Transcription closed and open complex formation coordinate expression of genes with a shared promoter region

Antti Häkkinen,[1] Samuel M. D. Oliveira,[1] Ramakanth Neeli-Venkata,[1] and Andre S. Ribeiro[1]

[1]*BioMediTech Institute and Department of Signal Processing,*
Tampere University of Technology, PO box 553, 33101, Tampere, Finland

Many genes are spaced closely, allowing coordination without explicit control through shared regulatory elements and molecular interactions. We study the dynamics of a stochastic model of a gene-pair in a head-to-head configuration, sharing promoter elements, which accounts for the rate-limiting steps in transcription initiation. We find that only in specific regions of the parameter space of the rate-limiting steps is orderly co-expression exhibited, suggesting that successful cooperation between closely spaced genes requires the co-evolution of compatible rate-limiting step configuration. The model predictions are validated by *in vivo* single-cell, single-RNA measurements of the dynamics of pairs of genes sharing promoter elements. Our results suggest that, in *E. coli*, the kinetics of the rate-limiting steps in active transcription can play a central role in the dynamics of pairs of genes sharing promoter elements.

I. INTRODUCTION

Closely-spaced gene-pairs abound in genomes of all life forms, from human [1, 2] to prokaryotes [3, 4]. Further, they are highly conserved [2, 5], suggesting that they yield functionalities with selective advantages.

Gene-pairs can be arranged head-to-head (transcriptionally divergent), with their transcription start sites (TSS) closely located, sharing promoter elements such as transcription factor binding sites [1, 2, 4]. Head-to-tail (tandem) and tail-to-tail (convergent) overlapping gene-pairs are also found, allowing interference between RNA polymerases (RNAP) [6] and/or with transcription factors [7, 8]. Each configuration can vary in several parameters, such as distance between TSSs, which affect transcription of the component genes [4, 9–11], allowing co-regulation without explicit control mechanisms. The multitude of naturally occurring configurations suggests that each yields distinct selective advantages.

While some configurations have been identified and their ubiquity established by models and measurements [2, 11–13], the range of possible behaviors and advantages as a gene regulation mechanism remain largely uncharacterized. Such characterization would benefit understanding the array of tasks that organisms such as *Escherichia coli* perform using closely-spaced promoters, as opposed to individual genes or genes connected by transcription factors.

One aspect not yet considered is the existence of multiple rate-limiting steps during transcription initiation [14, 15]. As only some of these steps are physically involved in the gene-pair interactions, we expect the nature of the rate-limiting steps of each promoter to affect the dynamics of closely-spaced configurations. Importantly, the durations of the open complex formation of a strong and a weak promoter can differ from little to up to two orders of magnitude [16] and live cell single-RNA measurements suggest that different promoters are rate-limited at different stages of transcription initiation [15, 17–20]. As such, it is plausible that promoters whose initiation kinetics are similar in mean duration but whose rate-limiting step structures differ will feature different dynamics in the bidirectional configuration.

Here, we study the dynamics of a stochastic model of a gene-pair in a head-to-head configuration sharing promoter elements (the most common closely-spaced gene-pair configuration [2, 5]) as a function of the rate-limiting step configuration of each gene. We analyze the models using analytical stochastic methods. Next, we validate the main findings by performing time-lapse microscopy measurements of individual genes and in pairs of genes sharing promoter elements, at the single-RNA level, in live *E. coli*.

II. METHODS

A. Models

Transcription in *E. coli* starts when an RNAP, recruiting the appropriate σ-factor, specifically binds to a promoter. This creates a closed complex of the RNAP and DNA, which can require several trials before stabilizing [21]. In strong promoters, this step is nearly irreversible [22]. The virtually irreversible open complex formation follows, consisting of e.g. DNA unwinding and compaction [23] and the RNAP clamp assembly [24].

We assume a variant of a model of transcription initiation of the overlapping promoters of the galactose operon in the absence of cAMP-CRP [3]. The transcribed promoter stochastically is selected based on the relative affinities between the two promoters and the RNAP, encoded in the forward rates of the closed complex formation of each promoter. After the selection, the remaining steps of transcription initiation occur at the promoter region [14]. The following stochastic chemical [25, 26] reactions are used to model this:

$$\begin{aligned} P_0 &\xrightarrow{k_1} I_1 \xrightarrow{k_3} P_0 + X \\ P_0 &\xrightarrow{k_2} I_2 \xrightarrow{k_4} P_0 + Y \end{aligned} \tag{1}$$

where P_0 represents a free promoter (unoccupied by an RNAP), I_1 and I_2 represent intermediate transcriptional complexes committed to transcribing genes 1 and 2, respectively, and X and Y represent the messenger RNA products (or, if they closely follow [27], proteins) of genes 1 and 2, respectively. A schematic is provided in Fig 2A, and a thorough analysis in the supplement.

If the genes did not share promoter elements, the intervals between productions of X (gene 1) would be [20]:

$$\begin{aligned} & {\tau_X}^{(1)} \sim \mathcal{E}^-(k_1, k_3) \\ \text{with} \qquad & \mathbb{E}\left[{\tau_X}^{(1)}\right] = {k_1}^{-1} + {k_3}^{-1} \\ & \mathbb{V}\text{ar}\left[{\tau_X}^{(1)}\right] = {k_1}^{-2} + {k_3}^{-2} \end{aligned} \tag{2}$$

where $\mathcal{E}^-(\lambda_1, \cdots, \lambda_n)$ represents a hypoexponential distribution with rates $\lambda_1, \cdots, \lambda_n$. Similarly, the production intervals for gene 2 would be ${\tau_Y}^{(1)} \sim \mathcal{E}^-(k_2, k_4)$.

These distributions are of low noise, as measured by the coefficient of variation (standard deviation over the mean), as this quantity equals unity for Poissonian production (exponential production intervals). The noise is determined by the ratio of k_1 and k_3. Regardless of the mean, it is minimized for steps of equal duration and maximized when a single step is rate-limiting. The dynamics of an individual gene is unaffected by the step order (i.e. interchanging k_1 and k_3 has no effect on ${\tau_X}^{(1)}$).

