BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1146–1153
doi:10.1093/bioinformatics/btt766

Gene expression

Advance Access publication January 2, 2014

Estimation of fluorescence-tagged RNA numbers from spot intensities

Antti Häkkinen, Meenakshisundaram Kandhavelu, Stefania Garasto and Andre S. Ribeiro*

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

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Present research on gene expression using live cell imaging and fluorescent proteins or tagged RNA requires accurate automated methods of quantification of these molecules from the images. Here, we propose a novel automated method for classifying pixel intensities of fluorescent spots to RNA numbers.

Results: The method relies on a new model of intensity distributions of tagged RNAs, for which we estimated parameter values in maximum likelihood sense from measurement data, and constructed a maximum a posteriori classifier to estimate RNA numbers in fluorescent RNA spots. We applied the method to estimate the number of tagged RNAs in individual live *Escherichia coli* cells containing a gene coding for an RNA with MS2-GFP binding sites. We tested the method using two constructs, coding for either 96 or 48 binding sites, and obtained similar distributions of RNA numbers, showing that the method is adaptive. We further show that the results agree with a method that uses time series data and with quantitative polymerase chain reaction measurements. Lastly, using simulated data, we show that the method is accurate in realistic parameter ranges. This method should, in general, be applicable to live single-cell measurements of low-copy number fluorescence-tagged molecules.

Availability and implementation: MATLAB extensions written in C for parameter estimation and finding decision boundaries are available under Mozilla public license at http://www.cs.tut.fi/%7ehakkin22/estrna/.

Contact: andre.ribeiro@tut.fi

Received on October 28, 2013; revised on December 9, 2013; accepted on December 25, 2013

1 INTRODUCTION

The processes of production and degradation of RNA and protein numbers are stochastic, which causes these numbers to differ between individuals (Elowitz *et al.*, 2002; Kaern *et al.*, 2005; Ozbudak *et al.*, 2002; Taniguchi *et al.*, 2010; Yu *et al.*, 2006) and to vary over time in individual cells, even under optimal stable conditions (Golding *et al.*, 2005; Kandhavelu *et al.*, 2012b). Present measurement techniques to study these processes in live cells rely on the usage of fluorescent proteins (Montero Llopis *et al.*, 2010; Raj *et al.*, 2008), and consequently on fluorescent microscopy and image processing methods, to extract the relevant statistical data.

One of the most accurate techniques to quantify gene expression dynamics *in vivo* consists of tagging target RNA molecules with multiple MS2d-GFP proteins, which makes the target RNA transcripts visible as bright spots (Golding and Cox, 2004), as soon as these are transcribed (Golding and Cox, 2004; Peabody, 1993). The present method of quantifying the RNA numbers in cells or in the fluorescent RNA spots relies on a manual selection of the intensity of a single-tagged RNA using the histogram of fluorescence intensities (Golding *et al.*, 2005). Following this selection, this method assumes that the peaks of spots intensities are concentrated at multiples of the expected intensity of a single RNA (Golding *et al.*, 2005), so as to quantify the number of RNA molecules in spots of varying fluorescence intensities.

This method of quantification is not optimal for several reasons. First, it relies on human intervention, and the manual selection can have a major effect in the results of the process of counting the RNA molecules. This hampers comparison between results because the selection of the intensity of the first peak varies between users. Second, the distribution of the number of RNA molecules composing the spots is not uniform, and thus the optimal quantification of these numbers is not achieved by simple rounding. Third, the variance in intensities of the spots is expected to increase with the number of tagged RNAs composing the spot. Fourth, the tagged RNA molecules tend to accumulate at the poles of the cells, and one cannot rely on spatial separation between them (Golding and Cox, 2004; Lloyd-Price *et al.*, 2012). Lastly, for large datasets, manually assisted quantification can be excessively laborious.

Here, we propose an automatic method that improves on the RNA quantification from fluorescence images. For this, we establish a mathematical model of the intensities, which is free from any assumption on the shape of the distribution of the RNAs, as this distribution is the subject of the study. Also, we propose methods to estimate the parameters of the model and construct a classifier. We then exemplify the usage of the method in determining the number of RNAs in clusters of MS2-GFP-tagged RNA molecules in *Escherichia coli* under various conditions. In particular, we compare results of analyzing cells with one of two target RNA constructs that differ in the number of binding sites for MS2-GFP proteins. In addition, the results are compared with a method that uses temporal information for the RNA quantification, and results from cells subject to different

*To whom correspondence should be addressed.

levels of induction of target RNA are compared with quantitative polymerase chain reaction (qPCR) measurements. Lastly, we study the performance of the classifier on simulated data, whose ground truth is known, and compare it with that of the previous method.

