CHALMERS | GÖTEBORG UNIVERSITY

PREPRINT 2007:20

A quantitative approach for Polymerase Chain Reaction based on a Hidden Markov Model

NADIA LALAM

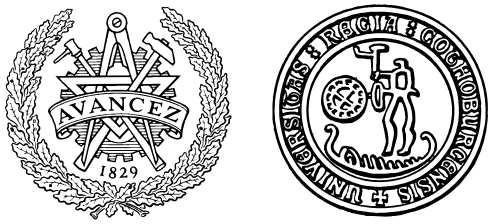
Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg Sweden 2007

Preprint 2007:20

A quantitative approach for Polymerase Chain Reaction based on a Hidden Markov Model

Nadia Lalam

CHALMERS | GÖTEBORG UNIVERSITY



Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, July 2007

Preprint 2007:20
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2007

A quantitative approach for Polymerase Chain Reaction based on a Hidden Markov Model

Nadia Lalam
Chalmers University of Technology
Department of Mathematical Statistics
SE-412 96 Göteborg, Sweden
(lalam@math.chalmers.se)

July 2007

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.

Key words and phrases: Data analysis; Hidden Markov Model; Monte Carlo Expectation Maximization algorithm; Polymerase Chain Reaction.

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 (Cortez et al., 2003), microbiology (Mackay, 2004), and gene expression analysis (Klein, 2002; Yuan et al., 2006) 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 (Livak, 1997; Ginzinger, 2002). 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 (User Bulletin 2, 2001). 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 (Raeymaekers, 1993; Liu and Saint, 2002; Tichopad et al., 2003; Zhao and Fernald, 2005).

Alvarez et al. (2007) 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 (Schlereth et al., 1998; Rutledge, 2004; Goll et al., 2006). Performing a biophysical analysis of the enzyme activity in the course of PCR, Stone et al. (2006) developed a deterministic model based on the reaction equations derived from the law of mass actions.

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 (Peccoud and Jacob, 1998), 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 (Jagers and Klebaner, 2003; Lalam et al., 2004).

Some models discern small and long molecules (Nedelman et al., 1992) and some models account for mutations affecting DNA sequences when they replicate

(Cariello et al., 1991; Olofsson and Shaw, 2002; Volles and Lansbury, 2005). 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. Also, this original approach allows to circumvent the use of standard calibration curves.

We present a quantitative procedure for analyzing an individual PCR amplification curve relying on a Hidden Markov Model (HMM) described in Section 2. We assume that the amplification curve is observed through a fluorescence-chemistry based method which is one of the main procedures used to record the kinetic accumulation of DNA molecules. Unknown parameters arising in this proposed formalism are determined using the maximum likelihood estimation method explained in Section 3. Usually, the implementation of the maximum likelihood estimators in the context of an HMM is done using the Expectation-Maximization (EM) algorithm as described in Section 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 Section 5.

2 Mathematical model

The amplification of the number of DNA molecules as PCR proceeds may be dynamically modelled using the branching process theory (Krawczak et al., 1989). 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 $1-p$. The quantity p is referred to as the replication or reaction efficiency. We will consider the exponential phase of PCR during which we make the classical assumption that p is constant (Livak, 1997) with $0 < p < 1$. 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 Stolovitzky and Cecchi (1996), 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 (Crockett and Wittwer, 2001; Mackay et al., 2002; Zipper et al., 2004), and we will consider here PCR data obtained with this type of chemistry.

We will make the classical assumption that the fluorescence signal emitted by the DNA molecules is proportional to the amount of these molecules (Livak, 1997). In addition, we will assume that the fluorescence data are obtained with additive Gaussian errors. These errors will be either assumed independent of the number of DNA molecules (case 1 below), or they will be 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 $t \in \{1, 2, \ldots, n-1\}$,

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

The process $\{F_t\}$ is assumed to be a sequence of conditionally independent random variables given the hidden branching process $\{X_t\}$. We will 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, see Ephraim and Merhav (2002).

Various models for the background noise have been proposed. Wilhelm et al. (2003) considered a constant background noise variance and they modelled the background noise mean by $\mu_t = a(1 - \exp\{-bt\}) + c$, where t is the replication cycle. Tichopad et al. (2003) and Goll et al. (2006) used a linear model $\mu_t = at+b$ with constant variance $\sigma_t^2 = \sigma^2$. 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, Lalam (2007) 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 (Ghahramani, 2001). The HMM proposed here is schematically represented in Figure 1.

