BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 69–75
doi:10.1093/bioinformatics/btu592

Gene expression

Advance Access publication September 3, 2014

Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data

Antti Häkkinen and Andre S. Ribeiro*

Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group, Department of Signal Processing, Tampere University of Technology, P.O. box 553, 33101 Tampere, Finland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MS2-GFP-tagging of RNA is currently the only method to measure intervals between consecutive transcription events in live cells. For this, new transcripts must be accurately detected from intensity time traces.

Results: We present a novel method for automatically estimating RNA numbers and production intervals from temporal data of cell fluorescence intensities that reduces uncertainty by exploiting temporal information. We also derive a robust variant, more resistant to outliers caused e.g. by RNAs moving out of focus. Using Monte Carlo simulations, we show that the quantification of RNA numbers and production intervals is generally improved compared with previous methods. Finally, we analyze data from live *Escherichia coli* and show statistically significant differences to previous methods. The new methods can be used to quantify numbers and production intervals of any fluorescent probes, which are present in low copy numbers, are brighter than the cell background and degrade slowly.

Availability: Source code is available under Mozilla Public License at http://www.cs.tut.fi/%7ehakkin22/jumpdet/.

Contact: andre.ribeiro@tut.fi

Received on June 3, 2014; revised on July 14, 2014; accepted on August 27, 2014

1 INTRODUCTION

Transcription, translation and degradation of RNA and proteins are stochastic processes (Elowitz *et al.*, 2002; Kaern *et al.*, 2005; Ozbudak *et al.*, 2002; Taniguchi *et al.*, 2010; Yu *et al.*, 2006). Their stochasticity results in phenotypic variability in monoclonal cell populations (Elowitz *et al.*, 2002; Pedraza and van Oudenaarden, 2005) and temporal phenotypic changes in cells (Golding *et al.*, 2005; Yu *et al.*, 2006). As such, to better understand phenotypic diversity, the mechanisms regulating RNA and protein numbers must be understood.

Presently, the only technique for quantifying RNA numbers and transcription dynamics in live *Escherichia coli* cells over time consists of tagging target RNA molecules with an array of fluorescent MS2-GFP proteins (Golding and Cox, 2004). Using this technique, the target RNA can be visualized as a bright spot (see e.g. Fig. 6) shortly after production (Golding and Cox, 2004; Golding *et al.*, 2005). This method allows studying transcription in the absence of several sources of stochasticity (Kandhavelu *et al.*, 2012a), such as RNA degradation (Taniguchi *et al.*, 2010) or dilution by cell division (Huh and Paulsson, 2011).

The quantification of transcription dynamics from fluorescence microscopy images requires estimation of the RNA numbers or the RNA production times using statistics extracted from microscopy images, such as temporal intensity signals. Some methods have been proposed for determining RNA numbers using cell and/or fluorescence spot intensities, either using manually assisted (Golding *et al.*, 2005) or automatic (Hakkinen *et al.*, 2014) techniques. However, these methods were designed to extract stationary RNA distributions from cell populations, and as such, they neglect any temporal information in the data. Recently, we introduced a method (Kandhavelu *et al.*, 2012b) to use such information to extract time intervals between consecutive RNA productions in individual cells. This method uses a piecewise-constant monotonic least-squares (LSQ) fit with an *F*-test to select the model order, after which a jump in the model curve is taken to correspond to the production of a target RNA. Unfortunately, this method does not extract the absolute RNA numbers.

Here, we propose a new method for automatic quantification of RNA numbers and RNA production intervals from intensity time series extracted from cells. Specifically, we consider two variants: one that uses LSQ costs, which can be derived using the central limit theorem, and one that uses least-deviations (LD) costs, which is a robust variant of the former. In particular, the latter was designed to counter outliers commonly observed in the intensity data, caused by e.g. RNA molecules moving out of the focal plane, at the cost of reduced accuracy. The new method was designed to exploit the temporal information in the data for improved accuracy, to not require post- and/or pre-processing for time interval extraction and to be free of any regularization in temporal domain to allow more accurate quantification of time intervals.

First, we present Monte Carlo simulations to demonstrate that the accuracy of the method is, in general, superior to the existing methods, in estimating both RNA numbers and RNA production intervals. Second, we apply our method and the previous methods on novel data extracted from time-lapse microscopy measurements of live *E.coli* cells expressing MS2-GFP and RNA target to show that, for large number of cells, statistically significant differences in the results can be detected between the new and previous methods, in both RNA numbers and RNA production time intervals.

*To whom correspondence should be addressed.

2 METHODS

2.1 Cells and plasmids

Escherichia coli strain DH5α-PRO was generously provided by I. Golding (University of Illinois) and contains two constructs: a PROTET-K133 medium-copy vector carrying a MS2-GFP reporter, controlled by $P_{LtetO-1}$, and the pIG-BAC single-copy vector coding for mRFP1-MS2-96bs RNA, whose expression is controlled by $P_{lac/ara-1}$ (Golding and Cox, 2004).

