Regulation of mean and noise of the \textit{in vivo} kinetics of transcription under the control of the \textit{lac/ara-1} promoter

Meenakshisundaram Kandhavelu$^a$, Jason Lloyd-Price$^a$, Abhishek Gupta$^a$, Anantha-Barathi Muthukrishnan$^a$, Olli Yli-Harja$^{a,b}$, Andre S. Ribeiro$^{a,+}$

$^a$ Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, 33101 Tampere, Finland

$^b$ Institute for Systems Biology, 1441N 34th St, Seattle, WA, 98103-8904, USA.

Running title: Regulation of the \textit{in vivo} kinetics of \textit{lac/ara-1}

$^+$ Corresponding author

Tel: +358408490788, Fax: +358331154989, Email: andre.ribeiro@tut.fi

Office TC336, Department of Signal Processing,

Tampere University of Technology,

P.O Box 553, 33101 Tampere, FINLAND
ABSTRACT

The kinetics of transcription initiation in *Escherichia coli* depends on the duration of two rate-limiting steps, the closed and the open complex formation. In a *lac* promoter variant, $P_{lac/ara-1}$, the kinetics of these steps is controlled by IPTG and arabinose. From *in vivo* single-RNA measurements, we find that induction affects the mean and normalized variance of the intervals between consecutive RNA productions. Transcript production is sub-Poissonian in all conditions tested. The kinetics of each step is independently controlled by a different inducer. We conclude that the regulatory mechanism of $P_{lac/ara-1}$ allows the stochasticity of gene expression to be environment-dependent.

**Keywords:** Transcription, *In vivo*, Stochasticity, Sub-Poissonian, Plasticity
INTRODUCTION

The distributions of phenotypes of monoclonal populations of *Escherichia coli* in a given environment are, to some extent, reproducible [1-3], although gene expression and other cellular processes are stochastic and many regulatory molecules exist in small numbers [4]. Much effort has been made to understand how mean and fluctuations in the numbers of RNA and proteins are regulated [4-7]. The observed mean and fluctuations of protein numbers appear to arise primarily from the kinetics of RNA numbers [4]. RNA steady-state levels are largely regulated by the kinetics of transcription [4,8], particularly in its multi-stepped initiation process [9],[10], rather than by, for example, the kinetics of RNA degradation [8].

*In vitro* studies indicate that, in general, the closed complex and the open complex formations [11], two processes that occur during transcription initiation, are the slowest (rate-limiting) steps of RNA production [12,13]. Relevantly, their kinetics is both sequence-dependent and affected by regulatory molecules [14,15], which may explain how different genes differ so widely in kinetics of expression [16].

The *lac* promoter and its variants are commonly used in studies of the kinetics of transcription [5,17-19]. Yet, much remains unknown about their regulatory mechanisms. *P*<sub>lac/ara-1</sub> [5] and *P*<sub>lar</sub> [6] are two such *P*<sub>lac</sub>-based promoters. Both contain the binding sites *I*<sub>1</sub> and *I*<sub>2</sub> for AraC upstream of the operator region (Fig 1), but otherwise differ. *In vitro* measurements of *P*<sub>lar</sub> kinetics suggest two major rate-limiting steps in initiation, namely, the closed and the open complex, followed by a faster step, the promoter clearance [5]. The *lac* repressor, LacI, is suggested to affect only the closed complex, and given this assumption, AraC appears to affect multiple steps [5]. However, this assumption [18] has recently been called into question [19]. In general, these studies suggest that kinetics of these steps is under strict control of the regulatory
molecules [5,10]. These results rely on in vitro measurements. It is unknown to what extent they apply in vivo. For one, cells are not well-stirred [20]. Also, the DNA structure may differ, among other possible differences. Finally, from in vitro studies, as these only assess the average duration of the steps in initiation, it is not possible to assess, e.g., the stochasticity in this process.

