Bioinformatics, 32(9), 2016, 1346–1352
doi: 10.1093/bioinformatics/btv744
Advance Access Publication Date: 31 December 2015
Original Paper

OXFORD

Gene expression

Characterizing rate limiting steps in transcription from RNA production times in live cells

Antti Häkkinen and Andre S. Ribeiro*

Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, P.O. box 553, 33101, Tampere, Finland

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on 10 August 2015; revised on 20 November 2015; accepted on 15 December 2015

Abstract

Motivation: Single-molecule measurements of live *Escherichia coli* transcription dynamics suggest that this process ranges from sub- to super-Poissonian, depending on the conditions and on the promoter. For its accurate quantification, we propose a model that accommodates all these settings, and statistical methods to estimate the model parameters and to select the relevant components.
Results: The new methodology has improved accuracy and avoids overestimating the transcription rate due to finite measurement time, by exploiting unobserved data and by accounting for the effects of discrete sampling. First, we use Monte Carlo simulations of models based on measurements to show that the methods are reliable and offer substantial improvements over previous methods. Next, we apply the methods on measurements of transcription intervals of different promoters in live *E. coli*, and show that they produce significantly different results, both in low- and high-noise settings, and that, in the latter case, they even lead to qualitatively different results. Finally, we demonstrate that the methods can be generalized for other similar purposes, such as for estimating gene activation kinetics. In this case, the new methods allow quantifying the inducer uptake dynamics as opposed to just comparing them between cases, which was not previously possible. We expect this new methodology to be a valuable tool for functional analysis of cellular processes using single-molecule or single-event microscopy measurements in live cells.
Availability and implementation: Source code is available under Mozilla Public License at http://www.cs.tut.fi/%7Ehakkin22/censored/.
Contact: andre.ribeiro@tut.fi or andre.sanchesribeiro@tut.fi
Supplementary information: Supplementary data are available at *Bioinformatics online*.

1 Introduction

In bacteria, transcription is the main regulatory mechanism of RNA numbers in the cell. This conclusion is supported by the lack of correlation between RNA numbers and their degradation rates (Bernstein *et al.*, 2002) and by the fact that most regulatory molecules modulate the transcription initiation process (McClure, 1985). Relevantly, the regulatory mechanisms of transcription allow wide adaptability, as the kinetics of this process varies widely between promoters, and for the same promoter under different conditions.

Live cell measurements in *Escherichia coli* suggest that depending on the promoter and the conditions, such as the presence/absence of repressor/activator molecules, RNA production can range from sub-Poissonian, that is, less uncertain than a Poisson process (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012), through Poissonian (Yu *et al.*, 2006), to super-Poissonian (Golding *et al.*, 2005; Taniguchi *et al.*, 2010). Such wide dynamic range is not likely achievable by a single mechanism and, in agreement, evidence suggests that this is a multi-step process (McClure, 1985).

Currently, the subprocesses that constitute transcription cannot be directly measured in live cells. However, it is possible to observe RNA production of individual promoters with single molecule resolution over time (Golding *et al.*, 2005). This information can be used to estimate the dynamical parameters of the subprocesses by the means of a stochastic model (Kandhavelu *et al.*, 2011). The success of this strategy requires accurate and unbiased statistical methods of data analysis as well as a model that can account for all possible dynamical regimes.

Previously, we made use of distributions of intervals between transcription events in various conditions in order to estimate in maximum likelihood (ML) sense, for each condition, the number and duration of the rate limiting steps in transcription (Kandhavelu *et al.*, 2011). Relevantly, unlike RNA numbers, these intervals are not affected by RNA degradation, or dilution due to cell division, and consequently allow more accurate quantification of the transcription process. However, the previous model of transcription does not cover all potential cases (e.g. super-Poissonian RNA production).

Here, we first propose a model that, by combining previous models responsible for different dynamical behaviors (McClure, 1985; Peccoud and Ycart, 1995), is capable of exhibiting behaviors ranging from sub- to super-Poissonian. Next, we present methods to estimate its parameters in ML sense. An advantage over previous methods (Kandhavelu *et al.*, 2011) is that the new methods also use information of the unobserved transcription events. Such additional information results in improved accuracy and can be used to correct the biases resulting from the limited the measurement time. The methods can also account for the discrete sampling, which is typical for a fluorescence microscopy measurement. The increased accuracy allows studying subprocesses with smaller time scales, while the lack of bias is essential in order to correctly estimate the parameters of the more noisy models and to compare them in an unbiased manner. The methods can also be used to provide features such as confidence in the estimated parameters, and we use statistical methods to select components of the model in/out, which can be used to determine if certain components are responsible for the observed dynamical behavior.