2 SYSTEM AND METHODS

2.1 Cells and plasmids

Escherichia coli strain DH5α-PRO was provided by I. Golding (University of Illinois) and contains two constructs: a PROTET-K133 medium-copy vector carrying a MS2d-GFP reporter, controlled by $P_{\mathrm{LtetO-1}}$, and the pIG-BAC single-copy vector coding for mRFP1-MS2-96bs RNA, whose expression is controlled by $P_{\mathrm{lac/ara-1}}$ (Golding and Cox, 2004).

The second construct was engineered by us, and it was designed to contain 48 binding sites (bs) for MS2-GFP. The target-gene vector pMK-BAC, containing $P_{\mathrm{lac/ara-1}}$ with a 48 MS2-GFP binding site array in a single-copy bacterial artificial chromosome (BAC), was constructed using standard molecular biology methods. $P_{\mathrm{lac/ara-1}}$-48bs was amplified with smaI restriction endonuclease from a BAC clone carrying a target gene $P_{\mathrm{lac/ara-1}}$-mRFP1-96bs. The primers (forward: 5′ CCCGGGGAAGACATGAGGATCA 3′ and reverse: 5′ CCCGG GTCAATTCTGTGTGAAATTG 3′) were designed to amplify the $P_{\mathrm{lac/ara-1}}$-48bs with smaI restriction site flanking regions. The amplicon and the BAC vector were subjected to smaI restriction digestion, followed by ligation of the amplified product. We obtained a single-copy F-based plasmid carrying the target region $P_{\mathrm{lac/ara-1}}$ with a 48bs array. This product was transformed into the competent *E.coli* strain DH5α-PRO. The recombinants were selected with antibiotic screening and confirmed with sequence analysis.

2.2 Microscopy measurements

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 of OD600 of 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). In all, 0.1% of L-arabinose (Sigma-Aldrich) and 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG, Fermentas) were used to fully induce the target RNA. In one case, IPTG was not added, so that the target gene remains only weakly induced. Cells were preincubated with arabinose at the same time as aTc. IPTG was added (if 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. Single time-point images were taken 1 h after induction by IPTG (if induced). For time series, images were taken 5 min after induction by IPTG, for 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, MS2d-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 The qPCR analysis of the target RNA

The target RNA was induced as described earlier in the text. Following induction, the cells were immediately fixed with RNAprotect bacteria reagent, followed by enzymatic lysis with Tris-EDTA lysozyme buffer (pH 8.3). From the lysed cells, total RNA was isolated with RNeasy RNA purification kit (Qiagen), according to the manufacturer's instructions. DNaseI treatment was performed to avoid DNA contamination. The complementary DNA (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 ribosomal RNA housekeeping gene for internal reference. The primers for the target mRNA (forward: 5′ TACGACGCCGAGGTCAAG 3′ and reverse: 5′ TTGTGGGA GGTGATGTCCA 3′) target the mRFP1 coding region and the reference gene 16S ribosomal RNA (forward: 5′ CGTCAGCTCGTGTTGTGAA 3′ and reverse: 5′ GGACCGCTGGCAACAAAG 3′). The experiment was performed using a Biorad MiniOpticon real-time PCR system (Biorad, Finland) with the following thermal cycling protocol: 40 cycles of 95°C for 10 s, 52°C for 30 s and 72°C for 30 s for each cDNA replicate. Reactions were performed in two experiments, each with two replicates per condition with a final reaction volume of 50 μl. Nonspecific signals and contamination were crosschecked using no reverse transcriptase and no template controls. PCR efficiencies of the reactions were >95%. The CFX Manager Software was used to calculate relative expression, whereas standard errors were calculated as in Livak and Schmittgen (2001).

2.4 Image processing

The individual frames of cells were analyzed as follows. First, the cells in the images were segmented using an automatic method (Chowdhury *et al.*, 2013). Next, the each cell intensity is fit to a surface, which is a quadratic polynomial of the distance from the cell border, in least-deviations sense, which is subtracted to obtain the foreground intensity. The foreground intensity is fit with a set of Gaussian surfaces, in least-deviations sense, with decreasing heights until the heights are in the 99% confidence interval of the background noise (estimated assuming a normal distribution and using median absolute deviation). The Gaussians are taken to represent spots, the volume under each representing the total spot intensity. Meanwhile, the volume under the whole foreground surface is taken to represent the total cell intensity. The time series analysis was performed as described in Kandhavelu *et al.* (2012b), from segmentation to spot intensity calculation and RNA estimation. The new procedure was found to perform similar but had lower noise in spot intensities than the method from Kandhavelu *et al.* (2012b).

3 ALGORITHM

3.1 Model of RNA spot intensities

The target RNA contains either 96 or 48 bs (earlier in the text) for MS2d-GFP molecules (Golding *et al.*, 2005). Not necessarily are all the binding sites occupied by GFPs at all moments, but it is reasonable to assume that a large number of the binding sites are occupied, as the MS2-GFP is highly abundant in the cells, and the spots are easily visible at almost all time. The observed intensity can also vary because of other reasons, such as variations of the molecule locations with respect to the focal plane.

