Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 3

A Bayesian Hierarchical Model for Quantitative Real-Time PCR Data

Turid Follestad, *Norwegian University of Science and Technology*
Tommy S. Jørstad, *Norwegian University of Science and Technology*
Sten E. Erlandsen, *Norwegian University of Science and Technology*
Arne K. Sandvik, *Norwegian University of Science and Technology & St. Olav's University Hospital*
Atle M. Bones, *Norwegian University of Science and Technology*
Mette Langaas, *Norwegian University of Science and Technology*

Recommended Citation:
Follestad, Turid; Jørstad, Tommy S.; Erlandsen, Sten E.; Sandvik, Arne K.; Bones, Atle M.; and Langaas, Mette (2010) "A Bayesian Hierarchical Model for Quantitative Real-Time PCR Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 9: Iss. 1, Article 3.
DOI: 10.2202/1544-6115.1427

A Bayesian Hierarchical Model for Quantitative Real-Time PCR Data

Turid Follestad, Tommy S. Jørstad, Sten E. Erlandsen, Arne K. Sandvik, Atle M. Bones, and Mette Langaas

Abstract

We present a Bayesian hierarchical model for quantitative real-time polymerase chain reaction (PCR) data, aiming at relative quantification of DNA copy number in different biological samples. The model is specified in terms of a hidden Markov model for fluorescence intensities measured at successive cycles of the polymerase chain reaction. The efficiency of the reaction is assumed to depend on the abundance of the target DNA through fluorescence intensities, and the relationship is specified based on the kinetics of the reaction. The model incorporates the intrinsic random nature of the process as well as measurement error. Taking a Bayesian inferential approach, marginal posterior distributions of the quantities of interest are estimated using Markov chain Monte Carlo. The method is applied to simulated data and an experimental data set.

KEYWORDS: Bayesian model, Markov chain Monte Carlo, polymerase chain reaction, real-time PCR

Author Notes: This work was supported by grants NFR 143250/140 and NFR 151991/S10 from the biotechnology and the functional genomics (FUGE) programs of the Norwegian Research Council. Turid Follestad and Sten Even Erlandsen were supported by the cross-disciplinary research project BIOEMIT at the Norwegian University of Science and Technology. The authors thank two anonymous referees for useful comments that helped to improve the paper.

1 Introduction

The real-time quantitative polymerase chain reaction (PCR) is a widely used technique for quantification of gene expression levels in a biological sample, in particular for low abundance genes. In contrast to microarray experiments, where the expression level of thousands of genes are measured simultaneously, real-time quantitative PCR is designed for targeted quantification of gene expression for a limited number of genes.

Starting from the biological sample, RNA is typically first isolated and reverse transcribed into complementary DNA. The initial target DNA is then amplified by the polymerase chain reaction. The PCR process consists of a series of repeated cycles, where at each cycle, a fraction of the target DNA is duplicated. The process is characterised by the efficiency, which, taking into account the intrinsic random nature of the process, can be defined as the probability that a molecule is duplicated at each cycle. In the initial phase of the reaction the efficiency is normally near one, while it decreases in the course of the reaction due to shortage of reaction material. Most commonly, the amount of target DNA at each cycle of the reaction is quantified by using fluorescence chemistry, and real-time PCR refers to cycle by cycle monitoring of the fluorescence intensities as the process proceeds. Examples of amplification curves, displaying the fluorescence intensities as a function of cycle, are shown in Section 4.2. For the first few cycles, denoted the baseline cycles, the fluorescence intensities corresponding to the copy numbers of the target DNA are normally not distinguishable from background fluorescence. At the other end, as the efficiency decreases, the curve is eventually levelling out, reaching a plateau.

An underlying assumption in most quantitative PCR (qPCR) approaches is that, disregarding the observational noise, the fluorescence intensity is proportional to the corresponding target DNA copy number. Consequently, whether or not the constant of proportionality is known, relative quantification of target DNA in two samples is possible, comparing the fluorescence intensities between the samples. However, in the presence of a series of baseline cycles, the fluorescence corresponding to the initial target DNA cannot be read directly from the amplification curve, and must be estimated from the curve.

Many currently available quantification methods are based on the assumption of a constant efficiency, implying exponential growth. In the approach by Livak and Schmittgen (2001), the efficiency is assumed to be equal to one, while in e.g. Gentle et al. (2001), Marino et al. (2003) and Cook et al. (2004), the assumption of exponential growth is utilised to estimate the efficiency from the amplification curve data. Strategies for identification of the cycles

assumed to exhibit exponential growth have been suggested by e.g. Gentle et al. (2001), Liu and Saint (2002a), and Tichopad et al. (2003). Relaxing the constant efficiency assumption allows for the inclusion of data for a larger subset of the cycles. One approach that has been proposed is to fit parametric curves to the fluorescence intensities as a function of cycle number (Liu and Saint, 2002b; Rutledge, 2004), or to the sample efficiency (computed from the ratio of the fluorescence intensities for successive cycles) as a function of intensity (Batsch et al., 2008; Alvarez et al., 2007). However, these are purely curve fitting methods, and do not model the PCR process as such.

Jagers and Klebaner (2003) propose to model the PCR as a branching process with an intensity dependent efficiency following the enzymological model for the PCR kinetics by Schnell and Mendoza (1997). The number of duplicates after each cycle is binomially distributed with probability given by the efficiency model. A modified version of the model is given by Lalam et al. (2004) and Lalam (2006), assuming that the efficiency is constant for the early cycles, but follows a damped version of the original efficiency model for intensities higher than a threshold value. The parameters of the efficiency model are estimated by conditional least squares, but no estimate of initial fluorescence for the target DNA is given.

In a relative qPCR experiment, the aim is to compare the gene expression levels of two or more conditions, and typically several replicates are made for each condition. We propose a Bayesian hierarchical model for relative qPCR, including observational noise, and estimating the model parameters based on data from all amplification curves jointly. With few exceptions, including Cook et al. (2004) and Batsch et al. (2008), this is in contrast to currently available methods for analysing qPCR data, which normally operate separately on each amplification curve. We quantify the uncertainty of the parameter estimates by their estimated posterior distributions, generated using Markov chain Monte Carlo methodology. Our approach is similar to that of Lalam (2007), but the latter is based on quantifying DNA from a single amplification curve, and on assuming constant efficiency. Our model is assessed by applying the approach to simulated data, and results from running the algorithm on an experimental data set are also presented.

2 Model specification

We specify the model for a relative qPCR experiment, where the aim is quantitative comparison of the gene expression levels for two or more conditions, often a treatment and a control group. In what follows, we refer to different

DOI: 10.2202/1544−6115.1427

experimental conditions as treatments. To control for non-biological effects that might influence the amplification process, the amplification curves for the gene of interest are usually contrasted with similar curves for a reference gene, that is expected not to be influenced by the treatments. We further consider the situation where the PCR is run in replicates for each treatment and gene combination. The model is specified for a qPCR experiment consisting of n amplification curves, representing replicated PCR runs for n_t treatments, and n_g genes for each treatment.

