Modeling and engineering promoters with pre-defined RNA production dynamics in *Escherichia coli*

Samuel M.D. Oliveira[1](0000-0002-6914-5529), Mohamed N.M. Bahrudeen[1](0000-0003-3717-0583), Sofia Startceva[1](0000-0002-2531-5530), Vinodh Kandavalli[1](0000-0001-8788-9399), and Andre S. Ribeiro[1](0000-0002-7255-5211)

[1] Laboratory of Biosystem Dynamics, BioMediTech Institute, Tampere University of Technology, Finland. P.O Box 553, 33101 Tampere, Finland.
andre.ribeiro@tut.fi

Abstract. Recent developments in live-cell time-lapse microscopy and signal processing methods for single-cell, single-RNA detection now allow characterizing the *in vivo* dynamics of RNA production of *Escherichia coli* promoters at the single event level. This dynamics is mostly controlled at the promoter region, which can be engineered with single nucleotide precision. Based on these developments, we propose a new strategy to engineer genes with predefined transcription dynamics (mean and standard deviation of the distribution of RNA numbers of a cell population). For this, we use stochastic modelling followed by genetic engineering, to design synthetic promoters whose rate-limiting steps kinetics allow achieving a desired RNA production kinetics. We present an example where, from a pre-defined kinetics, a stochastic model is first designed, from which a promoter is selected based on its rate-limiting steps kinetics. Next, we engineer mutant promoters and select the one that best fits the intended distribution of RNA numbers in a cell population. As the modelling strategies and databases of models, genetic constructs, and information on these constructs kinetics improve, we expect our strategy to be able to accommodate a wide variety of pre-defined RNA production kinetics.

Keywords: Model of Transcription Initiation, Synthetic Constructs, Rate-limiting Steps, Gene Engineering Framework.

1 Introduction

Several studies have determined that, in *Escherichia coli*, the main regulatory mechanisms of gene expression dynamics act at the stage of transcription initiation [1-9]. It has recently become possible to combine time-lapse live cell microscopy with single RNA detection techniques [6,10-13], synthetic biology techniques for gene engineering at the nucleotide level [4,6,9], stochastic models [14-18], and signal processing methods [19-21] to study how the dynamics of gene expression in *E. coli* is tuned by the kinetics of rate-limiting steps in transcription initiation [8,9,20,22].

Using this, we propose a new strategy for, from the specification of the desired dynamics of RNA production, mean and cell-to-cell variability in RNA numbers in individual cells, and the use of detailed stochastic modeling of transcription initiation [1,14,22], first, predict the necessary dynamics of the rate-limiting steps in transcription initiation. Next, select an existing promoter that best fit these specifications. Afterward, fine-tune the desired dynamics by single and double point mutations of the selected promoter, so as to engineer a synthetic promoter whose RNA production dynamics best fit the original specification. This fitting is analyzed at the single RNA level, by making use of MS2-GFP probes for detection of RNA numbers at the single-RNA level in live cells [9-11,23,24] and objective criteria to compare the dynamics of synthetic promoters with that of the stochastic model, which, here, aside from transcription dynamics, it also accommodates for RNA degradation and cell division.

The strategy has four main steps: i) Design a stochastic model that best fits the specifications using the modelling strategy proposed in [1,13,22]; ii) Select the promoter whose *in vivo* rate-limiting steps kinetics [23,24] best fits the model dynamics; iii) Engineer mutant promoters of the selected one (previous step) and probe their RNA production at the single-RNA, single-cell level [13]; and iv) Select the mutant promoter whose RNA numbers (mean and variability) best fit the model [1].

Here, we describe the strategy and the methods and present a case-study of the use of this strategy to obtain promoters with pre-defined transcription dynamics.

2 Methods

2.1 Stochastic Model of Transcription and cell division

The stochastic model transcription used here is based on multiple studies of transcription dynamics of individual genes [1,9,25]. The values set for each parameter were obtained from empirical data [9,26-30].

The multi-step transcription process of an active promoter, P_{ON}, is modeled by reaction (1) [31] and its repression mechanism by reaction (2).

In reaction (1), the closed complex (RP_c) is formed once an RNA polymerase (RNAp) binds to a free, active promoter [32]. Subsequent rate-limiting steps follow to form the open complex (RP_o) [31,32].

Finally, elongation starts [33], clearing the promoter. Elongation is not explicitly modeled since its time-length is much smaller than that of the rate-limiting steps in initiation [9]. Further, this process only affects noise in RNA production (mildly), not its mean rate [18]. In the end, an RNA is produced and the RNAp is released.

In the multi-step reaction (1), k_1 is the rate at which an RNAp finds and binds to promoter P, k_{-1} is the rate of reversibility of the closed complex, k_2 is the rate of open complex formation, and k_3 is the rate of promoter escape (expected to be much higher than all other rates, and thus assumed to be 'negligible' [9]):

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

The reaction in (1) should not be interpreted as elementary transitions. Namely, they represent the effective rates of the rate-limiting steps in the process, which is what defines the promoter strength [9]. Next, reaction (2) models the changes in the state of the promoter, from repressed (P_{OFF}.Rep) to free for transcription, i.e. active, due to the binding/unbinding of repressor proteins (Rep) to the promoter region:

$$P_{OFF}.\text{Rep} \underset{k_{off}*[\text{Rep}]}{\overset{k_{on}}{\rightleftarrows}} P_{ON} \tag{2}$$

In general, under full induction, the cell contains sufficient inducers to render all repressors "inactive" at all times (which can be modeled by having no Repressors in the cell). Finally, the single-step reaction (3) models RNA degradation [34]:

$$rna \xrightarrow{kd_{RNA}} \varnothing \tag{3}$$

We note that in this model, as our RNA probes (MS2-GFP coating of the target RNA, see section 2.4) cause the RNA to be non-degradable for a time longer than the observation time [9-11,23,24], reaction (3) is not included in the model.

Finally, we assume that our gene of interest is integrated into a single-copy plasmid (not anchored to the cell membrane). Thus, we assume that the accumulation of super-coiling caused by topological constraints is negligible [35].