2 Methods

Transcription intervals and activation time measurements in live *E. coli* cells were obtained using the MS2-GFP RNA-tagging system (Golding *et al.*, 2005). The cells contain a single-copy vector coding for the target RNA under the control of a specific target promoter: TetA (Muthukrishnan *et al.*, 2012), bacteriophage λ RM (Golding *et al.*, 2005), or arabinose BAD (Makela *et al.*, 2013). The target RNAs contain an array of 48 or 96 binding sites (depending on the construct) for the highly expressed MS2-GFP reporters to bind. This allows the target RNAs to be visualized using fluorescence microscopy right after their production. The systems were constructed previously (Golding *et al.*, 2005; Makela *et al.*, 2013; Muthukrishnan *et al.*, 2012), and more details of the measurements conducted here using these systems are given in the Supplementary material.

3 Algorithms

In this section, we propose a model of transcription initiation whose kinetics can, depending on parameters selection, range from sub- to super-Poissonian. Next, we describe how to extract time interval distributions from the model. Finally, we describe how to estimate the model parameters using the measurement data.

3.1 Model of transcription initiation

To allow transcription dynamics to range from sub- to super-Poissonian, we propose the model of elementary reactions:

$$\underbrace{P_{\mathrm{off}} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} P_{\mathrm{on}}}_{\text{on}-\text{off step}} \xrightarrow{k_1} I_1 \xrightarrow{k_2} \cdots \xrightarrow{k_{n-1}} I_{n-1} \xrightarrow{k_n} P_{\mathrm{on}} + E, \qquad (1)$$

where $P_{\mathrm{off}}, P_{\mathrm{on}}, I_j \in \mathbb{Z}_1$ represent different states of the promoter: inactive, active and some intermediate states of transcription initiation, respectively, while E represents the elongation complex. This model is designed to combine the active–inactive promoter model (Peccoud and Ycart, 1995) with a sequential model of transcription initiation (McClure, 1985; Saecker *et al.*, 2011). Note that any number of backward reactions for steps 1 through $n - 1$ is implicitly supported, since equal dynamics can be achieved by setting the rates of the model appropriately (see Supplementary material).

The on–off mechanism produces bursty RNA production, due to the random off periods (Peccoud and Ycart, 1995). When this mechanism dominates the dynamics, the intervals between transcript productions are highly noisy, resulting in super-Poissonian RNA production. The above model is appropriate regardless of the mechanism controlling the promoter on–off transitions as long as the state transitions occur with constant probability per unit time. Recent studies suggest that the dynamics of several promoters in *E. coli* may be dominated by such a mechanism (Chong *et al.*, 2014; Golding *et al.*, 2005; Taniguchi *et al.*, 2010).

In contrast, a sequential process of RNA production reduces noise, as it produces more regular intervals. *In vitro* measurements suggest the following sequence of events (Lutz *et al.*, 2001; McClure, 1985): first, an RNA polymerase must find and diffuse along the DNA template until finding the transcription start site (Saecker *et al.*, 2011). There, the polymerase forms a closed complex, and then goes through several isomerization steps, until completing the open complex formation (Saecker *et al.*, 2011). After escaping the start site, the complex elongates along the DNA template, clearing the promoter region. Recent *in vivo* measurements have shown that at least some promoters in *E. coli* are capable of exhibiting a dynamics consistent with this model for a wide range of conditions (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012).

In Equation (1), the steps which are much faster than the others can be neglected. Using the same argument, we can also take elongation complexes E to represent fully transcribed RNAs, as elongation takes tens of seconds (Herbert *et al.*, 2006), while interproduction intervals are in the order of hundreds of seconds (Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012). Regardless, elongation is not expected to affect the RNA production intervals on average, unless the initiation rate is so high that there is polymerase traffic (Rajala *et al.*, 2010). We also note that if the promoter states are unobservable (as is the case here), the order of sequential processes cannot be determined from the transcription dynamics.

3.2 Transcription interval distribution

Since transcription intervals can be split to sums of independent steps, the probability densities are easily manipulated in terms of their moment generating functions (MGFs). This is due to the fact that the MGF of a sum of independent variables is the product of the individual MGFs, and the MGF of a mixture is a weighted sum of the individual MGFs. The MGF of an exponential variate with a mean of ${k_i}^{-1}$ is:

$$M_i(t) = k_i\,(k_i - t)^{-1} \qquad (2)$$

which implies that the MGF of the on-off step (including on, off and the first reaction) is:

$$M_{\text{on}-\text{off}}(t) = k_1\,(k_{\text{on}} - t)\,(p^- - t)^{-1}(p^+ - t)^{-1}$$
$$\text{with} \quad p^\pm = \frac{k_{\text{off}} + k_1 + k_{\text{on}}}{2} \pm \frac{\sqrt{(k_{\text{off}} + k_1 - k_{\text{on}})^2 + 4\,k_{\text{off}}\,k_{\text{on}}}}{2} \qquad (3)$$

