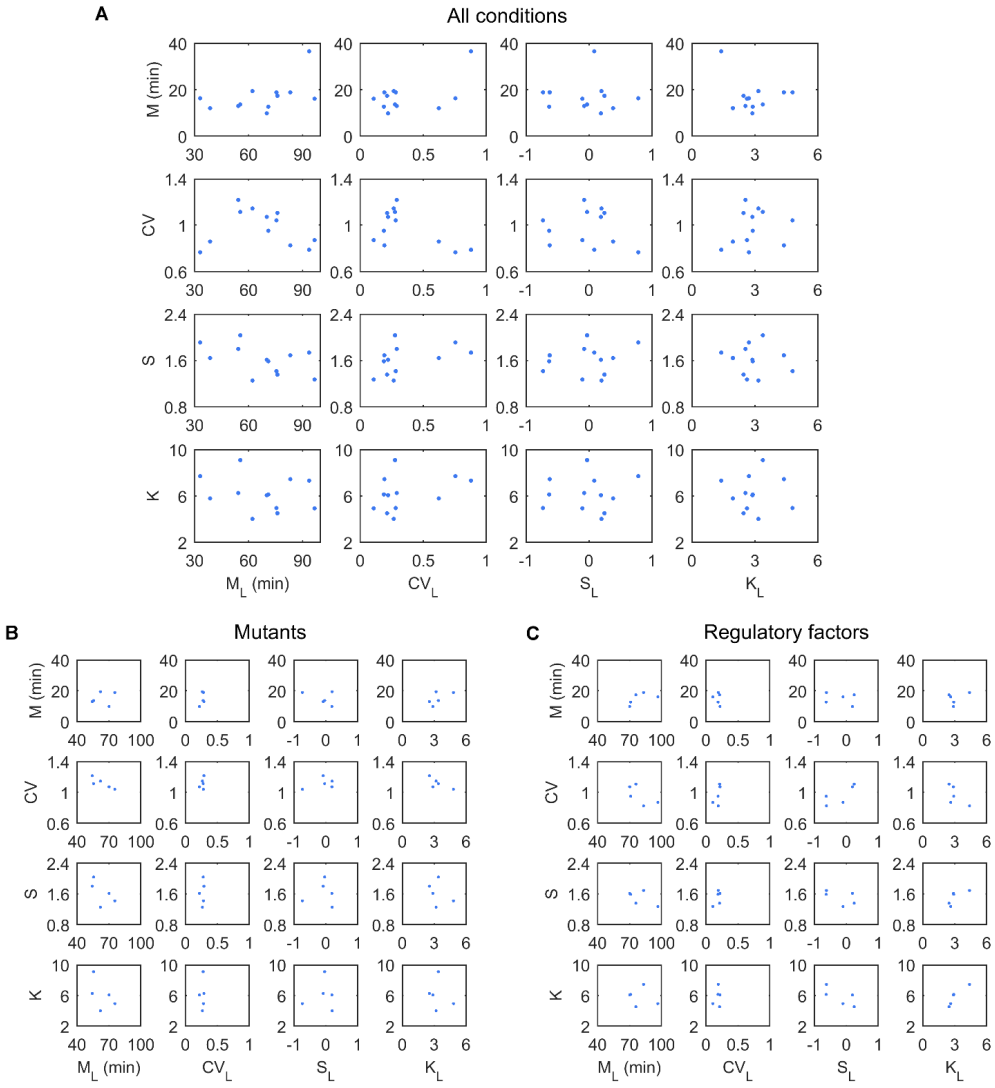


Figure S4. Related to Figure 1. Mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) of the distribution of intervals between consecutive RNA production events in individual cells plotted against the mean (M_L), coefficient of variation (CV_L), skewness (S_L) and kurtosis (K_L) of the corresponding distributions of single-cell lifetimes. Shown are (**A**) all conditions, (**B**) the 'Mutants' subset, (**C**) the 'Regulatory factors' subset.