If the amounts of light detected from a single GFP are independent and identically distributed, and the binding of MS2-GFP molecules are independent, occurring with a constant probability, the amount of light detected from a single-tagged

RNA should follow a binomially weighted mixture of sums of the amounts of lights detected from single MS2-GFPs. Regardless, if this is strictly true, given a large number of independent sources of light, the light detected from a single-tagged RNA can be well approximated by a normal distribution (central limit theorem), if the signal-to-noise ratio is low. If the signal-to-noise ratio was high, it would be possible to estimate the GFP numbers instead.

Letting the light detected from a single-tagged RNA to follow normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean of μ and variance of σ^2, the light emitted by k-tagged RNAs will be distributed according to $\mathcal{N}(k\,\mu, k\,\sigma^2)$, if the light emitted by the tagged RNAs are independent of one another. If the probability for finding a cluster of k-tagged RNAs is given by α_k, the mixture density takes the form as follows:

$$f_{\mathcal{M}}(x \mid \mu, \sigma^2, \alpha_1, \cdots, \alpha_\infty) = \sum_{k=1}^{\infty} \alpha_k f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2) \tag{1}$$

$$= \sum_{k=1}^{\infty} \frac{\alpha_k}{\sqrt{2\pi\,k\,\sigma^2}} \exp\left(-\frac{(x - k\,\mu)^2}{2\,k\,\sigma^2}\right) \tag{2}$$

where $f_{\mathcal{N}}(x \mid \mu, \sigma^2)$ is the density of a normal distribution with a mean of μ and variance of σ^2.

We do not wish to impose any constraints (model) on the distribution defined by α_k, which represents the RNA distribution, as this distribution is the subject of the study.

3.2 Parameter estimation

The form of the density of Equation (1) makes finding closed-form estimates for the parameters hard. We solve the problem by applying expectation maximization (EM) algorithm (Dempster *et al.*, 1977). The EM algorithm iteratively estimates new parameters $\boldsymbol{\theta}'$ using the 'incomplete' (observed) data $\boldsymbol{y}$, the 'complete' data $\boldsymbol{x}$ and the current parameter estimates $\boldsymbol{\theta}$ by maximizing:

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \mathbb{E}[\log f(\boldsymbol{x} \mid \boldsymbol{\theta}') \mid \boldsymbol{y}, \boldsymbol{\theta}] \tag{3}$$

where $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is the complete data density.

We denote the parameters by $\boldsymbol{\theta} \doteq (\mu, \sigma^2, \alpha_1, \cdots, \alpha_N)$. The log-likelihood function for $\boldsymbol{\theta}$ given the intensity observations $\boldsymbol{y} \doteq (y_1, \cdots, y_M)$ and the RNA numbers $\boldsymbol{k} \doteq (k_1, \cdots, k_M)$ is as follows:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{M} \log\left\{ \alpha_{k_i} f_{\mathcal{N}}(y_i \mid k_i\,\mu, k_i\,\sigma^2) \right\} \tag{4}$$

$$= \sum_{i=1}^{M} \log\alpha_{k_i} - \frac{1}{2}\log(2\pi k_i) - \log\sigma - \frac{(y_i - k_i\,\mu)^2}{2\,k_i\,\sigma^2} \tag{5}$$

where $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{k})$ is the complete data. The distribution of the missing parameter K_i under the parameters $\boldsymbol{\theta}$ is given by the following equation:

$$w_{k_i} \doteq \mathbb{P}[K_i = k_i \mid \boldsymbol{y}, \boldsymbol{\theta}] = \frac{\alpha_{k_i} f_{\mathcal{N}}(y_i \mid k_i\,\mu, k_i\,\sigma^2)}{\sum_{k'=1}^{N} \alpha_{k'} f_{\mathcal{N}}(y_i \mid k'\,\mu, k'\,\sigma^2)} \tag{6}$$

which yields the following form for $Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$:

$$Q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = \sum_{i=1}^{M} \sum_{k_i=1}^{N} w_{k_i}\, \ell(\boldsymbol{\theta}' \mid (y_i, k_i)) \tag{7}$$

The parameters $\boldsymbol{\theta}'$ maximizing Equation (7) can be found by finding the roots of the partial derivatives and verifying that the obtained point is a global maximum. The estimators are as follows:

$$\hat{\alpha}'_k = \frac{1}{M} \sum_{i=1}^{M} w_{k_i} \tag{8}$$

$$\hat{\mu}' = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M} \sum_{i=1}^{M} y_i \tag{9}$$

$$\hat{\sigma}'^2 = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{N} w_{k_i}\,(y_i - k\,\mu)^2 \tag{10}$$

where the variance estimator involves the true parameter μ. If the estimator $\hat{\mu}'$ is substituted for μ, Bessel's correction should be applied (or the variance will be underestimated on average).