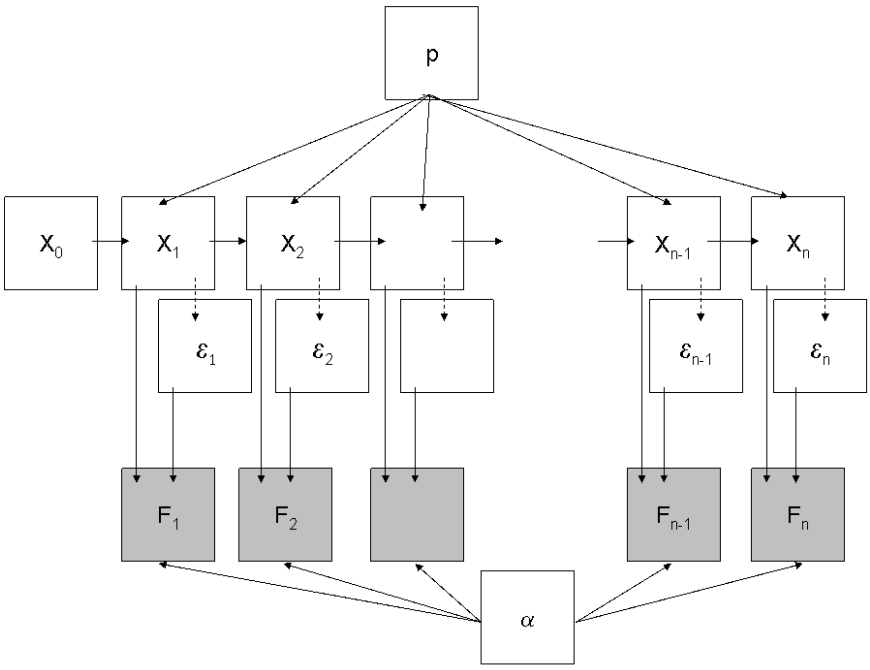


Figure 1: Graphical representation of model (1) as a dynamic Bayesian network model. A full line arrow shows direct dependence between two elements. Arrows in dashed lines, accounting for the fact that the distributions of ε_t conditionally to X_t are parts of the model, are present only in case 2. The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

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 $\theta_\varepsilon = (\mu, \sigma^2)$.

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 reaction 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, ..., 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: 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, ..., 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 maximum likelihood estimator (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

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

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 $i \leq j \leq 2i$,

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

For $j > 2i$ or $0 \leq j < i$, $a_{ij} = 0$. 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

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

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

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

The maximum likelihood estimator 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 maximum likelihood estimators is mainly performed with the Expectation-Maximization (EM) algorithm (Cappé et al, 2005).

4 EM algorithm

The EM algorithm (Dempster et al., 1977) is the tool of choice to calculate the MLE in an HMM. The EM algorithm is also known as the Baum-Welch algorithm (Baum et al., 1970), 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 definition of an auxiliary function, the expected log-likelihood of the complete (hidden and observed) data for the given observation of $F_{1:n}$ and any pair of parameters $\widetilde{\theta}$ and θ_k:
E-step

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

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

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

The log-likelihood $\ell_n(\theta)$ is such that $\ell_n(\theta) = Q(\theta, \theta_k) - H(\theta, \theta_k)$, where

$$H(\theta, \theta_k) = E_{\theta_k}\{\log p(X_{1:n}|F_{1:n}; \theta)|F_{1:n}\}.$$

Dempster et al. (1977) noted 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 (Wu, 1983; McLachlan and Krishnan, 1997). 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 (Baum et al., 1970).

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:

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

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

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

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 $t = 1, \ldots, n-1$ (Rabiner, 1989).

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

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

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 Monte Carlo EM (MCEM) algorithm introduced by Wei and Tanner (1990). 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 also complicates the procedure. As a consequence, we propose to carry out a 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

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

Given X_t, the binomial distribution 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 $X_t(1-p) \geq 5$.

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

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

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, ..., 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),

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

with

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

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

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

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

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

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

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

As a rule of thumb, Wei and Tanner (1990) advocate to increase M as iteration k increases.