Given this model, we define τ_{prior} as the mean expected time for a successfully closed complex formation, which depends on the mean-time and number of attempts to initiate an open complex formation (which depends on the RNAp intracellular concentration). Meanwhile, the remaining time to produce an RNA, τ_{after}, includes the steps following commitment to open complex formation (e.g. isomerization [36]), and *prior* to transcription elongation. The mean time interval between consecutive RNA productions (Δt_{active}) of a fully active promoter is thus given by:

$$\Delta t_{active} = \tau_{prior} + \tau_{after} \tag{4}$$

Relevantly, τ_{after} does not depend on the RNAp intracellular concentration [36]. This is of significance in that, e.g., changes in this concentration will only affect τ_{prior} and thus, will only partially affect Δt_{active}. Based on this, we simplify the model, so as to be in accordance with the sensitivity of the measurements of rate-limiting steps (see below), as follows. From (1), we assume the following approximate model:

$$P_{ON} + RNAp \xrightarrow{k_1^*} RP_c \xrightarrow{k_2} RP_o \xrightarrow{k_3} RNAp + P_{ON} + rna \tag{5}$$

where: $k_1^* = \tau_{prior}^{-1}$, $k_2 = \tau_{after}^{-1}$, and k_3 = fast (i.e. 'negligible' in that it does not act as a rate-limiting step in RNA production [9].

In addition, aside from transcription, note that cell division has a major effect on RNA numbers due to 'dilution', as the RNAs are partitioned in the two daughter cells. Here, we assume a near-perfect process of partitioning [37] as the RNAs are expected to be randomly distributed in the cytoplasm. Namely, we assume that, when the num-

ber of RNAs is even, each daughter cell receives half of them. If the number is odd, one daughter cell receives (half + 0.5) and the other receives (half - 0.5).

This model assumes only one copy of the promoter in each cell at any given time. This approximation is made possible by the slow division time of the bacteria strain used here (see section 2.3). Specifically, in our measurement conditions, it was established that these cells spend no more than 11 ± 1.2% of their lifetime with two copies of the target promoter (in agreement with previous measurements [9]).

2.2 Stochastic Simulations

Simulations are performed by SGNS [15], a simulator of chemical reaction systems based on the Stochastic Simulation Algorithm [38] and the Delay Stochastic Simulation Algorithm [22]. It thus allows simulating multi-delayed reactions within hierarchical, interlinked compartments that can be created, destroyed and divided at runtime. During cell division, molecules are near-evenly segregated into the daughter cells. Each model cell consists of reaction (5) along with the rate constants values (see Results section) and the initial number of each of its component molecules.

Our model (described in section 2.1, reaction 5), uses the following parameter values: $k_1^* = 390^{-1}\ s^{-1}$, $k_2 = 210^{-1}\ s^{-1}$. Given these, we expect $\Delta t = 600$ s, and $\tau_{after} / \Delta t = 0.35$. We begin simulations with 300 cells containing no RNAs. Each cell contains 1 promoter and 1 RNAp molecule. These numbers were shown to be able to reproduce realistically the RNA production kinetics of $P_{Lac\text{-}Ara\text{-}1}$ in [9]. Also, we set a mean cell division time of 1200 s (for simplicity assumed to be constant), and we analyze the RNA numbers in the cells at the end of the simulation time (3600 s). Finally, the partitioning of RNA molecules in cell division is performed as described in section 2.1.

2.3 Strain, Cell Growth, and Stress Conditions

E.coli strain used is DH5α-PRO (identical to DH5α-Z1) [39], and its 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*. Plasmids construction and transformation were done by using standard molecular cloning techniques (see section 2.4). From single colonies on LB agar plates, cells were cultured in LB medium with the appropriate concentration of antibiotics and incubated overnight at 30 °C and 250 rpm. The overnight cultures were then diluted to an initial optical density (OD_{600}) of 0.05 in fresh LB medium, with a culture volume of 5 ml supplemented with the antibiotics, kanamycin for the reporter gene, and chloramphenicol for the target gene (Sigma-Aldrich, USA). Cells along with antibiotics were then incubated at 37 °C with a 250 rpm agitation until reaching an OD600 of ~0.3.

Next, to induce the expression of the reporter MS2-GFP proteins, 100 ng/ml of aTc (Sigma-Aldrich, USA) was added and cells were incubated at 37 °C for 30 min with 250 rpm agitation. Then, cells were incubated at 37 °C (Innova® 40 incubator, New Brunswick Scientific, USA) for 15 min with agitation, before activating the target gene. Following full induction of the target gene (1mM IPTG and 0.1% L-arabinose, Sigma-Aldrich, USA), cells were incubated for 1 additional hour at 37 °C, *prior* to image acquisition. Partial induction of the target gene was achieved by adding to the

media either only 1mM IPTG or 0.1% L-arabinose. Finally, oxidative and acidic stresses were induced by adding, respectively, 0.6mM of H2O2 and 150mM of MES to the culture for 1 hour along with the induction of target gene during cell exponential phase, as described in [40].

2.4 Single-RNA detection system in a single-copy F-plasmid

We detect individual RNA molecules using a fluorescent MS2 tagging system that, using confocal microscopy, allow sensing integer-valued RNA numbers in individual cells, as soon as they are produced [10,11]. From these numbers in individual cells over time, we characterize the dynamics of transcription initiation of the promoter of interest [9,23,24]. For this, we obtain intervals between consecutive RNA production events in individual cells (here defined as 'Δt'). Also, we obtain the distribution of RNA numbers, and from them, calculate the mean (M) and coefficient of variation (CV) of RNA numbers in individual cells [4,26].

This technique uses an RNA coding sequence of multiple MS2 binding sites [10,11]. Here, we engineered an RNA with 48 MS2 binding sites, with unique restriction enzymes, validated by sequencing. The construction of the single-RNA, single-protein fluorescent probe was done into two steps. First, a promoter region, a coding region for a fluorescent protein (mCherry), and the RNA with binding sites for MS2d-GFP proteins were independently synthesized *de novo* (GeneScript, USA). Second, using GenEZ™ molecular cloning, these sequences were ligated until forming the sequence of the 'target gene'. Next, they were cloned into a single-copy F-plasmid (GeneScript, USA). This probe informs on the kinetics of RNA production, at the single cell level (Figure 1).

In our case-study, the 'target gene' is controlled by a $P_{Lac\text{-}Ara\text{-}1}$ promoter controlling the expression of a mCherry fluorescent protein, followed by an array of 48 binding sites for MS2d-GFP. Figure 1 shows the complete single-RNA detection system, composed of the 'target gene' and the 'reporter gene'. The latter is on a multi-copy plasmid carrying the $P_{L\text{-}tetO1}$ promoter controlling the expression of the fused fluorescent protein 'MS2d-GFP'. This protein rapidly binds to the MS2 binding sites of the target RNA, making it visible as a fluorescent 'spot' under fluorescence microscopy in less than 1 minute, provided sufficient MS2d-GFP proteins in the cell (Figure 3). In one of the strains engineered, the target promoter is P_{tetA}. In this case, we replaced the $P_{L\text{-}tetO1}$ promoter controlling the expression of MS2d-GFP by the P_{BAD} promoter.