2.2 Microscopy

Cells were grown in Miller lysogeny broth (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 OD600 of 0.3–0.5 was reached. To attain full induction of the MS2-GFP reporter, cells were incubated with 100 ng/ml of anhydrotetracycline (aTc, from IBA GmbH). In all, 0.1% of L-arabinose (Sigma-Aldrich) and 1 mM of Isopropyl-β-D-thiogalactopyranoside (IPTG, Fermentas) were used to induce the target RNA. Cells were pre-incubated with arabinose at the same time as aTc. IPTG was added 1 h after aTc, and cells were incubated for 5 min.

Microscopy was performed using a Nikon Eclipse (TE-2000-U, Nikon, Tokyo, Japan) inverted confocal laser scanning microscope. Cells were imaged in a thermal chamber set to 37°C. Images were taken 5 min after induction by IPTG, for a duration of 2 h, once per minute. For imaging, a few microliter of culture were placed between a coverslip and a slab of 1% agarose containing LB along with the appropriate concentrations of inducers. When both the reporter and the target RNA are present in the cells, MS2-GFP proteins bind to the target RNA, forming a bright fluorescent spot (Golding *et al.*, 2005). The RNA becomes visible during, or shortly after, elongation (Golding and Cox, 2004).

2.3 Image processing

Cells were detected from the microscope images using a semi-automatic method described in Kandhavelu *et al.* (2012b). First, a mask is manually painted over the area that a cell occupied during the time series. Principal component analysis is then used to obtain the dimensions and orientation of the cells from the fluorescence distribution within each mask. Next, target RNA spots were automatically segmented using kernel density estimation with a Gaussian kernel. Cell background-subtracted spot intensities were then calculated and summed for each cell to produce the total spot intensity within each cell, which was used to quantify RNA numbers.

3 ALGORITHM

3.1 Overview of the method

As each RNA is tagged by a large number (up to 96) of MS2-GFP molecules (Golding *et al.*, 2005), the intensities detected from a single RNA are expected to be well approximated by a normal distribution (central limit theorem). Consequently, sums of intensities of k RNA spots (e.g. in a cell) can be assumed to be normally distributed, with mean and variance k times that of the individual RNA, provided that the components are sufficiently independent (Hakkinen *et al.*, 2014).

Another assumption exploited by our method is that the MS2-GFP-tagged RNA molecules are virtually immortal during a cell lifetime [verified in Muthukrishnan *et al.* (2012)]. Because of this, one can assume that the RNA numbers form a non-decreasing series over time. For that, for example, the data before and after cell division must be treated as separate series.

We propose two variants of our method that use the same strategy but different intensity models. The LSQ variant uses normally distributed errors, as discussed above, but with constant (as opposed to linear) variance. We experimented with the linear variance model, but in the zero-noise limit, it approaches to the constant variance one, and the results were similar. Also, if additional noise sources affecting all the spots in an equal manner (regardless of their RNA numbers) are present, the model should be affine, i.e. somewhere between constant and linear.

The LSQ variant is similar to the method introduced in Kandhavelu *et al.* (2012b), which uses a piecewise-constant monotonic LSQ fit with an F-test to select the model order, after which a jump in the fit curve is taken to correspond to a production of an RNA. However, there the jump sizes are regulated by the means of the F-test (and the monotonicity constraint). Moreover, the F-test also causes regularization in the temporal direction. In our method, regularization is performed via constant jump size and the monotonicity constraint, to avoid regularization in the temporal domain.

Meanwhile, the LD variant was conceived to mitigate the problem that the RNA intensity time series sometimes contain 'holes', caused by, for example, RNA spots moving out of the focal plane (see e.g. the lower panel of Fig. 6). This variant uses the median as the location estimator, making it more robust than LSQ. If no such outliers are present, the LSQ should be preferred, as it is expected to be more accurate.

Regardless of the variant, the method operates as follows:

(1) A curve (without quantization) is fit to the intensity time series. This groups the related samples to reduce uncertainty, extracting the temporal information from the data. If the data are not temporal, this step is effectively a no-operation.

(2) Jump size (and other parameters such as uncertainty) is estimated from the fit pieces. We provide a new set of estimators for this step, but the one proposed in Hakkinen *et al.* (2014) could be used instead.

(3) A quantized curve is fit to the time series, given the parameters, enforcing the quantization to the fit. This is used to provide the RNA numbers.