which is described in more detail in the Supplementary material. This indicates that any MGF of the model must have the form:

$$M(t) = G\prod_{i=1}^{U}(z_i - t)^{u_i}\prod_{i=1}^{V}(p_i - t)^{-v_i} = \sum_{i=1}^{V}\sum_{v=1}^{v_i}\frac{R_i^{(v)}}{(p_i - t)^v}, \qquad (4)$$

where the latter is the partial-fraction decomposition of $M(t)$. The residues $R_i^{(v)}$ can be computed e.g. using: $z - t = (z - p) + (p - t)$ and $1/(p - t)/(q - t) = -1/(p - q)/(p - t) + 1/(p - q)/(q - t)$. By noting that the decomposition specifies a linear combination of the MGFs of sums of exponential variates, the probability density function (PDF) is recovered as:

$$f(x) = \sum_{i=1}^{V}\sum_{v=1}^{v_i}\left(\frac{R_i^{(v)}}{p_i{}^v}\right)\frac{p_i{}^v}{\Gamma(v)}\,x^{v-1}\,e^{-p_i\,x}, \qquad (5)$$

where the parenthesized term is the mixing weight and the remaining term is the PDF of a sum of v exponential variates, with a mean of $p_i{}^{-1}$ each. Here, $\Gamma(w)$ denotes the factorial of $w - 1$. In general, the mixing weights are not convex, so this is not a proper mixture density.

Manipulating the MGF can be also used to perform other useful operations: the survival function can be obtained by adding a pole at $t = 0$, subtracting the t^{-1} term, and taking the inverse transform as above. Also, differentiation might be easier to perform on the MGF, which is useful for evaluating gradients and/or Hessians for optimization or for confidence estimation. For example, differentiation with respect to a k_i^{-1} with $i > 1$ can be achieved by adding a zero at $t = 0$, a pole at $t = k_i$ and multiplying the residues by k_i.

3.3 ML estimation

As the measurements have finite length and discrete sampling (see Fig. 1), the true intervals between RNA production are not exactly known. An interval can be only observed if it fits in to the measurement window, and as such, in each cell, the last interval will not be observed due to the end of the measurement period. Neglecting these unobserved intervals will result in underestimation of the interval durations. Meanwhile, the discrete sampling implies that each interval contains some uncertainty about its exact duration, which should be communicated to the estimator. For example, the true interval between the second and third productions in Figure 1 is known to be 10–12 units long (interval-censoring). Meanwhile, the true interval between the third and fourth productions is not <9 units long (right-censoring). These two modes of uncertainty are called interval- and right-censoring, as the true value is bounded to an interval or to the right (on the real line) of some observation. More precise definitions can be found e.g. in Turnbull (1976).

Provided that the intervals T_i between transcription events are independent, the probability of observing a sequence of intervals $(t_1, \ldots, t_m)$ in a time series of length L is given by

$$\begin{aligned}&\mathbb{P}[\,S \simeq (t_1, \ldots, t_m)\,] = \mathbb{P}[\,T_1 \simeq t_1\,] \cdots \mathbb{P}[\,T_m \simeq t_m\,]\\ &\mathbb{P}[\,T_{m+1} > L - (t_1 + \cdots + t_m)\,]\;\mathbb{I}[\,t_1 + \cdots + t_m \le L\,],\end{aligned} \qquad (6)$$

where the notation abuses $\mathbb{P}[\,X \simeq x\,] = \mathbb{P}[\,x \le X < x + \partial x\,]$ for infinitesimal ∂x and $\mathbb{I}[\,\cdot\,]$ is the indicator function. Here, m and L need not to be constant, but can be realizations of independent random variables. This implies that a ML estimator can be written in the following form:

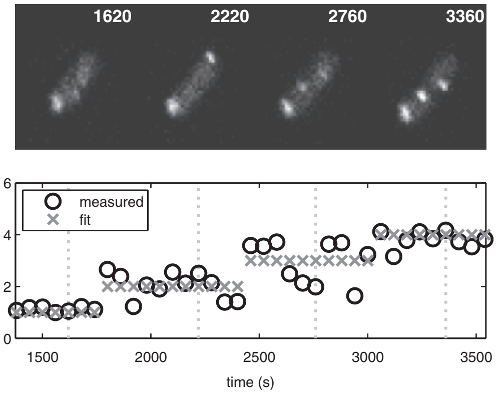

Fig. 1. Example data from MS2-GFP-tagged RNA measurements with the tetA promoter. Upper panel: fluorescence microscopy images of a cell at different time points, as indicated by the time stamp (in seconds). Lower panel: extracted intensities and estimated RNA numbers of the cell shown in the upper panel. The vertical lines represent the time points in the upper panel