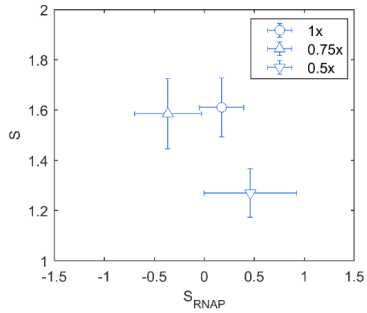


Figure S5. Related to Figure 1. Skewness of the Δt distribution (S) measured from a fully induced $P_{lac/ara-1}$ promoter in various media conditions (see section 1.3 in main manuscript) plotted against the skewness of the single-cell RNAP fluorescence distribution. The RNAP fluorescence distributions are measured by microscopy (~400 cells per condition). Error bars denote SEM.

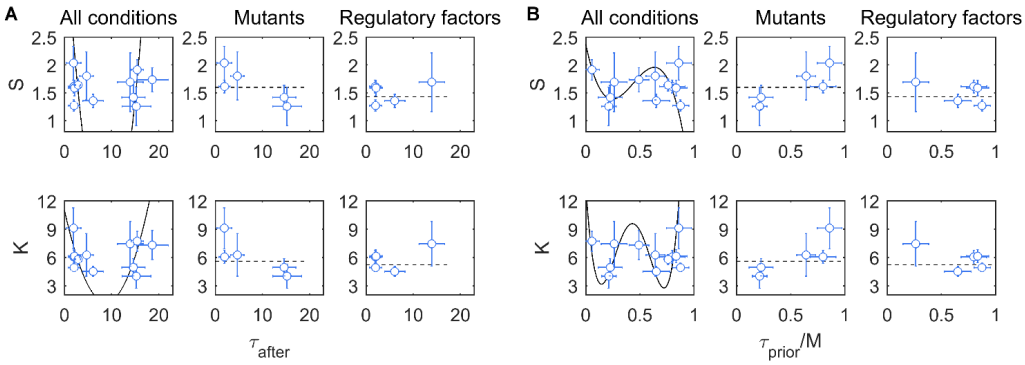


Figure S6. Related to Figure 3B and Supplementary Tables S5 and S6. Skewness (S) and kurtosis (K) of the distribution of intervals between consecutive RNA production events in individual cells do not show linear relationships with the fraction of time spent in events after commitment to the open complex formation (τ_{after}) nor with the mean fraction of time spent in the events prior to commitment to open complex formation (τ_{prior}/M, where M is the mean time between transcription events). Shown is (**A**) S and K as a function of τ_{after} and (**B**) S and K as a function of τ_{prior}/M, for all conditions and for the subsets of conditions 'Mutants' and 'Regulatory factors'. Error bars denote SEM. The best-fitting model is shown as a dashed line if it is a zero-degree polynomial and as a solid line if it is a polynomial of a higher degree. In plots where two separate lines are visible, the best fitting model is partially outside of the plot borders on the y-axis.

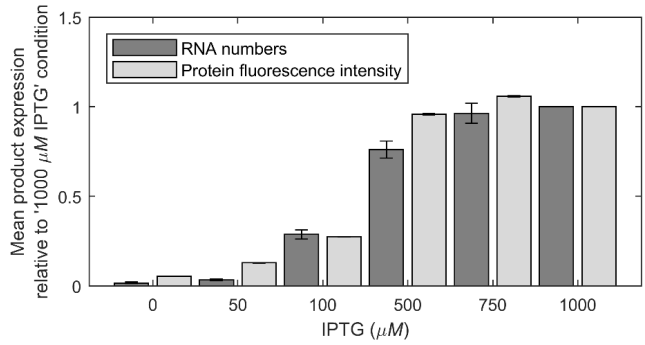


Figure S7. Related to Figure 4. Mean protein numbers of the target gene under the control of $P_{lac/ara-1}$ follow the corresponding average RNA numbers for increasing induction levels. Induction curve of $P_{lac/ara-1}$ as seen by observing the mean RNA numbers produced by the target promoter ($P_{lac/ara-1}$) in individual cells using microscopy (dark grey), and by observing the mean fluorescent intensity of proteins in individual cells from the same promoter using flow cytometry (light grey). In all conditions, cells are subject to 1% of arabinose. Data obtained by microscopy was collected 60 min after induction of the target gene, while data obtained by flow cytometry were collected 90 min after induction of the target gene. In both measurements, the values are shown relative to the value obtained in the condition '1000 μM IPTG' of the corresponding measurement. Error bars denote the standard error of the ratio.

