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. Ribeiro1
1BioMediTech 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:

\[
P_0 \xrightarrow{k_1} I_1 \xrightarrow{k_2} P_0 + X
\]

\[
P_0 \xrightarrow{k_1} I_2 \xrightarrow{k_2} P_0 + Y
\]  \hspace{1cm} (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]:

$$\tau_X^{(1)} \sim \mathcal{E}^-(k_1, k_3)$$

with

$$E[\tau_X^{(1)}] = k_1^{-1} + k_3^{-1}$$

$$\Var[\tau_X^{(1)}] = k_1^{-2} + k_3^{-2}$$

(2)

where $\mathcal{E}^-(\lambda_1, \ldots, \lambda_n)$ represents a hypoexponential distribution with rates $\lambda_1, \ldots, \lambda_n$. 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 the coefficient of variation (standard deviation over the mean), as this quantity equals unity for Poissonian production intervals for gene 2 would be $\tau_Y^{(1)} \sim \mathcal{E}^-(k_2, k_4)$.

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]:

$$\mu_Z \equiv \lim_{t \to \infty} \frac{E[Z(t)]}{t} = E[\tau_Z]^{-1}$$

$$\eta_Z \equiv \lim_{t \to \infty} \frac{\Var[Z(t)]}{t} = \Var[\tau_Z] E[\tau_Z]^{-3}$$

(3)

i.e. the mean number of RNAs produced per unit time ($\mu_Z$) equals the inverse interval mean, while the Fano factor (variance over the mean) of the RNA numbers ($\eta_Z$) 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*+ *traB*<sub>T14</sub> Δ*lacZ*<sub>AM16</sub> hsdR<sub>R</sub>514 Δ*araBAD*Δ<sub>H33</sub> Δ*rhaBAD*Δ<sub>L78</sub>) [30], which contains the constitutive promoters $P_{lac+}$ and $P_{araC}$ producing, respectively, LacI repressors [31] and AraC repressors. As this strain does not contain the tet<sup>R</sup> gene responsible for encoding TetR repressors, any gene downstream to a $P_{tetA}$ promoter is expressed constitutively.

We constructed five target systems on a single-copy pBELO plasmid. The first plasmid features the $P_{lacO3O1}$ 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 $P_{lacO3O1}$ promoter being replaced by the following promoters: (i) $P_{BAD}$ promoter; (ii) $P_{lacO3O1-tetA}$ dual-tandem promoter; (iii) $P_{lacO3O1-BAD}$ dual-tandem promoter; and (iv) $P_{lacO3O1-lacO3O1}$ dual-directional promoter. All strains aside from its target system also contain either a medium-copy plasmid pZAS25 with the reporter gene $P_{ara}$-MS2-GFP or a low-copy plasmid pZS12 with the reporter gene $P_{lac}$-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 $P_{lacO3O1}$, $P_{lacO3O1-tetA}$, and $P_{lacO3O1-lacO3O1}$ is regulated by the repressor LacI and the inducer isopropyl-β-D-1-thiogalactopyranoside (IPTG). Meanwhile, the activity of $P_{BAD}$ is regulated by the repressor AraC and the inducer L-arabinose. Finally, the activity of $P_{lacO3O1-BAD}$ 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 glucose (0.4% final concentration) and the appropriate antibiotics. Cells were left in the incubator until reaching OD<sub>600</sub> of about 0.25. For the pZAS25-$P_{ara}$-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-$P_{lac}$-MS2-GFP reporter plasmid were incubated in the same way and were activated with 1 mM IPTG. Next, for the activation of $P_{lacO3O1}$, $P_{lacO3O1-tetA}$, and $P_{lacO3O1-lacO3O1}$ target plasmids, specific concentrations of IPTG (either 5 µM or 1 mM) were added to the culture. For activating the $P_{BAD}$ or $P_{lacO3O1-BAD}$ 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...
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:

$$I_2 \xrightarrow{k_4} P_0 \xrightarrow{k_1} I_1 \xrightarrow{k_3} P_0 + X$$

(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]:

$$\tau_X \sim \mathcal{G}(k_4, k_2, k_1, k_3)$$

$$E[\tau_X] = \left(1 + \frac{k_2}{k_1}\right) k_1^{-1} + k_3^{-1}$$

$$\text{Var}[\tau_X] = \left(\left(1 + \frac{k_2}{k_1}\right)^2 + 2 \frac{k_1}{k_4}\frac{k_2}{k_4}\right) k_1^{-2} + k_3^{-2}$$