Two studies reported in vivo measurements of the kinetics of P_{lac/ara-1} [21,22]. In these, the RNA is visible shortly after transcription occurs by tagging them with MS2d-GFP proteins [23]. By detecting when the target RNAs first appear, it was possible to measure the intervals between consecutive transcription events. The distribution of these intervals is mostly determined by the kinetics of initiation and its shape shows that the process of transcript production is sub-Poissonian under weak and medium induction by both inducers, IPTG and arabinose [22]. Finally, assuming that initiation consists of a sequence of exponentially-distributed steps, the number and durations of the underlying rate-limiting steps were inferred [22]. Evidence was found for at least two rate limiting steps during initiation of P_{lac/ara-1}.

The individual effects of the inducers on the in vivo kinetics of initiation remain unclear [19]. Relevantly, if the duration of each step can be regulated by these molecules in an independent fashion it should be possible, to some extent, to regulate both the mean and degree of fluctuations in RNA numbers [24,25]. Here, we explore this possibility and address the following question: how are the durations of the sequential steps in initiation of P_{lac/ara-1} affected by each of the regulatory molecules? To answer this question, we measure intervals between consecutive transcription events under induction by IPTG alone, by arabinose alone, and by both inducers. From the results, we propose an explanatory model of the in vivo regulatory mechanism of P_{lac/ara-1}.
MATERIALS AND METHODS

Cells, plasmids

_E. coli_ strain DH5α-PRO (identical to DH5α-Z1) was generously provided by I. Golding of the University of Illinois, and contains two constructs: (i) PROTET-K133 carrying P<sub>LtetO-1</sub>-MS2d-GFP [23], and (ii) a pIG-BAC (P<sub>lac/ara-1</sub>-mRFP1-MS2-96bs) vector, carrying a 96 binding site array under the control of P<sub>lac/ara-1</sub> [23].

qPCR analysis of target RNA

Gene expression was induced as described above. Cells were then immediately fixed with RNAProtect bacteria reagent followed by enzymatic lysis with Tris-EDTA lysozyme buffer (pH 8.3). Total RNA was isolated using RNeasy kit (Qiagen) according to the manufacturer’s instructions. DNaseI treatment was performed to avoid DNA contamination. cDNA was synthesized (Fermentas, Finland) from 1 μg of RNA with iScript Reverse Transcription Supermix according to the manufacturer’s instructions. The cDNA templates with final concentration of 10 ng/μl were added to the qPCR master mix which contained iQ SYBR Green supermix (Fermentas, Finland) with primers for the target and reference genes at a final concentration of 200 nM. We used the _16S rRNA_ housekeeping gene for internal reference. The primers for the target mRNA (Forward: 5' TACGACGCCGAGGTCAAG 3' and Reverse: 5' TTGTGGGAGGTGATGTCCA 3') target the region of _mRFP1_ and for the reference gene _16S rRNA_ (Forward: 5' CGTCAGCTCGTGTTGTGAA 3' and Reverse: 5' GGACCGCTGGCAACAAAG 3'). The qPCR experiment was performed using a Biorad MiniOpticon Real time PCR system (Biorad, Finland). The following thermal cycling protocol was used: 40 cycles of 95 °C for 10 s, 52 °C for 30 s, and 72 °C for 30 s for each cDNA replicate. These reactions were performed in three experiments, each with three replicates per
condition with a final reaction volume of 50µl. No-RT controls and no-template controls were used to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were greater than 95%. The data from CFX ManagerTM Software was used to calculate the relative gene expression while the standard error was calculated according to Livak method [26].