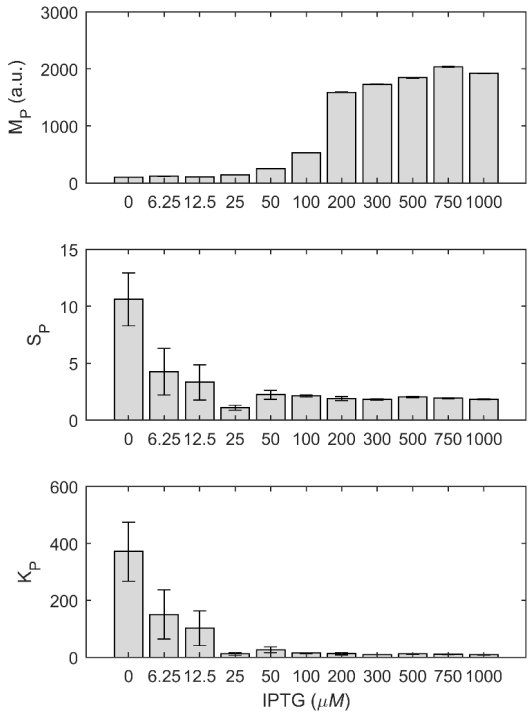


Figure S8. Related to Figure 4. The skewness (S_P) and kurtosis (K_P) of the single-cell distributions of protein expression levels differ with induction strength but can also have significantly different values for similar mean (M_P) expression levels. (Top) M_P, (middle) S_P and (bottom) K_P of the single-cell distributions of protein expression levels, as expressed under the control of $P_{lac/ara-1}$ for changing induction strength. Data were collected by flow cytometry, 90 min after induction of the target gene. Error bars denote SEM. In the top figure, the error bars are too small to be visible. In the regime of weak induction (25 or less μM IPTG), S_P and K_P show significant, consistent changes, although M_P does not exhibit significant changes. Meanwhile, above 25 μM IPTG concentration, the opposite occurs.