Regardless of the configuration, the mean and variance of the production intervals are linked to that of the produced RNAs. In the long-term (infinite time), the mean and variance of produced RNA per unit time are [28]:

$$\begin{aligned} \mu_Z &\doteq \lim_{t\to\infty} \frac{\mathbb{E}[\,Z(t)\,]}{t} = \mathbb{E}[\tau_Z]^{-1} \\ \eta_Z\,\mu_Z &\doteq \lim_{t\to\infty} \frac{\mathbb{V}\text{ar}[\,Z(t)\,]}{t} = \mathbb{V}\text{ar}[\tau_Z]\ \mathbb{E}[\tau_Z]^{-3} \end{aligned} \tag{3}$$

i.e. the mean number of RNAs produced per unit time (μ_Z) equals the inverse interval mean, while the Fano factor (variance over the mean) of the RNA numbers (η_Z) equals the squared coefficient of variation of the production intervals. The cell phenotype is also affected by other processes, such as RNA degradation and dilution due to cell division. Regardless, the mean and noise of the produced RNA numbers are directly linked to the phenotype (details in supplement) [29], so we expect our results to hold qualitatively in the presence of other processes.

B. Cells, plasmids, chemicals, and growth conditions

We used *E. coli* strain BW25113 (*lacI*$^+$ *rrnB*$_{\text{T14}}$ *ΔlacZ*$_{\text{WJ16}}$ *hsdR514* *ΔaraBAD*$_{\text{AH33}}$ *ΔrhaBAD*$_{\text{LD78}}$) [30], which contains the constitutive promoters P$_{\text{lacI+}}$ and P$_{\text{araC}}$ producing, respectively, LacI repressors [31] and AraC repressors. As this strain does not contain the *tetR* gene responsible for encoding TetR repressors, any gene downstream to a P$_{\text{tetA}}$ promoter is expressed constitutively.

We constructed five target systems on a single-copy pBELO plasmid. The first plasmid features the P$_{\text{lacO3O1}}$ promoter controlling the production of an RNA molecule coding for a red fluorescent mCherry protein followed by 48 binding sites for the MS2-GFP protein (mCherry-48BS). The other four systems are modified versions of the first, with the P$_{\text{lacO3O1}}$ promoter being replaced by the following promoters: (i) P$_{\text{BAD}}$ promoter; (ii) P$_{\text{lacO3O1-tetA}}$ dual-tandem promoter; (iii) P$_{\text{lacO3O1-BAD}}$ dual-tandem promoter; and (iv) P$_{\text{lacO3O1-lacO3O1}}$ dual-bidirectional promoter. All strains aside from its target system also contain either a medium-copy plasmid pZA25 with the reporter gene P$_{\text{ara}}$-MS2-GFP or a low-copy plasmid pZS12 with the reporter gene P$_{\text{lac}}$-MS2-GFP. These plasmids are responsible for producing the fusion protein MS2-GFP, both producing an abundance of MS2-GFP when activated as detailed below. The reporter plasmids were generously provided by Orna Amster-Choder (Hebrew University of Jerusalem, Israel) [32], and Philippe Cluzel (Harvard University, USA) [33], respectively. The activity of the promoters P$_{\text{lacO3O1}}$, P$_{\text{lacO3O1-tetA}}$, and P$_{\text{lacO3O1-lacO3O1}}$ is regulated by the repressor LacI and the inducer isopropyl β-D-1-thiogalactopyranoside (IPTG). Meanwhile, the activity of P$_{\text{BAD}}$ is regulated by the repressor AraC and the inducer L-arabinose. Finally, the activity of P$_{\text{lacO3O1-BAD}}$ is regulated by both repressors (LacI and AraC) and both inducers (IPTG and L-arabinose).

Cells were grown overnight in lysogeny broth (LB) medium supplemented with appropriate antibiotics (34 µg/ml of chloramphenicol, 50 µg/ml of ampicillin, and 50 µg/ml of kanamycin) with shaking at 250 rpm. We made subcultures, by diluting the stationary-phase culture into fresh M9 medium supplemented with glycerol (0.4% final concentration) and the appropriate antibiotics. Cells were left in the incubator until reaching OD$_{600}$ of about 0.25. For the pZA25-P$_{\text{ara}}$-MS2-GFP reporter plasmid activation, 0.4% of L-arabinose was added to the culture, which was then incubated at 37 °C for 60 minutes. Cells containing the pZS12-P$_{\text{lac}}$-MS2-GFP reporter plasmid were incubated in the same way and were activated with 1 mM IPTG. Next, for the activation of P$_{\text{lacO3O1}}$, P$_{\text{lacO3O1-tetA}}$, and P$_{\text{lacO3O1-lacO3O1}}$ target plasmids, specific concentrations of IPTG (either 5 µM or 1 mM) were added to the culture. For activating the P$_{\text{BAD}}$ or P$_{\text{lacO3O1-BAD}}$ target plasmids, 0.1% of L-arabinose was added. For the latter, similar concentrations of IPTG (5 µM or 1 mM) were added as well. Inducer-activated cells were then left in the incubator for 90 minutes, prior to microscopy observation.

C. Microscopy and Image analysis

Cells were visualized using a Nikon Eclipse (Ti-E, Nikon) inverted microscope equipped with a 100× Apo

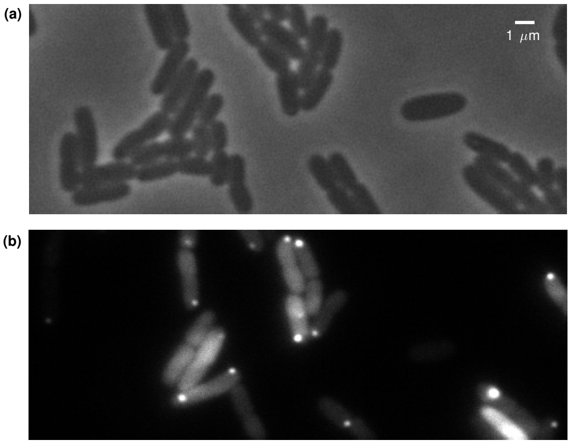

FIG. 1. Example images of live *E. coli* expressing GFP-tagged RNAs. (A) Phase contrast image of the live *E. coli* with the $P_{\mathrm{lacO3O1\text{-}tetA}}$ construct taken after 1 hour of induction with 1 mM IPTG induction at 37 °C. (B) HILO image visualizing the abundant GFP inside the same *E. coli* cells and the target RNA bound by an array of GFPs appearing as bright spots.