2.1 Full model

The model relies on the assumption that during the polymerase chain reaction for an experimental unit i, a molecule has a probability $p_{i,k}$ to duplicate at each cycle k, independently for each molecule. In what follows, we refer to the PCR process for each experimental unit i as reaction i. Let $n_{i,k}$ denote the number of target DNA molecules for reaction i at cycle k. The stochastic model for the kinetics can be written

$$n_{i,k} = n_{i,k-1} + z_{i,k}, \quad i = 1, \ldots, n, \ k = 1, \ldots, m_i, \tag{1}$$

where

$$z_{i,k} \mid n_{i,k-1}, p_{i,k} \sim \text{Binom}(n_{i,k-1}, p_{i,k}), \tag{2}$$

and m_i is the number of cycles for reaction i. For large $n_{i,k}$ the binomial distribution can be approximated by a normal distribution, such that

$$n_{i,k} \mid n_{i,k-1}, p_{i,k} \sim \mathcal{N}(n_{i,k-1}(1 + p_{i,k}), n_{i,k-1}p_{i,k}(1 - p_{i,k})), \\ i = 1, \ldots, n, \ k = 1, \ldots, m_i. \tag{3}$$

The number of molecules after each cycle is measured in terms of the corresponding fluorescence intensity. We adopt the commonly made assumption that the fluorescence is proportional to the number of molecules, but will assume in addition that the fluorescence is measured with additive noise. Let $x_{i,k} = \gamma n_{i,k}$, such that $x_{i,k}$ represents noise-free fluorescence. Conditionally on the reaction history, the variable $x_{i,k}$ will then also be approximately normal, with mean and variance given by

$$\begin{aligned} \text{E}(x_{i,k} \mid n_{i,k-1}, p_{i,k}) &= \gamma n_{i,k-1}(1 + p_{i,k}), & (4) \\ \text{Var}(x_{i,k} \mid n_{i,k-1}, p_{i,k}) &= \gamma^2 n_{i,k-1} p_{i,k}(1 - p_{i,k}). & (5) \end{aligned}$$

Substituting $x_{i,k-1}/\gamma$ for $n_{i,k-1}$, we get the model

$$x_{i,k} \mid x_{i,k-1}, p_{i,k}, \gamma \sim \mathcal{N}(x_{i,k-1}(1 + p_{i,k}), \gamma x_{i,k-1} p_{i,k}(1 - p_{i,k})), \\ i = 1, \ldots, n, \ k = 1, \ldots, m_i \tag{6}$$

for $x_{i,k}$. We further assume that the initial fluorescence $x_{i,0}$ for curve i is normally distributed with gene and treatment dependent means μ_{g_i,t_i}, and with variance $\kappa^{-1}\mu_{g_i,t_i}$ proportional to the mean. Here, g_i and t_i denote the considered gene (g) and treatment (t) corresponding to amplification curve i. That is,

$$x_{i,0} \mid \mu_{g_i,t_i}, \kappa \sim \mathcal{N}(\mu_{g_i,t_i}, \kappa^{-1}\mu_{g_i,t_i}), \quad i = 1, \ldots, n. \tag{7}$$

Our model for the efficiency $p_{i,k}$ of the reaction is motivated by the enzymological model for the PCR kinetics presented in Schnell and Mendoza (1997). Following their model, the efficiency $p_{i,k}$, going from cycle $k-1$ to k, and with $n_{i,k-1}$ molecules at cycle $k-1$, is

$$p_{i,k} = \frac{K_{g_i,t_i}}{K_{g_i,t_i} + n_{i,k-1}}, \tag{8}$$

where K_{g_i,t_i} is a reaction constant, assumed to depend on gene and treatment. We restrict our model to the subset of cycles for which (8) can be assumed to be a reasonable description of the process. Substituting $x_{i,k-1}/\gamma$ for $n_{i,k-1}$, we get

$$\begin{aligned} \text{logit}(p_{i,k}) &= \log\left(\frac{p_{i,k}}{1-p_{i,k}}\right) && (9)\\ &= \log\left(\frac{\gamma K_{g_i,t_i}}{\gamma K_{g_i,t_i} + x_{i,k-1}} / (1 - \frac{\gamma K_{g_i,t_i}}{\gamma K_{g_i,t_i} + x_{i,k-1}})\right) \\ &= \log(\gamma K_{g_i,t_i}) - \log(x_{i,k-1}). && (10) \end{aligned}$$

We consider $\alpha_{g_i,t_i} = \log(\gamma K_{g_i,t_i})$ as unknown constants to be estimated from the observed amplification curve. In addition, we will allow the coefficient of the $\log(x_{i,k-1})$ term to be different from -1, and denote this coefficient by β_{g_i,t_i}. This generalisation is motivated by results from linear model fits of the logarithm of empirical efficiencies against observed log-fluorescence for experimental data sets (not shown), indicating that the fitted slopes can deviate from -1. The empirical efficiencies used in these model fits are computed from observed fluorescence intensities $y_{i,k}$ by $y_{i,k}/y_{i,k-1} - 1$. Introducing the coefficients α_{g_i,t_i} and β_{g_i,t_i}, we get the generalisation

$$p_{i,k} = \frac{\exp(\alpha_{g_i,t_i})}{\exp(\alpha_{g_i,t_i}) + (x_{i,k-1})^{-\beta_{g_i,t_i}}} \tag{11}$$

of the model in (8). By adding normally distributed noise to the generalised model on logit scale, we arrive at the following model for the efficiency $p_{i,k}$:

$$\text{logit}(p_{i,k}) \mid x_{i,k-1}, \alpha_{g_i,t_i}, \beta_{g_i,t_i}, \tau_p \sim \mathcal{N}(\alpha_{g_i,t_i} + \beta_{g_i,t_i} \log(x_{i,k-1}), \tau_p^{-1}). \tag{12}$$

DOI: 10.2202/1544−6115.1427

As before, the indices g_i and t_i are used to represent the gene and treatment corresponding to curve i, respectively.

Finally, we assume that the fluorescence intensity is observed with additive normally distributed noise, and that the distribution for the observed fluorescence $y_{i,k}$ for cycle k of curve i is given by

$$y_{i,k} \mid x_{i,k}, \tau_y \sim \mathcal{N}(x_{i,k}, \tau_y^{-1}), \quad i = 1, \ldots, n, \ k = 1, \ldots, m_i. \tag{13}$$

In (13) we have implicitly assumed that $y_{i,k}$ represents background corrected data. In principle, background correction could be included in the model and estimated jointly with the remaining parameters, but we will assume that the amplification curves are background corrected in a preprocessing stage. The full hierarchical model is illustrated in Figure 1.