**Time-lapse single-molecule fluorescence microscopy**

Cells were grown in Miller LB medium, supplemented with antibiotics according to the specific plasmids. Cells were grown overnight at 37 °C with aeration, diluted into fresh medium and allowed to grow at 37 °C until an optical density of OD_{600} ≈ 0.3-0.5 was reached. To attain full induction of the MS2d-GFP reporter, cells were incubated with 100 ng/ml of anhydrotetracycline (aTc, from IBA GmbH). 0-0.1% of L-arabinose (Sigma-Aldrich) and 0-1 mM of Isopropyl-β-D-thiogalactopyranoside (IPTG, Fermentas) were used to induce the target RNA. For complete activation of the *ara* system in the conditions where arabinose and IPTG were added, cells were pre-incubated with arabinose at the same time as aTc. In cases where IPTG was added, the IPTG was added one hour after aTc, and incubated for 5 min, with image acquisition beginning thereafter. In the condition with arabinose alone, 0.1% arabinose was added one hour after aTc and incubated for 5 min, followed by imaging. Cells were imaged under the microscope in a thermal chamber set to 37 °C.

For imaging, a few µl of culture were placed between a cover-slip and a slab of 1% agarose containing LB along with the appropriate concentrations of inducers. When the reporter and target RNA are co-expressed, MS2d-GFP binds to the target RNA, forming a bright fluorescent spot. The RNA becomes visible during, or shortly after elongation [23]. Images of cells were taken from each slide every minute over two hours [7]. Microscopy was performed
using a Nikon Eclipse (TE-2000-U, Nikon, Tokyo, Japan) inverted confocal laser-scanning microscope. Example movies are provided in supplementary material.

**Image processing**

Cells were detected from the microscope images using the semi-automatic method described in [21]. First, a mask was manually painted over the area that a cell occupied during the time series [21]. Principal component analysis was then used to obtain the dimensions and orientation of the cells from the fluorescence distribution within each mask. Target RNA spots were automatically segmented using density estimation [27] with a Gaussian kernel and Otsu’s thresholding [28]. Cell background-corrected spot intensities were then calculated and summed for each cell to produce the total spot intensity within each cell. The time series of this value was fit to a monotone piecewise-constant function by least squares where the number of terms was selected using the F-test with p-value 0.01. Each jump was then taken to correspond to the production of a single RNA molecule. An example of this process is shown in Fig 4.

**Inference of sequential steps**

The number and durations of sequential steps in transcription initiation were inferred as in [22] by fitting the distribution of intervals between consecutive production events to a multi-step model, where each step is exponentially-distributed. That is, for a given number of steps, $d$, the measured distribution was fit to the probability density function, $\pi(x \mid \mu_1, \ldots, \mu_d)$, of the sum of $d$ exponential variables with mean durations $\mu_i$ ($i = 1, \ldots, d$). The mean duration of each of these exponentials is selected such that they maximize the likelihood function

$$L = \prod_i \pi(x_i \mid \mu_1, \ldots, \mu_d),$$

where $x_i$ are the samples of the measured distribution. The number of steps was selected using a likelihood ratio test between a $d$ and $d+1$ step model with p-value 0.01. That is, we add steps
until it does not result in a significant improvement of the fit. For a more detailed description of this methodology please refer to [22].

To assess the robustness of the inference procedure for an inferred set of mean step durations, we computed the standard deviation of the step durations inferred from time intervals that were sampled from the distribution of time intervals prescribed by the inferred model (shown as error bars in Fig 3). That is, for an inferred model with \( d \) steps, \( N \) simulated time intervals were sampled by summing \( d \) exponentially distributed variables with the same means as the inferred model, where \( N \) is the number of intervals obtained from the measurements. The inference process was applied to this sampled distribution, and the standard deviations of each of the \( i \) longest durations were calculated, i.e., the variance of the shortest step was calculated from the distribution of the shortest inferred steps.

RESULTS

We assess the kinetics of \textit{in vivo} transcript production by detecting in live cells when each target RNA molecule first appears, and from this information, determine the intervals between consecutive productions of RNA molecules. Distributions of these intervals can be used to assess the kinetics of the rate-limiting steps during transcription initiation, namely, their number and expected durations [21,22].