TIRF (1.49 NA, oil) objective. Cells and fluorescent spots within were imaged by Highly Inclined and Laminated Optical sheet (HILO) microscopy, using an EMCCD camera (iXon3 897, Andor Technology), a 488 nm argon laser (Melles-Griot), and an emission filter (HQ514/30, Nikon). Phase-contrast images were acquired by a CCD camera (DS-Fi2, Nikon). The software for image acquisition was NIS-Elements (Nikon, Japan). An example of each channel is shown in Fig 1.

We performed time-lapse fluorescence and phase-contrast imaging of the cells (the latter for cell segmentation and lineage construction). For this, 8 µl of cells were placed on a microscope slide between a coverslip and a M9 glycerol agarose gel pad. During image acquisition, cells were constantly supplied with fresh media containing IPTG and L-arabinose, at the same concentration as when in liquid culture, by a micro-perfusion peristaltic pump (Bioptechs) at 0.3 ml/minute. Images were captured for 5 hours, once per minute in the case of fluorescence and once per 5 minutes in the case of phase-contrast. During image acquisition, cells were kept in a temperature-controlled chamber (FCS2, Bioptechs) at optimal temperature (37 °C).

Time series microscopy images were processed as in [34] by, first, aligning consecutive images so as to maximize the cross-correlation of fluorescence intensities. Next, we annotated manually the region occupied by each cell in the time series. Afterwards, the location, dimension, and orientation of each cell in each frame is obtained by principal component analysis, assuming that fluorescence inside the cell is uniform [35]. Cell lineages were then extracted using CellAging, based on overlapping areas in consecutive frames [35]. Next, the intensity of each cell is fit with a surface (quadratic polynomial of the distance from the cell border) in least-deviations sense [36]. This surface represents the cellular background intensity which is subtracted to obtain the foreground intensity. Next, the foreground intensity is fit with a set of Gaussian surfaces, in least-deviations sense, with decreasing heights until the heights are in the 99% confidence interval of the background noise (estimated assuming a normal distribution and using median absolute deviation) [36]. The Gaussians represent fluorescent RNA spots, and the volume under each represent the total spot intensity. Finally, as MS2-GFP-tagged RNA lifetimes are much longer than cell division times [37], the cellular foreground intensity will be an increasing curve, with each jump corresponding to the appearance of a novel tagged RNA. The moments when a jump occurs are estimated using a specialized curve fitting algorithm [18]. The intervals between jumps in individual cells correspond to time intervals between consecutive RNA production events.

III. RESULTS AND DISCUSSION

A. Analytical distributions of production time intervals

From the perspective of the production kinetics of X alone, the reaction system of Eq (1) is equivalent to:

$$I_2 \underset{k_2}{\overset{k_4}{\rightleftharpoons}} P_0 \xrightarrow{k_1} I_1 \xrightarrow{k_3} P_0 + X \tag{4}$$

which is potentially a highly noisy process [20, 38]. While the expression of gene 1 might not be noisy on its own, its expression is perturbed by the transcription machinery occupying the shared promoter region for expression of gene 2, introducing (random) temporal gaps in the expression.

Let $\mathcal{G}(\cdot)$ denote the distribution of consecutive productions of X in Eq (4). The mean and variance of the time intervals between the productions of X are given by [20]:

$$\begin{aligned} \tau_X &\sim \mathcal{G}(k_4, k_2, k_1, k_3) \\ \mathbb{E}[\tau_X] &= \left(1 + \tfrac{k_2}{k_4}\right) {k_1}^{-1} + {k_3}^{-1} \\ \mathbb{V}\mathrm{ar}[\tau_X] &= \left(\left(1 + \tfrac{k_2}{k_4}\right)^2 + 2\,\tfrac{k_1}{k_4}\,\tfrac{k_2}{k_4}\right) {k_1}^{-2} + {k_3}^{-2} \end{aligned} \tag{5}$$

while, due to the symmetry of the model, the production intervals of Y are $\tau_Y \sim \mathcal{G}(k_3, k_1, k_2, k_4)$.

By comparing Eq (2) with Eq (5) we find that, regardless of the parameters, in a bidirectional configuration, the mean and variance of the time intervals between RNA productions of each gene are increased. Consequently, while ${\tau_X}^{(1)}$ is always sub-Poissonian, τ_X can exhibit either sub- or super-Poissonian behavior.

RNA production according to the model is exemplified in Fig 2B, and the expected interval distribution in Fig 2C. While the production intervals of each gene are often somewhat regular, as indicated by the bulk of the

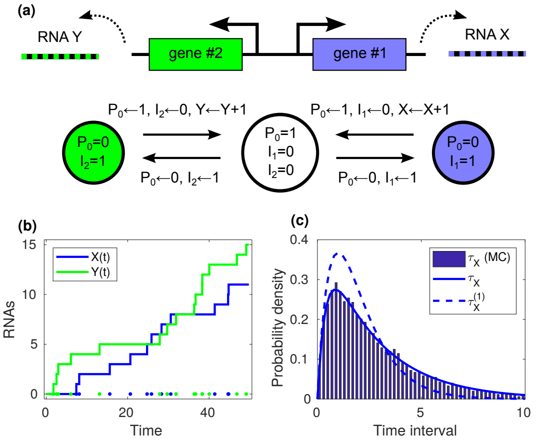

FIG. 2. Model schematic and simulated examples. (A) Schematic of gene-pair in a head-to-head configuration: genes 1 and 2 produce RNAs X and Y, respectively. The shared promoter can be in three-states: free, or occupied for transcription initiation of gene 1 or 2. (B) Produced RNA numbers over time in a single Monte Carlo simulation. The dots denote the moments when RNAs were produced. (C) Distribution of intervals between consecutive productions of X in 10,000 simulations. The parameter values are $(k_1, k_2, k_3, k_4) = (1, 1, 1, 1)$. Here, τ_X and $\tau_X{}^{(1)}$ have a mean (variance) of 3 (7) and 2 (2), respectively.

distribution, large outliers are present due to the temporal gaps, which coincide with the transcriptional activity of the other gene (cf. Fig 2B).