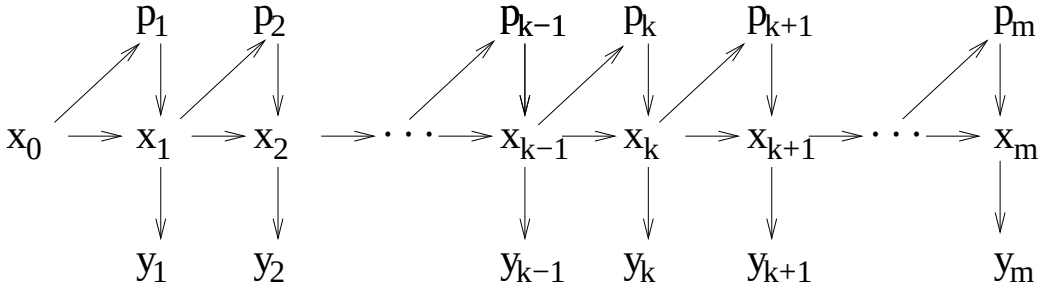

Figure 1: *Graphical description of the full hierarchical model for one experimental unit. Here, y, x and p are used to denote the observed fluorescence intensities, the noise-free intensities, and the efficiency, respectively. The subscript i for the experimental unit is suppressed, and the cycle independent parameters of the model are not shown.*

In a typical experiment $n_g = 2$ and $n_t = 2$, and the main parameter of interest is the ratio of the mean initial abundance for a treated and a control group for the gene of interest, adjusted for the corresponding ratio for a reference gene. In terms of the model parameters, this ratio, which we denote by R_{adj}, is

$$R_{adj} = \frac{\mu_{2,2}/\mu_{2,1}}{\mu_{1,2}/\mu_{1,1}}, \tag{14}$$

where the first subscript (1 or 2) denotes the reference gene and the gene of interest, respectively, and the second subscript (1 or 2) denotes the control and treated groups.

2.2 Simplified model

Observe that in the full model, stochastic growth is represented both by the model (12) for the efficiencies $p_{i,k}$, and in the normal model (6) for the fluorescence intensities, conditionally on $p_{i,k}$. Fitting the full model of Section 2.1 using the Markov chain Monte Carlo (MCMC) approach described in Section 3.2 to simulated data sets, we experienced that mixing and convergence of the MCMC algorithm were slow. Using a single-site MCMC algorithm, this is not surprising, in light of the correlation structure inherent in the model. Further, some of the parameters of the model appear to be hard to identify from the data. This applies in particular to the scaling factor γ in (6), representing the link between the fluorescence intensity and the number of molecules.

We therefore propose a simplified version of the full model, ignoring the noise in the model for the intensities $x_{i,k}$ as specified in (6). As a consequence, the scaling factor γ is eliminated from the model. Attempts to quantify the scaling factor from amplification data (see e.g. Rutledge (2004) and Goll et al. (2006)), indicate that the order of magnitude of γ is around 10^{-10}, and that the factor is small relative to typical fluorescence intensities. From (6) we find that in the full model, the conditional standard deviation for $x_{i,k}$ is less than $\sqrt{\gamma/x_{i,k-1}}$ times the conditional mean. Consequently, we expect that the conditional variance will be negligible, and the simplified model to be a good approximation to the full model. It turns out that the resulting model gives reasonable mixing and convergence for the corresponding MCMC algorithm.

By substituting the conditional prior mean $x_{i,k-1}(1+p_{i,k})$ for $x_{i,k}$ in the likelihood (13) and in the model (12) for the efficiency, the simplified model can be summarised as

$$y_{i,k} \mid x_{i,0}, p_{i,1}, \ldots, p_{i,k}, \tau_y \sim \mathcal{N}\left(x_{i,0}\prod_{j=1}^{k}(1+p_{i,j}), \tau_y^{-1}\right), \tag{15}$$

$$\operatorname{logit}(p_{i,k}) \mid \alpha_{g_i,t_i}, \beta_{g_i,t_i}, x_{i,0}, p_{i,1}, \ldots, p_{i,k-1}, \tau_p \sim \mathcal{N}\left(\alpha_{g_i,t_i} + \beta_{g_i,t_i}\log\left(x_{i,0}\prod_{j=1}^{k-1}(1+p_{i,j})\right), \tau_p^{-1}\right), \tag{16}$$

$$x_{i,0}|\mu_{g_i,t_i}, \kappa \sim \mathcal{N}\left(\mu_{g_i,t_i}, \kappa^{-1}\mu_{g_i,t_i}\right), \tag{17}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, m_i$.

DOI: 10.2202/1544−6115.1427

3 Parameter estimation

We take a Bayesian approach, and estimate marginal posterior distributions for the unknown quantities of the simplified model of Section 2.2 by Markov chain Monte Carlo (MCMC). The model specification is based on the assumption that a valid range of cycles corresponding to the chosen model for $p_{i,k}$, given in (16), has been identified. We first describe how the valid cycle window is selected for an experimental data set, and then describe the sampling algorithm.

3.1 Selection of the cycle window

In the course of a PCR run, reaction conditions will change. Due to shortage of reaction materials, the amplification curve will eventually level out, meaning that the efficiency approaches zero. The limiting mean growth rate corresponding to the model (16) is zero only if $\beta_{g_i,t_i} < -1$, implying that not all data points of an amplification curve can be included in the analysis. We select the valid cycle window by identifying the inflection point of the amplification curve from a four-parameter sigmoid curve fitted to the data, and choose the closest cycle prior to this inflection point as the final cycle m_i for each curve i. In addition, we have observed that even after background correction, there are trends for the early cycles that cannot be explained by the amplification process. The data for these cycles are discarded from the analysis. We first use the method in Tichopad et al. (2003) to identify the final baseline cycle, and then we discard data for the cycles prior to and including this cycle. Taking this approach, we will normally discard more data than necessary, and selecting a maximal valid cycle window represents a topic for further study.

3.2 The MCMC algorithm

To complete the model specification, we assign prior distributions to the cycle independent parameters of the model. We denote the vector of these parameters by $\boldsymbol{\theta}$, such that $\boldsymbol{\theta} = (\tau_y, \tau_p, \kappa, \{\mu_{g,t}\}, \{\alpha_{g,t}\}, \{\beta_{g,t}\})^T$. The priors are summarised in Table 1. The mean and precision of the normal priors for $\alpha_{g,t}$ and $\beta_{g,t}$, and the precision of the truncated normal prior for $\mu_{g,t}$, are computed by first specifying what we believe is a reasonable range for each of the these parameters, and then selecting the parameters of the distribution such that a prior probability of approximately 0.95 is assigned to that range. The precision parameters are all assigned non-informative priors.

