J. Math. Biol. (2009) 59:517–533
DOI 10.1007/s00285-008-0238-3

Mathematical Biology

A quantitative approach for polymerase chain reactions based on a hidden Markov model

Nadia Lalam

Received: 9 August 2007 / Revised: 17 October 2008 / Published online: 5 December 2008

Abstract Polymerase chain reaction (PCR) is a major DNA amplification technology from molecular biology. The quantitative analysis of PCR aims at determining the initial amount of the DNA molecules from the observation of typically several PCR amplifications curves. The mainstream observation scheme of the DNA amplification during PCR involves fluorescence intensity measurements. Under the classical assumption that the measured fluorescence intensity is proportional to the amount of present DNA molecules, and under the assumption that these measurements are corrupted by an additive Gaussian noise, we analyze a single amplification curve using a hidden Markov model (HMM). The unknown parameters of the HMM may be separated into two parts. On the one hand, the parameters from the amplification process are the initial number of the DNA molecules and the replication efficiency, which is the probability of one molecule to be duplicated. On the other hand, the parameters from the observational scheme are the scale parameter allowing to convert the fluorescence intensity into the number of DNA molecules and the mean and variance characterizing the Gaussian noise. We use the maximum likelihood estimation procedure to infer the unknown parameters of the model from the exponential phase of a single amplification curve, the main parameter of interest for quantitative PCR being the initial amount of the DNA molecules. An illustrative example is provided.

Keywords Data analysis · Hidden Markov model · Molecular biology · Monte Carlo expectation maximization algorithm · Polymerase chain reaction

This research was financed by the Swedish foundation for Strategic Research through the Gothenburg Mathematical Modelling Centre.

N. Lalam (✉)
Department of Mathematical Statistics,
Chalmers University of Technology, 412 96 Göteborg, Sweden
e-mail: lalam@math.chalmers.se

Mathematics Subject Classification (2000) 62M05 · 62P10 · 92C40

1 Introduction

Polymerase chain reaction (PCR) has emerged as one of the main tool to amplify the number of a specific fragment of target DNA molecules. This technique has many applications in virology [8], microbiology [34], and gene expression analysis [24,55] to name a few. As concerning the latter application, PCR is preceded by a reverse transcription step, and is referred to as RT-PCR, in order to create DNA templates from mRNA templates.

The quantitative approach of PCR (respectively RT-PCR) aims at determining the initial amount of the DNA (respectively mRNA) molecules present in a biological sample. Several quantification procedures are available in the literature. The most popular one is based on a calibration curve constructed from many amplification curves of a so-called standard [20,32]. Alternative methods relying on a single amplification curve have been proposed. This enables one to reduce costs and to increase throughput analysis because reaction tubes no longer need to be used for the standard curve samples. It may also eliminate the adverse effect of any dilution errors made in creating the standard sample curves [50]. These methods using a single reaction set-up are from very various kinds, and they may be based on either deterministic or stochastic models. Some methods rely on consecutive observations from the exponential phase above the background noise. This phase is identified and modelled by a deterministic geometric series for which the number of DNA molecules X_t, present at replication cycle t, is assumed to be defined by $X_t = X_0(1+p)^t$, where $p \in (0,1)$ is the replication efficiency from the exponential phase [31,40,49,56]. In [1], the authors proposed to use consecutive observations assumed to follow a similar geometric series with a replication efficiency varying with the amount of accumulated molecules.

Other methods based on deterministic models consist in fitting sigmoidal functions for the amplification curve constituted by observations of the amount of replicated molecules from both the exponential and the non-exponential phases [21,42,43]. Using a biophysical analysis of the enzyme activity in the course of PCR, a deterministic model, based on the reaction equations derived from the law of mass actions, was developed in [47].

Some methods account for the randomness inherent to DNA amplification. Stochastic models for the DNA amplification based on the theory of branching processes have been developed for quantitative PCR. They either rely on observations from the exponential phase above the background noise, using then a Galton-Watson branching process model [38], or they rely on observations above the background noise from both the exponential and the non-exponential phases, using then a population-size-dependent branching process model [22,26].

Some models discern small and long molecules [36] and some models account for mutations affecting DNA sequences when they replicate [7,37,51]. But here, we will not take these two features into account.

The main motivation of our study is to provide a tractable statistical method to analyze a single amplification curve based on a sound mathematical model. This

method takes into consideration the stochasticity inherent to the DNA amplification and the stochasticity inherent to the collecting of PCR measurements. We consider PCR experimental data observed through a fluorescence-chemistry based method which is one of the main procedures used to record the kinetic accumulation of DNA molecules.

We present in Sect. 2 a quantitative procedure for analyzing an individual PCR amplification curve relying on a hidden Markov model (HMM). Unknown parameters arising in the proposed formalism are determined using the maximum likelihood estimation method explained in Sect. 3. Usually, the implementation of the maximum likelihood estimators (MLE) in the context of an HMM is done using the Expectation-Maximization (EM) algorithm as described in Sect. 4. In our present model, because the underlying Markov chain has an infinite state space, the EM algorithm is not applicable. Instead, we propose to use a Monte Carlo EM (MCEM) algorithm when considering an approximated model specified in Sect. 5. The method is illustrated in Sect. 6.

2 Mathematical model

The amplification of the number of DNA molecules as PCR proceeds may be dynamically modelled using the branching process theory [25]. PCR is formed by the succession of replication cycles. At each replication cycle, a DNA molecule is either replicated successfully with probability p, or is not replicated with probability $1-p$. We consider the exponential phase of PCR during which we make the classical assumption that p is constant [32] with $0 < p < 1$. We exclude from our analysis the extreme theoretical cases $p = 0$ and $p = 1$ which are of no use in practical PCR experiments: $p = 0$ means that no molecule ever replicates, and $p = 1$ means that all molecules always replicate. Let X_0 be the initial number of DNA molecules, and let X_t be the number of DNA molecules present at replication cycle t. Denote by $Y_{t,i}$ the number of descendant molecules from molecule i from cycle t. If molecule i replicates correctly, then $Y_{t,i} = 2$ with probability p, and $Y_{t,i} = 1$ otherwise with probability $1-p$. We will assume that the offspring $Y_{t,i}$ are all independent and identically distributed (i.i.d.). The number of DNA molecules present at cycle $t+1$ equals then

$$X_{t+1} = \sum_{i=1}^{X_t} Y_{t,i}, \text{ with}$$
$$P(Y_{t,i} = 2) = p = 1 - P(Y_{t,i} = 1).$$

The Markovian process $\{X_t\}$ is a Galton-Watson branching process. Following [46], we will particularly rely on the fact that $\{X_t\}$ satisfies

$$X_{t+1} = X_t + \text{Bin}(X_t, p)$$

because a sum of X_t independent random variables $Y_{t,i} - 1$ distributed as a Bernoulli(p) random variable follows a Binomial(X_t, p) distribution.