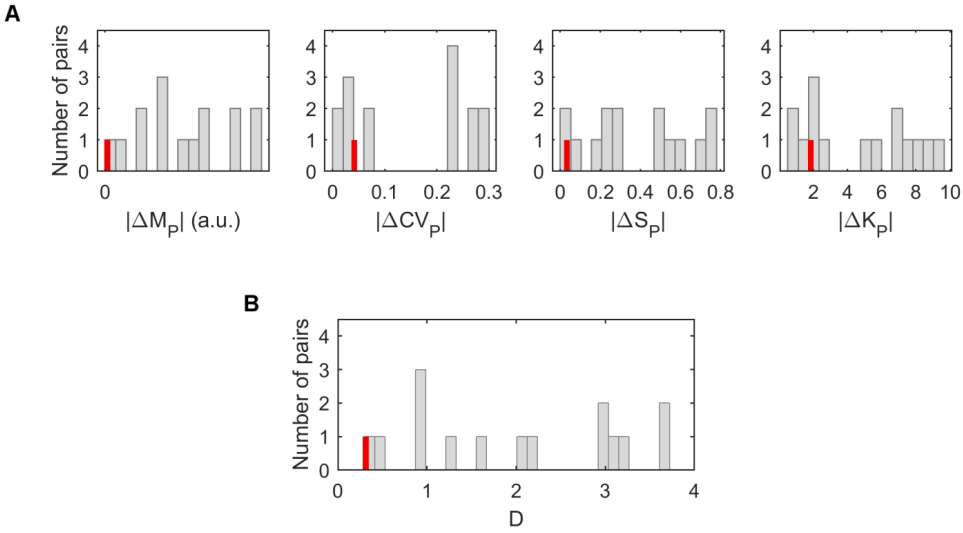


Figure S9. Related to Figure 4. Activation of the MS2-GFP reporter does not affect significantly the single-cell distribution of protein expression levels. (A) Numbers of pairs of conditions (grey bars) with given values of, respectively, the absolute differences in mean ($|\Delta M_P|$), coefficient of variation ($|\Delta CV_P|$), skewness ($|\Delta S_P|$) and kurtosis ($|\Delta K_P|$) of the single-cell distributions of protein expression levels. The conditions considered are LA, LA(IPTG), Mut1, Mut2, Mut3 and Mut4 (see Table 1 in main manuscript). Meanwhile, the red bar marks the values for these differences between the pair of measurements in the LA condition with and without activating the reporter. (B) Distance D between the values of M, CV, S and K (equation S4) for the same pairs of conditions as in (A). The red bar holds the value 0.31, while the grey bar further to the left holds the value 0.33.

Supplementary Tables

Table S1. Related to Figure 2B and Table 2. Two-tailed *p*-values obtained by testing, for each pair of conditions, the null hypothesis (H_0) that the difference in the mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) of the distribution of time intervals between consecutive RNA production events between the two conditions equals zero, using a 2-sample z-test. In cases where the *p*-value ≤ 0.05, the H_0 is rejected (highlighted with italics). In cases where the *p*-value > 0.05, the H_0 cannot be rejected.

M	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	0.16	*< 0.001*	*< 0.01*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(75)		*< 0.001*	*< 0.01*	*< 0.001*	*< 0.001*	0.28	0.90	*< 0.001*	0.06	*< 0.001*	*< 0.001*	0.28	*< 0.001*
LA(50)			0.19	0.17	*< 0.001*	*< 0.01*	0.18	0.12	0.36	0.75	*< 0.001*	*< 0.001*	*< 0.001*
LA(ara)				0.49	*< 0.001*	*< 0.01*	0.06	1.00	0.90	0.24	0.86	*< 0.01*	*< 0.01*
LA(IPTG)					*< 0.001*	*< 0.001*	0.07	0.42	0.58	0.31	0.07	*< 0.001*	*< 0.001*
LA(oxi)						*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
Mut1							0.77	*< 0.01*	0.12	*< 0.01*	*< 0.001*	*0.05*	*< 0.001*
Mut2								*< 0.01*	0.13	0.16	*< 0.01*	0.68	*< 0.001*
Mut3									0.90	0.17	0.84	*< 0.001*	*< 0.001*
Mut4										0.40	0.98	*< 0.01*	0.06
tetA											*< 0.01*	*< 0.001*	*< 0.001*
tetA(st)												*< 0.001*	*< 0.001*
BAD													*< 0.001*
CV	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	*< 0.01*	*< 0.001*	*< 0.01*	0.43	*< 0.001*	0.49	0.28	0.68	0.62	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(75)		*< 0.01*	0.23	*< 0.01*	*< 0.001*	*< 0.01*	*0.05*	0.25	0.20	*< 0.001*	*< 0.001*	*< 0.01*	*< 0.01*
LA(50)			0.66	*< 0.001*	*< 0.01*	*< 0.001*	*< 0.01*	*< 0.01*	0.06	*< 0.01*	*< 0.001*	0.65	0.25
LA(ara)				*< 0.01*	0.71	*< 0.01*	*< 0.01*	0.08	0.07	0.56	0.46	0.76	0.97
LA(IPTG)					*< 0.001*	0.91	0.42	0.41	0.80	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(oxi)						*< 0.001*	*< 0.01*	*< 0.001*	*< 0.01*	0.60	0.31	0.06	0.31
Mut1							0.47	0.42	0.84	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
Mut2								0.24	0.71	*< 0.001*	*< 0.001*	*< 0.01*	*< 0.01*
Mut3									0.52	*< 0.001*	*< 0.001*	*< 0.01*	*< 0.01*
Mut4										*< 0.01*	*< 0.01*	*0.05*	*< 0.01*
tetA											0.70	*< 0.01*	0.14
tetA(st)												*< 0.01*	*< 0.01*
BAD													0.47

(see the rest of the table on the next page)