Lastly, we note that in our case the EM iteration is guaranteed to converge to a maximum of the likelihood function (Wu, 1983). However, this maximum is not guaranteed to be the global maximum (Wu, 1983). For this reason, it is required that either the initial parameter values for the EM algorithm are close to the optimal values or multiple initializations are used.

An alternative iterative algorithm with reduced complexity can be derived by assuming some values of k_i estimating the parameters and by assigning new k_i by classifying the data under new estimate of the parameters. This algorithm is similar to the k-means clustering algorithm and is referred here as 'hard' EM, as opposed to the previous being 'soft' EM. Assuming k_i, the parameter estimates become (by letting $w_{k_i} = \mathbb{I}\{\,k_i = k\,\}$):

$$\hat{\alpha}'_k = \frac{1}{M} \sum_{i=0}^{M} \mathbb{I}\{\,k_i = k\,\} \tag{11}$$

$$\hat{\sigma}'^2 = \left(\sum_{k=1}^{N} k\,\hat{\alpha}'_k\right)^{-1} \frac{1}{M} \sum_{i=0}^{M} (y_i - k_i\,\mu)^2 \tag{12}$$

and for $\hat{\mu}'$ as in Equation (9). In these, $\mathbb{I}\{\cdot\}$ is the indicator function (unity if the condition is true, and zero otherwise), i.e. the estimator $\hat{\alpha}'_k$ is the fraction of items with k_i equal to k. With these, only $\mathcal{O}(M + N)$ work is required per iteration, instead of $\mathcal{O}(MN)$ of the soft EM algorithm, which is significant for large N.

Regardless of the algorithm used, an initial parameter estimate is required. We found the following scheme to be appropriate: (i) sort the observed data y_i and partition it into N bins, (ii) assign $k_i = j$ if y_i is in the jth bin and (iii) estimate initial parameters using Equations (9), (11) and (12). A good partitioning depends on the true values of α_k. We found that equidistant partitioning in y_i is simple, parameter-free and appeared to yield results equivalent to more complicated schemes (e.g. multiclass Otsu's method). It is noted that hard boundaries (as in the above scheme) cause the variance to be initially underestimated, if the overlapping of the clusters is large.

The parameter N can be found by finding the parameter estimates for several values of N and by selecting the model with least order that does fit significantly better than the lower-order models. To determine the significance, we use a likelihood ratio test, where the log-likelihoods ℓ_a and ℓ_b of models of orders a and b are obtained, respectively, and the statistic $-2(\ell_a - \ell_b)$ is computed. For $M \to \infty$, this statistic follows a χ^2 distribution, with $b - a$ degrees of freedom (Wilks, 1938), from which a P-value can be computed. If the P-value is smaller than a given significance level, the higher-order model should be favored over the lower-order one. We note that this procedure rarely selects a model of too high order (as determined by the significance level), but for small sample sizes it might lack evidence to select the appropriate high-order model.

3.3 Spot classification

Next, we construct a classifier that is used to estimate the number of RNAs in a cluster based on the intensity of the cluster. We use maximum a posteriori decision rule, i.e a class k, which is the most probable, is to be associated with an intensity value x:

$$C(x) \doteq \arg\max_k \mathbb{P}[K = k \mid x] \qquad (13)$$

$$= \arg\max_k \alpha_k f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2) \qquad (14)$$

where $\mathbb{P}[K = k \mid x]$ represents the posterior probability, and α_k the priors and $f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2)$ the likelihood functions of each class (the equality owing to the Bayes rule).

The classification can be performed by evaluating the term to be maximized for each k. Alternatively, a range of intensity values can be associated with each k. Possible decision boundaries can be obtained from the equation:

$$\alpha_a f_{\mathcal{N}}(x^* \mid a\,\mu, a\,\sigma^2) = \alpha_b f_{\mathcal{N}}(x^* \mid b\,\mu, b\,\sigma^2) \qquad (15)$$

$$\Rightarrow x^* = \pm\sqrt{a\,b\,(\mu^2 + \sigma^2 \log(a^{-1}\,b\,\alpha_a^2\,\alpha_b^{-2}))} \qquad (16)$$

If the decision boundaries x^* do not exist, the density of the higher order envelopes the density of the lower-order one. If so, the lower-order class must not be associated with any intensity. Alternatively, even though the decision boundaries exist, it might be that a lower-order density is enveloped by multiple higher-order ones. A procedure starting from the highest-order density and proceeding to the lowest can determine the decision boundaries in $\mathcal{O}(N)$ time, which enables the lower complexity of the hard EM algorithm. The decision boundaries are symmetric around zero, so the classification can be performed on $|x|$, rather than x.