(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, $\tau_X$ 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, \(\tau_X\) 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):

\[
\tau_{X+Y} \sim E(k_1 + k_2) + E^+ \left(\frac{k_1}{k_3+k_4}, k_3, k_4\right)
\]

\[
E[\tau_{X+Y}] = \left(1 + \frac{k_1 + k_2}{k_3+k_4}\right) (k_1+k_2)^{-1}
\]

\[
\text{Var}[\tau_{X+Y}] = \left(1 + \left(\frac{k_1 + k_2}{k_3+k_4}\right)^2 + 2\frac{k_1}{k_3}+\frac{k_2}{k_4} + 2\frac{k_1}{k_3}+\frac{k_2}{k_4}\right) (k_1+k_2)^{-2}
\]

where \(E(\lambda)\) is an exponential distribution with rate \(\lambda\) and \(E^+(p_1, \ldots, p_{n-1}, \Lambda_1, \ldots, \Lambda_n)\) is a hyperexponential distribution with mixing probabilities \(p_1, \ldots, p_n\) and rates \(\Lambda_1, \ldots, \Lambda_n\). 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 \(\rho_{XY}\) 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

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 \(\lambda, q_{12}, q_{13}, q_{24}\) was found to be insightful. Here, \(\lambda = \mu_X + \mu_Y\) is a timescale parameter (mean total production rate) and \(q_{ij} = k_i / k_j\) 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 \(1 / (1+q_{24})\), respectively, for gene 1. Specifically, if \(q_{13} > 1 (q_{13} < 1)\), then \(k_3 > k_2\) 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 \(\mu_Y = \lambda^{-1} / (1 + q_{24})\). 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.

<table>
<thead>
<tr>
<th>Region</th>
<th>Condition</th>
<th>Noise (\eta_X)</th>
<th>Noise (\eta_Y)</th>
<th>Corr. (\rho_{XY})</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(q_{24} &gt; 1, q_{24} &gt; q_{13})</td>
<td>(&gt; 1)</td>
<td>(&lt; 1^{*})</td>
<td>(&gt; 0)</td>
</tr>
<tr>
<td>A</td>
<td>(q_{13} &gt; q_{24}, q_{13} &gt; 1)</td>
<td>(&gt; 1)</td>
<td>(&gt; 1)</td>
<td>(&gt; 0)</td>
</tr>
<tr>
<td>B</td>
<td>(q_{13}, q_{24} &lt; 1)</td>
<td>(= 1)</td>
<td>(= 1)</td>
<td>(= 0)</td>
</tr>
<tr>
<td>C</td>
<td>(q_{13} \sim q_{24} &gt; 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 0)</td>
</tr>
<tr>
<td>D</td>
<td>(q_{13} \sim 1, q_{24} &lt; 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 0)</td>
</tr>
<tr>
<td>E</td>
<td>(q_{13} \sim q_{24} \sim 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 1)</td>
<td>(&lt; 0)</td>
</tr>
</tbody>
</table>
Correlation of $X$ and $Y$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Noise in $X + Y$

<table>
<thead>
<tr>
<th>Noise in $X$</th>
<th>Noise in $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.8</td>
</tr>
</tbody>
</table>

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 $5/9 + 2q_{12}/9$ for gene 1 and $5/9 + 2q_{12}^{-1}/9$ 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 in-
individually 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\textsubscript{lacO3O1}, P\textsubscript{tetA}, and P\textsubscript{BAD} (cf. Fig S1A). The other constructs feature pairs of genes sharing promoter elements. One of these constructs is P\textsubscript{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\textsubscript{lacO3O1-tetA} or a P\textsubscript{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\textsubscript{lacO3O1}, P\textsubscript{tetA} [17], and P\textsubscript{BAD}, one can predict the kinetics of P\textsubscript{lacO3O1-lacO3O1}, P\textsubscript{lacO3O1-tetA}, and P\textsubscript{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 lacO301-tetA construct at 5 µM IPTG, and p-values $< 4.614 \times 10^{-3}$ for the lacO301-BAD construct at 1000 µM IPTG, indicating that the alternatives are not likely for lacO301 at 5 µM and BAD. The step order for lacO301 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 lacO301-lacO301 and lacO301-tetA, and into region D for the lacO301-BAD construct. At low IPTG, the lacO301-lacO301 transits into region C, as both promoters are modulated by the changes in the inducer concentration, while the lacO301-tetA and lacO301-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 modulates 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 parameter 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.