S	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	0.89	*< 0.01*	0.88	0.12	0.62	0.19	0.67	0.43	0.33	0.17	0.40	0.85	*< 0.01*
LA(75)		0.06	0.85	0.20	0.57	0.18	0.64	0.51	0.37	0.16	0.54	0.75	0.06
LA(50)			0.43	0.57	*0.05*	*< 0.01*	0.23	0.53	0.97	*< 0.01*	0.17	*< 0.01*	0.89
LA(ara)				0.53	0.94	0.57	0.87	0.63	0.49	0.69	0.69	0.93	0.41
LA(IPTG)					0.13	*< 0.01*	0.32	0.79	0.78	*< 0.01*	0.46	0.07	0.52
LA(oxi)						0.42	0.89	0.30	0.24	0.53	0.29	0.70	*0.05*
Mut1							0.66	0.10	0.09	0.73	0.08	0.22	*< 0.01*
Mut2								0.43	0.32	0.81	0.47	0.72	0.22
Mut3									0.69	0.08	0.81	0.35	0.49
Mut4										0.09	0.54	0.29	0.99
tetA											*< 0.01*	0.20	*< 0.01*
tetA(st)												0.29	0.16
BAD													*< 0.01*

K	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	0.94	0.18	0.57	0.08	0.46	0.19	0.93	0.34	0.16	0.19	0.26	0.76	*< 0.01*
LA(75)		0.15	0.59	0.07	0.49	0.20	0.96	0.31	0.15	0.21	0.22	0.70	*< 0.01*
LA(50)			0.30	0.55	0.14	0.06	0.57	0.99	0.50	*< 0.01*	0.82	0.23	0.19
LA(ara)				0.23	0.96	0.61	0.72	0.32	0.20	0.92	0.33	0.49	0.16
LA(IPTG)					0.09	*< 0.01*	0.46	0.68	0.71	*< 0.01*	0.42	0.10	0.53
LA(oxi)						0.51	0.70	0.19	0.10	0.83	0.17	0.35	*< 0.01*
Mut1							0.37	0.08	*< 0.01*	0.57	0.07	0.14	*< 0.01*
Mut2								0.59	0.39	0.56	0.61	0.84	0.35
Mut3									0.55	*< 0.01*	0.89	0.43	0.40
Mut4										*< 0.01*	0.43	0.20	0.95
tetA											*< 0.01*	0.10	*< 0.01*
tetA(st)												0.35	0.13
BAD													*< 0.01*

Table S2. Related to Figure 2C-D. Percentage of the time intervals between consecutive RNA production events in individual cells (Δt intervals) longer than a given threshold in each condition.

Threshold / Condition	2M	3M	4M	5M	6M
LA	16.3	6.3	2.2	0.7	0.3
LA(75)	13.9	4.7	1.5	0.7	0
LA(50)	11.8	3.2	0.8	0.1	0
LA(ara)	10.6	3.5	1.8	0	0
LA(IPTG)	17.7	7.1	1.7	0.5	0
LA(oxi)	11.0	3.0	0.5	0.2	0
Mut1	13.8	6.2	2.5	1.2	0.5
Mut2	17.0	6.4	2.1	2.1	0
Mut3	16.3	6.2	1.6	0	0
Mut4	21.1	7.9	2.6	0	0
tetA	9.4	3.2	0.9	0	0
tetA(st)	11.4	2.9	0	0	0
BAD	12.1	4.0	0.8	0.1	0
BAD(st)	12.9	3.5	0.3	0	0

Table S3. Related to Figure 3A. Mean time length spent in the events prior to commitment to open complex formation (τ_{prior}) and in the events following the commitment to open complex formation (τ_{after}) for each condition, along with their SEM. Also shown, for each condition, is the mean fraction of time between transcription events that is spent in the events prior to commitment to open complex formation (τ_{prior}/M, where M is the mean time between transcription events), along with its SEM.