In practical PCR experiments, the numbers of DNA molecules as they replicate are not directly accessible. The current method mainly used to measure the amount of

DNA molecules as PCR proceeds relies on fluorescence chemistry [9,16,33,57], and we consider here PCR data obtained with this type of chemistry.

We make the classical assumption that the fluorescence signal emitted by the DNA molecules is proportional to the amount of these molecules [32]. In addition, we assume that the fluorescence data are obtained with additive Gaussian errors. These errors are either assumed independent of the number of DNA molecules (case 1 below), or they are assumed to have a variance depending on the number of DNA molecules (case 2). Therefore, under these assumptions, the fluorescence-chemistry based observation of the number of DNA molecules as they replicate during the exponential phase of PCR may be described by the following HMM: for all $t \in \{1, 2, \ldots, n-1\}$,

$$\begin{cases} X_{t+1} = X_t + \mathrm{Bin}(X_t, p), \\ F_t = \alpha X_t + \varepsilon_t, \text{ with} \\ \text{case 1: } \varepsilon_t \sim N(\mu_t, \sigma_t^2), \text{ or} \\ \text{case 2: } \varepsilon_t | X_t \sim N(\mu_t, \sigma^2 X_t). \end{cases} \tag{1}$$

The initial number X_0 of DNA molecules is assumed constant. The process $\{F_t\}$ is assumed to be a sequence of conditionally independent random variables given the hidden branching process $\{X_t\}$. We consider two different cases. In case 1, X_t and ε_t are independent, the background errors $\{\varepsilon_t\}$ are independent Gaussian random variables with μ_t, respectively σ_t^2, being the mean, respectively the variance, of ε_t. In case 2, the distribution of ε_t conditionally to X_t is assumed Gaussian with mean μ_t and with variance $\sigma_t^2 = \sigma^2 X_t$.

In the HMM terminology, the process $\{X_t\}$ is referred to as the regime, and $\{F_t\}$ as the observational process. For a comprehensive review on HMM's, the interested reader is referred to [13].

Various models for the background noise have been proposed. The authors in [53] considered a constant background noise variance and modelled the background noise mean by $\mu_t = a(1 - \exp\{-bt\}) + c$, where t is the replication cycle. A linear model $\mu_t = at + b$ with constant variance $\sigma_t^2 = \sigma^2$ was used in [21] and [49]. These proposals for the background noise mean do not rely on any biophysical justification concerning the fluorescence signal measurements, but they are rather based on visual inspection of fluorescence data from so-called no template controls which do not contain any DNA to amplify. Measurements from no template controls, which typically consist in four replicates, provide information on the errors from the fluorescence measuring device. It would seem more natural to assume a constant background level, and this is what we will do here.

Perfoming a simulation study in [28], the author investigated model (1) in the particular case 1 with $\mu_t = 0$ and $\sigma_t^2 = \sigma^2$ using a Bayesian framework.

HMM's are a particular instance of graphical models, they are namely dynamic Bayesian network models [18]. The HMM proposed here is schematically represented in Fig. 1.

Within model (1), we assume that the background noise is normally distributed with mean μ_t and variance σ_t^2. We will consider that the mean and variance of the errors ε_t depend on an unknown finite-dimensional parameter denoted by θ_ε. For example, assuming that $\mu_t = \mu$ and $\sigma_t^2 = \sigma^2$ yields $\theta_\varepsilon = (\mu, \sigma^2)$.

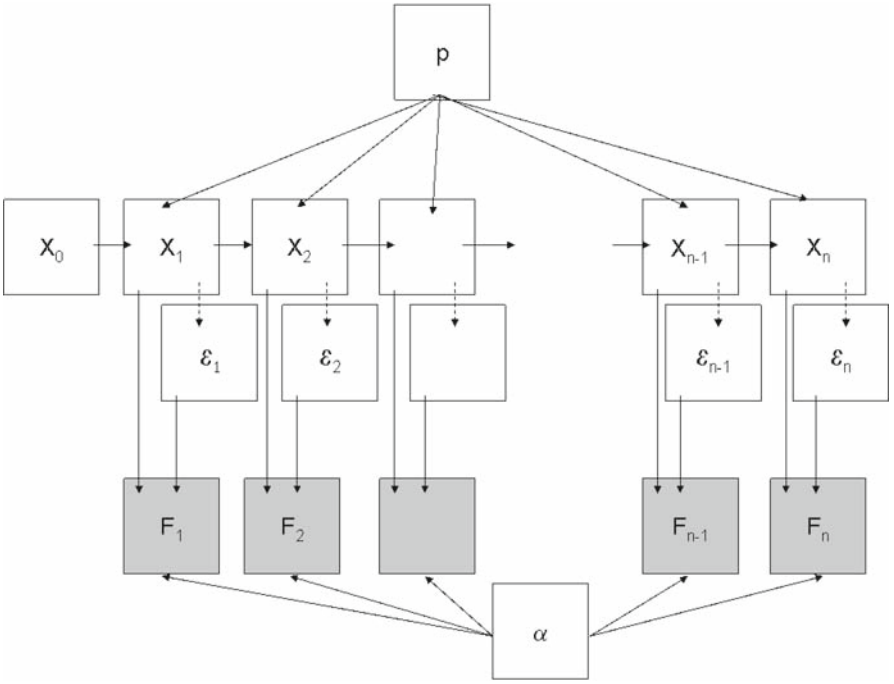

Fig. 1 Graphical representation of model (1) as a dynamic Bayesian network model. A *full line arrow* shows direct dependence between two elements. *Arrows in dashed lines*, accounting for the fact that the distributions of ε_t conditionally to X_t are parts of the model, are present only in case 2. The observable random variables $F_1, F_2, \ldots, F_n$ are in *grey*. The elements p, X_0 and α are deterministic constants, the other elements are random variables

We aim at estimating the unknown parameters of the model from the amplification process and from the observational process. The unknown parameters of the amplification process are the initial number of the DNA molecules X_0 and the replication efficiency p of the PCR exponential phase. The unknown parameter of the observational scheme is the parameter θ_ε characterizing the mean and variance from the Gaussian noise. In case 1, we will in particular consider $\mu_t = \mu$ and $\sigma_t^2 = \sigma^2$; in case 2, we will consider $\mu_t = \mu$. In both cases, the parameter θ_ε reads then $\theta_\varepsilon = (\mu, \sigma^2)$. But the method presented here may also be applied to more general parametric forms for μ_t and σ_t^2. In addition, for the model to be identifiable, we assume that the scale parameter α between the fluorescence level intensity and the number of DNA molecules is known.