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}}\;\mathbb{P}[\,t_1 \in [x_1, y_1], \ldots, t_{m+1} \in [x_m, y_m]\,|\,\boldsymbol{\theta}\,]\\ &= \arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{m+1}\log\left(\,F(y_i\,|\,\boldsymbol{\theta}) - F(x_i\,|\,\boldsymbol{\theta})\,\right),\end{aligned} \qquad (7)$$

where x_i and y_i are the (possibly infinite) known bounds for t_i, and $F(x)$ is the cumulative density function (CDF). The ordinary ML estimator is recovered at the limit $y_i \to x_i$, since $F(y_i\,|\,\boldsymbol{\theta}) - F(x_i\,|\boldsymbol{\theta}) \to f(x_i\,|\,\boldsymbol{\theta})\,(y_i - x_i)$, where $f(x)$ is the PDF, and $(y_i - x_i)$ is constant with respect to $\boldsymbol{\theta}$.

In general, the times from the beginning of the measurement to the first production might not have the appropriate distribution, and they cannot be used in the estimator. An exception to this occurs when it is known that the transcription process starts at the same time as the measurement. Another exception occurs when the transcription intervals are exponential, in which case the first production has the appropriate distribution due to the memorylessness of the exponential distribution.

If interval-censoring of the consecutive intervals is used, the samples are not independent, since the error terms of the consecutive measurements might be correlated. However, if the production intervals are much longer than the sampling intervals, a condition necessary for accurate estimation anyway, these correlations tend to be negligible. Despite violating this assumption, we found interval-censoring to improve the estimator performance considerably, as shown below.

In some cases, simple solutions to the ML problem exist (see Supplementary material). However, the general ML problem requires numerical methods. Also, the ML surface is not guaranteed to be concave, nor even unimodal. However, it tends to be well behaved in practice, especially for larger samples (the usual properties of an ML estimator apply). Due to this, we perform 100 restarts with random starting point. We used the Nelder-Mead method (Nelder and Mead, 1965) for optimization, since it appeared to perform well and is fast. We also experimented with the Broyden–Fletcher–Goldfarb–Shanno method, with either exact or finite difference derivatives, but it produced similar results and was significantly slower.

4 Results

4.1 Applying the methods on Monte Carlo simulations

Throughout this article, we use ‘exp’ and ‘seq-*n*’ to denote models of 1 or *n* sequential exponential steps and no on–off mechanism, respectively, and ‘onoff’, and ‘onoff-*n*’ to denote the full models with 1 or *n* steps, respectively. The parameters of each model is given as a vector $({k_{\mathrm{on}}}^{-1}, {k_{\mathrm{off}}}^{-1}, {k_1}^{-1}, \ldots, {k_n}^{-1})$ with ${k_2}^{-1} \le \cdots \le {k_n}^{-1}$, since the order of k_i is exchangeable.

We performed Monte Carlo simulations using the following models based on previous live *E. coli* measurements: exp with a mean of 2750 s with 60 cells sampled every 180 s for 3300 s (Yu *et al.*, 2006), seq-2 with means of 712 and 716 s with 40 cells sampled every 60 s for 7200 s (Kandhavelu *et al.*, 2011), seq-3 with means of 109, 254 and 254 s with 113 cells sampled every 60 s for 3600 s (Muthukrishnan *et al.*, 2012), and onoff with ${k_{\mathrm{on}}}^{-1} = 360$ s, ${k_{\mathrm{off}}}^{-1} = 1020$ s and ${k_1}^{-1} = 102$ s with 100 cells sampled every 20 s for 4800 s (Chong *et al.*, 2014). We constructed a hypothetical onoff-2 model based on the onoff model, by setting ${k_1}^{-1} = {k_2}^{-1} = 51$ s, since there are no live *E. coli* measurements supporting the more complex on–off models. However, such models have been used in eukaryotic context (Blake *et al.*, 2003).

The shapes of the model distributions are shown in Figure 2, and model statistics are shown in Supplementary Table S1. This table shows the parameters, the mean and standard deviation (sd) of the transcription intervals, their noise, as determined by the squared coefficient of variation, and the differential entropy, which is useful in interpreting the entropy-based statistics. In addition, the number of cells and the sampling (samples × sampling interval) is shown, which apply for the time series simulations.

4.1.1 Model selection and effects of sample sizes

First, we generated 100, 500 or 1000 intervals from each model, and fit the data with the following models: exp, seq-2, seq-3, onoff, onoff-2 and onoff-3. A network of likelihood ratio (LR) tests is used to choose the preferred model, such that the least complex model which cannot be rejected at a significance level of 0.01 is selected. Supplementary Table S2 shows the statistics for the most frequently selected model. Statistics were gathered from 1000 simulations in each case.