The unknown quantities to be estimated are the vector $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,m_i})$

$\mu_{g,t} \sim \mathcal{N}(0, 5.0^2) I(\mu_{g,t} > 0),\ g = 1, \ldots, n_g,\ t = 1, \ldots, n_t$
$\kappa \sim \text{Gamma}(0.1, 0.00001)$
$\alpha_{g,t} \sim \mathcal{N}(10.0, 7.5^2),\ g = 1, \ldots, n_g,\ t = 1, \ldots, n_t$
$\beta_{g,t} \sim \mathcal{N}(-1.5, 0.5^2),\ g = 1, \ldots, n_g,\ t = 1, \ldots, n_t$
$\tau_p \sim \text{Gamma}(0.5, 0.0005)$
$\tau_y \sim \text{Gamma}(0.5, 0.0005)$

Table 1: *Specification of the prior distributions for the cycle independent parameters. Here, n_g and n_t denote the number of genes and treatments, respectively.*

of efficiencies for each amplification curve $i = 1, \ldots, n$, the initial fluorescence, $x_{i,0},\ i = 1, \ldots, n$, and the parameter vector $\boldsymbol{\theta}$. The vector $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,m_i})$ of fluorescence intensities for each amplification curve can be computed from $x_{i,0}$ and $\boldsymbol{p}_i$ as $x_{i,k} = x_{i,0} \prod_{j=1}^{k}(1 + p_{i,j}),\ k = 1, \ldots, m_i$. We let $\boldsymbol{y}_i$ denote the vector of corresponding observed fluorescence for curve i.

The MCMC algorithm works by sampling from a Markov chain with the distribution of interest as stationary distribution, in our problem the joint posterior distribution. Except for the parameters $\alpha_{g,t}$ and $\beta_{g,t}$, we sample each parameter at a time from the full conditional distribution given the remaining parameters. If the full conditional distribution is non-standard, we apply a Metropolis-Hastings step. We then first generate a value from a proposal distribution given the current sample, and then accept or reject this value with a probability depending on the full conditional distribution and the proposal distribution.

The full conditional distributions of the precisions τ_p and τ_y and the parameter κ are Gamma distributions, and these full conditional distributions can be sampled from directly. The two parameters $\alpha_{g,t}$ and $\beta_{g,t}$ for each pair (g, t) are sampled jointly from their binormal full conditional distribution.

For the parameters $\mu_{g,t}$, as well as for the efficiencies $p_{i,k}$ and the initial fluorescence $x_{i,0}$, the full conditional distributions are non-standard, and we apply a Metropolis-Hastings step for each of these parameters. We use normal proposal distributions centred at the current sample. The standard deviations of the proposal distributions are tuned by repeatedly running a series of 5000 samples, and adjusting the standard deviations such that the acceptance probabilities are approximately between 0.2 and 0.5.

The sampling routine is implemented in C.

DOI: 10.2202/1544−6115.1427

4 Results

To illustrate our approach, we present results from applying the algorithm for the simplified model of Section 2.2 to simulated data and an experimental data set.

4.1 Results for a simulated data set

To assess the performance of the model and the MCMC algorithm, we run the algorithm on a data set simulated from the full model of Section 2.1. We simulate a set of data with $n = 12$ samples, representing three replicates of each pair of $n_g = 2$ genes and $n_t = 2$ treatments, using the parameter values given in Table 2. The values for $\alpha_{g,t}$ are calculated as $\alpha_{g,t} = \text{logit}(0.999999) - \beta_{g,t} \log(\mu_{g,t})$, using the selected values for $\beta_{g,t}$ and $\mu_{g,t}$. The total number of cycles for the twelve curves are $m_i = 20$, for $(g_i, t_i) = (1, 1)$ and $(g_i, t_i) = (1, 2)$, $m_i = 17$ for $(g_i, t_i) = (2, 1)$ and $m_i = 18$ for $(g_i, t_i) = (2, 2)$. These numbers are selected such that 5-6 cycles of the amplification curves are beyond the cycle range that is dominated by the observational noise. The simulated amplification curves are shown in Figure 2.

$\mu_{1,1} = 0.025$	$\alpha_{1,1} = 10.127$
$\mu_{1,2} = 0.025$	$\alpha_{1,2} = 10.127$
$\mu_{2,1} = 0.2$	$\alpha_{2,1} = 12.206$
$\mu_{2,2} = 0.1$	$\alpha_{2,2} = 11.513$
$\kappa = 1000$	$\beta_{g,t} = -1.0,\ \forall(g,t)$
$\tau_y = 1/40^2$	$\tau_p = 64.0$
$R_{adj} = \dfrac{\mu_{2,2}/\mu_{2,1}}{\mu_{1,2}/\mu_{2,1}} = 0.5$	$\gamma = 0.001$

Table 2: *Model parameters for the simulated data set.*

The algorithm was run for 20 mill. iterations, taking approximately 4h 30min of CPU time on a 2.66 GHz Unix system. In Figure 3, trace plots of samples from the MCMC algorithm are shown for a subset of the parameters. After thinning to every 4000th iteration to reduce autocorrelation for successive iterations, we discard the first 500 iterations as burn-in iterations. Estimated posterior means and 95% credibility intervals for the parameters were computed from the remaining 4500 iterations, and the results are summarised in Table 3 and Figure 4, together with true values. We observe that the parameter values used to generate the data set are reproduced.

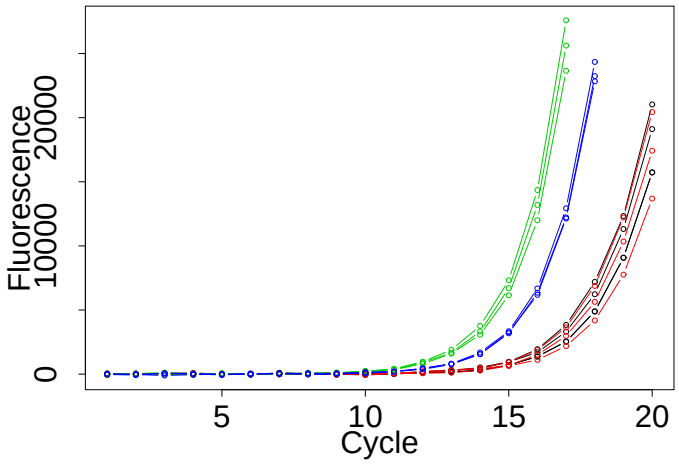

Figure 2: *Amplification curves for the simulated data set. Each colour represents triplicates of a gene and treatment combination* (g, t)*: Black* $= (1, 1)$*, red* $= (1, 2)$*, green* $= (2, 1)$*, and blue* $= (2, 2)$*.*

We have also run the algorithm on several other simulated datasets, with different values for τ_p, τ_y, $\alpha_{g,t}$, $\beta_{g,t}$, and $\mu_{g,t}$. Trace plots for τ_p and density plots for R_{adj} are shown for two of these data sets in Figure 5. These results, and others not shown, indicate that the MCMC algorithm converges, and that the parameter estimates are reproduced.