We will rely on the observed realizations of $F_1, F_2, \ldots, F_n$ from the exponential phase of a single amplification curve in order to infer $\theta = (X_0, p, \theta_\varepsilon)$. To this end, we will use the maximum likelihood approach.

Remark 1 When considering case 1, one may use data from no template controls in order to infer the parameter θ_ε from the Gaussian noise by the maximum likelihood procedure. One may then use the observations of $F_1, F_2, \ldots, F_n$ to infer $\theta = (X_0, p)$, with θ_ε fixed to its estimated value based on the no template controls data.

3 Maximum likelihood estimation

Let us introduce a few notations which are useful to define the likelihood of the observations to be maximized for deriving the MLE of the true value of the parameter θ in model (1).

The initial distribution of the underlying Markovian process $\{X_t\}$ is denoted by $\pi = (\pi_j : j \in \mathbb{N})$ and satisfies

$$\begin{aligned}\pi_j &= P(X_1 = j)\\ &= P(\text{Bin}(X_0, p) = j - X_0)\\ &= C_{X_0}^{j-X_0} p^{j-X_0}(1-p)^{2X_0-j} \quad \text{with } X_0 \le j \le 2X_0.\end{aligned}$$

We will assume that $X_0 \neq 0$, that is the biological sample contains effectively DNA molecules to amplify. If $X_0 = 0$, then $X_t = 0$ for all $t \in \mathbb{N}$.

The transition matrix $A = (a_{ij})$ of $\{X_t\}$ is such that, for $i \le j \le 2i$,

$$\begin{aligned}a_{ij} &= P(X_{t+1} = j | X_t = i)\\ &= P(X_t + \text{Bin}(X_t, p) = j | X_t = i)\\ &= P(\text{Bin}(i, p) = j - i)\\ &= C_i^{j-i} p^{j-i}(1-p)^{2i-j}.\end{aligned}$$

For $j > 2i$ or $0 \le j < i$, $a_{ij} = 0$. The Markovian process $\{X_t\}$ is said to be homogeneous since a_{ij} does not depend on t.

The conditional density $b(\cdot|x_t)$, or emission distribution in the HMM terminology, is given by

$$b(f_t|x_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{1}{2\sigma_t^2}(f_t - \alpha x_t - \mu_t)^2\right\}.$$

Let us write $F_{1:n} = (F_1, \ldots, F_n)$ and $X_{1:n} = (X_1, \ldots, X_n)$. The likelihood of observing $F_{1:n}$, under the parameter value θ, equals

$$\begin{aligned}P(F_{1:n}|\theta) &= \sum_{x_{1:n}} P(F_{1:n}|x_{1:n}, \theta) P(x_{1:n}|\theta)\\ &= \sum_{x_{1:n}} P(F_1|x_1, \theta) \prod_{t=1}^{n-1} [P(F_{t+1}|x_{t+1}, \theta)] P(x_1|\theta) \prod_{t=1}^{n-1} P(x_{t+1}|x_t, \theta)\\ &= \sum_{x_{1:n}} \left[\prod_{t=1}^{n} b(F_t|x_t)\right] \pi_{x_1} \prod_{t=1}^{n-1} a_{x_t x_{t+1}}. \qquad (2)\end{aligned}$$

The MLE of the true parameter value has no closed analytical expression. Its derivation should be numerically performed, but the direct maximization of the likelihood (2)

is computationally demanding. In the context of HMM's, the derivation of MLE is mainly performed with the EM algorithm [6].

4 EM algorithm

The EM algorithm [10] is the tool of choice to calculate the MLE in an HMM. The EM algorithm is also known as the Baum-Welch algorithm [4], or forward-backward algorithm, in the case of classical finite state space HMM's. It provides a computationally efficient iterative method for local maximization of the log-likelihood function

$$\ell_n(\theta) = \log P(F_{1:n}|\theta).$$

Starting from some initial parameter values, the EM procedure iterates between a step that fixes the current parameters and computes posterior probabilities over the hidden states (the E-step) and a step that uses these probabilities to maximize the expected log-likelihood of the observations as a function of the parameters (the M-step).

More precisely, suppose that an estimate θ_k of the parameter θ is available at the end of the k-th iteration of the algorithm. Let $\widetilde{\theta}$ denote some other estimate of θ. The EM algorithm follows from the use of an auxiliary function defined from the conditional expectation of the log-likelihood of the complete (hidden and observed) data with parameter $\widetilde{\theta}$ for the given observation of $F_{1:n}$ and the current value θ_k in the following way:

E-step

$$Q(\widetilde{\theta}, \theta_k) = E\{\log P(X_{1:n}, F_{1:n}, \widetilde{\theta}|F_{1:n}, \theta_k)\}, \tag{3}$$

where Q is a function of the parameter $\widetilde{\theta}$, given the current parameter estimate θ_k and the observation of $F_{1:n}$. An updated estimate of θ at iteration $k+1$, denoted by θ_{k+1}, is obtained as follows:

M-step

$$\theta_{k+1} = \text{argmax}_{\widetilde{\theta}} Q(\widetilde{\theta}, \theta_k).$$

It was noted in [10] that the inequality $\ell_n(\theta_{k+1}) \geq \ell_n(\theta_k)$ holds if θ_{k+1} maximizes $Q(\theta, \theta_k)$ with respect to θ.