For the purposes of evaluating the classifier performance, the expected accuracy (ACC) can be computed:

$$\mathbb{E}[\mathbb{I}\{C(X) = K\}] \qquad (17)$$

$$= \sum_{k=1}^{N} \alpha_k \int_{[x_k^-,\, x_k^+] \cup [-x_k^+,\, -x_k^-]} f_{\mathcal{N}}(x \mid k\,\mu, k\,\sigma^2)\,\delta x \qquad (18)$$

where x_k^- and x_k^+ are such that $\forall\, x, 0 \leq x_k^- \leq x < x_k^+ : C(x) = k$. The integral does not have a closed form solution, but it is the Gauss error function, and can be evaluated numerically. Lastly, we note that this quantity applies asymptotically if the model is true and the estimated parameters are correct. Nevertheless, it is likely useful to evaluate how hard the estimation problem is.

3.4 Means of comparison with the previous method

The previous method of RNA quantification from the spot intensity histogram, here called rounding method, relied on manual inspection of the intensity distribution (Golding *et al.*, 2005). Namely, the location of the first peak in the distribution of intensities is selected by an expert, after which the intensities are divided by this value to obtain the RNA numbers in each spot and cell (Golding *et al.*, 2005). The discretization is achieved by rounding, which can result in suboptimal choice of thresholds for the classifier accuracy.

Given our model of spot intensities, the expected accuracy of the rounding classifier can be computed using Equation (17) with $x_k^{\pm} = (k \pm \frac{1}{2})\,\Delta$ (with the exception that $x_1^- = 0$ and $x_N^+ = \infty$), where Δ is the location of the first peak. For comparison, we find Δ such that the accuracy is maximized. This classifier is not realizable, as the true parameters must be known, but it serves as the upper limit of performance of the classifier. Alternatively, it is possible to use the parameter estimation procedure proposed with the classification of the rounding method, which has the advantage that finding Δ can be automated. However, on average, such an automated method cannot perform better than the method proposed.

4 RESULTS

4.1 Estimating the number of MS2-GFP-tagged RNAs

We first used our method to estimate the number of MS2-GFP-tagged RNA molecules in live *E.coli* cells. In one case, cells contained an RNA coding for 96 bs for MS2-GFP (Golding *et al.*, 2005). In the other case, cells contained a different construct, with only 48 bs. In both cases, cells were induced with 1 mM of IPTG and 0.1% of arabinose, and images were taken 60 min after induction. In theory, we expect the intensity of tagged RNAs to be halved in the second set of measurements.

From both sets of images, we extracted the total pixel intensity of each cell and of each spot. This procedure yielded 269 and 155 samples of spot and cell intensities, respectively, from cells with the 96 bs construct, and 443 and 242 from cells with the 48 bs construct.

For each construct, we assumed that all tagged RNAs exhibit the same fluorescence level when measured from either spot or cell intensities and can be represented by a distribution with the same mean and variance (μ and σ^2). Such mean and variance only differ between the two constructs. With this constraint, one can use both datasets (cell and spot intensities), to jointly estimate the parameters of the model, for each construct. For this, we modified the estimator to account the joint estimation of the two sets of data and estimated the parameters in each case. The distribution of measured intensities along with the estimated distributions is shown in Figure 1, and the values of the parameter estimates are given in Table 1.

First, one would expect the mean of intensities detected from the 48 bs RNAs to be half of the one detected from the 96 bs

RNAs. However, from Table 1, the estimated ratio of their means of intensities is 0.83, rather than one half. One likely explanation for this is that the number of functional binding sites in the two RNA constructs differs from the intended numbers (particularly in the case of the longer construct). Meanwhile, the ratio between the variances is expected to be similar to the ratio between the means. Instead, it equals 0.69, which deviates from the measured value, likely due to deviations in noise levels arising from non-linear changes in intensities with the number of binding sites. Nevertheless, we kept the estimated parameters values (Table 1), rather than imposing the constraint on the values, as they allow a significantly better fit than would be achieved by constraining the ratios to one half, as determined by a likelihood-ratio test ($P < 2.1 \times 10^{-4}$).

Because the induction level of the target gene is the same in the two cases (96 and 48 bs constructs), one also expects similar RNA-per-spot (first and third row of Table 1) and RNA-per-cell distributions (second and fourth row) in both cases. The obtained values of $\widehat{\boldsymbol{\alpha}}$ suggest that this expectation appears to be correct. In agreement, we found no evidence that the parameters in Table 1 fit significantly better than in a case where each of the RNA-per-spot and RNA-per-cell distributions are constrained to be equivalent, as determined by a likelihood ratio test ($P > 0.54$).

Next, we used the parameters obtained from the spot intensities and cell intensities to estimate the number of RNAs in each cell. Results are shown in Table 2. We note that the procedure of estimating the RNA numbers using the classification of intensities of each spot is expected to yield better results than estimating the RNA numbers from the total intensities of each cell (cf. accuracy in Table 2), as the problem is easier and the sample size is larger, but it requires an accurate detection of the spots inside the cells.

