Estimation of the reaction efficiency in Polymerase Chain Reaction

Nadia Lalam[1], EURANDOM
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

Polymerase Chain Reaction (PCR) is largely used in molecular biology for increasing the copy number of a specific DNA fragment. The succession of 20 replication cycles makes it possible to multiply the quantity of the fragment of interest by a factor of one million. The PCR technique has revolutionized genomics research. Several quantification methodologies are available to determine the DNA replication efficiency of the reaction which is the probability of replication of a DNA molecule at a replication cycle. We elaborate a quantification procedure based on the exponential phase and the early saturation phase of PCR. The reaction efficiency is supposed to be constant in the exponential phase, and decreasing in the saturation phase. We propose to model the PCR amplification process by a branching process which starts as a Galton-Watson branching process followed by a size-dependent process. Using this stochastic modelling and the conditional least squares estimation method, we infer the reaction efficiency from a single PCR trajectory.

Key words and phrases: Polymerase Chain Reaction; Branching process; Reaction efficiency estimation.

1 Introduction

Polymerase Chain Reaction (PCR, Mullis and Faloona, 1987) is an in vitro enzymatic reaction capable of amplifying the number of copies of a specific DNA fragment. Mullis was awarded the 1993 Nobel Prize in chemistry for PCR. This technique is very commonly used in molecular biology since it is a rapid method which makes it possible to detect low abundance of DNA (Mullis et al., 1994). Protocols that quantify rare nucleic acids are increasingly used (Demidov and Broude, 2004). The ability to monitor the DNA molecules quantity as they accumulate thanks to fluorescence-based detection methods has led to a strong impetus in quantitative analyses of PCR (Bustin, 2003). Quantitative PCR (Q-PCR) which aims at determining the initial amount of specific DNA, known as the target, present in a biological sample has many applications in virology (Palmer et al., 2003) or genes expression studies (Ginzinger, 2002; Pfaffl et al., 2004).

PCR is a DNA amplification technology formed by the repetition of typically

[1]Tel: +31 40 247 81 12; Fax: +31 40 247 81 90; Email: lalam@eurandom.tue.nl

30 to 50 replication cycles. The number of copies of the target DNA molecules is doubled at most at each cycle. But in practice, the probability that a molecule will be successfully duplicated after one amplification cycle, known as the efficiency of the reaction, is less than one. The precise determination of the efficiency is required in most quantification methodologies of the initial amount of DNA molecules (Bustin, 2003). The beginning of PCR is characterized by an exponential increase in target molecules. Then, because of a depletion of reaction components or because of a decline in polymerase activity or because of both (Liu and Saint, 2002), the reaction efficiency decreases and eventually ceases leading to a saturation phase decomposed into a linear phase and a plateau phase. Also, in the course of PCR, mutations may occur along the DNA replicated fragments (Krawczak et al., 1989). As this is common in most of the studies related to the determination of the replication efficiency, we will neglect here the copying errors in replication, that is we will assume that all replicated molecules are identical to the initial template target molecules.

In the literature, the theory of Galton-Watson branching processes in discrete time, the time step being a replication cycle, has been extensively used to model the exponential phase of the amplification process. Sun (1995) and Weiss and Von Haeseler (1995) examined replication errors of the DNA polymerase enzyme. Ignoring these copying errors in replication, Stolovitzky and Cecchi (1996) studied the number of cycles during which the PCR amplification process undergoes the exponential phase and may therefore be modelled by a Galton-Watson branching process. In this setting, Peccoud and Jacob (1996, 1998) built conditional least squares estimators of the reaction efficiency. Piau (2001) investigated PCR from a probabilistic perspective. Branching processes provide general population dynamic models, and are used in the modelling and analysis of many biological phenomena (Jagers, 1975; Haccou et al., 2005). Relying on the enzymological approach of PCR performed by Schnell and Mendoza (1997), Jagers and Klebaner (2003) modelled the DNA amplification process by defining a size-dependent branching process with the following replication efficiency

$$p(N_n) = \frac{K}{K + N_n}, \tag{1}$$

where K is a Michaelis-Menten constant type of the reaction, and N_n is the number of DNA molecules present at replication cycle n.

In this paper, we present a statistical procedure to estimate the reaction efficiency from a single amplification trajectory by using a stochastic modelling of the PCR amplification process. The model generalizes the one proposed by Jagers and Klebaner (2003). The PCR amplification process is modelled by a size-dependent branching process which enables one to take into account the stochastic variability of the reaction and both phases of the amplification process, that is the exponential

phase and the saturation phase.