<table>
<thead>
<tr>
<th>Promoter</th>
<th>IPTG (µM)</th>
<th>$k_1^{-1}, k_2^{-1}$ (s)</th>
<th>$k_3^{-1}, k_4^{-1}$ (s)</th>
<th>Mean (s)</th>
<th>Sd (s)</th>
<th>Noise</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lacO301</td>
<td>1000</td>
<td>362.3</td>
<td>737.3</td>
<td>1099.6</td>
<td>821.5</td>
<td>0.558</td>
<td>0.446</td>
</tr>
<tr>
<td>lacO301</td>
<td>5</td>
<td>25.8</td>
<td>1236.8</td>
<td>1262.6</td>
<td>1237.0</td>
<td>0.960</td>
<td>0.273</td>
</tr>
<tr>
<td>tetA</td>
<td>-</td>
<td>287.4</td>
<td>385.5</td>
<td>672.9</td>
<td>480.9</td>
<td>0.511</td>
<td>0.075</td>
</tr>
<tr>
<td>BAD</td>
<td>-</td>
<td>1036.7</td>
<td>333.7</td>
<td>1370.4</td>
<td>1089.1</td>
<td>0.632</td>
<td>0.059</td>
</tr>
<tr>
<td>lacO301-tetA</td>
<td>1000</td>
<td>-</td>
<td>-</td>
<td>702.2</td>
<td>638.8</td>
<td>0.828</td>
<td>0.604</td>
</tr>
<tr>
<td>lacO301-tetA</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>1111.6</td>
<td>1089.1</td>
<td>0.960</td>
<td>0.164</td>
</tr>
<tr>
<td>lacO301-lacO301</td>
<td>1000</td>
<td>-</td>
<td>-</td>
<td>1659.3</td>
<td>1437.0</td>
<td>0.750</td>
<td>0.112</td>
</tr>
<tr>
<td>lacO301-lacO301</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>2205.9</td>
<td>2119.3</td>
<td>0.923</td>
<td>0.971</td>
</tr>
<tr>
<td>lacO301-BAD</td>
<td>1000</td>
<td>-</td>
<td>-</td>
<td>866.8</td>
<td>612.9</td>
<td>0.500</td>
<td>0.099</td>
</tr>
<tr>
<td>lacO301-BAD</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>1274.7</td>
<td>1248.9</td>
<td>0.960</td>
<td>0.698</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Promoter</th>
<th>IPTG (µM)</th>
<th>Mean (s)</th>
<th>Sd (s)</th>
<th>Noise</th>
<th>Correlation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lacO301-lacO301</td>
<td>1000</td>
<td>1836.9</td>
<td>1685.2</td>
<td>0.842</td>
<td>-0.188</td>
<td>0.168</td>
</tr>
<tr>
<td>lacO301-lacO301</td>
<td>5</td>
<td>2499.3</td>
<td>2486.5</td>
<td>0.990</td>
<td>-0.010</td>
<td>0.931</td>
</tr>
<tr>
<td>lacO301-tetA</td>
<td>1000</td>
<td>701.4</td>
<td>616.1</td>
<td>0.772</td>
<td>-0.166</td>
<td>0.603</td>
</tr>
<tr>
<td>lacO301-tetA</td>
<td>5</td>
<td>1190.4</td>
<td>1212.9</td>
<td>1.038</td>
<td>+0.172</td>
<td>0.123</td>
</tr>
<tr>
<td>lacO301-BAD</td>
<td>1000</td>
<td>901.3</td>
<td>731.4</td>
<td>0.659</td>
<td>-0.070</td>
<td>0.141</td>
</tr>
<tr>
<td>lacO301-BAD</td>
<td>5</td>
<td>1240.0</td>
<td>1230.9</td>
<td>0.985</td>
<td>+0.105</td>
<td>0.620</td>
</tr>
</tbody>
</table>

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 gap caused by the transcription events of the opposite gene. This programs the behavior intricately: a similar gap caused by the sharing of elements.
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, imply-
ing 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 mul-
tiple transcription initiation sites.