As expected (and necessary if the methods are appropriate), the RNA statistics (Table 2) are similar in the four cases (even though classifying spots are expected to yield better results than classifying cells). Also, the results are in agreement with previous measurements using the RNA target for 96 bs (Kandhavelu *et al.*, 2012b) in terms of mean and variance of RNA numbers.

4.2 Comparison with a time series method

To validate our results, we compared our method with a previously introduced method (Kandhavelu *et al.*, 2012b) to extract RNA statistics from time series data (Kandhavelu *et al.*, 2012a, b; Muthukrishnan *et al.*, 2012). Unlike our method, this method uses temporal information, i.e. time series of intensities in the cells, which allows better accuracy of RNA counting, but makes it unsuitable for analysis of individual frames of cell populations.

We collected images taken for 2 h, separated by 1 min intervals, of cells subject to the same media and induction as in the previous cases. Then, we made use of the method from Kandhavelu *et al.* (2012b) to extract the RNA statistics at ~60 min. The results are shown in Table 3. Visibly, these results are similar to those obtained by our method (cf. Table 2).

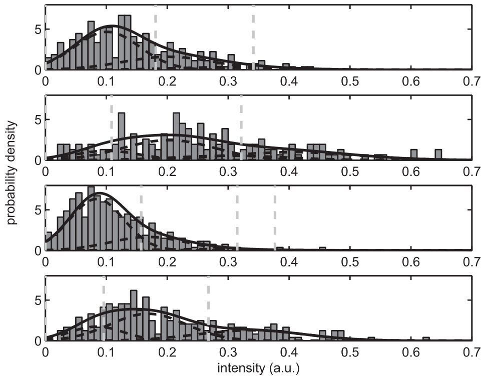

Fig. 1. Distribution of intensities from MS2-GFP-tagged RNA measurements. Panels from top to bottom: spots (96), cells (96), spots (48) and cells (48). The gray bars represent the measured intensity histograms, the solid black lines the estimated distributions, the dashed black lines their components and the dashed gray lines the decision boundaries

Table 1. Estimated model parameters from the MS2-GFP-tagged RNA measurements

Case	M	$\hat{N}$	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{\boldsymbol{\alpha}}$
Spots (96)	269	4	0.101	0.053	(062, 0.30, 0.07, 000)
Cells (96)	155	4	0.101	0.053	(015, 0.46, 0.12, 026)
Spots (48)	443	4	0.084	0.044	(070, 0.26, 0.03, 001)
Cells (48)	242	4	0.084	0.044	(019, 0.52, 0.00, 029)

Note: The table shows the number of samples M, the estimated model order $\hat{N}$ (Section 2), the estimated parameters mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of the intensity of one RNA and the vector of probabilities $\widehat{\boldsymbol{\alpha}}$. The value of $\hat{\alpha}_k$ is the estimated probability that one has k RNA molecules in a spot or cell (depending on the case).

Table 2. Estimated distribution of RNAs per cell from the MS2-GFP-tagged RNA measurements

Symbol	Case	M'	μ_r	σ_r^2	ACC
A	Spots (96)	182	1.97	1.55	0.78
B	Cells (96)	182	2.07	1.68	0.68
C	Spots (48)	270	2.02	1.46	0.82
D	Cells (48)	270	2.10	1.48	0.77

Note: The table shows number of cells M', mean μ_r and variance σ_r^2 of RNA numbers per cell and the expected accuracy (ACC).

To verify the agreement between the results using the two methods, we performed two-sample Kolmogorov–Smirnov permutation tests with the null hypothesis of the RNA numbers extracted from different methods and/or cases (A through F in Tables 2 and 3) are generated by an equal distribution, which assesses if the distributions are significantly different. All tests were done with 10^6 permutations, and the results are shown in Table 4.

From Table 4, we find that the results obtained from the images of cell populations using the spot intensities provided similar results to those extracted from time series (i.e. the null hypothesis cannot be rejected when comparing cases A, C, E and F). This indicates that our method performs consistently with the time series method. The same does not hold for the results extracted from the cell populations using cell intensities in all cases, where statistical differences are detected with higher sample sizes (B or D versus C or E). This further suggests that the spot intensities should be used in favor of the cell intensities for more accurate quantification of the RNA numbers. Generally, and confirming the results of the previous section, we found no evidence that the results extracted from the 96 and 48 bs constructs are statistically different.

4.3 Comparison with qPCR measurements

We compared the fold-change in RNA numbers estimated using our method with those obtained by qPCR. For this, we used the 96 bs construct with induction levels of 1 mM (Table 2) and of 0 mM of IPTG (not shown), the former resulting in a higher expression rate. The estimated mean RNA numbers in the case of 0 mM of IPTG using our method were 0.68 and 0.63, using the spot and cell intensities, respectively. These result in expression ratios of 0.345 and 0.303. The expression ratio obtained by qPCR is 0.305 with a standard deviation of 0.024. Both numbers estimated using our method are within the 90% confidence interval of the qPCR measurement, indicating a strong agreement.