By combining again *de novo* synthesis of DNA fragments with common molecular cloning and DNA assembling techniques, this new fluorescent probe can be further modified in the promoter region, RBS, and in regions between arrays of 12 MS2 binding sites, since pre-defined 'cutting points' were inserted on those regions.

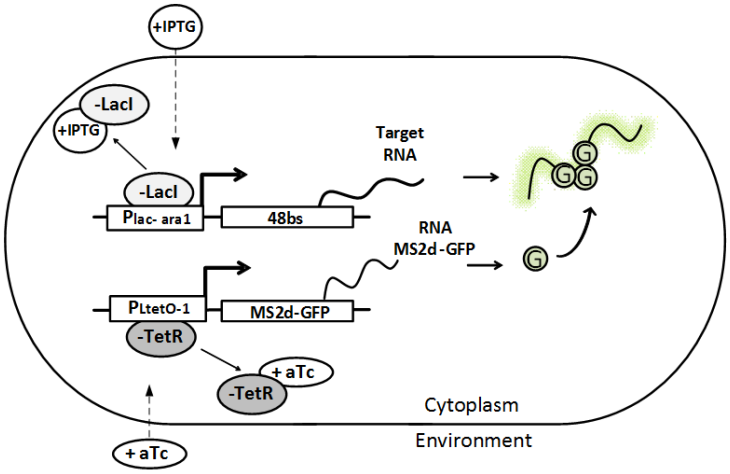

Fig 1. Schematics of the engineered strain DH5α-PRO with the 'target gene' and its RNA tagging system, along with the intake system of one of the inducers of the 'target gene', IPTG. When in the cytoplasm, IPTG neutralizes the overexpressed LacI repressors by forming inducer-repressor complexes (LacI-IPTG). This allows the $P_{Lac\text{-}ara\text{-}1}$ promoter to express RNAs that include the array of 48 MS2d-binding sites. Meanwhile, MS2d-GFP expression is controlled by a $P_{L\text{-}tetO\text{-}1}$ promoter, which is regulated by TetR repressor, produced by its native promoter in *E. coli*'s chromosome, and the inducer anhydrotetracycline (aTc). Once an individual RNA molecule is produced, multiple tagging MS2d-GFP proteins (referred to as G) rapidly bind to it forming a visible bright spot under a confocal microscope [6,23]. The tagging of MS2d-GFP molecules provides the RNA a long lifetime, with constant fluorescence, beyond the observation time of the measurement (see Figure 3) [4].

2.5 Relative RNAp Quantification

To achieve different RNAp concentrations in cells, we altered their growth conditions as in [5]. For this, we used modified LB media which differed in the concentrations of some of their components. The media used are denoted as m x, where the composition per 100 ml are: m g tryptone, $m/2$ g yeast extract and 1 g NaCl (pH = 7.0). E.g. 0.25x media has 0.25 g tryptone and 0.125 g yeast extract per 100 ml.

Relative RNAp concentrations were measured using *E. coli* RL1314 cells with fluorescently-tagged β' subunits. These were grown overnight in the respective media. A pre-culture was prepared by diluting cells to an OD_{600} of 0.1 with a fresh specific medium and grown to an OD_{600} of 0.5 at 37 ºC 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). As a control, we also measured the relative RNAp concentrations in RL1314 cells under a confocal microscope (see section 2.7). Relative RNAp concentrations were estimated from the mean fluorescence of cells growing in each media. We found no differences using either method.

2.6 Mutant Target Promoters

We engineered 4 mutant target promoters from the original $P_{Lac\text{-}Ara\text{-}1}$ (referred to as 'control'). Figure 2 shows these sequences, including the control. As also shown in the Results section, single- and double-point mutations in the -35 and -10 promoter elements can affect the transcription initiation rate-limiting steps kinetics [2,36].

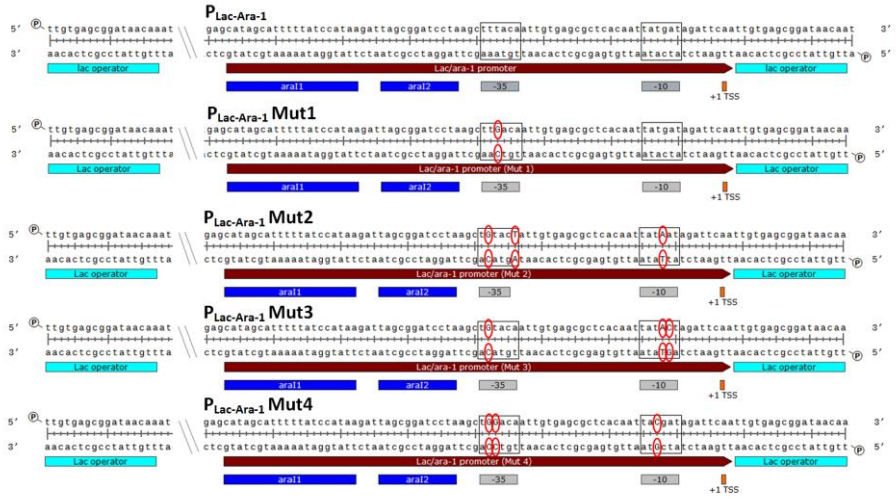

Fig 2. Schematic representation of the target promoter's sequences: The -35 and -10 promoter elements are shown in black boxes. The transcription start site (+1 TSS) are marked in orange. Operator sites are marked as cyan and blue. In the mutants, specific nucleotide changes in the -35 and -10 regions are marked as red circles.

2.7 Microscopy and Image Analysis

Cells with the target and reporter genes were grown as above. After, cells were pelleted and re-suspended in ~100 µl of the remaining media. 3 ml of cells were placed on a 2% agarose gel pad of LB medium and kept in between the microscope slide and a coverslip. Cells were visualized by a Nikon Eclipse (Ti-E, Nikon) inverted microscope with a 100x Apo TIRF (1.49 NA, oil) objective. Confocal images were taken by a C2+ (Nikon) confocal laser-scanning system. MS2-GFP-RNA fluorescent spots and RNAp-GFP were visualized by a 488 nm laser (Melles-Griot) and an HQ514/30 emission filter (Nikon). Phase contrast images were taken by an external phase contrast system and DS-Fi2 CCD camera (Nikon). Phase contrast and confocal images were taken once and simultaneously by Nis-Elements software (Nikon).