3.2 Curve fitting

Piecewise-constant curve fitting can be done in polynomial time using a dynamic programming technique. Let $d(x, y)$ be some metric and $(x_i)_{i=1}^{n}$ be some sequence of length n, x_i denoting its ith element. Let $D_{a,b}^{k}$ be the distance between a substring $(x_i)_{i=a}^{b}$ of the input and a k-piece model:

$$D_{a,b}^{k} \doteq \sum_{i=a}^{p_1} d(x_i, \mu_1) + \ldots + \sum_{i=p_{k-1}+1}^{b} d(x_i, \mu_k) \tag{1}$$

which is to be minimized for $a = 1$, $b = n$ and some $k \leq n$:

$$\min_{\substack{p_1,\ldots,p_{k-1},\\ \mu_1,\ldots,\mu_k}} D_{1,n}^{k} = \min_{p_{k-1}} \left(\min_{\substack{p_1,\ldots,p_{k-2},\\ \mu_1,\ldots,\mu_{k-1}}} D_{1,p_{k-1}}^{k-1} + \min_{\mu_k} D_{p_{k-1}+1,n}^{1} \right) \tag{2}$$

which can be computed by memoizing the minima of each $D_{1,b}^{k}$ and $D_{a,b}^{1}$. Given that $D_{1,b}^{k-1}$ have already been minimized, it takes $\mathcal{O}(n^2)$ time to minimize $D_{1,b}^{k}$ for all b. If the minima of $D_{a,b}^{1}$ are computed for all a, b in time $T(n)$, $D_{1,n}^{k}$ can be minimized in

$T(n)+\mathcal{O}(k\,n^2)$ time and $\mathcal{O}(n)$ space. In the process, the solutions for all b and for all orders up to k are obtained. To fit the data with a monotonic function, the solutions violating the monotonicity constraint can be ignored in the minimization process.

In particular cases, a link between the choice of the metric $d(x, y)$ and maximum likelihood estimation can be established. If one assumes some piecewise-constant curve, corrupted by additive independent zero-mean normal distributed errors, a curve fit by the above procedure using squared error metric $d(x, y)=(x-y)^2$ corresponds to the maximum likelihood estimate of the signal. Similarly, there is a link between using absolute error metric $d(x, y)=|x-y|$ and Laplace distributed errors.

We implemented LSQ and LD metrics. The LSQ problem can be trivially implemented in $T(n)=\mathcal{O}(n^2)$ time, by computing running averages of the data. Similarly, the LD problem can be solved by computing running median, and can be implemented in $T(n)=\mathcal{O}(n^2\log n)$ time using priority queues with $\mathcal{O}(\log n)$ update times. In practice, we have been using these methods for up to 10 000 samples on a standard personal computer.

3.3 Jump size estimation

We estimate the jump size using a maximization of approximate likelihood in the low-noise limit. This method is expected to work when the noise-to-signal ratio of the data is reasonably small (see results for examples).

The probability density function of an equiprobable mixture of normal distributions with means at $k\,\mu$ for $k\in\mathbb{Z}$ and variance of σ^2 is as follows:

$$f(x)=\sum_{k=-\infty}^{\infty}\frac{C}{\sqrt{2\pi\,\sigma^2/w}}\exp\left(-\frac{(x-k\,\mu)^2}{2\,\sigma^2/w}\right) \tag{3}$$

where x is the intensity and w is used to scale the uncertainty of x. The constant C is selected such that the density integrates to unity. Constructing the infinite summation as a limit suggests that scaling of $C\propto\mu$ is appropriate. In the limit $\sigma^2\to 0$, the density $f(x)$ can be approximated (disregarding the scale) by the following:

$$g(x)=\frac{\mu}{\sqrt{2\pi\,\sigma^2/w}}\exp\left(-\frac{(x-K(x)\,\mu)^2}{2\,\sigma^2/w}\right) \tag{4}$$

For n independent and identically distributed samples, the approximate likelihood $\prod_{i=1}^{n}g(x_i)$ will be maximized when the squared coefficient of variation σ^2/μ^2 is minimized. This can be found by finding the roots of the partial derivatives of the function and verifying that they are maxima, yielding the following estimators:

$$\hat{\mu}=\frac{\sum_{i=1}^{n}w_i\,x_i^2}{\sum_{i=1}^{n}w_i\,K(x_i)\,x_i} \tag{5}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}w_i\,(x_i-K(x_i)\,\mu)^2 \tag{6}$$

The $K(x)$ that minimizes the approximation error is $K(x)=\lfloor x/\mu\rceil$, where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer. As $K(x_i)$s depend on μ, the problem is not yet solved. However, for each choice of $K(x)$ there is a range of associated μs involved. The problem can be solved by finding the μ and the likelihood for each set of $K(x)$, which is feasible if the values of $K(x)$ are small. Specifically, the solution can be computed incrementally in $\mathcal{O}(k\,n\log n)$ time and $\mathcal{O}(n)$ space using a priority queue with $\mathcal{O}(\log n)$ update times, where k is a bound for $K(x)$.