Sherman et al. (1999), Fort and Moulines (2003), and Cappé et al. (2005) studied convergence conditions for the MCEM procedure. Sherman et al. (1999) emphasized 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. Levine and Casella (2001) studied the Monte Carlo error inherent to the MCEM algorithm.

In order to simulate a realization x of the hidden data $X_{1:n}$ conditionally to $F_{1:n}$ and to some parameter θ, we propose to 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. Gilks et al. (1996) provide an introduction to MCMC methods. Jones and Hobert (2001) investigated the problem of assessing the convergence of an MCMC scheme to the target distribution.

For θ given, one may update X_1, ..., X_n conditionally on $F_{1:n}$ by relying on the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). 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 relative 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 relative to the signal arising from the DNA molecules. In order to take this feature into account, 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, we propose to 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 $1 \leq t \leq q$. We 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 Subsection 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 maximum likelihood estimator, as the number of observations n tends to infinity, have been investigated (Leroux, 1992; Bickel et al., 1998). 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 Concluding remarks

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

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 (Liu and Saint, 2002; Swillens et al., 2004). 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 (Jagers and Klebaner, 2003; Lalam, 2006).

Because fluorescence data are measurements of intensity levels, a possible line of investigation consists in performing a data preprocessing before statistical analysis, e.g. log-transformation of the data, similar to microarray data studies (Sebastiani et al., 2003).

Acknowledgements: 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. This research was financed by the Gothenburg Mathematical Modelling Centre.

References

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

[2] Baum, L. E., Petrie, T., Soules, G., Weiss N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, 41, 164–171.

[3] Bickel, P. J., Ritov, Y., Rydén, T. (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, *The Annals of Statistics*, 26, 1614-1635.

[4] Cappé, O., Moulines, E., Rydén, T. (2005) Inference in Hidden Markov Models, Springer.

[5] Cariello, N. F., Swenberg, J. A., Skopek, T. R. (1991) Fidelity of Thermococcus litoralis DNA polymerase (Vent^{TM}) in PCR determined by denaturing gradient gel electrophoresis, *Nucleic Acids Research*, 19, 4193–4198.

[6] Cortez, K. J., Fischer, S. H., Fable, G. A., Calhoun, L.B., Childs, R. W., Barrett, A. J., Bennett, J. E. (2003) Clinical trial of quantitative real-time Polymerase Chain Reaction for detection of cytomegalovirus in peripheral blood of allogeneic hematopoietic stem-cell transplant recipients, *The Journal of Infectious Diseases*, 188, 967–972.

[7] Crockett, A. O., Wittwer, C. T. (2001) Fluorescein-labeled oligonucleotides for real-time PCR: using the inherent quenching of deoxyguanosine nucleotides, *Analytical Biochemistry*, 290, 89–97.

[8] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

[9] Ephraim, Y., Merhav, N. (2002) Hidden Markov processes, *IEEE Transactions on Information Theory*, 48, 1518–1569.

[10] Fort, G., Moulines, E. (2003) Convergence of the Monte Carlo Expectation Maximization for curved exponential families, *The Annals of Statistics*, 31, 1220–1259.

[11] Gelfand, A. E., Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85, 398–409.

[12] 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, *Clinical Chemistry*, 43, 752–758.

[13] Geman, S., Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

[14] Ghahramani, Z. (2001) Introduction to Hidden Markov Models and Bayesian Networks, *International Journal of Pattern Recognition and Artificial Intelligence*, 15, 9–42.

[15] Gilks, W. R., Richardson, S. Spiegelhalter, D. J. E. (1996) Markov Chain Monte Carlo in practice. Chapman and Hall.

[16] Ginzinger, D. G. (2002) Gene quantification using real-time quantitative PCR: An emerging technology hits the mainstream, *Experimental Hematology*, 30, 503–512.

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

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

[19] Jones, G. L., Hobert, J. P. (2001) Honest exploration of intractable probability distributions via Markov Chain Monte Carlo, *Statistical Science*, 16, 312–334.

[20] Klein, D. (2002) Quantification using real-time PCR technology: applications and limitations, *Trends in Molecular Medicine*, 8, 257–260.

[21] Krawczak, M., Reiss, J., Schmidtke, J, Rosler, U. (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis, *Nucleic Acids Research*, 17, 2197–2201.