The two steps of the EM algorithm are alternated until the change in the parameters is small. The EM algorithm is proved to converge as the number of iterations k tends to infinity with a fixed number of observations n under some mild assumptions [35,54]. In practice, the algorithm may converge to a local maximum of the likelihood surface of the HMM. A common practice is then to start the EM optimization algorithm from several parameter values.

Maximization of the auxiliary function $Q(\theta, \theta_k)$ for a given sequence $F_{1:n}$ results in re-estimation formulas for the parameter θ. In the case of Gaussian emission distribution and finite state space Markov chain, explicit formulas are available and based on the forward and backward densities [4]. Define the forward density by

$\alpha(x_t, f_{1:t}) = p(x_t, f_{1:t})$ representing the joint density of X_t and the sequence F_1 to F_t, and define the backward density by $\beta(f_{t+1:n}|x_t)$ representing the conditional density of F_{t+1} to F_n given X_t. For $t = 1, \ldots, n$, one has

$$\begin{aligned} p(x_t, f_{1:n}) &= p(x_t, f_{1:t}, f_{t+1:n}) \\ &= p(x_t, f_{1:t}) p(f_{t+1:n}|x_t) \\ &= \alpha(x_t, f_{1:t}) \beta(f_{t+1:n}|x_t). \end{aligned}$$

The forward and backward densities satisfy the following recursions:

$$\alpha(x_t, f_{1:t}) = b(f_t|x_t) \sum_{x_{t-1}} \alpha(x_{t-1}, f_{1:t-1}) a_{x_{t-1}\, x_t}, \quad \text{for all } 2 \leq t \leq n$$

with $\alpha(x_1, f_1) = \pi_{x_1} b(f_1|x_1)$, and

$$\beta(f_{t+1:n}|x_t) = \sum_{x_{t+1}} \beta(f_{t+2:n}|x_{t+1}) a_{x_t\, x_{t+1}} b(f_{t+1}|x_{t+1}), \quad \text{for all } n-1 \geq t \geq 1$$

with $\beta(f_{n+1:n}|x_n) = 1$. Recursions rely on the conditional independence of $(F_1, \ldots, F_t)$ and $(F_{t+1}, \ldots, F_n)$ given X_t, for $t = 1, \ldots, n-1$ (see [39]).

The conditional probability density function $p(x_t|f_{1:n})$, for all $1 \leq t \leq n$, can be calculated as

$$p(x_t|f_{1:n}) = \frac{\alpha(x_t, f_{1:t}) \beta(f_{t+1:n}|x_t)}{\sum_{x_t} \alpha(x_t, f_{1:t}) \beta(f_{t+1:n}|x_t)},$$

and the conditional probability density function $p(x_{t-1}, x_t|f_{1:n})$, for all $2 \leq t \leq n$, satisfies

$$p(x_{t-1}, x_t|f_{1:n}) = \frac{\alpha(x_{t-1}, f_{1:t-1}) \beta(f_{t+1:n}|x_t) a_{x_{t-1}x_t} b(f_t|x_t)}{\sum_{i=1}^{\infty} \sum_{j=i}^{2i} \alpha(i, f_{1:t-1}) \beta(f_{t+1:n}|j) a_{ij} b(f_t|j)}.$$

These quantities appear in the expression of the auxiliary function Q to use in the EM algorithm. The expression of (3) reads here

$$\begin{aligned} Q(\widetilde{\theta}, \theta_k) &= E\{\log P(X_{1:n}, F_{1:n}, \widetilde{\theta}|F_{1:n}, \theta_k)\} \\ &= \sum_{j=1}^{\infty} P(X_1 = j|F_{1:n}, \theta_k) \log \pi_j 1_{\{X_0 \leq j \leq 2X_0\}} \\ &\quad + \sum_{i=1}^{\infty} \sum_{j=i}^{2i} \sum_{t=2}^{n} P(X_{t-1} = i, X_t = j|F_{1:n}, \theta_k) \log a_{ij} \\ &\quad + \sum_{j=1}^{\infty} \sum_{t=1}^{n} P(X_t = j|F_{1:n}, \theta_k) \log b(F_t|X_t = j). \end{aligned}$$

As a consequence, it is not possible to use the exact EM algorithm because it is not feasible to compute forward and backward densities for an infinite number of values. Even if the underlying branching process is restricted to take its values in a finite set, say $\{1, 2, \ldots, X_{\max}\}$, the value of $X_{\max}$ would be very large because X_n grows exponentially fast: for example, if $X_0 = 100$ and $p = 0.8$, if one considers 20 observations, then $X_{20} \leq X_0(1+p)^{20}$ entails that $X_{\max} = 1.275\,10^7$. Such a large value for $X_{\max}$ prevents us from using the exact EM algorithm. We will rather use a MCEM algorithm introduced in [52]. The principle of this algorithm is to replace the E-step by a Monte Carlo integration procedure. Also, we will use an approximation of the likelihood because this will lead to more tractable computations. The approximation will consist in replacing the binomial distribution in (1) by a Gaussian distribution. If one uses the exact likelihood, then the unknown quantity X_0 appears in a combinatorial term and this complicates the maximization step. In addition, in the case of the exact likelihood when considering model (1), one should constrain the underlying Markov chain in such a way that $X_t \leq X_{t+1} \leq 2X_t$, and this would also complicate the procedure. As a consequence, we propose to carry out an MCEM algorithm in an approximated model.

5 MCEM algorithm in the approximated model

5.1 Principle

In order to render the estimation procedure more tractable, we will consider the approximated model

$$\begin{cases} X_{t+1} = X_t + N(X_t p, X_t p(1-p)), \\ F_t = \alpha X_t + \varepsilon_t, \text{ with} \\ \text{case 1: } \varepsilon_t \sim N(\mu_t, \sigma_t^2), \text{ or} \\ \text{case 2: } \varepsilon_t | X_t \sim N(\mu_t, \sigma^2 X_t). \end{cases} \tag{4}$$

Given X_t, the binomial distribution $\mathrm{Bin}(X_t, p)$ from (1) may be reasonably approximated by the normal distribution $N(X_t p, X_t p(1-p))$ if $X_t p \geq 5$ and $X_t(1-p) \geq 5$. Typically, these two inequalities hold when considering realistic values for p and X_t. Indeed, p belongs usually to the range $[0.7; 0.95]$, and X_0 usually varies between a few dozens and a few thousands, and $X_t \geq X_0$ for all $t \in \mathbb{N}$.