4.2 Results for an experimental data set

The model was fitted to an experimental data set comparing gene expression in rats treated with Octreotide long-acting release (LAR) to untreated controls. The gene of interest, KLF4, was contrasted with the reference gene β-actin, and the amplifications were run in triplicates for each of the four treatment and gene combinations, in total $n = 12$ reactions. The experimental protocol of the data set is described in Appendix A. The aim of the study was not a thorough analysis of the specific data set, but to illustrate application of the method to an experimental data set.

The data were background corrected by first manually selecting a window of 6-7 observations that were apparent baseline cycles, and subtracting the average fluorescence of these cycles from all observations. The twelve background corrected amplification curves are shown in the left panel of Figure 6. For

DOI: 10.2202/1544−6115.1427

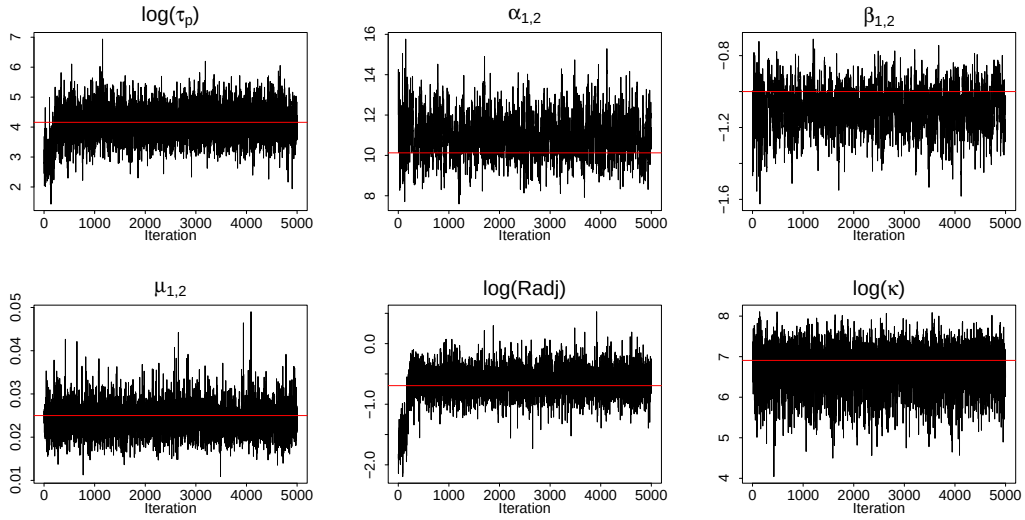

Figure 3: *Trace plots for some of the parameters for the simulated data set. Every 4000th iteration is shown. The horizontal lines indicate the true values.*

Parameter	Posterior mean	95% credibility interval		True value
		Lower limit	Upper limit	
τ_p	73.8	22.8	194	64
$\alpha_{1,1}$	10.0	8.42	11.9	10.1
$\beta_{1,1}$	-0.99	-1.2	-0.803	-1
$\alpha_{1,2}$	11.0	9.15	13.0	10.1
$\beta_{1,2}$	-1.09	-1.32	-0.887	-1
$\alpha_{2,1}$	12.3	8.04	17.1	12.2
$\beta_{2,1}$	-1.00	-1.51	-0.546	-1
$\alpha_{2,2}$	13.3	10.4	16.7	11.5
$\beta_{2,2}$	-1.19	-1.56	-0.88	-1
τ_y	0.000555	0.000447	0.000675	0.000625
κ	910	246	1940	1000
$\mu_{1,1}$	0.0269	0.0204	0.0349	0.025
$\mu_{1,2}$	0.0241	0.0178	0.0315	0.025
$\mu_{2,1}$	0.209	0.189	0.231	0.2
$\mu_{2,2}$	0.10	0.0863	0.115	0.1
R_{adj}	0.547	0.35	0.82	0.5

Table 3: *Estimated posterior means and 95% credibility intervals for the model parameters of the simulated data set.*

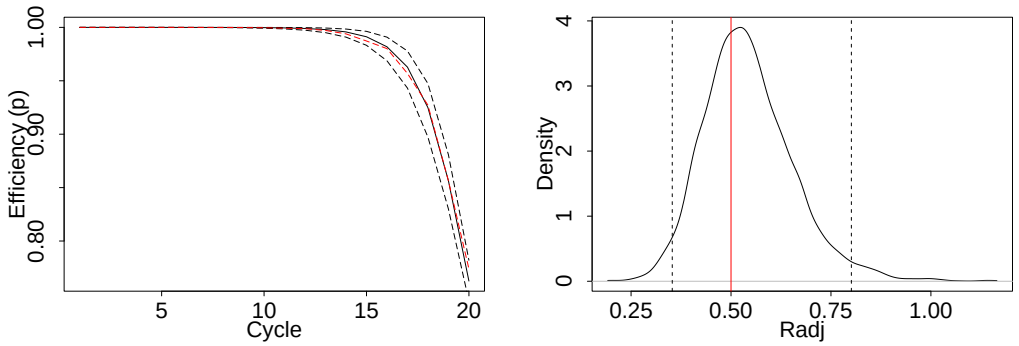

Figure 4: *Estimated posterior means (left, full line) for the efficiency p for all cycles for one reaction with $(g, t) = (1, 2)$, and posterior density plot for R_{adj} (right), for the simulated data set. The black dashed lines indicate 95% credibility intervals, and the red lines the realisations for p and the true value of R_{adj}.*