Table 3. RNA statistics estimated using time series method (Kandhavelu *et al.*, 2012b) at 60 min after induction

Symbol	Case	M'	μ_r	σ_r^2
E	Cells (96), time series	252	2.09	1.51
F	Cells (48), time series	107	2.02	1.41

Note: The table shows number of cells M' and the mean μ_r and variance σ_r^2 of RNA numbers per cell.

Table 4. Comparison between the RNA distributions extracted different methods and/or data

Symbol	B	C	D	E	F
A	0.027	0.632	0.022	0.696	0.952
B	–	0.00991	0.547	0.00457	0.022
C	–	–	0.00911	0.501	0.872
D	–	–	–	0.00375	0.029
E	–	–	–	–	0.913

Note: *P*-values of the Kolmogorov-Smirnov permutation test under the null hypothesis that a pair of samples of RNA numbers from different cases (A through F) come from the same distribution. A low *P*-value (i.e. less than 0:01) indicates that the RNA distributions are likely inequal.

4.4 Applying the method on simulated data

Lastly, we assessed the performance of the parameter estimation and classification using data from Monte Carlo (MC) simulations of our model [Equation (1)] so that the ground truth is known.

First, we computed the expected accuracy of the classification procedure for different parameters values, which assesses the asymptotic performance of the classifier if the parameters are well estimated. This is shown in Figure 2. The accuracy can be good for either low-mean levels, where the problem is simpler, as most of time the number of RNAs is equal to unity, or for low-noise levels, as in this case the distribution consists of distinct peaks. More importantly, for typical RNA levels in *E.coli* [i.e. 1–10 (Bernstein *et al.*, 2002)], the accuracy is >0.8 for noise levels < 0.25 (which are in agreement with our results in Table 1). For high RNA levels and/or high noise levels, the accuracy deteriorates, as the distribution no longer exhibits distinct peaks.

We found that the expected accuracy is similar with slightly non-Poissonian data (not shown) but generally better with sub-Possonian and worse with super-Poissonian. For typical RNA levels and low-noise levels (<0.1), the accuracy remains >0.9 even for geometric distributed RNA numbers. The rounding method has comparable performance only for limited noise levels, and its performance is more sensitive to the RNA distribution.

With a finite sample, the parameter estimates are not necessarily correct, which causes errors in the classification. We tested the method for various samples sizes, and the results for noise level of $\sigma\,\mu^{-1} = 0.25$ and Poissonian RNA numbers with a mean of $\lambda = 2.5$ are shown in the top panel of Figure 3. For small samples (e.g. 10), the performance is not comparable with the asymptotic performance, whereas for sample sizes $>10^3$ the mean accuracy exceeds the theoretical accuracy of the rounding method (0.8049).

We also used noise level $\sigma\,\mu^{-1} = 0.5$ and a bimodal RNA distribution of $\alpha = (030, 0.01, 0.01, 0.66, 0.01, 001)$ (these are observed e.g. in the case of genes integrated into circuits such

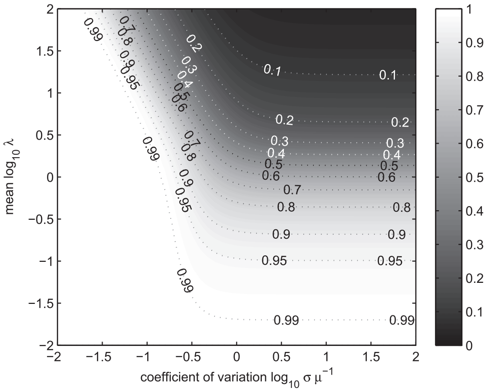

Fig. 2. Expected accuracy of the classification problem for Poisson-distributed RNA numbers. Surface plot of the expected accuracy of the classifier for Poisson-distributed RNA numbers, as a function of noise $\sigma\,\mu^{-1}$ (coefficient of variation) and RNA mean level λ. Light shades of gray represent high accuracy, whereas dark shades represent low

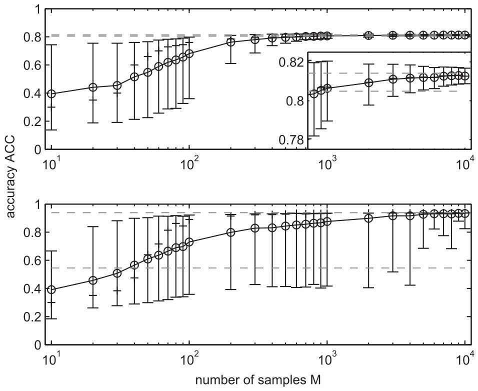

Fig. 3. Distributions of accuracies for various samples sizes M obtained using MC simulations. Top panel: Poisson distributed α, with $\lambda = 2.5$ and $\sigma\mu^{-1} = 0.25$. Bottom panel: bimodal distribution of α and $\sigma\mu^{-1} = 0.5$. Circles represent means, pluses medians and whiskers upper and lower standard deviations from the mean. The dashed lines are the expected accuracy for our method (higher value) and for the rounding method (lower value)

as a toggle switch), shown in the bottom panel of Figure 3. The expected accuracies are 0.9394 and 0.5457, suggesting that our method is appropriate, but the rounding method is not.