For large μ, $K(x)\to 0$ and the squared coefficient of variation, $\hat{\sigma}^2/\mu^2$ is around $\frac{1}{n}\sum_{i=1}^{n}w_i\,x_i^2/\mu^2$. On the other hand, for $\mu\to 0$, $x_i-K(x_i)\,\mu$ becomes a uniform random variable in $[-\frac{1}{2},\frac{1}{2})$ and $\hat{\sigma}^2/\mu^2\to\frac{1}{n}\sum_{i=1}^{n}\frac{1}{12}w_i$. Equating the two, an upper bound for the search can be obtained, to avoid the trivial solution of the large μ model. We have not yet devised any better stopping condition than to stop the search when $K(x)$ become sufficiently large [e.g. 100 if one considers typical RNA numbers in *E.coli* (Taniguchi *et al.*, 2010)].

A similar procedure can be applied for Laplace distributed rather than normal distributed data. The estimator for the absolute deviation b is as follows:

$$\hat{b}=\frac{1}{n}\sum_{i=1}^{n}w_i\,|x_i-K(x_i)\,\mu| \tag{7}$$

where $|\cdot|$ denotes the absolute value. The estimator for the location parameter μ is the multiplicative inverse of the $w_i\,x_i$ weighted median of $K(x_i)/x_i$, which is the minimizer of the coefficient of mean deviation $\hat{b}/\mu$.

As the Laplace distribution also has a density that is symmetric about the mode and decreases away from it, the $K(x)$ that minimizes the approximation error is as in the LSQ case. The asymptotic behavior for $\mu\to 0$ is that $\hat{b}/\mu\to\frac{1}{n}\sum_{i=1}^{n}\frac{1}{4}w_i$, and for large μ, the statistic $\hat{b}/\mu$ approaches a value of $\frac{1}{n}\sum_{i=1}^{n}w_i\,|x_i|/\mu$, which can again be used to obtain an upper bound for the search.

3.4 Quantization

The procedure of obtaining a fit curve with quantization, which is the third step in our method, depends on the choice of the error metric. In the case of squared error metric, a fit signal with quantization can be obtained from the fit performed without quantization. This is possible, as for $d(x, y)=(x-y)^2$ and some $Q(\mu)$:

$$\sum_{i=a}^{b}d(x_i,Q(\mu))=\sum_{i=a}^{b}d(x_i,\mu)+d(\mu,Q(\mu)) \tag{8}$$

if μ is a minimizer of $\sum_{i=a}^{b}d(x_i,\mu)$. The minimizer of $d(\mu,Q(\mu))$ is $Q(\mu)=\lfloor\mu/q\rceil\,q$, where q is the chosen quantization level, i.e. in the LSQ case, quantization can be performed by rounding the fit signal to the nearest multiple of the estimated jump size.

The same does not hold for the absolute error metric. For finding the quantized result, we use the curve fitting procedure again with an additional constraint, which imposes the quantization.

4 RESULTS

4.1 Monte Carlo simulations

We performed Monte Carlo simulations with various parameters of a general simple model of appearance of molecules inside cells.

In this model, molecules are generated with some intervals, whose durations are exponentially distributed. For reasons described in the methods, no degradation is modeled. Finally, the molecule numbers are corrupted by adding zero-mean independent and identically normal distributed noise.

First, we generated the true curves using exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ and then corrupted them by adding normally distributed noise with a coefficient of variation of $\sigma\mu^{-1}$ of 0.5, 1 or 2. A total of 100 series were used for analysis in each case, which was performed using the LSQ estimator. Figure 1 shows example time series that were generated by sampling every 10 s for 2 h. For quantifying the performance of the method, we estimated its accuracy, i.e. the proportion of correct RNA estimates in all estimates. The measured accuracies from each simulation were 0.992, 0.965 and 0.880 for $\sigma\mu^{-1}$ of 0.5, 1 and 2, respectively. In this setting, the accuracy remains good even for moderate noise levels, such as for $\sigma\mu^{-1}=2$, despite the fact that the jump size estimator was derived in the zero-noise limit assumption.

Next, in Figure 2, we show the results of testing for sampling frequency of $(1\,\text{min})^{-1}$ with noise level of $\sigma\mu^{-1}=1$ and using both the LSQ and LD estimators. In addition, samples were zeroed independently with a 25% probability to corrupt the data further (bottom case in Fig. 2), so as to test the resistance of the LD estimator to 'holes' (for reference, the LSQ estimator breaks around 1% of zeros in this case). Other parameters were as specified in the previous paragraph. The measured accuracies were 0.834, 0.808 and 0.621, for LSQ and LD without zeroing, and LD with zeroing, respectively. In these cases, the performance is worse, as less data are provided to the method, but the results remain likely useful.