Overall, these results suggest that, in E. coli, the ki-
netics 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 pro-
moters, 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 pro-
moters 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 expres-
sion programs of E. coli.

**ACKNOWLEDGMENTS**

The work was supported by the Alfred Kordelin Found-
dation (AH), the Vilho, Yrjo and Kalle Vaisala Foun-
dation (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, de-
cision to publish, or preparation of the manuscript. The
authors declare that they have no conflict of interest.

**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 = ∞) RNA polymerase
starved null models.

<table>
<thead>
<tr>
<th>Promoter</th>
<th>IPTG (µM)</th>
<th>Mean (s)</th>
<th>Sd (s)</th>
<th>Noise</th>
<th>Max. p-value (R = 2)</th>
<th>P-value (R = ∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lacO3O1-lacO3O1</td>
<td>1000</td>
<td>1099.6</td>
<td>821.5</td>
<td>0.558</td>
<td>4.116 × 10⁻⁴</td>
<td>6.601 × 10⁻⁵</td>
</tr>
<tr>
<td>lacO3O1-lacO3O1</td>
<td>5</td>
<td>1262.6</td>
<td>1237.0</td>
<td>0.960</td>
<td>9.981 × 10⁻³</td>
<td>8.386 × 10⁻³</td>
</tr>
<tr>
<td>lacO3O1-tetA</td>
<td>1000</td>
<td>417.5</td>
<td>303.5</td>
<td>0.529</td>
<td>2.127 × 10⁻³</td>
<td>4.267 × 10⁻⁴</td>
</tr>
<tr>
<td>lacO3O1-tetA</td>
<td>5</td>
<td>439.0</td>
<td>358.5</td>
<td>0.667</td>
<td>7.289 × 10⁻⁵</td>
<td>5.049 × 10⁻⁶</td>
</tr>
<tr>
<td>lacO3O1-BAD</td>
<td>1000</td>
<td>610.1</td>
<td>468.9</td>
<td>0.591</td>
<td>6.325 × 10⁻⁶</td>
<td>1.456 × 10⁻⁷</td>
</tr>
<tr>
<td>lacO3O1-BAD</td>
<td>5</td>
<td>657.1</td>
<td>588.7</td>
<td>0.803</td>
<td>3.464 × 10⁻³</td>
<td>5.124 × 10⁻⁴</td>
</tr>
</tbody>
</table>

[1] N. Adachi and M. R. Lieber, Bidirectional gene or-
ganization: a common architectural feature of the human
genome, Cell 109, 807 (2002).
Schroeder, R. P. Otillar, and R. M. Myers, An abundance of
bidirectional promoters in the human genome, Genome
[3] M. Herbert, A. Kolb, and H. Buc, Overlapping prom-
Svenningsen, K. Sneppen, S. Krishna, and S. Senseby, Direct
and indirect effects in the regulation of overlapping prom-
Analysis of genomic context: prediction of functional
associations from conserved bidirectionally transcribed
[6] E. M. Prescott and N. J. Proudfoot, Transcriptional col-
liision between convergent genes in budding yeast, Proc.
[7] B. P. Cullen, K. E. Shearwin, and J. B. Egan, Trans-
criptional interference between convergent promoters
caused by elongation over the promoter, Mol. Cell 14,
and K. E. Shearwin, Potent transcriptional interference
by pausing of RNA polymerases over a downstream pro-
Genome wide study of NF-Y type CCAAT boxes in uni-
directional and bidirectional promoters in human and
Ribeiro, and H. T. Jacobs, NF-Y influences directional-
ity of transcription from the bidirectional Mrps12/Sarsm
promoter in both mouse and human cells, BBA Gene
O. Yli-Harja, J. M. Fonseca, and A. S. Ribeiro, Dynam-
ics of transcription of closely spaced promoters in Es-
cherichia coli, one event at a time, J. Theor. Biol. 301,