When approximating the binomial distribution by its normal counterpart, the transition probability of $\{X_t\}$ reads

$$\begin{aligned} P(X_{t+1} = j | X_t = i) &= P(N(X_t p, X_t p(1-p)) = j - X_t | X_t = i) \\ &= \frac{1}{\sqrt{2\pi i p(1-p)}} \exp\{-\frac{1}{2ip(1-p)}(j - (1+p)i)^2\} \\ &= \tilde{a}_{ij}, \text{ say.} \end{aligned}$$

The initial distribution satisfies

$$P(X_1 = j) = \tilde{a}_{X_0 j} = \tilde{\pi}_j, \text{ say.}$$

Within model (4), we will use the MCEM algorithm. Instead of computing the quantity $Q(\widetilde{\theta}, \theta_k)$ with θ_k the current parameter estimate, one simulates M realizations $x^1, \ldots, x^M$ of the hidden data $X = (X_1, \ldots, X_n) = X_{1:n}$ conditionally on the observable $F_{1:n}$ and given the current estimate θ_k, and then one approximates $Q(\widetilde{\theta}, \theta_k)$ by

$$\widehat{Q}_M(\widetilde{\theta}, \theta_k) = \frac{1}{M} \sum_{m=1}^{M} \log P(x^m, F_{1:n}, \widetilde{\theta}),$$

where, in view of formula (2),

$$P(x^m, F_{1:n}, \widetilde{\theta}) = \left[\prod_{t=1}^{n} b(F_t | x_t^m) \right] \tilde{\pi}_{x_1^m} \prod_{t=1}^{n-1} \tilde{a}_{x_t^m x_{t+1}^m}$$

with

$$b(F_t | x_t^m) = \begin{cases} \frac{1}{\sqrt{2\pi\widetilde{\sigma}^2}} \exp\{-\frac{1}{2\widetilde{\sigma}^2}(F_t - \alpha x_t^m - \widetilde{\mu})^2\} \text{ in case 1,} \\ \frac{1}{\sqrt{2\pi\widetilde{\sigma}^2 x_t^m}} \exp\{-\frac{1}{2\widetilde{\sigma}^2 x_t^m}(F_t - \alpha x_t^m - \widetilde{\mu})^2\} \text{ in case 2.} \end{cases}$$

After re-arranging the terms, in case 1, $P(x^m, F_{1:n}, \widetilde{\theta})$ equals

$$\frac{1}{(2\pi\tilde{\sigma})^n} \frac{1}{\sqrt{\tilde{X}_0 \prod_{t=1}^{n-1} x_t^m}} \frac{1}{(\sqrt{\tilde{p}(1-\tilde{p})})^n} \exp\left\{ -\frac{1}{2\tilde{\sigma}^2} \sum_{t=1}^{n} (F_t - \alpha x_t^m - \tilde{\mu})^2 \right.$$
$$\left. - \frac{1}{2\tilde{X}_0 \tilde{p}(1-\tilde{p})} (x_1^m - (1+\tilde{p})\tilde{X}_0)^2 - \frac{1}{2} \sum_{t=1}^{n-1} \frac{1}{x_t^m \tilde{p}(1-\tilde{p})} (x_{t+1}^m - (1+\tilde{p}) x_t^m)^2 \right\},$$

and in case 2, $P(x^m, F_{1:n}, \widetilde{\theta})$ equals

$$\frac{1}{(2\pi\tilde{\sigma})^n \prod_{t=1}^{n-1} x_t^m} \frac{1}{\sqrt{\tilde{X}_0 x_n^m}} \frac{1}{(\sqrt{\tilde{p}(1-\tilde{p})})^n} \exp\left\{ -\frac{1}{2\tilde{\sigma}^2} \sum_{t=1}^{n} \frac{1}{x_t^m} (F_t - \alpha x_t^m - \tilde{\mu})^2 \right.$$
$$\left. - \frac{1}{2\tilde{X}_0 \tilde{p}(1-\tilde{p})} (x_1^m - (1+\tilde{p})\tilde{X}_0)^2 - \frac{1}{2} \sum_{t=1}^{n-1} \frac{1}{x_t^m \tilde{p}(1-\tilde{p})} (x_{t+1}^m - (1+\tilde{p}) x_t^m)^2 \right\}.$$

The parameter update θ_{k+1} of the k-th iteration of the MCEM algorithm is given by an ordinary M-step applied to $\widehat{Q}_M$:

$$\theta_{k+1} = \text{argmax}_{\widetilde{\theta}} \widehat{Q}_M(\widetilde{\theta}, \theta_k).$$

As a rule of thumb, it is advocated in [52] to increase M as iteration k increases.

Convergence conditions for the MCEM procedure were studied in [6,14], and [45]. It was emphasized in [45] that increased confidence in an MCEM procedure can be obtained by running the procedure with different starting values for the parameters and by checking the nature of the limit points using the Louis method. The Monte Carlo error inherent to the MCEM algorithm was investigated in [30].

In order to simulate a realization x of the hidden data $X_{1:n}$ conditionally to $F_{1:n}$ and to some parameter θ, one may rely on a Markov Chain Monte Carlo (MCMC) sampling scheme. MCMC methods consist in generating a Markov chain whose stationary distribution is the target distribution of interest. After some burn-in time, the realizations of this Markov chain may be viewed as realizations of sampling from the desired distribution. [19] provides an introduction to MCMC methods. The problem of assessing the convergence of an MCMC scheme to the target distribution was investigated in [23].

For θ given, one may update $X_1, \ldots, X_n$ conditionally on $F_{1:n}$ by relying on the Gibbs sampler [15,17]. This sampling scheme is based on the full conditionals of the distribution of interest. It consists in drawing sequentially a realization of a variable according to the distribution of this variable conditionally to all the other variables held fixed. The variables are first assigned arbitrary initial values, and the Markov chain is simulated until it converges to its stationary distribution. More precisely, for θ given, denote the distribution of interest by $\mathcal{L}(X_{1:n}|F_{1:n})$. Consider that the full conditional distributions $\mathcal{L}_i(X_i|F_{1:n}) = \mathcal{L}(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, F_{1:n})$ are available. Gibbs sampling aims at approximating $\mathcal{L}$ when generations from the $\mathcal{L}_i$ are possible. It provides an alternative generation scheme based on successive generations from the full conditional distributions as follows:

Step 1. Set initial values $X_{1:n}^{(0)} = (X_1^{(0)}, \ldots, X_n^{(0)})$.