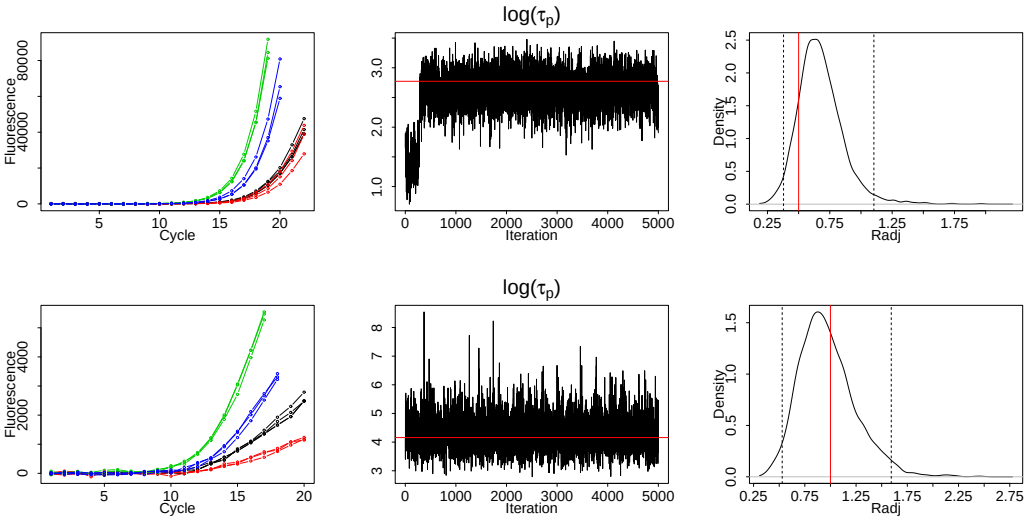

Figure 5: *Amplification plots, trace plots for the precision τ_p, and density plots for R_{adj}, including 95% credibility intervals (dashed lines) and true values (red lines), for two simulated datasets. The data sets are equal to the one presented in Table 2, with the exceptions $\tau_p = 16$ and $\tau_y = 1/80^2$ (top), and $\mu_{1,1} = 0.05$, $\alpha_{1,1} = 6.214$, $\alpha_{1,2} = 5.521$, $\alpha_{2,1} = 7.601$, $\alpha_{2,2} = 6.908$, and $R_{adj} = 1$ (bottom).*

DOI: 10.2202/1544−6115.1427

each of the amplification curves, the approach described in Section 3.1 was used to select the cycle window corresponding to the efficiency model (16). The selected cycle windows for the twelve amplification curves are illustrated in the right panel of Figure 6.

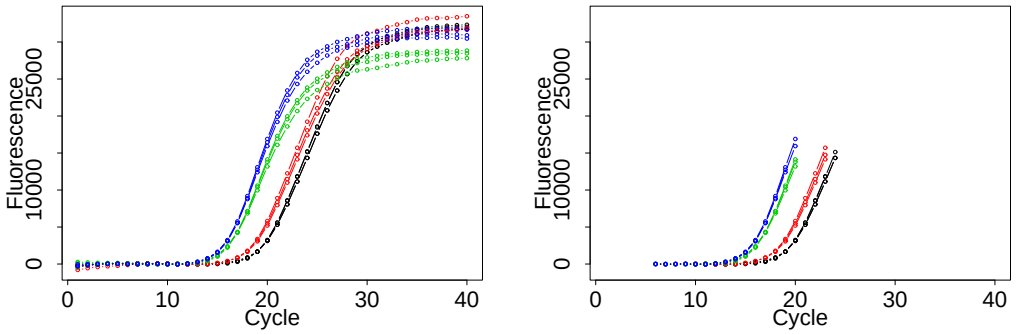

Figure 6: *Amplification curves for treated (red: KLF4, blue: reference gene) and untreated (black: KLF4, green: reference gene) rats for the Octreotide LAR data set. The right panel shows the cycle windows used in the analysis.*

The simplified model was fitted using the MCMC algorithm described in Section 3.2, running the algorithm for 50 mill. iterations. The convergence of the algorithm was monitored by visual inspection of trace plots. The mixing was found to be reasonable, and the algorithm was considered to have converged. Resulting trace plots for a few model parameters are shown in Figure 7. After thinning to every 10000th iteration, the initial 500 iterations are considered burn-in iterations, leaving 4500 iterations for further analysis.

Estimated posterior means and 95% credibility intervals for the model parameters are listed in Table 4. Similar results for the the efficiencies $p_{i,k}$ for one amplification curve, and for the initial fluorescence $x_{i,0},\ i = 1, \ldots, n$ for the individual curves, are illustrated in Figures 8 and 9. The efficiencies are estimated to decrease from nearly 1 to between 0.28 and 0.43 for the twelve curves. The precision τ_p of the efficiency model (16) is estimated to be relatively high, and the posterior variability on logit-scale decreases with cycle. The latter seems reasonable since, relative to the fluorescence intensities, the noise of the amplification curves is expected to be largest close to the baseline cycles. From Figure 9 we observe that the credibility intervals for the gene and treatment dependent means of the initial fluorescence are of similar width as the corresponding intervals for the individual initial values. This might seem counter-intuitive, but we should keep in mind that the estimates are based on

only three replicates within each gene and treatment group.

The estimated marginal posterior density of the main parameter of interest, the ratio R_{adj} (14), is shown in the bottom right panel of Figure 9. The gene of interest, KLF4, is estimated to be up-regulated in the treated group compared to the control group. The posterior mean gives a point estimate of 1.50, and using the sample quantiles, we arrive at a 95% credibility interval between 1.29 and 1.75. The estimated posterior probability that R_{adj} exceeds 1 is $P(R_{adj} > 1 \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = 0.9999$. This means that the posterior probability that the gene of interest is up-regulated in the treated rats is estimated to 0.9999.

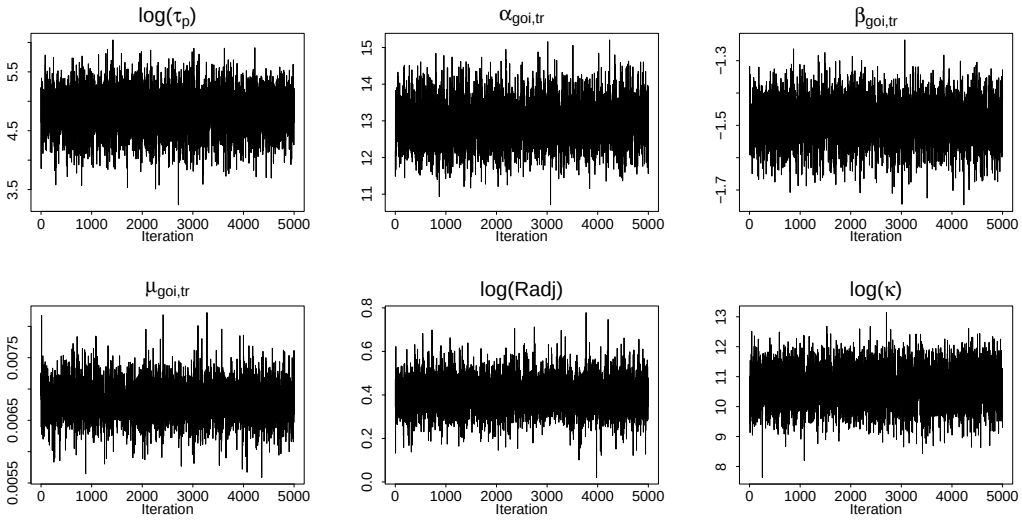

Figure 7: *Trace plots for some of the parameters for the Octreotide LAR data set. Every 10000th of 50 mill. iterations are shown. Here, 'goi' refers to the gene of interest, and 'tr' to the treated group.*

5 Discussion

We have presented a Bayesian hierarchical model for quantitative real-time PCR data, based on a generalisation of the branching process model of Jagers and Klebaner (2003). The model allows for fluorescence intensity dependent efficiency, and incorporates the intrinsic random nature of the reaction process as well as measurement error.