For time series imaging, a peristaltic pump provided a continuous flow of fresh LB media (supplemented with inducers for the target and reporter genes and chemicals for stresses, at appropriate concentrations) to the cells, at the rate of 0.3 ml/min, through the thermal chamber (CFCS2, Bioptechs, USA). The temperature was kept as desired (at either 30, 37 or 39°C) by a cooling/heating microfluidic system, which

provides a continuous deionized water flow at a stable temperature (with no contact with cells) into the thermal chamber.

After image acquisition, cells were detected from phase contrast images as in [9,23]. Phase contrast and fluorescence images were aligned using cross-correlation maximization and then cells were automatically segmented from phase contrast images using CellAging [41], followed by manual correction. Cell lineages were determined by overlapping areas of the segments between consecutive images.

The number of RNA molecules in individual cells and their corresponding production rates were obtained. Since the lifetime of an MS2-GFP-tagged RNA is much longer than cell division times [3,10,42], the cellular foreground intensity is expected to always increase (by 'jumps'), with a jump in intensity corresponding to the appearance of a new tagged RNA (Figure 3). The position of the jumps, thus the time interval between them, are estimated by applying a specialized curve fitting algorithm. The observed time intervals, which are related to the moment of two consecutive RNA productions, are extracted, and the intervals that occur after the last observed production event are rendered right censored. Because the observed time intervals tend to be short ones, i.e. lacking longer intervals, the right censored procedure is applied to improve the accuracy and avoid underestimating time interval durations [43].

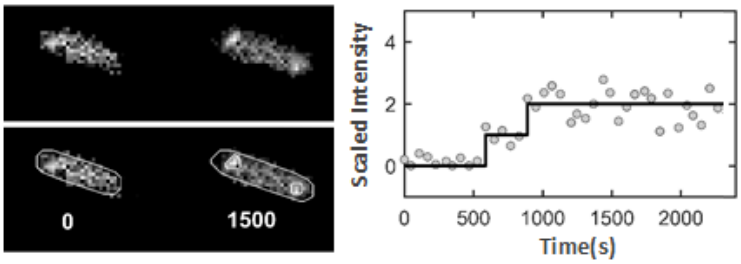

Fig 3. (Left) Example images of an *Escherichia coli* cell expressing MS2-GFP and target RNA, taken by confocal microscopy (Top). (Middle) Segmented cells and RNA-MS2-GFP spots within. (Right) Time series of the scaled intensity of the two spots in the cell shown at the top, along with a monotone piecewise-constant fit (black line) [43].

Finally, for fluorescent RNAp studies, RNAp abundance was quantified from the total fluorescence intensity extracted from the fluorescence microscopy images.

2.8 Extracting the duration of the rate-limiting steps in transcription

This method established in [9] and used in [4,23,24,26] is based on the assumption that, increasing in the concentration of active RNAp leads to an increase in the rate of RNA production, in accordance with the model in reaction (5) and validated by recent measurements *in vivo* (see e.g. [9]).

Visibly, from reaction (5), this increase is due to the increased rate of the steps *prior* to commitment to the open complex formation, while the rate of the subsequent steps remains unaltered [36]. Note that, given this, as one further increases the amount of RNAp concentration, the model assumes that, at some point, the duration of the

first step becomes negligible, when compared to the time-length of the second step. In such regime, one expects the rate of transcription to equal the inverse of the rate of the steps after commitment to the open complex formation [9,36].

Given the above, by conducting measurements of transcription rates at different intracellular concentrations of RNAp, it is possible, by linear fitting, to infer the duration of the steps after commitment to the open complex formation [9,36]. For this, from microscopy images, using the 'jump detection' method described above, we first obtained the mean duration of the time intervals (Δt) between consecutive RNA production events in individual cells.

Next, to estimate τ_{after} / Δt, we plot the inverse of the RNA production rate (Δt) against the inverse of the relative RNAp concentrations, for various conditions differing in the concentration of RNAp in the cells relative to the control (these are such that cell growth rate is unaltered, as described in [9]). From this, one obtains a Lineweaver–Burk plot [44] and then fits a line to the data points to obtain the estimated rate of RNA production for "infinite" RNAp concentration (i.e. for an amount of RNAp sufficient for the steps *prior* to open complex formation to have negligible duration). This method is valid if the increase in the rate of RNA production is linear with the increase in RNAp concentration, within the range of conditions used [9] (which was shown to be true in [9,26].

Given the model of transcription (eq. 4), one can write the mean time interval between consecutive RNA productions (Δt) as:

$$\Delta t = \tau_{prior} + \tau_{after} \tag{6}$$

τ_{prior} includes the time taken by multiple attempts to form a stable closed complex, whose kinetics depends on the RNAp intracellular concentrations, whereas τ_{after} does not depend on RNAp intracellular concentrations. As such, after a change in the RNAp concentration, since only τ_{prior} is affected, the new mean time interval is:

$$\Delta t^{new} = \tau_{prior}{}^{new} + \tau_{after} \tag{7}$$

Where $\tau_{prior}{}^{new} = S^{-1} \times \tau_{prior}$ with, $S = \frac{[RNAp]^{new}}{[RNAp]}$. From this, one can write:

$$\frac{\Delta t^{new}}{\Delta t} = \frac{S^{-1} \times \tau_{prior} + \tau_{after}}{\tau_{prior} + \tau_{after}} \tag{8}$$

Assuming a condition where cells contain an infinite concentration of RNAp, S^{-1} becomes null and equation 8 can be rewritten as:

$$\frac{\tau_{after}}{\tau_{prior} + \tau_{after}} = \frac{\Delta t^{new}(RNAp = \infty)}{\Delta t} \tag{9}$$

Given this, from measurements of Δt from single-cell time-lapse microscopy measurements and the corresponding RNAp concentrations in a few conditions differ-

ing in intracellular RNAp concentrations, one can extrapolate τ_{after} / Δt. From the value of Δt, one can use equation (9) to obtain, τ_{after}, and subsequently, τ_{prior}.

Knowing these values, it is possible to then simulate a model (reaction (5)), which, if accounting for RNA dilution due to cell division, is expected to provide estimations of the expected mean and CV of RNA numbers in individual cells, at any moment *t* following the induction of the target promoter.