Condition	τ_{prior} ± SEM (min)	τ_{after} ± SEM (min)	τ_{prior}/M ± SEM
LA	7.8 ± 1.2	2.0 ± 1.2	0.80 ± 0.12
LA(75)	10.6 ± 1.4	2.1 ± 1.3	0.83 ± 0.10
LA(50)	14.1 ± 1.3	2.0 ± 1.2	0.87 ± 0.08
LA(ara)	5.0 ± 2.3	13.9 ± 2.7	0.26 ± 0.12
LA(IPTG)	11.3 ± 2.2	6.1 ± 2.2	0.65 ± 0.12
LA(oxi)	17.9 ± 3.4	18.6 ± 3.4	0.49 ± 0.09
Mut1	11.8 ± 1.7	1.9 ± 1.6	0.86 ± 0.11
Mut2	8.3 ± 1.9	4.7 ± 1.4	0.64 ± 0.09
Mut3	4.3 ± 2.1	14.6 ± 2.5	0.23 ± 0.11
Mut4	4.2 ± 1.5	15.2 ± 3.1	0.21 ± 0.07
tetA	0.9 ± 1.2	15.4 ± 1.3	0.06 ± 0.07
BAD	9.2 ± 0.6	2.9 ± 0.5	0.76 ± 0.04

Table S4. Related to Figure 3B. One-tailed *p*-values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees n and m. The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of τ_{prior}, estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where τ_{prior} does not change between conditions is denoted as $n = 0_{inv}$. For *p*-values ≤ 0.05, we assumed that the model of degree m fits the data significantly better than the model of degree n.

Data set	S			K		
	n	m	*p*-value	n	m	*p*-value
All conditions	0	1	0.01	0	1	0.11
	1	2	0.54	0	2	0.15
	1	3	0.03	0	3	0.25
	3	4	0.98	0	4	0.02
	3	5	0.66	4	5	0.93
	3	6	0.84	4	6	0.94
	3	7	0.65	4	7	0.70
	3	8	0.68	4	8	0.83
	3	9	0.65	4	9	0.92
	3	10	0.75	4	10	0.96
	3	11	0.83	4	11	0.98
	0_{inv}	3	< 0.001	0_{inv}	4	< 0.001
Mutants	0	1	0.05	0	1	0.03
	1	2	0.86	1	2	0.77
	1	3	0.98	1	3	0.86
	1	4	0.97	1	4	0.95
	0_{inv}	1	< 0.001	0_{inv}	1	< 0.001
Regulatory factors	0	1	0.01	0	1	0.04
	1	2	0.70	1	2	0.85
	1	3	0.92	1	3	0.72
	1	4	0.90	1	4	0.70
	0_{inv}	1	< 0.001	0_{inv}	1	< 0.001

Table S5. Related to Supplementary Figure S6A. One-tailed *p*-values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees n and m. The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of τ_{after}, estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where τ_{prior} does not change between conditions is denoted as n = 0_{inv}. For *p*-values ≤ 0.05, we assumed that the model of degree m fits the data significantly better than the model of degree n.

Data set	S			K		
	n	m	*p*-value	n	m	*p*-value
All conditions	0	1	0.17	0	1	0.53
	0	2	< 0.01	0	2	< 0.01
	2	3	0.99	2	3	0.75
	2	4	1.00	2	4	0.30
	2	5	1.00	2	5	0.43
	2	6	1.00	2	6	0.59
	2	7	1.00	2	7	0.56
	2	8	1.00	2	8	0.68
	2	9	1.00	2	9	0.78
	2	10	1.00	2	10	0.86
	2	11	1.00	2	11	0.91
	0_{inv}	2	< 0.001	0_{inv}	2	< 0.001
Mutants	0	1	0.15	0	1	0.08
	0	2	0.24	0	2	0.19
	0	3	0.41	0	3	0.34
	0	4	0.41	0	4	0.30
	0_{inv}	0	< 0.001	0_{inv}	0	< 0.001
Regulatory factors	0	1	0.65	0	1	0.38
	0	2	0.13	0	2	0.09
	0	3	0.21	0	3	0.10
	0	4	0.14	0	4	0.19
	0_{inv}	0	< 0.001	0_{inv}	0	< 0.001

Table S6. Related to Supplementary Figure S6B. One-tailed *p*-values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees n and m. The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of τ_{prior}/M, estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where τ_{prior} does not change between conditions is denoted as n = 0_{inv}. For *p*-values ≤ 0.05, we assumed that the model of degree m fits the data significantly better than the model of degree n.