As the marginals shown in Eq (5) fail to capture the co-expression of the two genes, further analysis is necessary. The time between consecutive productions by either gene, i.e. a jump in $X(t)+Y(t)$, is (detailed in the supplement):

$$\begin{aligned}
\tau_{X+Y} &\sim \mathcal{E}(k_1+k_2) + \mathcal{E}^+\left(\tfrac{k_1}{k_1+k_2}, k_3, k_4\right) \\
\mathbb{E}[\,\tau_{X+Y}\,] &= \left(1+\tfrac{k_1}{k_3}+\tfrac{k_2}{k_4}\right)(k_1+k_2)^{-1} \qquad (6)\\
\mathbb{V}\mathrm{ar}[\,\tau_{X+Y}\,] &= \left(1+\left(\tfrac{k_1}{k_3}-\tfrac{k_2}{k_4}\right)^2 + 2\,\tfrac{k_1}{k_3}\,\tfrac{k_2}{k_3} + 2\,\tfrac{k_1}{k_4}\,\tfrac{k_2}{k_4}\right)(k_1+k_2)^{-2}
\end{aligned}$$

where $\mathcal{E}(\lambda)$ is an exponential distribution with rate λ and $\mathcal{E}^+(p_1, \cdots, p_{n-1}, \lambda_1, \cdots, \lambda_n)$ is a hyperexponential distribution with mixing probabilities $p_1, \cdots, p_n$ and rates $\lambda_1, \cdots, \lambda_n$. Again, this distribution can feature either sub- or super-poissonian behavior, depending on its parameter values. By combining Eq (3), (5), and (6), one can determine the asymptotic covariance and the (Pearson) correlation ρ_{XY} between the produced RNA numbers $X(t)$ and $Y(t)$ (detailed in the supplement).

B. Noise and correlation in the transcription kinetics of genes in a head-to-head configuration

TABLE I. Noise and correlation in RNA production kinetics in the different regions of the parameter space of head-to-head configuration. Here, $\sim 1^-$ ($\sim 1^+$) denotes weakly sub- (super-) Poissonian behavior (noise of about 1), while ~ 1 denotes that both behaviors are possible. Finally, $< 1^*$ indicates that < 1 holds at least for one of the genes, possibly for both.

Region	Condition	Noise η_X	Noise η_Y	Corr. ρ_{XY}
A	$q_{24} > 1,\ q_{24} > q_{13}$	> 1	$\sim 1^-$	> 0
A	$q_{13} > q_{24},\ q_{13} > 1$	$\sim 1^-$	> 1	> 0
B	$q_{13},\ q_{24} < 1$	~ 1	~ 1	$\sim 0^-$
C	$q_{13} \sim q_{24} > 1$	$< 1^*$	$< 1^*$	< 0
D	$q_{13} \sim 1,\ q_{24} < 1$	< 1	$\sim 1^+$	< 0
D	$q_{13} < 1,\ q_{24} \sim 1$	$\sim 1^+$	< 1	< 0
E	$q_{13} \sim q_{24} \sim 1$	$< 1^*$	$< 1^*$	< 0

Based on the above, we first analyzed how the noise and correlation in the transcription kinetics of a head-to-head configuration depends on the dynamics of the individual genes. For this, the parameterization λ, q_{12}, q_{13}, q_{24} was found to be insightful. Here, $\lambda \doteq \mu_{X+Y}$ is a timescale parameter (mean total production rate) and $q_{ij} \doteq k_i \,/\, k_j$ denote ratios of rates of two reactions. Further, q_{12} controls the bias, i.e. the expression ratio of each gene: for large (small) q_{12}, gene 1 (gene 2) is expressed more frequently. Finally, q_{13} and q_{24} control the relative durations of closed and open complex formation, which equal $1 \,/\, (1+q_{13})$ and $q_{13} \,/\, (1+q_{13})$, respectively, for gene 1. Specifically, if $q_{13} > 1$ ($q_{13} < 1$), then $k_1 > k_3$ and the gene is limited at the open (closed) complex formation.

The RNA number means are controlled by the bias and the scale: $\mu_X = \lambda^{-1}\, q_{12} \,/\, (1+q_{12})$ and $\mu_Y = \lambda^{-1} \,/\, (1+q_{12})$. As such, the stage at which the transcription kinetics of each gene is rate-limited does not affect the mean number of produced RNAs. Meanwhile, the noise and correlation exhibit complex behavior, which can be divided into a few regions. The regions and their properties are shown in Table I (and Table S-I). The noise of each gene and the correlation coefficient are shown in Fig 3 and Fig 4A, and their analytical forms in the supplement.

Region A: For $q_{24} > 1$, $q_{24} > q_{13}$, the expression of gene 2 is most limited at the open complex formation, while that of gene 1 is more symmetric. As such, the promoter region is mostly occupied, and gene 1 must express either fast or rarely. In the former case, there is a burst of production of proteins X after each Y, so the expression of the two genes is positively correlated, and while gene 2 is Poissonian, solely controlled by its open complex formation process, gene 1 is highly noisy as the geometric burst of RNA is separated by the gaps created by the other gene. In the latter case, the expression of gene 1 is controlled by uniform random productions and the correlation vanishes. Specifically, in the latter case, the noise of gene 1 is $1+2\,q_{12}$ (super-Poissonian) and gene 2 is Poissonian. The correlation for large q_{12} is $\sqrt{1/2}$, which is maximal for the configuration, while for small q_{12} the correlation vanishes. The part $q_{13} > q_{24}$, $q_{13} > 1$

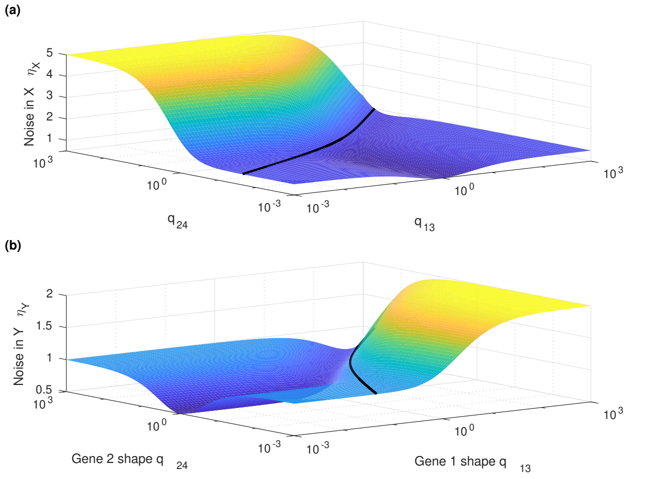

FIG. 3. Noise in the RNA production in a head-to-head configuration as a function of their relative durations of closed and open complex formations. (A) gene 1 and (B) gene 2. The black curves denote unity, and $q_{12}=2$.