2.9 Assessing the similarity between the rate-limiting steps kinetics of the model and the rate-limiting steps kinetics of the promoter of interest

To determine the best fitting promoter to the desired rate-limiting steps kinetics, we calculate as follows the Euclidean distance between the vectors (Δt, τ_{after}) of the constructs and of our preferred values (Δt_0, $\tau_{after,0}$), as determined by the model:

$$D = \sqrt{\alpha \times (\Delta t - \Delta t_0)^2 + \beta \times (\tau_{after} - \tau_{after,0})^2} \qquad (10)$$

Here, for simplicity, the 'weights' α and β are set to 1, which implies that 'similar importance' is given to fitting the values of Δt_0, $\tau_{after,0}$ (as they have the same order of magnitude). Other methods of calculating this distance could be used, depending on the importance of fitting Δt_0 and $\tau_{after,0}$, by tuning the values of α and β.

2.10 Assessing the similarity between the RNA numbers in synthetic mutant promoters and the desired RNA numbers at the single cell level.

To best fit the mean rate of production and noise in RNA production (here assessed by the CV of RNA numbers in individual cells), we assess the "goodness of fit" of our construct dynamics to the "desired dynamics", by calculating the Euclidian distance between (M, CV) of our 'best fit' construct to the desired (M_0, CV_0) as follows:

$$D_{(M,CV)} = \sqrt{\alpha \times (M - M_0)^2 + \beta \times (CV - CV_0)^2} \qquad (11)$$

Here, for simplicity, the 'weights' α and β are set to 1, which implies that 'more importance' is given to fitting the value of M_0, since, as seen in the Results section (Table 2), the values of *M* are 1 order of magnitude higher than the values of *CV*. Other methods of calculating this distance could be used, depending on the importance of fitting M_0 and CV_0, by tuning the values of α and β.

3 Results and Conclusions

To show how the framework performs and assess its performance (which depends on our library of genes available), we first created a hypothetical specification of a gene with a given RNA production dynamics that would result in a mean number of produced RNA molecules in each cell (M(RNA)) equal to 2.5 and a coefficient of variation (CV(RNA)) equal to 0.6, after 60 minutes of induction.

We selected these criteria with prior knowledge that they should be more suited by a $P_{Lac\text{-}Ara\text{-}1}$ promoter than, e.g. P_{BAD} or P_{tetA} [26]. The fact that the methodology chooses the former rather than the latter two (see below), is a means of assessing its effectiveness. We further note that, in this example, the construct is to be implemented on a single-copy plasmid in *E. coli* DH5α-PRO cells grown at 37 ºC (Methods, section 2.3), and that the intended distribution is to be reached 1 hour after induction of the target gene, responsible for producing the RNA target the MS2-GFP proteins.

Finally, we note that one could overstep the modeling stage, by instead testing a large number of promoters and mutant ones, until satisfying the specifications. The purpose of modeling is to assist in the selection of the 'most promising' promoter, so as to minimize time not only in the genetic constructs but perhaps more importantly, in the measurements (including image analysis, etc.) that are required to determine whether a promoter is suitable for the goals.

3.1 Design of a stochastic model that best fits the intended distribution of RNA numbers in individual cells

We first calculate the necessary values of τ_{prior} and τ_{after} in transcription initiation that would produce such RNA numbers, in the conditions defined, in accordance with the model of transcription assumed. First, the value of Δt is obtained from the expected mean RNA numbers in the cells, taking into consideration the cells division rate in the pre-established conditions. For this, we used the following formula:

$$\Delta t = D \times \left(1 - 2^{\left(\frac{-t}{D}\right)}\right) \times \left(M \times \log 2\right)^{-1} \qquad (12)$$

where D is the mean cell division time and M is M(RNA) in a population observed at time *t* after induction of the target gene. We measured D to be ~20 minutes and we have set our measurement duration (*t*) to 60 minutes. From these values, using equation (12), we find that we require a model whose mean interval between consecutive RNA production events (Δt) equals 10.1 minutes.

Next, and most importantly, we tuned the values of k_1^* and k_2 such that CV(RNA) equals 0.6 (bounded by the requirement that Δt = 10.1 min).

For this, using SGNS [15], we performed simulations of 300 model cells per condition, each with, at the start, 1 promoter and no RNA. From these, we found that our goal (Δt = 10.1 min and CV = 0.6 after 1 hour of simulation time and accounting for cell divisions as described in sections 2.1 and 2.2) can be achieved in good approximation by setting $[RNAP]*k_1^*$ and k_2 such that $\tau_{after} / \Delta t$ equals 0.35. There are several solutions that fit these criteria. Here, we set $[RNAP]*k_1^* = 390^{-1}\ s^{-1}$ and $k_2 = 210^{-1}\ s^{-1}$.

3.2 Select a known promoter whose kinetics best fits the expected rate-limiting steps kinetics

Having established the model, we next searched for a best fitting promoter. This could be done using a pre-defined library of promoters whose rate-limiting steps kinetics has been previously dissected [9,26].

Here, as an example, we dissected *de novo* the rate-limiting steps kinetics of RNA production of 3 promoters ($P_{Lac\text{-}Ara\text{-}1}$, $P_{tetA,}$ and P_{BAD}) from microscopy measurements (Methods). We also measured the duration of τ_{after} for each of these promoters, using τ plots (section 2.8). Results are shown in Table 1.

For this, we first inserted each of the promoters of interest in the single-copy plasmid carrying the RNA coding for 48 MS2-GFP binding sites as described in section 2.4. Aside from inserting this plasmid, we also inserted the multi-copy plasmid coding for MS2-GFP (section 2.4). In the case of P_{tetA}, the expression of MS2-GFP in the multi-copy plasmid was controlled by a P_{BAD} promoter (Methods).

We then performed microscopy measurements in order to measure Δt and $\tau_{after}/\Delta t$ (Methods), from which we obtain also τ_{after}, for each of these 3 promoters (Methods, sections 2.5 and 2.8). These values are also shown in Table 1.

To determine the best fitting promoter (section 2.9), applying equation (10), we calculated the Euclidean distance between the vectors (Δt, τ_{after}) of the constructs and the pre-established values obtained by the stochastic model. Results are shown in Table 1.

Table 1. The promoter name, the mean (Δt) and τ_{after} as measured by microscopy, and the Euclidean distance between each promoter and model RNA production kinetics.