Step 2. Obtain a new value $X_{1:n}^{(j)} = (X_1^{(j)}, \ldots, X_n^{(j)})$ from $X_{1:n}^{(j-1)}$ through successive generation of values

$$
\begin{aligned}
X_1^{(j)} &\sim \mathcal{L}(X_1|X_2^{(j-1)}, \ldots, X_n^{(j-1)}, F_{1:n}) \\
X_2^{(j)} &\sim \mathcal{L}(X_2|X_1^{(j-1)}, X_3^{(j-1)}, \ldots, X_n^{(j-1)}, F_{1:n}) \\
&\vdots \\
X_n^{(j)} &\sim \mathcal{L}(X_d|X_1^{(j-1)}, \ldots, X_{n-1}^{(j-1)}, F_{1:n}).
\end{aligned}
$$

Step 3. Return to Step 2 until convergence is reached.

5.2 Improvement of the estimation method when the early observations are very noisy

The estimation method that we propose is applicable if the Gaussian noise ε_t in (4) is moderate relatively to the signal αX_t coming from the DNA molecules. In most practical experiments, the early observations are swamped by the measurement noise and, as more and more DNA molecules accumulate, the measurement error becomes smaller relatively to the signal arising from the DNA molecules. In order to take this

feature into account in case 1, we suggest the following adaptation of the estimation method presented above. The early observations contain information on the noise error, whereas subsequent observations provide information on the parameters defining the amplification process. Therefore, one may split the data $F_1, \ldots, F_n$ in such a way that the early observations are used to infer the parameter θ_ε from the Gaussian noise, and the rest of the observations is used to infer (X_0, p). We may use $F_1, \ldots, F_q$, with $q < n$ such that αX_t is negligible relatively to ε_t for $1 \leq t \leq q$, and we proceed by maximum likelihood estimation for inferring $\theta_\varepsilon = (\mu, \sigma^2)$ assuming that the observations come from i.i.d. realizations from a Gaussian distribution $N(\mu, \sigma^2)$ since αX_t is negligible relatively to ε_t for the considered F_t. We may use $F_{h+1}, \ldots, F_n$, with $h + 1 > q$, in order to derive X_h and p based on the MCEM algorithm described in Sect. 5.1 with replacing $F_{1:n}$, $X_{1:n}$ and $\theta = (X_0, p, \theta_\varepsilon)$ by $F_{h+1:n}$, $X_{h+1:n}$, and $\theta = (X_h, p)$, respectively in the notations, and by setting θ_ε to its estimated value based on $F_1, \ldots, F_q$. An estimator of X_0 may then be defined by the estimate of $X_h/(1 + p)^h$ based on the relationship $E(X_h/(1 + p)^h) = X_0$.

5.3 Theoretical properties of the estimators

Within the framework of general HMM's, consistency and asymptotic normality of the MLE, as the number of observations n tends to infinity, have been investigated [5,29], when the Markovian regime is stationary which is a classical assumption. Inferential properties of non-stationary hidden Markov chain models were studied in [2] in the finite state space case when considering a deterministic initial distribution. In our case of interest presented in Sect. 2, the Markovian regime is a non-stationary branching process with infinite state space. The authors from [11] investigated maximum likelihood inference for non-stationary HMM's. They provided consistency and asymptotic normality results for a MLE of the Euclidean parameter upon which the transition kernel of the Markov chain and the conditional distribution of the observations depends. In [12], the authors proved asymptotic properties of the MLE in a possibly non-stationary autoregressive process with Markov regime.

However, these asymptotic properties are of little use in the context of real-time PCR data as one has at hand typically a few dozens of observations.

6 Illustration of the method

An example of application of the method is given. The experimental data at the author's disposal are unfortunately not suitable for the applicability of the proposed methodology. Indeed, with these experimental data, the conversion factor between fluorescence units and numbers of molecules is not known, whereas the presented method assumes that this quantity is given. As a consequence, we propose to illustrate the method with synthetic data. Simulated data are obtained using version 6.0 of Mathematica (Wolfram Inc.). The parameters of the simulation are: $\theta = (X_0, p, \mu, \sigma) = (50, 0.75, 0, 0.01)$. The conversion factor α equals 10^{-8}. The values $X_0 = 50$ and $p = 0.75$ are chosen as they represent realistic parameter values in PCR experiments. The mean and the

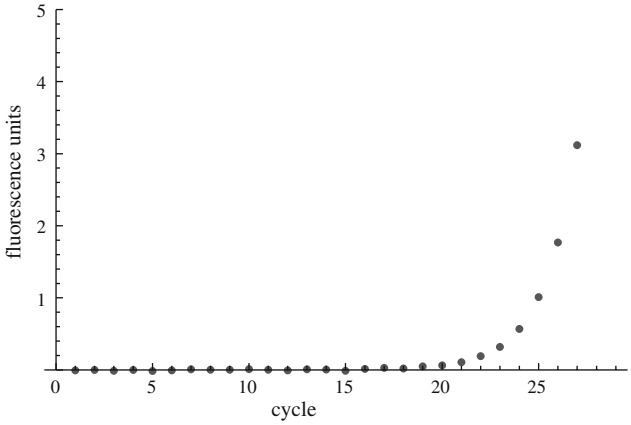

Fig. 2 Simulation of fluorescence data $f_1, \ldots, f_n$ according to case 1 from model (1) with the parameter values $(X_0, p, \mu, \sigma) = (50, 0.75, 0, 0.01)$, with the conversion factor $\alpha = 10^{-8}$, and with the number of replication cycles $n = 30$

standard deviation of the Gaussian noise are taken in such a way that, together with the given conversion factor α, they allow to reproduce the profile from the exponential phase of real PCR data. See for example [21] for profiles of real data. The total number of replication cycles n equals 30. With these quantities, we simulate data from the HMM (1) under case 1. The simulated data are drawn in Fig. 2

We apply the MCEM method described in Sect. 5 with $M = 20$ trajectories of the Markov Chain and with 30 iterations of the EM step. Furthermore, we impose some restrictions on the search space of the parameter in order to make the estimation procedure more efficient:

- the estimator of X_0, denoted by $\widehat{X}_0$, should belong to the range $[X_0^{\text{down}}, X_0^{\text{up}}]$, where we take $X_0^{\text{down}} = 10$ and $X_0^{\text{up}} = 100$;
- the estimator of p, denoted by $\widehat{p}$, should belong to $(0, 1)$;
- the estimator of μ, denoted by $\widehat{\mu}$, should belong to $[\mu^{\text{down}}, \mu^{\text{up}}]$, where we take $\mu^{\text{down}} = -1$ and $\mu^{\text{up}} = 1$;
- the estimator of σ, denoted by $\widehat{\sigma}$, should belong to $]0, \sigma^{\text{up}}]$, where we take $\sigma^{\text{up}} = 5$.