We indicate in section 2 the approximate relationship upon which most current quantification methodologies are based to estimate the reaction efficiency. In section 3, we define a size-dependent branching process modelling of the two phases of the amplification process based on the notion of saturation (Lalam et al., 2004). We assume that there exists a saturation threshold S such that the reaction efficiency $p(N_n)$ at cycle n is a decreasing function of $\mathcal{S}(N_n)/S = \max(N_n/S, 1)$ with $\max(a, b)$ being equal to the maximum between a and b. Therefore, the reaction efficiency is modelled as being constant as long as $N_n < S$, where PCR undergoes its exponential phase, and the reaction efficiency is assumed to decrease when $N_n \geq S$, where PCR is in its saturation phase. The function $p(\cdot)$ is such that the model we propose generalizes (1) and leads to good data fits (Lalam, 2003; Lalam et al., 2004). Furthermore, $p(\cdot)$ is chosen such that we can apply theoretical asymptotic results related to the estimation of the offspring mean of a general size-dependent branching process by the conditional least squares method (Lalam and Jacob, 2004). Note that the asymptotics consist in letting n go to infinity such that N_n goes to infinity. In the PCR setting, even if the replication cycle n is of the order of a few dozens, the corresponding number of molecules N_n is very large due to the exponential phase of the PCR amplification so that theoretical asymptotic results in Lalam and Jacob (2004) may be applied. We define and study the conditional least squares estimators of the reaction efficiency based on a single PCR amplification trajectory. We estimate parameters of the reaction efficiency model, and also the cycle of the end of the exponential phase by using the conditional least squares method. Although our theoretical results are asymptotic and although we rely on a few successive observations to infer the reaction efficiency, we obtain accurate estimators with simulated or real-time PCR data. This is explained by the strong law of large numbers leading to a good precision of the observations at the end of the exponential phase and in the saturation phase. Finally, we conclude the paper by a discussion of the presented results to determine the reaction efficiency. Mathematical technicalities are deferred to the appendix.

2 Mathematical model currently used

The release of systems allowing to collect kinetic PCR data as they are generated during the amplification has revolutionized Q-PCR (Higuchi et al., 1992): at each replication cycle, a measurement of the fluorescence emitted by the accumulated DNA molecules is performed. As noted by Rutledge (2004), the fluorescence chemistry is currently widely used to monitor the amount of DNA molecules amplified by PCR.

Current Q-PCR protocols rely on the exponential phase of the PCR ampli-

fication trajectory[2] (Bustin, 2003). The quantification is based on the classical assumption that the fluorescence measured at cycle n, denoted by F_n, is proportional to the number of DNA molecules N_n, the present DNA molecules being measured thanks to the fluorescence they emit (Kang et al., 2000). Current quantification methodologies rely on the approximation

$$N_n \simeq (1+p)^n N_0, \tag{2}$$

where N_0 is the number of DNA molecules initially present before amplification by PCR, and p is the reaction efficiency assumed constant during the exponential phase (see chapter 3 of Bustin, 2003). The geometric series model (2) does not take into account the stochasticity of the amplification process. Furthermore, most of the current quantification procedures use only one observation per amplification trajectory, assumed to be in the exponential phase, and need many of them. For example, the predominant Q-PCR methodology is a threshold-based procedure requiring the elaboration of standard curves (Rutledge and Côté, 2003). This relies on observations of several PCR trajectories at a replication cycle, known as the fractional cycle, at which the PCR trajectories have identical amounts of replicated DNA molecules. This common amount of DNA molecules is set above the background noise of the DNA quantity measuring device. Van et al. (2005) described a method for determining the fractional cycle relying on the study of the second derivative of the fluorescence signal with respect to the replication cycle. A standard curve is constructed by amplifying known initial amounts of dilutions of a standard assumed to have the same efficiency as the target. Relying on the approximate relationship (2) and on the assumption that the amount of fluorescence signal is proportional to the number of present molecules, the fractional cycle measured for each dilution is proportional to the logarithm of the initial amount of the target. The reaction efficiency is then obtained by regression analysis. This efficiency is assumed to be the same for all the standard dilutions and the target but some authors noticed that this assumption may be questionable (Raeymakers, 1995). Alvarez et al. (2000) conducted a simulation study of the influence that reaction efficiency differences between the target and the standard templates have on target quantification.

Note that the fractional cycle is also called threshold cycle, denoted by C_t, in ABI PRISM^{TM} literature (Applied Biosystems, Foster City, CA, USA), whereas it is called crossing point, denoted by CP in LightCycler^{TM} literature (Roche Applied Science, Indianapolis, IN, USA). See Wong and Medrano (2005) for more detail.