It is noted that the tagging of RNA molecules by MS2-GFP has been shown to perturb the natural system in two ways. First, the tagging ‘immortalizes’ the RNA molecules (at least for the duration of the measurements reported here) [7,21,22]. Second, the spatial localization of the tagged RNAs is likely to differ from the unperturbed system [29]. These two perturbations do not interfere with our conclusions, since we only rely on when the tagged RNAs first appear, and not
their location or their total amounts in the cell at any given moment. In fact, in general, the tagged RNAs tend to appear in the midcell region, where the F-plasmid is located, and then slowly move to a cell pole, reaching it before the next target RNA is produced[21,22,30]. In that sense, this facilitiates the detection of appearances of new RNA molecules, and thus the measurement of intervals between consecutive transcription events.

We measured the in vivo transcript production under five conditions (Table I). After induction (Methods), cells were imaged once per minute for two hours. The relative differences in mean RNA production rates between conditions were confirmed by qPCR (Table II). The results show a graded response to the concentration of inducers, consistent with [6,21].

Table I shows the mean, variance and normalized variance (CV$^2$) of the distribution of intervals in each induction regime. The CV$^2$ is, in all conditions, smaller than 1, indicating that transcript production by P$_{lac/ara-1}$ is always sub-Poissonian [25]. The CV$^2$ differs between conditions, due to the change in shape of the distribution (Fig 2).

The above conclusion regarding the kinetics of transcript production is only valid if the intracellular concentration of inducers is constant during the measurements. This can be tested for, since if it is not constant, the distributions of intervals obtained in the first and last hour of the measurements should differ. For each induction condition, the Kolmogorov-Smirnov test was unable to differentiate between the two distributions (all p-values > 0.5), supporting the assumption of approximate steady state transcript production.

We next infer the number and mean durations of the sequential steps in initiation from each of the distributions, as proposed in [22] (Methods). However, since this method is only valid if consecutive intervals between productions are approximately independent, we first test for this condition. Events in elongation, e.g. pauses [31], could contribute to the variance of
elongation times. High variance in elongation times would introduce correlations between consecutive intervals. We measured the Pearson correlation from 211 pairs of consecutive intervals, and found it to be 0.07, implying that if there is a correlation, it is negligible (p-value = 0.31).

The results of the inference of the kinetics and number of the steps are shown in Table I and Fig 3. In all conditions, the likelihood ratio test (Methods) indicates that the interval distributions are better-fit by two exponentially-distributed sequential steps than one (all p-values < 2×10^{-4}). A third step does not improve the fit sufficiently to reject the two-step model (all p-values > 0.12). To determine if the changes in the durations of the steps are significant, we estimated the robustness of the inference procedure (Methods), shown as error bars (Fig 3).

From Fig 3, as induction by IPTG is decreased, one step becomes longer while the other does not change significantly. On the other hand, when decreasing induction by arabinose, at least one step increases in duration and the two steps become similar in duration. This implies that IPTG and arabinose induce \( P_{\text{luc/ara-1}} \) by different mechanisms.

While the inference method does not provide the order of the steps, the points in Fig 3 can be connected using the following reasoning. First, between (1, 0.1) to (1, 0), there is no other way to connect the steps. From (1, 0.1) to (0, 0.1), the robustness of the inference procedure suggests that the duration of one of the steps does not change with induction by IPTG (the shorter interval). By connecting these, the remaining points can only be connected as they are.

**DISCUSSION**

Recently, transcription initiation of \( P_{\text{luc/ara-1}} \) was shown to be multi-stepped, and sub-Poissonian under weak and medium induction by both of its inducers, IPTG and arabinose [22].
Here, we characterized the individual effects of each of these inducers on the *in vivo* kinetics of $P_{lac/ara-1}$, namely on the distributions of the intervals between consecutive transcription events. The kinetics is sensitive to induction, as it affects both the mean and variance of the distribution of intervals.