The results depend on the initial parameter value used as a starting point for the MCEM algorithm. With the initial parameter value $(45, 0.5, 0.1, 0.1)$, the results are the following: $\widehat{\theta} = (\widehat{X}_0, \widehat{p}, \widehat{\mu}, \widehat{\sigma}) = (24.06, 0.507, 0.736, 1.972)$.

As the results are far from being satisfying, we follow the procedure described in Sect. 5.2. On the one hand, we rely on $F_1, \ldots, F_q$, with $q = 15$ to infer $\theta_\varepsilon = (\mu, \sigma)$ using the MLE, i.e., $\widehat{\mu} = \frac{1}{q}\sum_{k=1}^{q} F_k$ and $\widehat{\sigma} = \sqrt{\frac{1}{q}\sum_{k=1}^{q}(F_k - \widehat{\mu})^2}$. On the other hand, we rely on $F_{h+1}, \ldots, F_n$, with $h = 15$ to compute the estimator $(\widehat{X}_h, \widehat{p})$ using the MCEM procedure and deducing an estimator of X_0 with the formula $\widehat{X}_h/(1 + \widehat{p})^h$. The obtained results are, using the same starting value for (X_0, p) as above, that is $(45, 0.5)$: $\widehat{\theta} = (\widehat{X}_0, \widehat{p}, \widehat{\mu}, \widehat{\sigma}) = (44.439, 0.499, -1.247\,10^{-3}, 7.691\,10^{-3})$, which are better than the previous results except for the estimate of p. In order to improve the estimation of the efficiency p, one may either impose stronger restrictions on the

search range for this parameter, or use an other method to infer p as suggested in [27] or [41] for example in combination with the MCEM scheme for the other parameters.

7 Concluding remarks

We have described how fluorescence PCR data might be analyzed using an HMM accounting for the stochastic amplification of DNA molecules during the exponential phase, and accounting for the observation of the process with Gaussian errors.

Several quantitative methods are available to analyze PCR experiments. Because these quantitative methods rely on different parts of amplification curves, it is difficult to compare them quantitatively but it is relevant to perform a qualitative comparison. The standard curve-based method is usually used for quantitative PCR. It is based on the assumption that there exists a linear relationship between the threshold cycle at the exponential phase and the logarithm of the amount of molecules. It relies on one data-point from amplification curves of several known dilutions of a standard which has to be designed and validated. The generation of a standard curve is based on the strong assumption that the efficiencies of each dilution sample are equal, which may be questionable [3]. Another method relies on regression based on consecutive observations from the exponential phase of a single reaction set-up [41]. These observations are assumed to be above the background level in such a way that, typically, the first 15–25 observations are not accounted, and the following 4–8 observations are considered. Using a stochastic model based on the theory of branching processes, and relying on a single amplification curve, the reaction efficiency is inferred by conditional least squares estimators based on consecutive observations of the exponential phase in [38], or on consecutive observations spanning from the exponential phase to the early plateau in [27]. But in these approaches, the noise inherent to the PCR data is not explicitly accounted for. Some quantification procedures rely on a fitting of an individual amplification curve by an S-shaped function [21], but these procedures assume a deterministic evolution of the number of molecules with respect to the replication cycle, which is a very strong approximation. The main advantage of the quantitative approach presented here is that it allows one to consider both the uncertainty from the amplification process (intrinsic uncertainty) and the uncertainty from the measurement device (observational uncertainty).

The PCR exponential phase is followed by a linear phase and a plateau for which there is a decrease in PCR efficiency, possibly explained by a decline in DNA polymerase activity or a depletion of certain reaction components [31,48]. It would be challenging to extend the proposed study to account for data belonging to the linear and plateau phases of PCR for which the accumulation of DNA molecules may be modelled by a population-size-dependent branching process [22,27].

Because fluorescence data are measurements of intensity levels, a possible line of investigation consists in performing a data preprocessing step before statistical analysis, e.g., log-transformation of the data, similar to microarray data studies [44].

Acknowledgments The author deeply acknowledges the helpful comments from anonymous referees which led to strong improvements of the manuscript. The author is grateful to professor Peter Jagers for helpful discussions. The author thanks professor Tobias Rydén for suggesting the use of the MCEM in the

approximated normal model. The author extends her thanks to professor Mikael Kubista and doctor Jochen Wilhelm for useful suggestions about the issue of fluorescence signal measurements.