In the table, log-likelihood (LL) quantifies how well the estimated model fits the data. The differential Kullback–Leibler divergence (KLD) from the true model to the estimated one measures the information lost when the estimated model is used to approximate the true model, which will, as opposed to the likelihood, penalize overfitting. The spatial median (med) of the parameter estimates represents a typical estimate. Finally, the choice frequency indicates how often this model is selected in favor of the others. The alternative model (alt model) indicates the second most frequently selected model.

The results indicate that the information lost with the on–off models is an order of magnitude greater, which is expected, since they are noisier. The likelihood values suggest that the estimators behave as is expected from an ML estimator, despite the numerical optimization procedure. For the exp and seq-2 models, 100 samples appear to be sufficient to identify the appropriate model for >90% of the time. For 500 or 1000 samples, this is true for all but the onoff model, in which it occurs more than 80% of the time. Finally, we note that especially with 100 samples, the multi-parameter models have biases in the parameter estimates, such as toward/away from equal values in the sequential models. Fortunately, these biases vanish with larger sample sizes, as is typical for ML estimates, so the problem can be mitigated by collecting more samples.

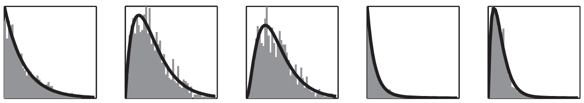

Fig. 2. Probability densities for the interval distributions of the models. The black curves show the asymptotic distribution, and the gray bars the histogram of 1000 random samples. From left to right, the models are: exp, seq-2, seq-3, onoff and onoff-2

4.1.2 Advantages of censoring

Next, we generated time series of RNA numbers to simulate our measurement settings. The data extracted from these series were fit with the appropriate model using full censoring, without interval-censoring (i.e. disregarding discrete sampling effects), without right-censoring (i.e. without correcting for the unobserved samples) or with neither mode of censoring. For the sequential exponential models, the last method corresponds to a previous method (Kandhavelu *et al.*, 2011).

The results are shown in Supplementary Table S3. Again, the likelihood quantifies how well the estimated model fits the data, the divergence how well the estimated model corresponds to the true model, and the median of the estimates represents a typical parameter estimate. The likelihood values are not comparable between the different modes of censoring. While these results demonstrate the applicability of the methods under typical settings, the different models cannot be compared as the relative sampling settings of the models differ widely (cf. Samples in Supplementary Table S3). To allow such comparison, we also simulated each model for 5μ time units, sampling every $5^{-1}\mu$ units, where μ is the mean production interval. These results are shown in Table 1.

The results indicate that the lack of right-censoring will result in drastic underestimation of the transcription interval duration, especially in the noisy cases, such as with the on–off models. Also, without right-censoring, the variance is generally underestimated in a similar manner. In the tested settings, the effects of discrete sampling were found comparatively weaker, but this is likely mitigated by the relatively frequent sampling (sampling rates are around 10-fold to that of transcription). However, accounting for it offers slight improvements in the accuracy in all cases.

In summary, it is essential to apply both modes of censoring for the high-noise models, while for sequences of exponentials neither mode of censoring is critical. However, in all cases the variant with full censoring performs best, as expected. Also, if the true model is not known but model selection will be performed, it is again necessary to use censoring to avoid biases toward selecting a less noisy model.

4.2 Applying the methods on transcription interval measurements

4.2.1 Transcription kinetics of the tetA promoter

We used MS2-GFP RNA-tagging and the methods with full and no censoring to analyze transcript production intervals in live *E. coli* (see Fig. 1). This allows determining if the two methods result in significantly different results with true measurement data. For this, we performed experiments, as described in Section 2. Again, the model is selected using a network of LR tests and the following models: exp, seq-2, seq-3, onoff, onoff-2, onoff-3.

First, we measured the RNA production in a construct where the target gene is controlled by the tetA promoter. A previous study reports that the dynamics are explained by a sequence of two or three

Table 1. Performance when using different modes of censoring in Monte Carlo simulations