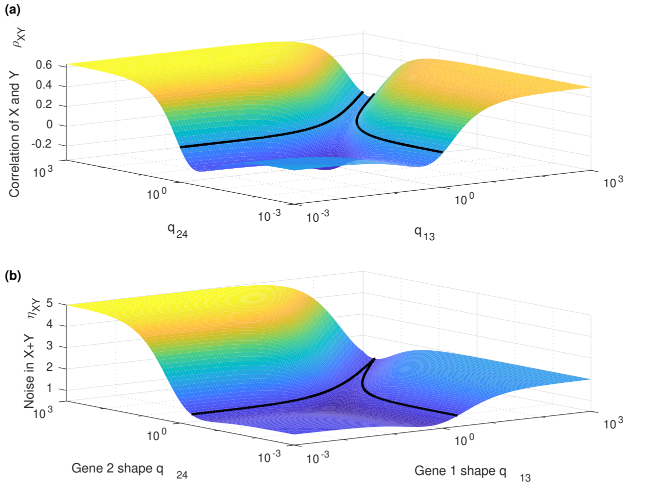

FIG. 4. Correlation and total noise (tandem configuration) as a function of the relative durations of closed and open complex formation. (A) Correlation between the RNA production kinetics of two head-to-head genes. (B) Noise of RNA production of a gene with two initiation sites. The black curves denote zero or unity, and $q_{12}=2$.

is symmetric. Note that the bias q_{12} controls the upper bounds for noise and correlation.

Region B: Here, both genes are limited at the closed complex formation. Thus, the promoter region is rarely occupied, as the expression is limited by an RNAP finding the gene and initiating transcription. This causes the expression of both genes to be Poissonian, as each is limited by a single step, and uncorrelated, as their activities do not interfere at the promoter region.

Region C: Both genes are limited at the open complex formation, which makes them to alternate in occupying the shared promoter region. The noise is set by the bias q_{12}, which determines the gene more disturbed by the activity of the other. Specifically, the noise of gene 1 equals $1/2+q_{12}\,/\,2$ and the noise of gene 2 equals $1/2+{q_{12}}^{-1}\,/\,2$. As the genes inhibit each other by competing for the shared promoter region, the expression patterns are anticorrelated.

Region D: For $q_{13}\sim 1$, $q_{24}<1$, gene 2 is limited during the closed complex formation, so it does not block the shared promoter area. Meanwhile, gene 1 is limited at both stages, making its RNA production to be sub-Poissonian. The expression of gene 2, originally Poissonian, becomes affected by periods of inactivity as gene 1 employs the promoter, increasing the noise, as controlled by the bias, yielding noise of $1{+}{q_{12}}^{-1}\,/\,2$. The correlation is negative, as gene 1 inhibits the expression of gene 2. The part $q_{13}<1$, $q_{24}\sim 1$ is symmetric.

Region E: Both genes have similar closed and open complex formation durations, resulting in low noise in a non-bidirectional configuration. If their closed complex formation durations are similar (i.e. $q_{12}\sim 1$), both genes are of low noise ($\sim 7/9$) and their expression is anticorrelated ($\sim -2/7$), as they alternate in activity. Otherwise, one is of low noise ($\sim 5/9$), unaffected by the configuration, while the other is of high noise, with its expression being disturbed by the frequent gaps caused by the other. Specifically, the noise is $5/9+2\,q_{12}\,/\,9$ for gene 1 and $5/9+2\,{q_{12}}^{-1}\,/\,9$ for gene 2. The correlation is negative, with a maximum of $-2/7$ at $q_{12}=1$, and minima of $-1/\sqrt{10}$ at $q_{12}\to 0$ and $q_{12}\to\infty$.

In summary, for coupled gene activity, one (or both) genes must not be limited at the closed complex formation alone. When coupled, both genes are low noise only if both feature similar relative closed-to-open complex durations. In this case, their expression is likely anticorrelated. If the relative closed-to-open complex durations differ, one is of high noise and the other of low noise, while their expression is, surprisingly, positively correlated. While our analysis lacks processes other than transcription initiation, the presence of e.g. first-order degradation pulls the noise toward unity and the correlation toward zero, leaving the conclusions qualitative useful.

C. Noise in a gene with two initiation sites: model predictions and empirical validation

Next, we investigate the dynamics of a gene controlled by a promoter with two TSSs (cf. Fig S1C). This is common in *E. coli* [39] and more so in, e.g. plant mitochondria [40]. The configuration is readily accommodated by our model, by considering the dynamics of $X+Y$. As the mean and variance of $X+Y$ follow the mean and covariance of (X,Y), the results can be derived from those obtained in the previous section.

Fig 4B shows the noise for $X+Y$, representing the RNAs produced through either TSS. The noise is low only if both TSSs exhibit production dynamics with low noise, i.e. in the regions C, D, or E. Compared to individual TSSs, the RNA number fluctuations are lower, being suppressed by the negative correlation. If one TSS exhibits highly noisy production (region A), the RNA numbers become highly noisy, regardless of the dynamics of the other TSS. Finally, in region B, the production is exponential-like, as multiple TSSs only increase the RNAP to promoter binding affinity, which makes their dynamics indiscernible from that of a single TSS.

To validate our predictions, we observed transcription in live *E. coli* at the single-RNA level in various constructs. Three of the constructs feature synthetic genes whose production is controlled by a single promoter (specifically P_{lacO3O1}, P_{tetA}, and P_{BAD} (cf. Fig S1A). The other constructs feature pairs of genes sharing promoter elements. One of these constructs is $P_{\text{lacO3O1-lacO3O1}}$, with overlapping lacO3O1 promoters in the opposite strands (cf. Fig S1B), with the reporter being on a single side. In the other two such constructs, the expression is controlled by a $P_{\text{lacO3O1-tetA}}$ or a $P_{\text{lacO3O1-BAD}}$ dual-tandem promoters (cf. Fig S1C). In all these, the expression of the lacO3O1 promoter is modulated by the IPTG concentration, an inducer for the lac promoter[41]. Meanwhile, aTc concentration is held constant at 15 ng/ml , in order to trigger full expression of the tetA promoter. Similarly, L-arabinose concentration is held constant at 0.1% , to trigger full expression of the BAD promoter. In all cases, RNA production dynamics was measured by time-lapse microscopy imaging using MS2-GFP tagging (Methods).