The approach relies on the assumption that the model for the efficiency is valid in the selected cycle window, but it is not restricted to the specific model

DOI: 10.2202/1544−6115.1427

Parameter	Posterior mean	95% credibility interval	
		Lower limit	Upper limit
τ_p	130	59.7	241
$\alpha_{\text{goi,ctrl}}$	13.5	12.3	14.8
$\beta_{\text{goi,ctrl}}$	-1.55	-1.69	-1.41
$\alpha_{\text{goi,tr}}$	13.0	11.8	14.2
$\beta_{\text{goi,tr}}$	-1.49	-1.62	-1.36
$\alpha_{\text{ref,ctrl}}$	14.6	13.3	16.1
$\beta_{\text{ref,ctrl}}$	-1.66	-1.83	-1.51
$\alpha_{\text{ref,tr}}$	15.1	13.8	16.5
$\beta_{\text{ref,tr}}$	-1.69	-1.85	-1.55
τ_y	0.000183	0.000142	0.000229
κ	55100	12500	147000
$\mu_{\text{goi,ctrl}}$	0.00332	0.00294	0.0037
$\mu_{\text{goi,tr}}$	0.0068	0.00626	0.00737
$\mu_{\text{ref,ctrl}}$	0.0362	0.0346	0.0379
$\mu_{\text{ref,tr}}$	0.0496	0.0477	0.0518
R_{adj}	1.50	1.29	1.75

Table 4: *Estimated posterior means and 95% credibility intervals for the model parameters for the Octreotide LAR data set. Here, 'goi' and 'ref' denote the gene of interest and the reference gene, respectively, and 'tr' and 'ctrl' the treated and control groups.*

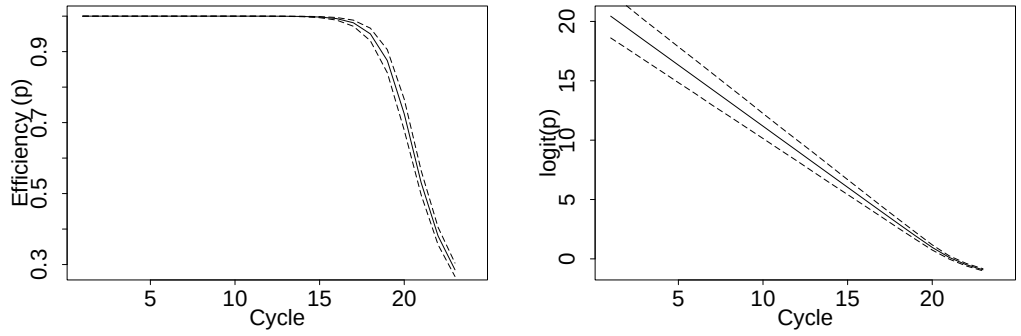

Figure 8: *Estimated posterior means (full line) and 95% credibility intervals (dashed lines) for the efficiency p on original scale and logit scale, for one of the reactions with treated cells and the gene of interest for the Octreotide LAR data set.*

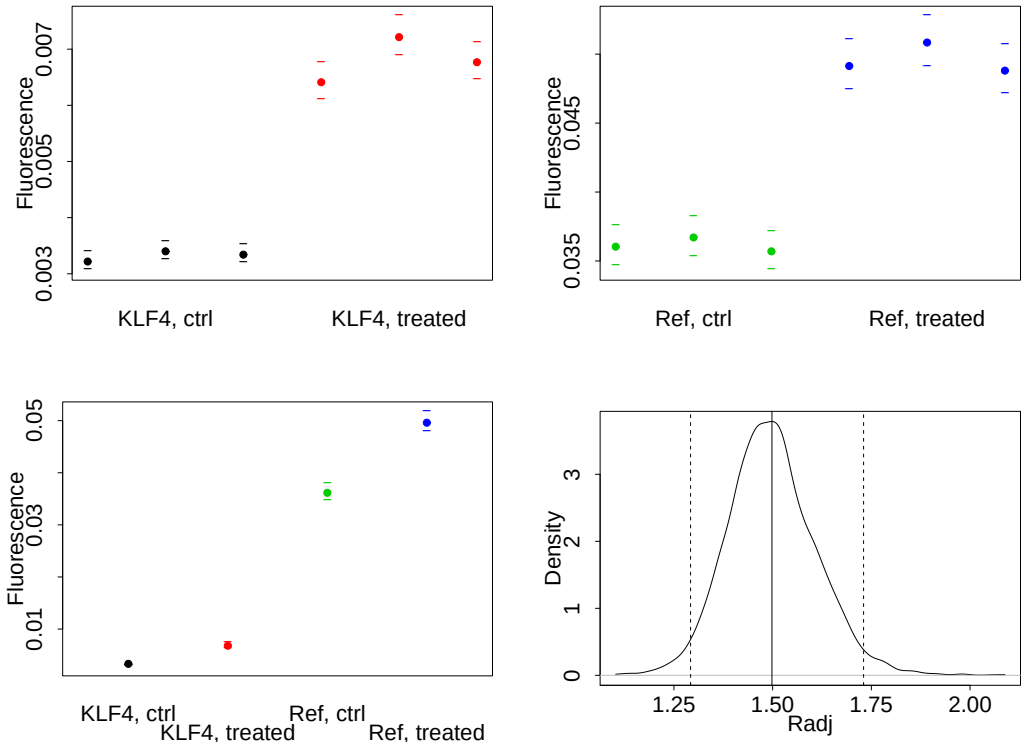

Figure 9: *Estimated posterior means and 95% credibility intervals for the initial fluorescence, $x_{i,0}$ (top panels), for the mean initial fluorescence, $\mu_{g,t}$ (bottom left panel), and for R_{adj} (bottom right panel), for the Octreotide LAR data set. The vertical lines indicate the posterior mean and 95% credibility interval for R_{adj}. Note the differences in range for the y-axes for the two top panels.*

given in (16). This model includes more than data from the exponential phase, but still a large fraction of the data has to be excluded from the fitting process. Further modifications to model (16) to also include measurements from the later part of the curve would be an interesting task to pursue. In principle, any model describing the relationship between efficiency and fluorescence intensity or cycle can be used, as long as a corresponding cycle window, representing the valid cycles for the model, can be identified. As an example, the damped efficiency model of Lalam et al. (2004) and Lalam (2006) will probably allow for more cycles to be included in the analysis.