References

1. Alvarez MA, Vila-Ortiz GJ, Salibe MC, Podhajcer OL, Pitossi FJ (2007) Model based analysis of real-time PCR from DNA binding dye protocols. BMC Bioinformatics 8:85
2. Bakry D, Milhaud X, Vandekerkhove P (1997) Statistics of hidden Markov chains, finite state space, nonstationary case. C R Acad Sci Paris Ser I 325(2):203–206
3. Bar T, Ståhlberg A, Muszta A, Kubista M (2003) Kinetic outlier detection (KOD) in real-time PCR. Nucleic Acids Res 31:e105
4. Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171
5. Bickel PJ, Ritov Y, Rydén T (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Ann Stat 26:1614–1635
6. Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer, New York
7. Cariello NF, Swenberg JA, Skopek TR (1991) Fidelity of *Thermococcuslitoralis* DNA polymerase (Vent^{TM}) in PCR determined by denaturing gradient gel electrophoresis. Nucleic Acids Res 19: 4193–4198
8. Cortez KJ, Fischer SH, Fable GA, Calhoun LB, Childs RW, Barrett AJ, Bennett JE (2003) Clinical trial of quantitative real-time polymerase chain reaction for detection of cytomegalovirus in peripheral blood of allogeneic hematopoietic stem-cell transplant recipients. J Infect Dis 188:967–972
9. Crockett AO, Wittwer CT (2001) Fluorescein-labeled oligonucleotides for real-time PCR: using the inherent quenching of deoxyguanosine nucleotides. Anal Biochem 290:89–97
10. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B 39:1–38
11. Douc R, Matias C (2001) Asymptotics of the maximum likelihood estimator for general hidden Markov models. Bernoulli 7(3):381–420
12. Douc R, Moulines E, Rydèn T (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. Ann Stat 32(5):2254–2304
13. Ephraim Y, Merhav N (2002) Hidden Markov processes. IEEE Trans Inform Theory 48:1518–1569
14. Fort G, Moulines E (2003) Convergence of the Monte Carlo expectation maximization for curved exponential families. Ann Stat 31:1220–1259
15. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
16. Gelmini S, Orlando C, Sestini R, Vona G, Pinzani P, Ruocco L, Pazzagli M (1997) Quantitative polymerase chain reaction-based homogeneous assay with fluorogenic probes to measure c-erbB-2 oncogene amplification. Clin Chem 43:752–758
17. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741
18. Ghahramani Z (2001) Introduction to hidden Markov models and Bayesian networks. Int J Pattern Recogn Artif Intell 15:9–42
19. Gilks WR, Richardson S, Spiegelhalter DJE (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London
20. Ginzinger DG (2002) Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol 30:503–512
21. Goll R, Olsen T, Cui G, Florholmen JR (2006) Evaluation of absolute quantitation by nonlinear regression in probe-based real-time PCR. BMC Bioinformatics 7:107
22. Jagers P, Klebaner F (2003) Random variation and concentration effects in PCR. J Theor Biol 224: 299–304
23. Jones GL, Hobert JP (2001) Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Stat Sci 16:312–334
24. Klein D (2002) Quantification using real-time PCR technology: applications and limitations. Trends Mol Med 8:257–260
25. Krawczak M, Reiss J, Schmidtke J, Rosler U (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. Nucleic Acids Res 17:2197–2201

26. Lalam N, Jacob C, Jagers P (2004) Modelling the PCR amplification process by a size-dependent branching process. Adv Appl Probab 36:602–615
27. Lalam N (2006) Estimation of the reaction efficiency in polymerase chain reaction. J Theor Biol 242:947–953
28. 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, article 10
29. Leroux BG (1992) Maximum likelihood estimation for hidden Markov models. Stoch Process Appl 40:127–143
30. Levine R, Casella G (2001) Implementations of the Monte Carlo EM algorithm. J Comput Graph Stat 10:422–439
31. Liu W, Saint DA (2002) A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal Biochem 302:52–59
32. Livak KJ (1997) ABI prism 7700 sequence detection system, user bulletin 2. PE applied biosystems
33. Mackay IM, Arden KE, Nitsche A (2002) Real-time PCR in virology. Nucleic Acids Res 30:1292–1305
34. Mackay IM (2004) Real-time PCR in the microbiology laboratory. Clin Microbiol Infect 10:190–212
35. McLachlan G, Krishnan T (1997) The EM algorithm and extensions. Wiley, London
36. Nedelman J, Heagerty P, Lawrence C (1992) Quantitative PCR: procedures and precisions. Bull Math Biol 54:477–502
37. Olofsson P, Shaw CA (2002) Exact sampling formulas for multi-type Galton-Watson processes. J Math Biol 45:279–293
38. Peccoud J, Jacob C (1998) Statistical estimations of PCR amplification rates. In: Ferré F (ed) Gene quantification. Birkhauser, New York, pp 111–128.
39. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
40. Raeymaekers L (1993) Quantitative PCR: theoretical considerations with practical implications. Anal Biochem 214:582–585
41. Ramakers C, Ruijter JM, Lekanne Deprez RH, Moorman AFM (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339:62–66
42. Rutledge RG (2004) Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. Nucleic Acids Res 32:e178
43. Schlereth W, Bassukas ID, Deubel W, Lorenz R, Hempel K (1998) Use of the recursion formula of the Gompertz function for the quantitation of PCR-amplified templates. Int J Mol Med 1:463–467
44. Sebastiani P, Gussoni E, Kohane IS, Ramoni MF (2003) Statistical challenges in functional genomics. Stat Sci 18:33–60
45. Sherman RP, Ho Y-YK, Dalal SD (1999) Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. Econom J 2:248–267
46. Stolovitzky G, Cecchi G (1996) Efficiency of DNA replication in the polymerase chain reaction. Biophysics 93:12947–12952
47. Stone E, Goldes J, Garlick M (2006) A two stage model for quantitative PCR, The University of Montana, Department of Mathematical Sciences, technical report 5-2006
48. Swillens S, Goffard J-C, Maréchal Y, de Kerchove d'Exaerde A, El Housni H (2004) Instant evaluation of the absolute initial number of cDNA copies from a single real-time PCR curve. Nucleic Acids Res 32:e56
49. Tichopad A, Dilger M, Schwarz G, Pfaffl M (2003) Standardized determination of real-time PCR efficiency from a single reaction set-up. Nucleic Acids Res 31:e122
50. User bulletin 2, ABI PRISM 7700 sequence detection system, applied biosystems, P/N 4303859B, stock no. 777802-002 (2001)
51. Volles MJ, Lansbury PT Jr (2005) A computer program for the estimation of protein and nucleic acid sequence diversity in random point mutagenesis libraries. Nucleic Acids Res 33:3667–3677
52. Wei GCG, Tanner MA (1990) A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J Am Stat Assoc 85:699–704
53. Wilhelm J, Pingoud A, Hahn M (2003) SoFAR: software for fully automatic evaluation of real-time PCR data. Biotechniques 34:324–332
54. Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103
55. Yuan JS, Reed A, Chen F, Stewart CN Jr (2006) Statistical analysis of real-time PCR data. BMC Bioinformatics 7:85

56. Zhao S, Fernald RD (2005) Comprehensive algorithm for quantitative real-time polymerase chain reaction. J Comput Biol 12:1047–1064
57. Zipper H, Brunner H, Bernhagen J, Vitzthum F (2004) Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. Nucleic Acids Res 32:e103