Using our models, we aim to predict the behavior of the pairs of genes sharing promoter elements, given knowledge of the behavior of the constituent genes when not sharing such elements. I.e., we test whether, from the measured dynamics of RNA production of P_{lacO3O1}, P_{tetA} [17], and P_{BAD}, one can predict the kinetics of $P_{\text{lacO3O1-lacO3O1}}$, $P_{\text{lacO3O1-tetA}}$, and $P_{\text{lacO3O1-BAD}}$.

For this, we first extracted the number of RNAs in each cell in the first and the last frame of the time series for all the constructs in each condition. These data were used to estimate the mean and standard deviation the production intervals, and the most likely (maximum likelihood fit) model of Eq (S1) for the single promoters. The estimated intervals are shown in Table II, along with the model parameters of Eq (S1) where applicable. A Wald test testing for a specific mean and standard deviation was used to compute a p-value to confirm that the model predicts the mean and variance of the RNA distributions.

The results in Table II indicate that changing IPTG concentration alters the noise of the lacO3O1 promoter in addition to changing its mean expression rate, which is expected to be due to changes in the open-to-closed complex duration ratio, in agreement with previous reports [19]. The p-values indicate that there is no evidence that any of the models fit the measurements poorly. We also extracted the intervals from the full time series for several of the cases (about 120 frames, one every second) to verify that they are correctly estimated (see Table S-II).

Next, using the above parameters, we constructed the models for the dual promoters. The obtained models are shown in Table III. The mean and standard deviation show an agreement with the empirical data, while the noise and correlation indicate that that promoters operate at different regions of the open-to-closed complex ratio space. The results indicate that the model predicts the behavior of the dual-promoter measurements well, and that the noise is modulated by the change in the coordination between the two promoters in the dual promoter construct.

As our methodology cannot identify which of the steps correspond to ${k_1}^{-1}$ and ${k_3}^{-1}$ in Table II, we also considered the alternative step ordering. The dual-promoter model fits had a p-values $< 3.964 \times 10^{-3}$ in for the lacO3O1-tetA construct at 5 µM IPTG, and p-values $< 4.614 \times 10^{-3}$ for the lacO3O1-BAD construct at 1000 µM IPTG, indicating that the alternatives are not likely for lacO3O1 at 5 µM and BAD. The step order for lacO3O1 at 1000 µM IPTG and tetA cannot be resolved from these data, but the alternatives result in a qualitatively similar dual-promoter models and p-values > 0.116. For the constructs containing these two promoters, we report the most likely models, all suggesting the order specified in Table II. These findings are also supported by prior evidence using a different methodology[19].

The fact that the measurements fall into the different regions of operation (see Fig 4 and Table I) is apparent in Fig S2, Fig S3, and Fig S4. Namely, the high IPTG condition falls into region E for the lacO3O1-lacO3O1 and lacO3O1-tetA, and into region D for the lacO3O1-BAD construct. At low IPTG, the lacO3O1-lacO3O1 transits into region C, as both promoters are modulated by the changes in the inducer concentration, while the lacO3O1-tetA and lacO3O1-BAD transit into (opposite directions) of region A. This explains the widely different noise levels in the measured intervals, which are well predicted by our models in each case.

Finally, we attempted to predict the mean, noise, and intervals in a dual-promoter measurement assuming that there were no interactions between the two promoters. The results in Table IV show that the associated model fails to explain the observed dual-promoter behavior. Note that the models are also unaffected by the (k_1, k_3) identifiability problem. While the mean and noise of the system consisting of two independent promoters trivially follow from their independent components, the time intervals of the combined production do not. In particular, the intervals are not independent. We also considered the possibility that while the promoters might have interactions, their expression levels may be altered by the other promoter utilizing the same finite pool of RNA polymerases. For this, we assumed that the number of RNA polymerases modulate the closed complex formation rate (i.e. $k_1 = R\,\tilde{k}_1$ where $\tilde{k}_1$ is the per-polymerase closed complex formation rate, and R represents an RNA polymerase), which will cause a slight reduction of the closed complex formation rate, as determined by the closed to open-to-closed complex duration

TABLE II. Estimated RNA production intervals for each of the promoter constructs. The table shows the promoter, induction, estimated paramater of model Eq (S1) for the single promoters, the estimated mean, standard deviation (sd), and noise (coefficient of variation) of the RNA production intervals, and the p-value of the test of model versus data.

Promoter	IPTG (µM)	${k_1}^{-1}, {k_2}^{-1}$ (s)	${k_3}^{-1}, {k_4}^{-1}$ (s)	Mean (s)	Sd (s)	Noise	P-value
lacO3O1	1000	362.3	737.3	1099.6	821.5	0.558	0.446
lacO3O1	5	25.8	1236.8	1262.6	1237.0	0.960	0.273
tetA	-	287.4	385.5	672.9	480.9	0.511	0.075
BAD	-	1036.7	333.7	1370.4	1089.1	0.632	0.059
lacO3O1-tetA	1000	-	-	702.2	638.8	0.828	0.604
lacO3O1-tetA	5	-	-	1111.6	1089.1	0.960	0.164
lacO3O1-lacO3O1	1000	-	-	1659.3	1437.0	0.750	0.112
lacO3O1-lacO3O1	5	-	-	2205.9	2119.3	0.923	0.971
lacO3O1-BAD	1000	-	-	866.8	612.9	0.500	0.099
lacO3O1-BAD	5	-	-	1274.7	1248.9	0.960	0.698

TABLE III. Models derived for the dual promoters from the individual promoter fits of Table II using the model with interactions during transcription initiation. The table shows the promoter/induction scheme, the mean and standard deviation (sd) of the RNA production intervals and the correlation between the RNA numbers assuming the derived models, and the p-value of the test of model versus data.

Promoter	IPTG (µM)	Mean (s)	Sd (s)	Noise	Correlation	P-value
lacO3O1-lacO3O1	1000	1836.9	1685.2	0.842	−0.188	0.168
lacO3O1-lacO3O1	5	2499.3	2486.5	0.990	−0.010	0.931
lacO3O1-tetA	1000	701.4	616.1	0.772	−0.166	0.603
lacO3O1-tetA	5	1190.4	1212.9	1.038	+0.172	0.123
lacO3O1-BAD	1000	901.3	731.4	0.659	−0.070	0.141
lacO3O1-BAD	5	1240.0	1230.9	0.985	+0.105	0.620

ratio of the other promoter. Any of these models (all R and all step orders) failed to explain the behavior of our dual-promoter measurements as well. The effects are most extreme for $R = 2$, but we verified that models for other R have no better fit. Our model is recovered at $R = 1$ and the independent model without polymerases is recovered at $R = \infty$.