In the full model in Section 2.1, we motivate the model for the fluorescence intensities by a normal approximation to a binomial distribution of the number of new copies at each cycle. This assumption can be questioned for small copy

DOI: 10.2202/1544−6115.1427

numbers combined with large efficiencies. However, in the simplified model of Section 2.2 the noise in (6) is ignored, and no such assumption is needed.

Due to the Markov structure of the model, the noise-free fluorescence $\boldsymbol{x}$ and the efficiencies $\boldsymbol{p}$ are highly auto- and cross-correlated in the full model. In the simplified model, auto-correlation for the efficiencies is still a source of slow mixing in the single-site MCMC algorithm. Block sampling algorithms are known to improve mixing and convergence properties, and blocking $\boldsymbol{x}$ and $\boldsymbol{p}$, as well as related hyperparameters, could be explored. However, a challenge of such an approach is to find joint proposal distributions that give reasonable acceptance probabilities.

The model is based on the assumption that the amplification curve data are background corrected in a preprocessing step, and the choice of background correction has potentially a strong impact on the results. In principle, background correction could be included in the model by adding a linear or non-linear term to the mean of the likelihood (15). However, parameter identifiability might become an issue in this case, and introducing background correction into the model remains a topic for further study.

We have focused on relative quantification, estimating the ratio R_{adj} (14). However, if an estimate of the scaling factor γ, relating fluorescence to DNA copy number, is available, the approach can in principle also be used to quantify the absolute amount of target DNA in a biological sample from the estimated initial fluorescence.

A Experimental protocol for the Octreotide LAR data

The animal experiments were approved by the Animal Welfare Committee of St.Olav's University Hospital. Two groups of female Sprague-Dawley rats (body weight 193-227 g) were used, one group received Octreotide LAR and a control group received the LAR vehicle. After 21 days the rats were anaesthesized, drained for blood and gastric oxyntic mucosa was isolated. (For a full description of the experimental procedure see Erlandsen et al. (2007)). Total RNA for qPCR was isolated using RNeasy Midi Kit (Qiagen, Valencia, CA). cDNA synthesis and qPCR were performed using iScript$^{\text{TM}}$cDNA Synthesis Kit and iQ$^{\text{TM}}$ SYBR$^{\text{®}}$ Green Supermix (Bio-Rad Laboratories, Hercules, CA), respectively, and the qPCR reactions were done on the Mx3000P$^{\text{TM}}$Real-Time PCR System (Stratagene, La Jolla, CA). The gene of interest and reference gene used in the experiment were the KLF4 (Unigene-ID Rn. 7719) and β-actin (Unigene-ID Rn. 94978). The gene specific primers used for KLF4

were:
forward primer: 5'-CTTGTGACTATGCAGGCTGT-3',
reverse primer: 5'-AGTGCCTGGTCAGTTCATCT-3'.
The primers for the reference gene were:
forward primer: 5'-CTGGCTCCTAGCACCATGA-3',
reverse primer: 5'-AGCCACCAATCCACACAGA-3'.
The PCR temperature profile was: once for 2 min at 95 °C (activation), then 40 reaction cycles of 20 sec at 95 °C (denaturation), 30 sec (optimised annealing temperature), 40 sec at 72 °C (synthesis) and finally 5 min at 72 °C (elongation) and meltingpoint determination.

References

Alvarez, M. J., Vila-Ortiz, G. J., Salibe, M. C., Podhajcer, O. L., and Pitossi, F. J. (2007). Model based analysis of real-time PCR data from DNA binding dye protocols. *BMC Bioinformatics*, 8:85.

Batsch, A., Noetel, A., Fork, C., Urban, A., Lazic, D., Lucas, T., Pietsch, J., Lazar, A., Schömig, E., and Gründemann, D. (2008). Simultaneous fitting of real-time PCR data with efficiency of amplification modeled as Gaussian function of target fluorescence. *BMC Bioinformatics*, 9:95.

Cook, P., Fu, C., Hickey, M., Han, E.-S., and Miller, K. S. (2004). SAS programs for real-time RT-PCR having multiple independent samples. *BioTechniques*, 37(6):990–995.

Erlandsen, S. E., Fykse, V., Waldum, H. L., and Sandvik, A. K. (2007). Octreotide induces apoptosis in the oxyntic mucosa. *Molecular and Cellular Endocrinology*, 264(1-2):188–196.

Gentle, A., Anastasopoulos, F., and McBrien, N. A. (2001). High-resolution semi-quantitative real-time PCR without the use of a standard curve. *BioTechniques*, 31(3):502–508.

Goll, R., Olsen, T., Cui, G., and Florholmen, J. (2006). Evaluation of absolute quantitation by nonlinear regression in probe-based real-time PCR. *BMC Bioinformatics*, 7:107.

Jagers, P. and Klebaner, F. (2003). Random variation and concentration effects in PCR. *Journal of Theoretical Biology*, 224(3):299–304.

DOI: 10.2202/1544−6115.1427

Lalam, N. (2006). Estimation of the reaction efficiency in polymerase chain reaction. *Journal of Theoretical Biology*, 242(4):947–953.

Lalam, N. (2007). Statistical inference for quantitative polymerase chain reaction using a hidden Markov model: A Bayesian approach. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 10.

Lalam, N., Jacob, C., and Jagers, P. (2004). Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Advances in Applied Probability*, 36(2):602–615.

Liu, W. and Saint, D. A. (2002a). A new quantitative method of real time reverse transription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. *Analytical Biochemistry*, 302(1):52–59.

Liu, W. and Saint, D. A. (2002b). Validation of a quantitative method for real time PCR kinetics. *Biochemical and Biophysical Research Communications*, 294(2):347–353.

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(4):402–408.

Marino, J. H., Cook, P., and Miller, K. S. (2003). Accurate and statistically verified quantification of relative mRNA abundances using SYBR Green I and real-time RT-PCR. *Journal of Immunological Methods*, 283(1):291–306.

Rutledge, R. G. (2004). Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of delevoping automated high-troughput applications. *Nucleic Acids Research*, 32(22):e178.

Schnell, S. and Mendoza, C. (1997). Enzymological considerations for a theoretical description of the quantitative competitive polymerase chain reaction (QC-PCR). *Journal of Theoretical Biology*, 184(4):433–440.

Tichopad, A., Dilger, M., Schwarz, G., and Pfaffl, M. W. (2003). Standardized determination of real-time PCR efficiency from a single reaction set-up. *Nucleic Acids Research*, 31(20):e122.