Data set	S			K		
	n	m	*p*-value	n	m	*p*-value
All conditions	0	1	0.06	0	1	0.41
	0	2	0.04	0	2	0.04
	2	3	< 0.01	2	3	0.76
	3	4	0.77	2	4	0.02
	3	5	0.95	4	5	0.77
	3	6	0.52	4	6	0.96
	3	7	0.69	4	7	0.99
	3	8	0.81	4	8	1.00
	3	9	0.89	4	9	1.00
	3	10	0.94	4	10	1.00
	3	11	0.97	4	11	1.00
	0_{inv}	3	< 0.001	0_{inv}	4	< 0.001
Mutants	0	1	0.13	0	1	0.07
	0	2	0.23	0	2	0.12
	0	3	0.26	0	3	0.23
	0	4	0.40	0	4	0.28
	0_{inv}	0	< 0.001	0_{inv}	0	< 0.001
Regulatory factors	0	1	0.12	0	1	0.55
	0	2	0.10	0	2	0.12
	0	3	0.16	0	3	0.14
	0	4	0.18	0	4	0.22
	0_{inv}	0	< 0.001	0_{inv}	0	< 0.001

Table S7. Related to Figure 4. One-tailed *p*-values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees n and m. The models are best-fitted to the values of mean (M_P), skewness (S_P) and kurtosis (K_P) of the distribution of protein numbers as a function of the corresponding features (M, S and K) of the distribution of time intervals between consecutive RNA production events in individual cells (Δt distribution), estimated in the conditions from the subset 'Mutants' and LA(IPTG) condition. A model where a feature of the Δt distribution does not change between conditions is denoted as n = 0_{inv}. For *p*-values ≤ 0.05, we assumed that the model of degree m fits the data significantly better than the model of degree n.

M_P vs M			S_P vs S			K_P vs K		
n	m	*p*-value	n	m	*p*-value	n	m	*p*-value
0	1	< 0.001	0	1	< 0.001	0	1	< 0.001
1	2	0.99	1	2	0.21	1	2	0.17
1	3	0.97	1	3	0.45	1	3	0.38
1	4	1.00	1	4	0.66	1	4	0.58
1	5	1.00	1	5	0.81	1	5	0.58
0_{inv}	1	< 0.001	0_{inv}	1	0.02	0_{inv}	1	0.04

Table S8. Related to Figure S3. Comparisons of the Δt distributions (normalized by the mean) of pairs of conditions (see Table 1) by a two-tailed Kolmogorov-Smirnov test. The table shows the *p*-values obtained from these tests, for each pair of conditions. In cases where the *p*-value ≤ 0.05, the H_0 that the Δt values normalized by the mean are from the same distribution is rejected (highlighted with italics).

	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	*< 0.001*	*< 0.001*	*< 0.01*	*< 0.01*	*< 0.001*	*< 0.01*	0.78	0.67	0.18	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(75)		*< 0.01*	0.28	*< 0.001*	*< 0.001*	*< 0.01*	0.23	0.18	0.05	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(50)			0.56	*< 0.001*	*< 0.001*	*< 0.01*	0.15	*< 0.01*	*< 0.01*	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.01*
LA(ara)				*< 0.01*	0.39	0.06	0.14	0.07	0.09	0.31	0.31	0.18	0.43
LA(IPTG)					*< 0.001*	0.09	0.74	0.67	0.49	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
LA(oxi)						*< 0.001*	*< 0.01*	*< 0.001*	*< 0.001*	0.23	0.83	*< 0.01*	*< 0.01*
Mut1							0.75	0.92	0.30	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
Mut2								0.65	0.78	*< 0.01*	*< 0.01*	*< 0.01*	*< 0.01*
Mut3									0.25	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.001*
Mut4										*< 0.001*	*< 0.001*	*< 0.001*	*< 0.01*
tetA											0.41	*< 0.01*	*< 0.001*
tetA(st)												*< 0.01*	*< 0.01*
BAD													0.14