We conclude that our model of closely-spaced promoters that assumes interactions between the promoters is the one that well predicts the measurements in each setting, for both the head-to-head and tandem constructs. Relevantly, our models reveal that the observed changes arise from changes in the coordination between the two coupled transcription start sites of our synthetic constructs.

IV. CONCLUSION

We analyzed a stochastic model of two genes in a head-to-head configuration as a function of whether each gene is rate-limited during the closed and/or open complex formation. Compared to individual genes, in the bidirectional configuration, the transcription activity is slower and noisier in both genes, as each gene interferes with the activity of the other, allowing two genes with sub-Poissonian dynamics to exhibit super-Poissonian dynamics when coupled. Importantly, provided information on the kinetics of the constituent promoters when not sharing promoter elements, the models were shown to be able to predict well the behavior of the pairs of the same genes when sharing promoter elements, implying that they capture accurately the effects of the complex interference caused by the sharing of elements.

We found that for such prediction to be accurate, the models have to account for the two-rate-limiting step kinetics of active transcription in *E. coli*. In particular, the time-length of such rate limiting steps, namely the closed and open complex formations, controls not only the expression rate and noise of each gene (as in isolated genes, see e.g. [15]), but also the kinetics of the temporal gaps caused by the transcription events of the opposite gene. This programs the behavior intricately: a similar rate-limiting step structure combined with a rate-limiting open complex formation is required for both genes to feature low noise; otherwise one tends to be highly noisy. Also, orderly systems tend to exhibit strong negative correlation, while the genes alternate expression, but the

TABLE IV. Null models derived for the dual independent promoters from the individual promoter fits of Table II. The table shows the promoter/induction scheme, and the mean and standard deviation (sd) of the RNA production intervals assuming the null models, and the p-value of the test of model versus data for maximally ($R = 2$) and minimally ($R = \infty$) RNA polymerase starved null models.

Promoter	IPTG (µM)	Mean (s)	Sd (s)	Noise	Max. p-value ($R = 2$)	P-value ($R = \infty$)
lacO3O1-lacO3O1	1000	1099.6	821.5	0.558	4.116×10^{-4}	6.601×10^{-5}
lacO3O1-lacO3O1	5	1262.6	1237.0	0.960	9.981×10^{-3}	8.386×10^{-3}
lacO3O1-tetA	1000	417.5	303.5	0.529	2.127×10^{-3}	4.267×10^{-4}
lacO3O1-tetA	5	439.0	358.5	0.667	7.289×10^{-5}	5.049×10^{-6}
lacO3O1-BAD	1000	610.1	468.9	0.591	6.325×10^{-6}	1.456×10^{-7}
lacO3O1-BAD	5	657.1	588.7	0.803	3.464×10^{-3}	5.124×10^{-4}

correlation can be lost or become positive if the open-to-closed complex formation time-lengths are incompatible. As such, not only the mean and variance of the durations of each stage, but also the mechanistic underpinnings, affect the dynamics of closely-spaced gene-pairs, implying that promoters with seemingly identical dynamics in isolation may differ widely in their dynamics in a closely-spaced configuration. Relevantly, as shown, the results generalize to the behavior of individual genes with multiple transcription initiation sites.

Overall, these results suggest that, in *E. coli*, the kinetics of the rate-limiting steps in active transcription needs to be considered for dissecting the dynamics of pairs of genes sharing promoter elements. In this regard, we find it to be striking that pairs of closely-spaced promoters, by tuning the kinetics of their closed and open complex formation (which are sequence dependent and, thus, evolvable) tunes the orderliness of the whole gene-pair. This new knowledge provides an important route to follow in the engineering of pairs of closely-spaced promoters with desired dynamics and contributes to a better understanding of the dynamics of natural pairs of closely spaced genes and their potential role in the gene expression programs of *E. coli*.

ACKNOWLEDGMENTS

The work was supported by the Alfred Kordelin Foundation (AH), the Vilho, Yrjo and Kalle Vaisala Foundation (SMDO), the Tampere University of Technology President's Graduate Programme (RN-V), the Jane and Aatos Erkko Foundation (610536; ASR), and Academy of Finland (295027 and 305342; ASR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare that they have no conflict of interest.

[1] N. Adachi and M. R. Lieber, Bidirectional gene organization: a common architectural feature of the human genome, Cell **109**, 807 (2002).

[2] N. D. Trinklein, S. Force Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers, An abundance of bidirectional promoters in the human genome, Genome Res. **14**, 62 (2004).

[3] M. Herbert, A. Kolb, and H. Buc, Overlapping promoters and their control in *Escherichia coli*: The *gal* case, Proc. Natl. Acad. Sci. U.S.A **83**, 2807 (1986).

[4] K. Moss Bendtsen, J. Erdossy, Z. Csiszovszki, S. Lo Svenningsen, K. Sneppen, S. Krishna, and S. Semsey, Direct and indirect effects in the regulation of overlapping promoters, Nucl. Acids Res. **39**, 6879 (2011).

[5] J. O. Korbel, L. J. Jensen, C. von Mering, and P. Bork, Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs, Nat. Biotechnol. **22**, 911 (2004).

[6] E. M. Prescott and N. J. Proudfoot, Transcriptional collision between convergent genes in budding yeast, Proc. Natl. Acad. Sci. U.S.A. **99**, 8796 (2002).

[7] B. P. Callen, K. E. Shearwin, and J. B. Egan, Transcriptional interference between convergent promoters caused by elongation over the promoter, Mol. Cell **14**, 647 (2004).

[8] A. C. Palmer, A. Ahlgren-Berg, J. B. Egan, I. B. Dodd, and K. E. Shearwin, Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter, Mol. Cell **34**, 545 (2009).

[9] A. Hakkinen, S. Healy, H. T. Jacobs, and A. S. Ribeiro, Genome wide study of NF-Y type CCAAT boxes in unidirectional and bidirectional promoters in human and mouse, J. Theor. Biol. **281**, 74 (2011).