We found that the *in vivo* kinetics of RNA production is sub-Poissonian not only under weak and medium induction [22], but also under full induction by both inducers and when either inducer is absent, though the degree of stochasticity differs between conditions. Two rate-limiting steps fit the data well in all conditions, though the durations of these steps vary widely between conditions. The data also indicates that the mean durations of these steps can be varied in a graded fashion by varying the concentrations of the inducers.

Our observations indicate that the regulatory mechanism of the kinetics of the sequential steps of $P_{lac/ara-1}$ permits the two regulatory molecules to control, in a complementary fashion, the mean and variance of the intervals between transcription events. This regulatory mechanism further allows this mean and variance to be controlled, to some extent, independently. This is of relevance, since the kinetics of these intervals will determine, to a great extent, both the mean and variance in RNA numbers [25], and thus, proteins [4].

The induction mechanism of $P_{lac/ara-1}$ allows it to exhibit considerably complex behaviours. For example, noise is reduced when induction by arabinose is neither minimum nor maximum. Moreover, provided that the model of the mechanism of initiation is accurate, the concentration of arabinose for which noise is minimal should differ based on the concentration of IPTG, since the $CV^2$ of the sum of two exponentially distributed variables is minimized when they have equal means.
We conclude that the induction mechanism of $P_{lac}$ allows the independent regulation of the durations of the two major rate-limiting steps, while maintaining the kinetics of RNA production sub-Poissonian. It is unknown whether this is a common feature of native promoters in $E. coli$. Genome-wide assessments of cell-to-cell diversity in RNA numbers suggest otherwise [16]. However, these results depend not only on the kinetics of transcription but also on the kinetics of RNA degradation [8] and RNA partitioning in cell division [32,33], while our results do not.

So far, there is no method to experimentally assess the shape of the distributions of the duration of intermediate steps in initiation. In vivo methods only assess intervals between consecutive RNA molecules while in vitro methods only assess the mean duration of intermediate steps, provided some assumptions (see e.g. [5]). It is thus worthwhile to discuss the assumption made here to infer the number and duration of the rate-limiting steps in initiation, particularly given that some of the events in initiation, such as the opening of the double strand for reading, are not elementary steps in a strict sense. Nevertheless, we assumed the simplest possible model, i.e., that the steps are elementary reactions of the form $A \rightarrow B$, with a constant probability of occurring per unit time. This entails that the distributions of intervals between events are exponential [34]. Given this assumption, we find evidence for two rate-limiting steps. This result is in agreement with results from in vitro studies of several promoters in $E. coli$ [14]. Further, the inferred distributions and the data are statistically indistinguishable from one another by the Kolmogorov-Smirnov test, which implies that there is no evidence to assume that the choice of shapes is wrong, at the level of precision of the present measurements.

The inference method cannot determine the temporal order of the rate-limiting steps, and thus we cannot conclude, for example, the correspondence between the inferred steps and the
closed or the open complex formations. Previous studies provide tentative clues. LacI appears to regulate the closed complex formation in P_{lac/ara-1} [18]. When we varied IPTG concentration we observed a significant change solely in the longer step. We thus expect that the shorter rate-limiting step in Table I corresponds to the open complex formation, while the other corresponds to the closed complex formation. In that case, Arabinose affects at least the open complex formation, which is consistent with in vitro measurements [5]. This identification of the steps is speculative, since a recent study has questioned previous conclusions on the effects of LacI [19]. While we cannot resolve this identification, we can confirm that there are two major steps, and that each is affected by a different inducer.

Nevertheless, our measurements inform on the variability of the intervals between transcript productions in all induction ranges, without the interference of events following transcription such as RNA degradation, which at the moment is not possible by other methods. They thus provide valuable information on the stochasticity of transcription. For example, we observed that by controlling the kinetics of the two rate-limiting steps independently, by different inducers, this mechanism enables the degree of stochasticity of RNA production to be a function of environmental conditions.