Model	exp	seq-2	seq-3	onoff	onoff-2
Samples mean (sd)	500 (22.1)	474 (16)	468 (13.7)	562 (29.1)	514 (20.1)
KLD mean (sd)	0.00161 (0.00214)	0.00184 (0.00227)	0.00273 (0.00242)	0.0247 (0.0762)	0.042 (0.0466)
LL mean (sd)	−792 (28.5)	−698 (20.1)	−647 (17.7)	−859 (35.2)	−750 (24.2)
Parameter med	2680	694, 714	162, 190, 257	319, 812, 96.7	464, 926, 51.7, 48.6
Samples mean (sd)	500 (22.1)	474 (16)	468 (13.7)	562 (29.1)	514 (20.1)
KLD mean (sd)	0.0019 (0.00252)	0.00282 (0.00321)	0.00385 (0.00356)	0.0325 (0.0485)	0.0455 (0.0483)
LL mean (sd)	−3590 (180)	−3070 (118)	−2680 (89.4)	−2700 (145)	−2360 (96.9)
Parameter med	2860	644, 820	105, 247, 278	566, 1120, 107	828, 999, 55.6, 48.1
Samples mean (sd)	401 (22)	374 (16)	368 (13.7)	465 (28.5)	416 (19.7)
KLD mean (sd)	0.0378 (0.0132)	0.0209 (0.0101)	0.0178 (0.00915)	0.178 (0.0669)	0.142 (0.0408)
LL mean (sd)	−3470 (179)	−2990 (118)	−2610 (90.4)	−2540 (144)	−2250 (98.7)
Parameter med	2120	620, 623	119, 214, 223	0.415, 112, 81.2	41.9, 559, 45.1, 44.3

Blocks from top to bottom: full censoring, no interval-censoring, and no censoring. The table shows the mean (sd) number of samples per time series, the mean (sd) KLD, the mean (sd) LL, and the spatial median of the parameter estimates. Units of time are in seconds, and the entropy-based measures are in nats

exponentials, depending on the conditions (Muthukrishnan *et al.*, 2012). Our measurements were conducted under full induction (15 ng/ml of anhydrotetracycline) (Muthukrishnan *et al.*, 2012), imaging the cells for every 1 min for a duration of 60 min.

The histogram of observed intervals collected from the measurements is shown in the upper panel of Supplementary Figure S1 along with the PDFs of the estimated models. The intervals were extracted by analyzing the time series of individual cells separately (as opposed to observing production intervals in the whole cell lineage). By pooling the observed intervals from multiple cells, we implicitly assume that there are no significant variations in the model parameters between the cells. In addition, the lower panel of Supplementary Figure S1 shows the Turnbull's CDF estimate (Turnbull, 1976), which is a non-parametric ML estimate of the CDF accounting for both modes of censoring. Note that the histogram only contains the observed samples, so it is expected to underestimate (overestimate) the probability for large (small) values. On the other hand, the Turnbull estimator is expected to represent well the true CDF.

Table 2 shows statistics for the two methods. The number of samples is different in the two cases, as the full censoring method also includes the right-censored samples. Here, both methods suggest a three-exponential model and rule out the possibility of an on–off mechanism as the primary regulator of the dynamics. In both cases, the estimated parameter standard deviation is ~2 min per parameter, with full censoring resulting in a slightly higher confidence on the parameters. With full censoring, there is about 1.3-fold increase in the mean, which is expected, since neglecting the right-censored data tends to result in underestimation of the durations. To confirm the statistical significance of this difference, we performed a one-sided *t*-test with a null hypothesis that the two means are equal, resulting in a *P*-value of 5.20×10^{-9}. Also, we found differences in the noise levels, but both methods suggest cv-squared of <0.5, favoring the three-exponential model.

To demonstrate that the full censoring method is immune to changes in the measurement duration, we repeated the estimation procedures such that the time series was split into two halves of 30 min each, from which the data were extracted separately. The results are shown in the lower block of Table 2, and they indicate that with the full censoring method, only the confidence in the parameter estimates is visibly affected, whereas the non-censoring method underestimates the mean and standard deviation (even more than when applied to the time series 60 min long).

Table 2. Statistics of the estimated models for the tetA promoter

Method	Full censoring	No censoring
Samples	362	254
Sel model	seq-3	seq-3
Parameters	131, 131, 522	109, 254, 254
Parameter sd	109, 109, 54.2	119, 138, 139
Est mean (sd)	784 (554)	617 (375)
Est cv-squared	0.499	0.370
Samples	345	175
Sel model	seq-3	seq-3
Parameters	131, 131, 522	171, 171, 171
Parameter sd	171, 173, 76	134, 136, 137
Est mean (sd)	784 (554)	514 (297)
Est cv-squared	0.499	0.333

Blocks from top to bottom: 60 min series and 30 min series. The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model

Finally, we studied whether our results are affected by the elongation process. First, to test if significant RNA polymerase traffic occurs (e.g. due to pausing), we estimated the correlation between consecutive transcription intervals. We found a correlation coefficient of 0.149 with a *P*-value of 0.196 in an LR test with a null model of uncorrelated normal data, indicating that the correlation is not significant. Next, we added a normal zero-mean noise term to the transcription model to simulate stochasticity resulting from chain elongation. The estimated sd of this noise term was 24.4 s (*P*-value of 0.574 in an LR test with a null model from Table 2), suggesting that such noise term is not significant at our resolution. As neither of the null hypotheses can be rejected, we conclude that there is no evidence that the dynamics of elongation affects our results. In addition, in agreement, we note that no differences have been found in the dynamics of RNA production between constructs with 48 and 96 of the MS2-GFP binding sites (Golding *et al.*, 2005; Hakkinen *et al.*, 2014).