Next we quantified the accuracies for the LSQ method using 1000 independent simulations. We used two sampling settings (10 s and 1 min intervals for a duration of 2 h) and exponential production intervals with a rate of $(15\,\text{min})^{-1}$. For the tests, various noise levels $\sigma\mu^{-1}$ and number of series were used. In addition to the LSQ method, we computed an upper bound for the accuracy of any method that uses only the intensity values and no temporal information–this represents a comparison with a previous method proposed in Hakkinen *et al*. (2014). The results for the 10 s sampling intervals are shown in Figure 3 and for 1 min sampling interval are shown in Figure 4.

For 10 s sampling interval, the accuracy remains high (>0.8) for noise levels up to $\sigma\mu^{-1}=2.5$ for large number of series (> 100), or up to $\sigma\mu^{-1}=2$ for smaller numbers (10). Using >1000 series provides no significant improvement in accuracy. For lower noise levels, the accuracy is generally high, regardless of the number of series. In this setting, the performance of a non-time series (non-ts) method is not comparable, as the noise levels are high and the vast amount of temporal information available is neglected.

The results for 1 min sampling interval exhibit less differences between the two methods in our setting. The accuracy remains high (>0.8) for noise levels up to $\sigma\mu^{-1}$ for large number of series (>100) and up to $\sigma\mu^{-1}=0.75$ for small numbers (10). The figure also shows that around 100 samples are sufficient, and even smaller sample sizes can be used for lower noise levels, in this setting. For high noise levels, both methods are likely to perform poorly, whereas for small noise levels, both methods work well. Importantly, the new method is advantageous for moderate noise levels.

We also varied other parameters, such as the shape of the production interval distribution. Changing the production interval to a more noisy distribution (e.g. gamma distribution with shape parameter <1) results in reduced accuracy, whereas changing it to a lesser noisy distribution results in improved accuracy. Similarly, using the LD classifier for data generated using a model with normal errors results in slightly reduced accuracy when compared with the LSQ method.

Finally, we tested the performance of using our method in estimating the production interval distribution from the results of the fit. For comparison, we also present results using a previous method (Kandhavelu *et al*., 2012b). The goodness was assessed by computing the Kullback–Leibler divergence of the true distribution from that of the intervals extracted after applying one of the methods—this is the criterion minimized by a

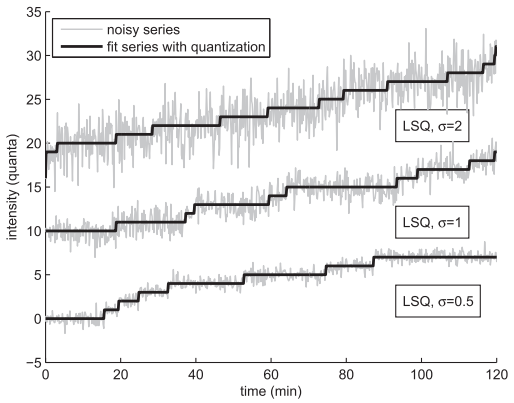

Fig. 1. Example results of the LSQ method with 100 series (first of which is shown) with a duration of 2 h and sampled every 10 s, for various noise levels. The series were generated using a jump size of $\mu = 1$ and exponential intervals with a rate of $(15\,\text{min})^{-1}$

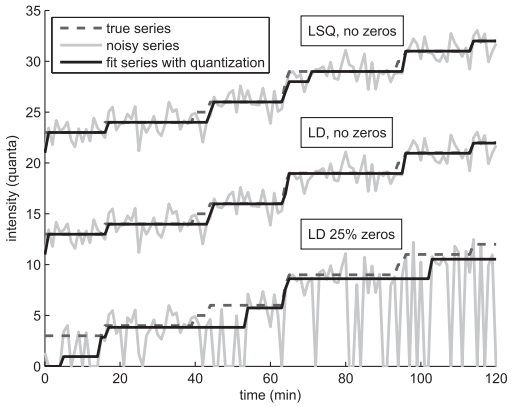

Fig. 2. Different methods tested on 100 series (first of which is shown), each with a duration of 2 h, sampled every 1 min. The series were generated using a jump size of $\mu = 1$, noise level of $\sigma = 1$ and exponential intervals with a rate of $(15\,\text{min})^{-1}$. In the bottom series, a sample is independently zeroed with a probability of 0.25. Most of the time, the true series is covered by the equal fit one