[22] Lalam, N., Jacob, C., Jagers, P. (2004) Modelling the PCR amplification process by a size-dependent branching process, *Advances in Applied Probability*, 36, 602–615.

[23] Lalam, N. (2006) Estimation of the reaction efficiency in Polymerase Chain Reaction, *Journal of Theoretical Biology*, 242, 947–953.

[24] 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.

[25] Leroux, B. G. (1992) Maximum likelihood estimation for hidden Markov models, *Stochastic Processes and their Applications*, 40, 127-143.

[26] Levine, R., Casella, G. (2001) Implementations of the Monte Carlo EM algorithm, *Journal of Computational and Graphical Statistics*, 10, 422–439.

[27] Liu, W., Saint, D. A. (2002) A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics, *Analytical Biochemistry*, 302, 52–59.

[28] Livak, K. J. (1997) ABI Prism 7700 Sequence Detection System, User Bulletin 2. PE Applied Biosystems.

[29] Mackay, I. M., Arden, K. E., Nitsche, A. (2002) Real-time PCR in virology, *Nucleic Acids Research*, 30, 1292–1305.

[30] Mackay, I. M. (2004) Real-time PCR in the microbiology laboratory, *Clinical Microbiology and Infection*, 10, 190–212.

[31] McLachlan, G., Krishnan, T. (1997) The EM algorithm and extensions. John Wiley and Sons.

[32] Nedelman, J., Heagerty, P., Lawrence, C. (1992) Quantitative PCR: Procedures and precisions, *Bulletin of Mathematical Biology*, 54, 477–502.

[33] Olofsson, P., Shaw, C. A. (2002) Exact sampling formulas for multi-type Galton-Watson processes, *Journal of Mathematical Biology*, 45, 279–293.

[34] Peccoud, J., Jacob, C. (1998) Statistical estimations of PCR amplification rates. In Gene Quantification. Ed. Ferré, F., Birkhauser, New-York, pp. 111–128.

[35] Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77, 257–286.

[36] Raeymaekers, L. (1993) Quantitative PCR: Theoretical considerations with practical implications, *Analytical Biochemistry*, 214, 582–585.

[37] Rutledge, R. G. (2004) Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications, *Nucleic Acids Research*, 32, e178.

[38] Schlereth, W., Bassukas, I. D., Deubel, W., Lorenz, R., Hempel, K. (1998) Use of the recursion formula of the Gompertz function for the quantitation of PCR-amplified templates, *International Journal of Molecular Medicine*, 1, 463–467.

[39] Sebastiani, P., Gussoni, E., Kohane, I. S., Ramoni, M. F. (2003) Statistical challenges in functional genomics, *Statistical Science*, 18, 33–60.

[40] Sherman, R. P., Ho, Y.-Y. K., Dalal, S. D. (1999) Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling, *Econometrics Journal*, 2, 248–267.

[41] Stolovitzky, G., Cecchi, G. (1996) Efficiency of DNA replication in the polymerase chain reaction, *Biophysics*, 93, 12947–12952.

[42] 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.

[43] 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 Research*, 32, e56.

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

[45] User Bulletin 2, ABI PRISM 7700 Sequence Detection System (2001) *Applied Biosystems*, P/N 4303859B, Stock No. 777802-002.

[46] Volles, M. J., Lansbury Jr, P. T. (2005) A computer program for the estimation of protein and nucleic acid sequence diversity in random point mutagenesis libraries, *Nucleic Acids Research*, 33, 3667–3677.

[47] Wei, G. C. G., Tanner, M. A. (1990) A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, 85, 699–704.

[48] Wilhelm, J., Pingoud, A., Hahn, M. (2003) SoFAR: Software for Fully Automatic evaluation of Real-time PCR data, *BioTechniques*, 34, 324–332.

[49] Wu, C. F. J. (1983) On the convergence properties of the EM algorithm, *The Annals of Statistics*, 11, 95–103.

[50] Yuan, J. S., Reed, A., Chen, F., Stewart Jr, C. N. (2006) Statistical analysis of real-time PCR data, *BMC Bioinformatics*, 7:85.

[51] Zhao, S., Fernald, R. D., (2005) Comprehensive algorithm for quantitative real-time Polymerase Chain Reaction, *Journal of Computational Biology*, 12, 1047–1064.

[52] 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 Research*, 32, e103.