4.2.2 Transcription kinetics of the λ RM promoter

Next, we analyzed measurements of the transcription intervals of the MS2-GFP-tagged target RNA controlled by the bacteriophage λ

Table 3. Statistics of the estimated models for the bacteriophage λ RM promoter

Method	Full censoring	No censoring
Samples	303	155
Sel model	onoff	onoff-2
Parameters	5840, 1730, 1140	782, 2450, 527, 25.5
Parameter sd	3020, 469, 134	532, 2880, 90.7, 11.0
Est mean (sd)	4990 (8360)	721 (864)
Est cv-squared	2.81	1.44
Est burst size (interval)	1.52 (7560)	4.65 (3350)
Est duty cycle	0.228	0.768

The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model. Also shown is the burst size, burst interval and the duty cycle of the estimated model

RM promoter. This construct is expected to result in a bursty, highly noisy expression (Golding *et al.*, 2005). In this case, the cells were imaged every 1 min for a duration of 120 min.

Again, the histogram of the observed intervals and the PDFs of the estimated models are shown in the upper panel of Supplementary Figure S2, and the Turnbull's CDF estimate and the CDFs of the models are shown in the lower panel. Similarly, Table 3 shows statistics from the two estimation procedures.

Both methods indicate that an on–off model is required to explain the measurements, and that the interval distribution is highly noisy (cv-squared above unity). However, in the case where the unobserved samples (which constitute ~50% of the samples) are neglected, both the mean and sd of the distribution are drastically (over 5-fold) underestimated. In terms of cv-squared, this results in a 2-fold underestimation of stochasticity in the transcription initiation process.

Finally, we computed the statistics of the bursts in the two estimated models, which are shown in Table 3. The model estimated using full censoring suggests that the noise results from a low-duty cycle (i.e. the gene being repressed most of the time), while the model estimated without censoring suggests that the noise is due to the large size of the bursts.

4.3 Applying the methods on measurements of external transcription activation times

Finally, we analyzed the activation dynamics of the arabinose BAD promoter to demonstrate that the methods generalize to other estimation problems. The analysis was performed by collecting both the time for each cell to produce the first RNA after introducing arabinose in the medium, and the subsequent time intervals between consecutive productions of transcripts. For this, 1% of L-arabinose was introduced at the start of the measurement, after which cells were imaged every 1 min for 120 min.

The time to produce the first RNA is expected to include the time for the cell to uptake sufficient arabinose to turn on the active arabinose uptake system, for the arabinose to form the complex activating the BAD promoter (Daruwalla *et al.*, 1981), and for the first transcript to be produced. As such, after Megerle *et al.* (2008) and Makela *et al.* (2013), we fit the following model

$$\underbrace{A_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} A_{\text{on}} \xrightarrow{k_1}}_{\text{activation}} P_{\text{on}} \underbrace{\xrightarrow{k_2} I_2 \xrightarrow{k_3} I_3 \xrightarrow{k_4} P_{\text{on}} + E}_{\text{transcription}}, \qquad (8)$$

where $A_{\text{on}}, A_{\text{off}} \in \mathbb{Z}_1$ represent the states that the internal arabinose concentration has or has not reached sufficient concentration to turn on the arabinose uptake mechanism, respectively, and $P_{\text{on}}, I_j \in \mathbb{Z}_1$ represent the states of the BAD promoter. Since the intervals were found to have low noise (cv-squared of 0.347), we model transcription with a sequence of three exponentials.

Table 4. Statistics of the estimated models for the BAD promoter

Distribution	First production	Activation	Transcription
Samples	599	–	345
Model	seq-5	seq-2	seq-3
Parameters	(see right)	33.2, 1580	9.64, 1110, 1750
Parameter sd	(see right)	124, 201	39.1, 469, 512
Est mean (sd)	4490 (2610)	1620 (1580)	2870 (2080)
Est cv-squared	0.338	0.960	0.521

The table shows the number of samples, the selected model (sel model), the estimated parameters and estimates of their standard deviation and the mean (est mean), standard deviation (est sd) and squared coefficient of variation (est cv-squared) resulting from the estimated model

In the above model, the times for the first RNA production follow an seq-5 model with the parameters (p^-, p^+, k_2, k_3, k_4), where $p^\pm$ are as in Equation (3), and the intervals of the subsequent productions follow a seq-3 model with the parameters (k_2, k_3, k_4). As the models share parameters, they are fit jointly to both data. The histogram of the observed data and the PDFs of the estimated models are shown in the upper panels of Supplementary Figure S4, and the Turnbull's CDF estimates and the CDFs of the models are shown in the lower panels. Table 4 shows statistics from the estimation procedure. The exponential-likeness of the activation process suggests that either k_{on}, k_1 or both must be fast (non-rate limiting). Meanwhile, in transcription, two of the rates k_2, k_3 and k_4 are rate limiting.