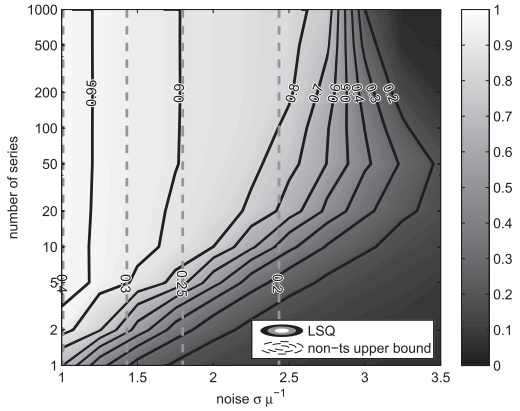

Fig. 3. Mean accuracy of the LSQ method applied to each of 1000 Monte Carlo simulations and the corresponding upper bound for a non-time series type method as a function of the noise $\sigma\mu^{-1}$ and the number of series. The plot was obtained with exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ for a duration of 2 h sampled every 10 s

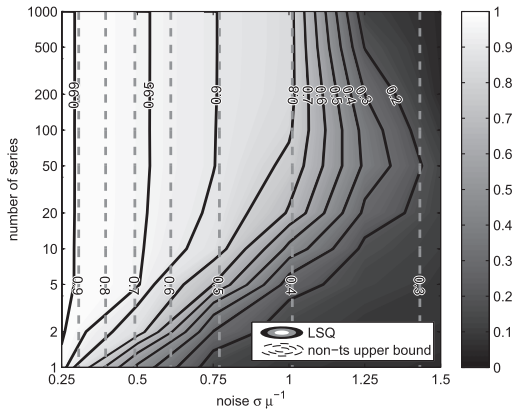

Fig. 4. Mean accuracy of the LSQ method applied to each of 1000 Monte Carlo simulations and the corresponding upper bound for a non-time series type method as a function of the noise $\sigma\mu^{-1}$ and the number of series. The plot was obtained with exponentially distributed production intervals with a rate of $(15\,\text{min})^{-1}$ for a duration of 2 h sampled every 1 min

maximum likelihood estimator. The divergence was computed using a parametric method.

For this, we first varied the shape and mean of the production interval distribution. This was done by using gamma distributed production intervals, with shape $\alpha \in \{0.5, 1, 2\}$, where $\alpha = 1$ yields an exponential distribution, and $\alpha < 1$ and $\alpha > 1$ yield less and more noisy distributions, respectively. The mean duration of the production interval was varied in the range $\beta^{-1} \in [1, 100]$ min. For each simulation, 100 series were generated for the duration of 2 h with sampling intervals of 1 min and noise level of $\sigma\mu^{-1} = 0.5$, and the results were averaged from 1000 simulations. The results for this case are shown in the upper panel of Figure 5. Next, we varied the noise $\sigma\mu^{-1} \in [0.1, 1]$ of the intensities along with the shape of the production interval distribution with a constant mean production rate of

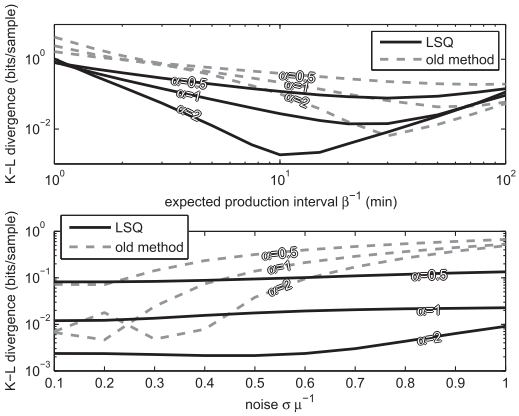

Fig. 5. Kullback–Leibler divergence estimated from 1000 Monte Carlo simulations for estimating the parameters of the interval distribution, as a function of mean production rate β^{-1} (upper panel) and as a function of intensity noise $\sigma\mu^{-1}$ (lower panel). The results were obtained with gamma distributed production intervals with shapes of $\alpha \in \{0.5, 1, 2\}$, for a duration of 2 h sampled every 1 min. Unless otherwise specified, the mean production rate is $\beta^{-1} = 15$ min and the noise is $\sigma\mu^{-1} = 0.5$

$\beta^{-1} = 15$ min. All other parameters were as specified in the previous case. These results are shown in the lower panel of Figure 5.

The results in Figure 5 suggest that, generally, the new method outperforms the previous method. The previous method was found to perform better with specific parameter ranges, particularly for more deterministic ($\alpha = 2$) production intervals, likely owing to tighter regularization. As expected, higher variance in the production interval results in reduced performance in all cases. Similarly, short or long production intervals result in reduced performance, as in the former case, both methods suffer from poor time resolution and in the latter from lack of observed intervals. In the case of low noise $\sigma\mu^{-1}$ in the intensities, both methods have similar performance; however, the performance of the new method is superior for moderate to high noise levels.