Promoter	Mean Δt (min)	τ_{after} (min)	Euclidean Distance to Model
Model	10.00	3.5	-
$P_{Lac\text{-}Ara\text{-}1}$	9.82	2.4	1.11
P_{tetA}	19.23	18.1	17.3
P_{BAD}	12.03	2.9	2.12

From Table 1, the best fitting promoter, of those tested, is $P_{Lac\text{-}Ara\text{-}1}$, given that it is the one whose Euclidean distance to the model transcription kinetics is minimal.

3.3 Design and engineering mutant promoter sequences with a fluorescent RNA-sensor and comparison of the construct and the model dynamics

Having selected the promoter ($P_{Lac\text{-}Ara\text{-}1}$) which best fits the desired values of Δt and τ_{after}, next, we fine tune our construct by engineering mutant promoters (from $P_{Lac\text{-}Ara\text{-}1}$) with differing kinetics. The mutant promoters differ in sequence from the original $P_{Lac\text{-}Ara\text{-}1}$ in the -35 and -10 regions (see section 2.6) as these were shown to control, to some extent, the kinetics of the rate-limiting steps in transcription initiation [9].

Our aim is to find a mutant promoter that can 'outperform' $P_{Lac\text{-}Ara\text{-}1}$ regarding the resulting single-cell RNA numbers (Mean and CV) when compared to these numbers obtained by the stochastic model. For this, we induced each of these mutant promoters as in the case of the original $P_{Lac\text{-}Ara\text{-}1}$ construct (Methods) and measured by microscopy the mean and CV of RNA numbers in individual cells, 1 hour after induction, at 37 ^{o}C (Methods). We also obtained these numbers from simulations of model cells. Next, we estimated uncertainties of these features (Mean and CV) using a non-parametric bootstrap method [45]. Results are shown in Table 2. To assess which construct best fits the model numbers, using equation (11), we calculated the Euclidean distance

between the vectors (M(RNA), CV(RNA)) of the constructs and the model (M_0(RNA), CV_0(RNA)) (section 2.10). These distances are also shown in Table 2.

Table 2. Shown are the promoter name, the mean number of RNAs in each cell (M) and the coefficient of variance (CV) of RNA numbers in individual cells 60 min. after induction of the target gene $P_{Lac\text{-}Ara\text{-}1}$ promoter, referred to as 'LA', and its four mutations (in order, here referred to as 'Mu1', 'Mu2', ' Mu3', and ' Mu4'), with cells grown in the same induction scheme and environment conditions (Full induction: 0.1% Arabinose, 1mM IPTG; 1x LB media, 37°C) (see sections 2.3 and 2.5). Also shown are M and CV from the simulations of the model, along with the Euclidean distance between each promoter's resulting RNA numbers in individual cells and the model RNA numbers. Error bars represent the standard error, calculated as the standard deviation of the bootstrapped distributions (200 cells each) from 1000 random resamples with replacement.

Condition	M(RNA)	CV (RNA)	Euclidian Distance to Model	KS test (mean *P* value)
LA (control)	3.59±0.16	0.63±0.04	1.35	< 0.01
Mut1	2.02±0.09	0.65±0.04	0.22	0.05
Mut2	1.54±0.06	0.56±0.02	0.71	< 0.01
Mut3	1.61±0.11	0.97±0.02	0.71	< 0.01
Mut4	1.23±0.03	0.39±0.02	1.04	< 0.01
Model	2.24±0.1	0.65±0.04	-	-

From Table 2, we find $P_{Lac\text{-}Ara\text{-}1}$ Mut1 to be the construct that best fits the model-based pre-established single-cell RNA numbers.

Finally, we assessed 'how well' the best fitting construct fits the original criteria set by the stochastic model. For this, we obtained the distribution of RNA numbers in individual model cells and in measurements of RNA numbers in individual cells, 1 hour after induction. Next, we performed Kolmogorov-Smirnov tests of statistical significance comparing the distributions of RNA numbers in model and measurements. In general, the higher the *P* value, the more 'similar' are the two distributions compared, as the more likely it is that the two sets of data are drawn from the same distribution. The results, for all constructs, are shown in Table 2. From these, we find that the RNA numbers distribution resulting from 'Mut1' has the highest probability of being drawn from the same distribution as the one drawn from the model data, confirming that this construct is the one that best fits the model criteria.

Further, its p-value equals 0.05, which in general allows assuming that the two sets of data are drawn from the same distribution and, thus, cannot be distinguished in a statistical sense. We thus conclude that the construct fits the necessary criteria.

3.4 Achieving the desired dynamics by changing promoter induction scheme and environment conditions

In general, genes are already equipped with several regulatory mechanisms (such as the repression mechanism modeled above) (see, e.g., [46]). Further, cells interact

with the environment, in that they have global regulatory mechanisms of gene expression, such as σ factors, and responsiveness to changes in the chemical composition of the media. As a result of these systems, cells express genes with a dynamics that is environment-dependent. One can make use of this, to further enhance the goodness of fit between the observed expression dynamics and the desired dynamics, using the same methodology as in the section above.

As an example, we measured the kinetics of transcription of the $P_{Lac\text{-}Ara\text{-}1}$ promoter ('LA') when subject to different induction schemes, when under oxidative and acidic stresses, at different temperatures, and at different media conditions. Results of these measurements are shown in Table 3.

Again, from the empirical data and model dynamics, we calculated the Euclidean distance between the kinetics at each measurement condition and the model dynamics. We also performed KS tests to assess the goodness of fit of each construct dynamics to the original criteria set by the model. Results are also shown in Table 3.

Table 3. Shown are the promoter, mean number of RNAs in each cell (M) and coefficient of variance (CV) of RNA numbers in individual cells 60 min. after induction of the target gene $P_{Lac\text{-}Ara\text{-}1}$ promoter, referred to as 'LA', when cells are grown in various environment conditions (see sections 2.3 and 2.5). Also shown are M and CV from the simulations of the model, along with the Euclidean distance between each promoter's resulting RNA numbers in individual cells and the model RNA numbers. Error bars represent the standard error, calculated as the standard deviation of the bootstrapped distributions (200 cells each) from 1000 random resamples with replacement.