In conclusion, in our assessment of the in vivo kinetics of P_{lac/ara-1}, we observed a sub-Poissonian multi-stepped mechanism of initiation, whose plasticity relies on the ability of the two inducers to affect the kinetics of the steps independently. Similar to this promoter, which was engineered from two native promoters [6], most native promoters in E. coli are regulated by more than one regulatory molecule [35]. We hypothesize that this may be a consequence of a need for plasticity of noise in gene expression, to facilitate survival in fluctuating environments [36].
ACKNOWLEDGEMENTS

The authors thank Antti Häkkinen for valuable advice and discussions. Work supported by the TUT President’s graduate programme (JLP, ABM) and the Finnish Funding Agency for Technology and Innovation (AG, OY-H). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES


Figure 1: Schematic of the induction of the $P_{lac/ara-1}$ promoter by arabinose and IPTG.
Figure 2: Distributions of intervals between transcription events (gray bars) when induced by 0.1% arabinose (top), 1 mM IPTG (bottom), and both (middle). The best-fit models with one (dotted line), two (solid line) and three (dashed line) exponentially-distributed steps in initiation are shown.
**Figure 3:** Mean durations of the inferred steps for each condition (circles). The standard deviations of the step duration inference for the same inferred means and the same number of samples are also shown (error bars, see Methods).

**Figure 4:** (A) Example image of *E. coli* cells expressing MS2d-GFP and target RNA, taken with confocal microscopy. (B) Segmentation of the image in (A). Gray areas show segmented cells while segmented spots are shown in white. (C) Time course of the total intensity of spots (gray) in the cell shown at top, including a cell division. The monotone piecewise-constant fits are shown in black.
Table I: *In vivo* transcript production kinetics under various induction conditions, and statistics on the time intervals between consecutive transcription events in individual cells.

<table>
<thead>
<tr>
<th>IPTG (mM)</th>
<th>arabinose (%)</th>
<th><em>In vivo</em> Mean RNA/h</th>
<th>Interval (\mu) (s)</th>
<th>Interval (\sigma) (s)</th>
<th>(\sigma^2/\mu^2)</th>
<th>Durations of steps (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
<td>0.3</td>
<td>108</td>
<td>1368</td>
<td>1128</td>
<td>0.68 (1122, 246)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>1.0</td>
<td>71</td>
<td>1300</td>
<td>989</td>
<td>0.58 (976, 325)</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>1.6</td>
<td>343</td>
<td>965</td>
<td>698</td>
<td>0.52 (574, 391)</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>1.1</td>
<td>185</td>
<td>1483</td>
<td>819</td>
<td>0.30 (741, 741)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.7</td>
<td>205</td>
<td>1587</td>
<td>1076</td>
<td>0.46 (793, 793)</td>
</tr>
</tbody>
</table>

Cells were induced by IPTG and L-arabinose. The mean number of RNA molecules produced by live cells per hour is shown for each condition (*In vivo* Mean RNA/h). The mean (\(\mu\)), standard deviation (\(\sigma\)) and normalized variance (\(\sigma^2/\mu^2\)) of the distribution of time intervals between subsequent production events in single cells are also shown. N denotes the number of such intervals observed in the experiments. In all cases, the distribution of these intervals was well-fit by a two exponentially-distributed step model with the means shown in the last two columns.

Table II: Comparison of mean transcript production when measured by qPCR and *in vivo* single-molecule measurements.

<table>
<thead>
<tr>
<th>IPTG (mM)</th>
<th>arabinose (%)</th>
<th>Relative Mean RNA (qPCR)</th>
<th><em>In vivo</em> Mean RNA/h</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
<td>0.49</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.63</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Cells were induced by IPTG and L-arabinose. The relative changes in mean RNA are shown as measured by qPCR (Relative Mean RNA). For comparison, the mean number of RNA molecules
produced by live cells per hour is also shown (In vivo Mean RNA/h). The qPCR relative means have been scaled to the measured in vivo mean RNA/h at maximum induction.