4.2 Statistics of tagged RNA numbers and intervals

Finally, we used our method to estimate statistics related to the production of MS2-GFP-tagged RNA molecules in live *E.coli* cells from time-lapse images obtained by confocal microscopy (see Section 2). Each cell contains one lac/ara-1 promoter expressing the target RNA, which consists of 96 binding sites for the MS2-GFP for visualization, preceded by a sequence coding for mRFP (see upper panel of Fig. 6). Target and reporter genes were induced as described in the methods, and images were taken 60 min after induction, sampled every 1 min for a duration of 2 h. The image analysis procedure (see Section 2.3) of three independent experiments performed in the same conditions produced 503 cells with a total of 24 466 intensity samples.

For comparison, we also used the two previous methods used in the previous section. The first, proposed in Hakkinen *et al.* (2014), does not use the temporal information as our method does. Consequently, we expected it to yield less accurate results, as suggested by results on the Monte Carlo simulations shown in the previous section. The other previous method was proposed

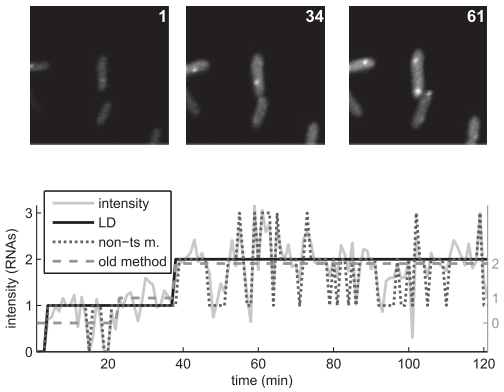

Fig. 6. Upper panel: Regions of confocal microscope images at 1, 34 and 61 min after starting the imaging. Lower panel: example intensity series and fit curves using the LD method and the two previous methods. The curves are shown in the space of RNA numbers estimated by the LD method; corresponding RNA numbers using the old method are shown on the right. The centermost cell of the images in the upper panel corresponds to that shown in the lower

and first used in Kandhavelu *et al.* (2012b). Again, based on the results from the previous section, we expect our method to be slightly more accurate than this method.

An example of a time series of cell fluorescence intensities extracted from the microscopy data is shown in the lower panel of Figure 6. The RNA curves fit using each of the three methods for the example series are also shown. Visibly, the intensity time series contains moderate amount of noise and occasional samples with small values (cf. the bottom curve in Fig. 2). Because of the latter observation, we opted to use the LD variant for the analysis.

The fit curves in the lower panel of Figure 6 exemplify how the three methods operate using different strategies. Without any pre- or post-processing, the non-ts method seems ill-suited. Also, the example shows how the LD method is regularized in intensity space and the previous method in temporal space. Nevertheless, the results from the two methods tend to generally agree with each other—both insert RNA productions when large increases in the intensity series occur.

We compared the values of RNA numbers extracted using the three methods from the data of the experiments. For the purposes of comparison, we extracted the RNA numbers averaged over each cell and each point in time. As these numbers appeared to be similar, we performed a Kolmogorov–Smirnov (K-S) test to assess whether the RNA numbers extracted using either of the previous methods differ from that of our LD method, in a statistical sense. In this test, the null hypothesis is that the two sets of data are generated from an equal distribution, and the alternative hypothesis is that the distributions are unequal. As the data are discrete and possibly correlated over time, the test was performed by permuting the series (rather than individual samples) for a total of 10^6 times. The results are summarized in Table 1.

For the obtained time series of intensities, we also extracted the RNA production intervals using both our LD method and the previous method designed for production interval extraction (Kandhavelu *et al.*, 2012b). The mean and the squared coefficient of variation of the extracted intervals are shown in Table 2.

Table 1. Statistics of RNA numbers

Method	Cells	Samples	Mean	K-S *P*-value
LD	503	24 466	1.43	N/A
Non-ts method	-"-	-"-	1.49	2.0×10^{-5}
Old method	-"-	-"-	1.44	2.8×10^{-4}

Notes: The table lists the number of cells, number of RNA samples, mean RNA numbers and *P*-value of the K-S test when comparing with the LD method.

Table 2. Statistics of time intervals between RNA productions

Method	Intervals	Mean (min)	Squared-CV	K-S *P*-value
LD	373	15.42	0.82	N/A
Old method	344	16.09	0.52	2.1×10^{-4}

Notes: The table lists the number, mean, and squared coefficient of variation (squared-CV) of the extracted intervals, and the *P*-value of the K-S test when comparing to the LD method.

Also, a K-S test was performed to assess the statistical significance of the differences between the interval distributions.