Condition	M(RNA)	CV (RNA)	Euclidian Distance to Model	KS test (mean *P* value)
LA 37°C 0.1% Ara, 1x LB	1.83±0.07	0.55±0.04	0.42	< 0.01
LA 37°C 1mM IPTG, 1x LB	1.85±0.08	0.62±0.04	0.39	< 0.01
LA 37 °C Oxidative, Full Ind., 1x LB	1.39±0.05	0.51±0.03	0.86	< 0.01
LA 37 °C Acidic, Full Ind., 1x LB	1.62±0.07	0.62±0.04	0.62	< 0.01
LA 30 °C, Full Ind., 1x LB	2.38±0.1	0.61±0.03	0.15	0.24
LA 39 °C, Full Ind., 1x LB	2.72±0.11	0.52±0.02	0.50	< 0.01
LA 37 °C, Full Ind., 0.5x LB	2.66±0.13	0.61±0.04	0.42	< 0.01
LA 37 °C, Full Ind., 2x LB	3.16±0.18	0.82±0.13	0.94	< 0.01
Model	2.24±0.1	0.65±0.04	-	-

Similarly to when testing different promoter sequences, we find that one condition ('LA 30 °C, Full Ind., 1x LB') fits the model in a statistical sense (*P* value equals 0.24).

4 Discussion

We proposed a new strategy for designing genes with a predefined dynamics of RNA production (mean and CV of the single-cell RNA numbers in a population). In short, based on the desired RNA numbers in individual cells, we first produce a stochastic model that informs on the necessary promoter initiation kinetics (rate-limiting steps kinetics). From that information, we search for a promoter that best fits this initiation kinetics from τ-plot measurements [9] and/or from a database already containing such information. Once reducing the state-space of our search to a 'best-possible' promoter, next we fine-tune its RNA production kinetics. Namely, mutant promoters are engineered from this best-possible promoter and their RNA production kinetics is analyzed at the single-cell, single-RNA level (namely, we obtain mean and variability in RNA numbers at the single cell level).

In addition to this, we showed that the RNA production dynamics can be further (or alternatively) tuned by changes in the promoter induction scheme (i.e. inducer concentrations) and environment conditions (e.g. temperature and media richness).

Combining these multiple strategies (mutations, multiple environmental conditions, and various induction schemes) will provide a significantly large array of various transcription dynamics, which will greatly enhance the chances of finding a 'best fitting' dynamics to that of the desired model.

At the moment, the main limitation of this strategy is the limited number of studies on the *in vivo* underlying kinetics of the rate-limiting steps of promoters in *E. coli* using live single-RNA sensitivity [6,7,11,23,24,26]. As the library of the genetic constructs whose dynamics have been analyzed with this level of detail increases, we expect the impact of our strategy to increase as well, as the chances to match a predefined kinetics of RNA production increase.

Other means to further improve this strategy include the development of more detailed models of transcription that incorporate phenomena such as promoter-proximal pausing, σ factor competition for RNA polymerase core enzymes, cell-cycle stage, tuning cell division rates [47] based on media richness, etc.

Overall, as the databases of models and the library of genetic constructs are enlarged, and as the techniques in synthetic biology, fluorescent microscopy, and stochastic modeling are enhanced, we will be able to accommodate an increasingly wide variety of predefined RNA production dynamics with enhanced precision.

We expect that the engineering of genes with tailored RNA production dynamics, while currently a rare practice (except in a few state-of-the-art basic scientific projects making use of Synthetic Biology), will become a key strategy in Biomedicine and Biotechnology for regulating cellular behavior for medicinal purposes, bioreactors output enhancement, etc.. Also, we expect it to become a key step in synthetic engineering of genetic circuits with tailored dynamics [48,49].

5 Acknowledgments

Work supported by Academy of Finland (295027 and 305342 to ASR), Jane and Aatos Erkko Foundation (610536 to ASR), Finnish Academy of Science and Letters (SO), TUT President Ph.D. grants (MB, SS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Jones, D.L., Brewster, R.C., Phillips, R.: Promoter architecture dictates cell-to-cell variability in gene expression. Science 346, 1533–1537 (2014).
2. Shih, M-C., Gussin, G.: Mutations affecting two different steps in transcription initiation at the phage λ PRM promoter, Proc. Natl. Acad. Sci. USA 80, 496-500 (1983).
3. Golding, I., Cox, E.C.: RNA dynamics in live *Escherichia coli* cells. Proc Natl Acad Sci USA 101(31), 11310–11315 (2004).
4. Goncalves, N.S.M., Startceva, S., Palma, C.S.D., Bahrudeen, M.N.M., Oliveira, S.M.D., Ribeiro, A.S.: Temperature-dependence of the single-cell kinetics of transcription activation in *Escherichia coli*. Phys. Biol. 15(2), 026007 (2017).
5. Liang, S., Bipatnath, M., Xu, Y.-C., Chen, S.-L., Dennis, P., Ehrenberg, M., Bremer, H.: Activities of constitutive promoters in *Escherichia coli*. J. Mol. Biol. 292, 19–37 (1999).
6. Ribeiro, A.S.: Kinetics of gene expression in live bacteria: from models to measurements, and back again. Can. J. Chem. 91(7), 487-494 (2013).
7. Mäkelä, J., Lloyd-Price, J., Yli-Harja, O., Ribeiro, A.S.: Stochastic sequence-level model of coupled transcription and translation in prokaryotes. BMC Bioinformatics 12, 121 (2011).
8. Häkkinen, A., Tran, H., Yli-Harja, O., Ribeiro, A.S.: Effects of rate-limiting steps in transcription initiation on genetic filter motifs, PLoS ONE 8(8), e70439 (2013).
9. Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J., Goncalves, N., Oliveira, S.M.D., Häkkinen, A., Ribeiro, A.S.: Dissecting the stochastic transcription initiation process in live *Escherichia coli*. DNA Res. 23(3), 203-214 (2016).
10. Peabody, D.S.: The RNA binding site of bacteriophage MS2 coat protein. EMBO J. 12, 595–600 (1993).
11. Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C.: Real-time kinetics of gene activity in individual bacteria. Cell. 123, 1025–1036 (2005).
12. Peabody, D.S.: Role of the coat protein-RNA interaction in the life cycle of bacteriophage MS2. Mol. Gen. Genet. 254, 358–364 (1997).
13. Fusco, D., Accornero, N., Lavoie, B., Shenoy, S.M., Blanchard, J.M., Singer, R.H., Bertrand, E.: Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. Curr. Biol. 13(2), 161-167 (2003).
14. Ribeiro, A.S., Zhu, R., Kauffman, S.A.: A General Modeling Strategy for Gene Regulatory Networks with Stochastic Dynamics. J. Comput. Biol. 13(9), 1630-1639 (2006).
15. Lloyd-Price, J., Gupta, A., Ribeiro, A.S.: SGNS2: A Compartmentalized Stochastic Chemical Kinetics Simulator for Dynamic Cell Populations. Bioinformatics 28, 3004-5 (2012).
16. Ribeiro, A.S., Häkkinen, A., Mannerström, H., Lloyd-Price, J., Yli-Harja, O.: Effects of the promoter open complex formation on gene expression dynamics. Phys. Rev. E 81(1), 011912 (2010).

17. Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O., Ribeiro, A.S.: Effects of Transcriptional Pausing on Gene Expression Dynamics. PLoS Comput. Biol. 6(3), e1000704 (2010).
18. Bahrudeen, M.N.M., Startceva, S., Ribeiro, A.S.: Effects of Extrinsic Noise are Promoter Kinetics Dependent. In: The 9th International Conference on Bioinformatics and Biomedical Technology on Proceedings, pp. 44-47. ICBBT 2017, Lisbon, Portugal (2017).
19. Häkkinen, A., Ribeiro, A.S.: Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. Bioinformatics 31(1), 69-75 (2015).
20. Häkkinen, A., Ribeiro, A.S.: Identifying Rate-Limiting Steps in Transcription from RNA Production Times in Live Cells. Bioinformatics 32(9), 1346-1352 (2016).
21. Häkkinen, A., Tran, H., Ingalls, B., Ribeiro, A.S.: Effects of multimerization on the temporal variability of protein complex abundance. BMC Syst. Biol. 7(Sup. 1), S3 (2013).
22. Ribeiro, A.S.: Stochastic and delayed stochastic models of gene expression and regulation. Math. Biosci. 223(1), 1-11 (2010).
23. Oliveira, S.M.D., Häkkinen, A., Lloyd-Price, J., Tran, H., Kandavalli, V., Ribeiro, A.S.: Temperature-Dependent Model of Multi-step Transcription Initiation in *Escherichia coli* Based on Live Single-Cell Measurements. PLoS Comput. Biol. 12, e1005174 (2016).
24. Mäkelä, J., Kandavalli, V., Ribeiro, A.S.: Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. Sci. Rep. 7, 10588 (2017).
25. Taniguchi, Y., Choi, P.J., Li, G.-W., et al.: Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533–538 (2010).
26. Kandavalli, V.K., Tran, H., Ribeiro. A.S.: Effects of σ factor competition are promoter initiation kinetics dependent. Biochim. Biophys. Acta (BBA)-Gene Regul. Mech. 1859, 1281–1288 (2016).
27. Mitarai, N., Sneppen, K., Pedersen, S.: Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J. Mol. Biol. 382, 236-245 (2008).
28. Bremer, H. and Dennis, P.P.: 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, 1553–1569 (1996).
29. Kennel, D., and Riezman, H.: Transcription and translation initiation frequencies of the *Escherichia coli* lac operon. J. Mol. Biol. 114(1), 1-21 (1977).
30. Cormack, B.P., Valdivia, R.H., Falkow, S.: FACS-optimized mutants of the green fluorescent protein (GFP). Gene 173(1), 33-38 (1996).
31. Saecker, R.M., Record, M.T. Dehaseth, P.L.: Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J. Mol. Biol. 412, 754–771 (2011).
32. Chamberlin, M.: The selectivity of transcription. Annu. Rev. Biochem. 43, 721–75 (1974).
33. deHaseth, P.L., Zupancic, M.L. and Record, M.T.: RNA polymerase promoter interactions: The comings and goings of RNA polymerase. J. Bacteriol. 180, 3019–3025 (1998).
34. Bernstein, J.A., Khodursky, A.B., Pei-Hsun, L., Lin-Chao, S. Cohen, S. N.: Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl Acad. Sci. USA 99, 9697–9702 (2002).
35. Chong, S., Chen, C., Ge, H., Xie, X.S.: Mechanism of Transcriptional Bursting in Bacteria. Cell 158, 314–326 (2014).
36. McClure, W.R.: Mechanism and control of transcription initiation in prokaryotes. Annu. Rev. Biochem. 54, 171–204 (1985).
37. Abhishekh, G., Lloyd-Price, J., Ribeiro, A.S.: *In silico* analysis of division times of Escherichia coli populations as a function of the partitioning scheme of non-functional proteins. In Silico Biol. 12, 9–21 (2014).

38. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81(25), 2340–2361 (1977).
39. Lutz, R., Bujard., H.: Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I 2 regulatory elements. Nucleic Acids Res. 25, 1203–1210 (1997).
40. Muthukrishnan, A.-B., Martikainen, A., Neeli-Venkata, R., Ribeiro, A.S.: n Vivo Transcription Kinetics of a Synthetic Gene Uninvolved in Stress-Response Pathways in Stressed *Escherichia coli* Cells. PLoS One 9, e109005 (2014).
41. Häkkinen, A., Muthukrishnan, A.B., Mora, A., Fonseca, J.M., Ribeiro, A.S.: CellAging: a tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. Bioinformatics 29(13), 1708–1709 (2013).
42. Tran, H., Oliveira, S.M.D., Goncalves, N.S.M., Ribeiro, A.S.: Kinetics of the cellular intake of a gene expression inducer at high concentrations. Mol. BioSyst. 11, 2579-2587 (2015).
43. Häkkinen, A., Ribeiro, A.S.: Characterizing rate-limiting steps in transcription from RNA production times in live cells. Bioinformatics 32(9), 1346-1352 (2016).
44. Lineweaver, H., Burk. D.: The determination of enzyme dissociation constants. J. Am. Chem. Soc. 56, 658–666 (1934).
45. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med. 19(9), 1141-64 (2000).
46. Schleif, R.: Regulation of the L-arabinose operon of *Escherichia coli*. Trends Gen. 16(12), 559–565 (2000).
47. Lloyd-Price, J., Tran, H., Ribeiro, A.S.: Dynamics of small genetic circuits subject to stochastic partitioning in cell division. J. Theor Biol. 356, 11-19 (2014).
48. Mannerström, H., Yli-Harja, O., Ribeiro, A.S.: Inference of kinetic parameters of delayed stochastic models of gene expression using a Markov chain approximation. EURASIP J Bioinform Syst Biol. 2011(1), 572876 (2011).
49. Oliveira, S.M.D., Chandraseelan, J.G., Häkkinen, A., Goncalves, N.S.M., Yli-Harja, O., Startceva, S., and Ribeiro, A.S.: Single-cell kinetics of the Repressilator when implemented in a single-copy plasmid. Mol. BioSyst. 11, 1939-1945 (2015).