Lastly, we note that our method is likely biased for finite samples. We found that for sample sizes <100, it generally overestimates the mean and the variance, which primarily results from underestimation of the order due to lack of evidence for selecting a high–order model with small samples. Regardless, even if the order was known, the maximum likelihood estimator is likely biased. However, these effects are negligible for larger sample sizes ($>10^2$), and e.g. the standard deviation of the parameter estimates exceeds the bias (Fig. 4).

5 DISCUSSION

We have presented a fully automatic method of quantification of RNA numbers from the intensities of the either fluorescent spots or cells. The method consists of a numerical maximum likelihood parameter estimation step (with one of the two proposed methods) followed by a maximum a posteriori classification.

We showed that the method proposed has several advantages. First, by being automated, it will allow an objective comparison of results from independent measurements. Second, our method is expected to have better accuracy than the previous method (Golding *et al.*, 2005), when the distribution of RNAs is non-uniform and/or, when the measurement noise is high. When the distribution is uniform and/or the noise-level is low, the solution converges to that of the previous method. Third, our method allows the estimation of its own accuracy. The theoretical analysis indicates that the finite sample biases of the maximum likelihood estimators are negligible and that the method is expected to perform well (~80% accuracy or above) in a typical setting, when the sample size is few hundred samples or more, which is typical for a single-cell study.

The method can be applied on various intensity distributions. We demonstrated its applicability on both the spot intensities and cell intensities extracted from live *E.coli* cells. The choice of the input is two-fold: spot intensities result in more accurate classification, but the detection of spots is required in addition to cell segmentation.

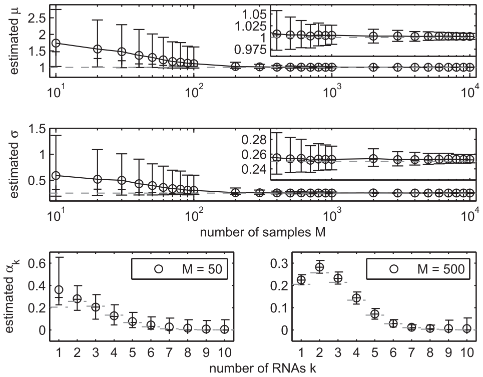

Fig. 4. Distributions of parameter estimates for various sample sizes M obtained using MC simulations. Estimated parameters $\hat{\mu}$ (top panel, true value $\mu = 1$) and $\hat{\sigma}$ (middle panel, true value $\sigma = 0.25$) for various samples sizes. Also shown is the distribution of estimated parameters $\hat{\alpha}_k$, for $M = 50$ (bottom left panel) and $M = 500$ (bottom right). Circles represent means, pluses represent medians and whiskers represent upper and lower standard deviations. The dashed lines are true parameter values

In theory, the method is applicable to any fluorescent or fluorescence-tagged molecules present in low-copy numbers, such as low-expression level fluorescent proteins. However, a proper counting requires a certain degree of separation between brightness levels (not much smaller than the one between e.g. the RNAs with 48 bs). In this regard, our tests showed that the method fails if the data are too noisy or the degree of clustering of the spots is too high. For example, we tested the method on confocal microscopy measurements of tsr-venus proteins coded in *E.coli*, but the signal-to-noise ratio was found to be too low, except for rare cases, where all cells would have one or two proteins. Overall, we expect that, as the methods of fluorescent tagging and microscope improve, our method will become more widely applicable, as it is automatic and allows comparing data from different sources, which is currently not possible.

Funding: Tampere City Science Foundation [to A.H.]; Academy of Finland [257603 to A.S.R.]; Tekes [1386/31/2012 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

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. USA*, **99**, 9697–9702.

Chowdhury,S. *et al.* (2013) Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE). *BMC Bioinformatics*, **14**, S8.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, **39**, 1–38.

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

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

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

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.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods*, **25**, 402–408.

Lloyd-Price,J. *et al*. (2012) Asymmetric disposal of individual protein aggregates in *Escherichia coli*, one aggregate at a time. *J. Bacteriol*., **194**, 1747–1752.

Montero Llopis,P. *et al*. (2010) Spatial organization of the flow of genetic information in bacteria. *Nature*, **466**, 77–81.

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.

Peabody,D.S. (1993) The RNA binding site of bacteriophage MS2 coat protein. *EMBO J*., **12**, 595–600.

Raj,A. *et al*. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877–879.

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

Wilks,S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Satist*., **9**, 60–62.

Wu,C.F.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat*., **11**, 95–103.

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