In summary, the three methods extracted similar but statistically distinct RNA numbers and durations between the productions of consecutive RNA molecules from the data. Also, in all cases, the results agree with those reported in Kandhavelu *et al.* (2012a). In particular, while the mean numbers these quantities appear similar, the K-S tests were able to detect significant differences at the level of resolution of the measurements. This, along with the evidence presented in the previous section, suggests that our method offers significant improvements over the methods previously used for such studies. Interestingly, the results using the LD method suggest slightly noisier shape of the distribution than previously reported (Kandhavelu *et al.*, 2012a). This is likely to be of relevance to studies of the mechanisms underlying transcription in live cells. Further, might suggest that our new method produces more accurate results by avoiding the regularization in the temporal domain.

5 DISCUSSION

We have presented two variants of a novel, more accurate method for the automatic quantification of RNA numbers and RNA production intervals from intensity time series extracted from images of live cells expressing fluorescently tagged RNA molecules. The new method exploits the temporal information in the data for improved accuracy, does not require post- and/or pre-processing for time interval extraction and has no regularization in the temporal domain.

One of the proposed variants uses LSQ costs, which can be derived using the central limit theorem. The other uses LD costs, which is a robust variant of the former, and is to be used when there is potential for outliers (e.g. spots transiently leaving the focal plane of the microscope). Meanwhile, the former is preferred when no such corruption is present (e.g. if using multiple slices along the *z*-axis at each time point).

We used Monte Carlo simulations to demonstrate that the accuracy of the new methods is, in general, superior to that of our two previously proposed methods (Hakkinen *et al*., 2014; Kandhavelu *et al*., 2012b), both in estimating the RNA numbers and the RNA production intervals. We also applied the new and the previous methods on novel data from time-lapse images of live *E.coli* cells expressing RNA target for MS2-GFP to show that, if the data contains large number of cells, statistically significant differences in the results can be detected, in both the RNA numbers and RNA production time intervals. In this regard, it should be noted that such 'large' numbers of cells are required to, e.g. compare changes in the dynamics of RNA production by a promoter when under different temperatures or levels of induction (Kandhavelu *et al*., 2012a; Makela *et al*., 2013).

Currently, MS2-GFP-tagging of RNA molecules is the only existing method for detecting RNA molecules, as they are produced in live cells. An accurate quantification of the copy numbers of these tagged RNA molecules and determination of the moments when they first appear are essential, particularly in studies of the dynamics of transcription in live cells [see e.g. Muthukrishnan *et al*. (2012)]. Such measurements are currently being used to assess, for example, the role of induction and repression mechanisms in regulating gene expression dynamics or the effects of environmental factors such as temperature on this dynamics. However, detection and quantification of individual, tagged RNA molecules, and thus our method, are valuable to other endeavors as well. For example, recent uses of the counting of such molecules as they appear in cells include a study of the dynamics of small genetic circuits (Chandraseelan *et al*., 2013) and a study of errors in the partitioning of RNA molecules in cell division (Lloyd-Price *et al*., 2012).

The two variants of the method proposed here should prove valuable in increasing the accuracy of these, as well as of other studies making use of fluorescent molecules, provided that the fluorescent molecules exist in small numbers in each cell, that their fluorescence is significantly above the background fluorescence and that they have a slow degradation rate when compared with dilution caused by cell division.

Funding: Work supported by Jenny and Antti Wihuri Foundation [to A.H.]; Academy of Finland [257603 to A.S.R.]; Tekes [40226/12 to A.S.R.]; and Fundacao para a Ciencia e Tecnologia [PTDC/BBB-MET/1084/2012 to A.S.R.].

Conflict of interest: none declared.

REFERENCES

Chandraseelan,J.G. *et al*. (2013) Temperature dependence of the LacI-TetR-CI repressilator. *Mol. Biosyst*., **9**, 3117–3123.

Elowitz,M.B. *et al*. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Golding,I. and Cox,E.C. (2004) RNA dynamics in live *Escherichia coli* cells. *Proc. Natl Acad. Sci. USA*, **101**, 11310–11315.

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.

Huh,D. and Paulsson,J. (2011) Random partitioning of molecules at cell division. *Proc. Natl Acad. Sci. USA*, **108**, 15004–15009.

Kaern,M. *et al*. (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet*., **6**, 451–464.

Kandhavelu,M. *et al*. (2012a) Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett*., **586**, 3870–3875.

Kandhavelu,M. *et al*. (2012b) Single-molecule dynamics of transcription of the lar promoter. *Phys. Biol*., **9**, 026004.

Lloyd-Price,J. *et al*. (2012) Probabilistic RNA partitioning generates transient increases in the normalized variance of RNA numbers in synchronized populations of *Escherichia coli*. *Mol. BioSyst*., **8**, 565–571.

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.

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.

Ozbudak,E.M. *et al*. (2002) Regulation of noise in the expression of a single gene. *Nat. Genet*., **31**, 69–73.

Pedraza,J.M. and van Oudenaarden,A. (2005) Noise propagation in gene networks. *Science*, **307**, 1965–1969.

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

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