The mean (sd) of the observed first production times and transcription intervals were 3880 s (1700 s) and 1700 s (1000 s), respectively, which agrees with those reported in Makela et al. (2013). Again, this suggests that neglecting the unobserved data results in slight underestimation of the mean and the variance. Regardless, the qualitative results, such as noise, reported in Makela et al. (2013) appear to hold.

It is worth noting that, to avoid artificial correlations between the first and the subsequent production times, Makela et al. (2013) used a windowing method to compare the activation dynamics in different conditions. In our method, such windowing is not needed, as the censored data is used. Because of this, in addition to using more information, our method also allows unbiased quantification of the different distributions, not just their comparison. Furthermore, our method can also be used to deconvolve the activation dynamics distribution, as shown in Table 4.

5 Discussion

We have proposed a model that combines a promoter on–off mechanism (Peccoud and Ycart, 1995) with a sequential process of transcription initiation (McClure, 1985), which allows explaining recent measurements of transcription dynamics under a wide range of conditions (Golding *et al.*, 2005; Kandhavelu *et al.*, 2011; Muthukrishnan *et al.*, 2012; Taniguchi *et al.*, 2010; Yu *et al.*, 2006), and established methods to estimate its parameters in maximum likelihood sense using transcription interval data.

The methods enable more accurate quantification of the transcriptional dynamics both in theory and in practice, as demonstrated by the Monte Carlo simulations, as well as testing if particular components of the model are responsible for the observed dynamics. In addition, we compared the methods with previous methods using measurement data from live *E. coli*, and showed that the new methods produce significantly different results and can provide new biological insight (e.g. on the underlying sources of noise in transcription). Finally, we demonstrated that the methods have a wider applicability on problems of similar nature, such as estimating the kinetics of external activation of a promoter.

In its present form, the proposed model should already be detailed enough to allow a genome-wide analysis of transcription and transcription activation of individual genes under a wide range of conditions in prokaryotes, which is necessary to understand how inducers and repressors regulate the dynamics of gene expression. Further, we believe that our methods can be extended to enable future studies of eukaryotic transcription dynamics and of translational dynamics at the single protein level.

Funding

The work was supported by the Jenny and Antti Wihuri Foundation [to A.H.] and Academy of Finland [257603 to A.S.R.].

Conflict of Interest: none declared.

References

Bernstein,J.A. *et al*. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 9697–9702.

Blake,W.J. *et al*. (2003) Noise in eukaryotic gene expression. *Nature*, **422**, 633–637.

Chong,S. *et al*. (2014) Mechanism of transcriptional bursting in bacteria. *Cell*, **158**, 314–326.

Daruwalla,K.R. *et al*. (1981) Energization of the transport systems for arabinose and comparison with galactose transport in *Escherichia coli*. *Biochem. J.*, **200**, 611–627.

Golding,I. *et al*. (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Hakkinen,A. *et al*. (2014) Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics*, **30**, 1146–1153.

Herbert,K.M. *et al*. (2006) Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, **125**(6), 1083–1094.

Kandhavelu,M. *et al*. (2011) *In vivo* kinetics of transcription initiation of the lar promoter in *Escherichia coli*. Evidence for a sequential mechanism with two rate-limiting steps. *BMC Syst. Biol.*, **5**, 149.

Lutz,R. *et al*. (2001) Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Res.*, **29**, 3873–3881.

Makela,J. *et al*. (2013) *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Res.*, **41**, 6544–6552.

McClure,W.R. (1985) Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, **54**, 171–204.

Megerle,J.A. *et al*. (2008) timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys. J.*, **95**, 2103–2115.

Muthukrishnan,A.B. *et al*. (2012) Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells. *Nucleic Acids Res.*, **40**, 8472–8483.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

Peccoud,J. and Ycart,B. (1995) Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, **48**, 222–234.

Rajala,T. *et al*. (2010) Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput. Biol.*, **6**, e.1000704.

Saecker,R.M. *et al*. (2011) Mechanism of bacterial transcription initiation. *J. Mol. Biol.*, **412**, 754–771.

Taniguchi,Y. *et al*. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.

Turnbull,B.W. (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B*, **38**, 290–295.

Yu,J. *et al*. (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.