[10] E. Zanotto, A. Hakkinen, G. Teku, B. Shen, A. S. Ribeiro, and H. T. Jacobs, NF-Y influences directionality of transcription from the bidirectional *Mrps12/Sarsm* promoter in both mouse and human cells, BBA Gene Regul. Mech. **1789**, 432 (2009).

[11] L. Martins, J. Makela, A. Hakkinen, M. Kandhavelu, O. Yli-Harja, J. M. Fonseca, and A. S. Ribeiro, Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time, J. Theor. Biol. **301**,

83 (2012).

[12] K. Sneppen, I. B. Dodd, K. E. Shearwin, A. C. Palmer, R. A. Schubert, B. P. Callen, and J. B. Egan, A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*, J. Mol. Biol. **346**, 399 (2005).

[13] C. Yan, S. Wu, C. Pocetti, and L. Bai, Regulation of cell-to-cell variability in divergent gene expression, Nat. Commun. **7**, 11099 (2015).

[14] W. R. McClure, Rate-limiting steps in RNA chain initiation, Proc. Natl. Acad. Sci. U.S.A. **77**, 5634 (1980).

[15] J. Lloyd-Price, S. Startceva, V. Kandavalli, J. Chandraseelan, N. Goncalves, S. M. D. Oliveira, A. Hakkinen, and A. S. Ribeiro, Dissecting the stochastic transcription initiation process in live *Escherichia coli*, DNA Res. **23**, 203 (2016).

[16] A. Revyakin, R. H. Ebright, and T. R. Strick, Promoter unwinding and promoter clearance by RNA polymerase: Detection by single-molecule DNA nanomanipulation, Proc. Natl. Acad. Sci. U.S.A. **101**, 4776 (2004).

[17] A.-B. Muthukrishnan, M. Kandhavelu, J. Lloyd-Price, F. Kudasov, S. Chowdhury, O. Yli-Harja, and A. S. Ribeiro, Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells, Nucl. Acids Res. **40**, 8472 (2012).

[18] A. Hakkinen and A. S. Ribeiro, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, Bioinformatics **31**, 69 (2015).

[19] V. K. Kandavalli, H. Tran, and A. S. Ribeiro, Effects of σfactor competition are promoter initiation kinetics dependent, BBA Gene Regul. Mech. **1859**, 1281 (2016).

[20] A. Hakkinen and A. S. Ribeiro, Characterizing rate limiting steps in transcription from RNA production times in live cells, Bioinformatics **32**, 1346 (2016).

[21] I. O. Vvedenskaya, H. Vahedian-Movahed, Y. Zhang, D. M. Taylor, R. H. Ebright, and B. E. Nickels, Interactions between RNA polymerase and the core recognition element are a determinant of transcription start site selection, Proc. Natl. Acad. Sci. U.S.A. **113**, E2899 (2016).

[22] M. T. Record, Jr., W. S. Reznikoff, M. L. Craig, K. L. McQuade, and P. J. Schlax, *Escherichia coli* RNA polymerase (eσ^{70}), promoters, and the kinetics of the steps of transcription initiation, in *Escherichia coli and Salmonella typhimurium: Cellular and molecular biology*, edited by F. C. Neidhart, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ASM Press, Washington, DC, 1996) 2nd ed., pp. 792–820.

[23] F. Wang and E. C. Greene, 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 (2011).

[24] M. Patrick, P. P. Dennis, M. Ehrenberg, and H. Bremer, Free RNA polymerase in *Escherichia coli*, Biochimie **119**, 80 (2015).

[25] D. A. McQuarrie, Stochastic approach to chemical kinetics, J. Appl. Probab. **4**, 413 (1967).

[26] D. T. Gillespie, Stochastic simulation of chemical kinetics, Annu. Rev. Phys. Chem. **58**, 35 (2009).

[27] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, Stochasticity in gene expression: From theories to phenotypes, Nat. Rev. Genet. **6**, 451 (2005).

[28] D. R. Cox, *Renewal theory* (Methuen, London, UK, 1962).

[29] J. M. Pedraza and J. Paulsson, Effects of molecular memory and bursting on fluctuations in gene expression, Science **319**, 339 (2008).

[30] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection, Mol. Syst. Biol. **2**, 1 (2006).

[31] C. B. Glascock and M. J. Weickert, Using chromosomal *lacI*Q1 to control expression of genes on high-copy-number plasmids in *Escherichia coli*, Gene **223**, 221 (1998).

[32] K. Nevo-Dinur, A. Nussbaum-Shochat, S. Ben-Yehuda, and O. Amster-Choder, Translation-independent localization of mRNA in *E. coli*, Science **331**, 1081 (2011).

[33] T. T. Le, S. Harlepp, C. C. Guet, K. Dittmar, T. Emonet, T. Pan, and P. Cluzel, Real-time rna profiling within a single bacterium, Proc. Natl. Acad. Sci. U.S.A. **102**, 9160 (2005).

[34] S. M. D. Oliveira, A. Hakkinen, J. Lloyd-Price, H. Tran, V. Kandavalli, and A. S. Ribeiro, Temperature-dependent model of multi-step transcription initiation in *Escherichia coli* based on live single-cell measurements, PLoS Comput. Biol. **12**, e1005174 (2016).

[35] A. Hakkinen, A.-B. Muthukrishnan, A. Mora, J. M. Fonseca, and A. S. Ribeiro, CellAging: a tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*, Bioinformatics **29**, 1708 (2013).

[36] A. Hakkinen, M. Kandhavelu, S. Garasto, and A. S. Ribeiro, Estimation of fluorescence-tagged RNA numbers from spot intensities, Bioinformatics **30**, 1146 (2014).

[37] I. Golding and E. C. Cox, Rna dynamics in live *Escherichia coli* cells, Proc. Natl. Acad. Sci. U.S.A. **101**, 11310 (2004).

[38] J. Peccoud and B. Ycart, Markovian modeling of gene-product synthesis, Theor. Popul. Biol. **48**, 222 (1995).

[39] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juarez, B. Contreras-Moreira, A. M. Huerta, J. Collado-Vides, and E. Morett, Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*, PLoS ONE **4**, e7526 (2009).

[40] R. L. Tracy and D. B. Stern, Mitochondrial transcription initiation: promoter structures and rna polymerases, Curr. Genet. **28**, 205 (1995).

[41] R. Lutz, T. Lozinski, T. Ellinger, and H. Bujard, Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator, Nucl. Acids Res. **29**, 3873 (2001).