[2]We denote by PCR amplification trajectory, or simply PCR trajectory, the observation of the successive DNA molecule quantities monitored at each replication cycle, that is the observation of the fluorescence counterpart of $\{N_k\}_{1\leq k\leq n_{max}}$ with n_{max} the total number of replication cycles performed. Recall that n_{max} ranges typically between 30 and 50.

Q-PCR requires expensive equipment and reagents. Recently, quantification procedures based on a single PCR trajectory have been proposed (Ramakers et al., 2003; Rutledge, 2004). This presents the advantage of reducing the costs of the PCR experiment. In this study, we propose an alternative method to infer the reaction efficiency from a single PCR trajectory using a particular class of branching processes. This stochastic modelling enables us to account for the variability inherent to the amplification by PCR.

3 Stochastic modelling

As in Jagers and Klebaner (2003), we assume that each molecule can give birth in the next cycle to two identical molecules if the replication succeeds or remains unchanged otherwise. The number of DNA molecules at cycle $n+1$ is given by the recursion formula

$$N_{n+1} = \sum_{i=1}^{N_n} Y_{n+1,i}, \tag{3}$$

where $Y_{n+1,i}$ is the number of offspring at cycle $n+1$ of the *i-th* molecule belonging to cycle n. The random variable $Y_{n+1,i}$ can take only two values: $Y_{n+1,i} = 2$ if molecule i present at cycle n has been successfully replicated, and $Y_{n+1,i} = 1$ otherwise. We assume that, N_n being given, the offspring $\{Y_{n+1,i}\}_i$ are independent and identically distributed (i.i.d.) random variables. The stochastic process $\{N_n\}_n$ thus defined is a branching process. See Haccou et al. (2005) for more detail on branching process theory. If all the random variables $\{Y_{n+1,i}\}_{n,i}$ have a common distribution such that the probability that $Y_{n+1,i} = 2$ is equal to the constant value p, then $\{N_n\}_n$ is a Galton-Watson branching process whose expectation satisfies

$$E(N_n) = (1+p)^n E(N_0). \tag{4}$$

This equality is similar to the approximation (2): relationship (2) is obtained from (4) where the mean of the random variable N_n (respectively N_0) is substituted by N_n (respectively N_0).

We assume here that the replication at a given cycle depends only on the reacting components initially introduced in the reaction tube and on the amount of molecules already synthesized at this cycle. Then, the process $\{N_n\}_n$ may be considered as a size-dependent branching process: the reaction efficiency at cycle n is a function of the number of molecules present at cycle n. We will denote by $p(N_n)$ the efficiency at cycle n.

The aim is estimation of the reaction efficiency $p(\cdot)$ where the whole amplification process is modelled by a size-dependent branching process. The advantage

of using also the saturation phase is that this phase is relatively much less noisy than the exponential phase, and it enables one to use more data for inference. The efficiency $p(N_n)$ is supposed to satisfy the following assumption: there exists a saturation threshold, denoted by $S \geq N_0$, such that, when $N_n < S$, the underlying branching process is considered as a Galton-Watson branching process with constant efficiency $p(N_n) = p$, whereas when $N_n \geq S$, the branching process is a near-critical size-dependent branching process with efficiency $p(N_n)$ decreasing to zero as n (and therefore N_n) increases. We assume that the efficiency $p(N_n)$ is a decreasing function of $\mathcal{S}(N_n)/S$ (recall that $\mathcal{S}(N_n)/S = \max(N_n/S, 1)$). More precisely, we consider the following parametric efficiency model introduced in Lalam et al. (2004):

$$p(N_n) = \begin{cases} \frac{K}{K+S} & \text{if} \quad N_n < S \\ (\frac{K}{K+N_n})(\frac{1+\exp(-C(N_n/S-1))}{2}) & \text{if} \quad N_n \geq S, \end{cases} \tag{5}$$

where K, S and C are unknown parameters in $\mathbb{R}_+^*$. This efficiency model fits into the exponential phase and the linear part of the saturation phase (Lalam et al., 2004). The quantity $K/(K+S) = p$ is the reaction efficiency of the exponential phase. The assumption ($K > 0$ and $S > 0$) yields $0 < p < 1$ which is consistent with real-time PCR experiments. Note that the Galton-Watson branching process modelling the PCR exponential phase is called supercritical because $p > 0$. Model (5) is expressed in the number of DNA molecules whereas the real-time PCR data are expressed in fluorescence units. We make the classical assumption that the fluorescence emitted by the present DNA molecules is proportional to the number of present DNA molecules (Bustin, 2003). Since the proposed efficiency model depends on the ratio $\mathcal{S}(N_n)/S$ for which the proportionality coefficient between fluorescence and DNA molecules simplifies, one can obtain efficiency estimators even when considering real-time measurements expressed in fluorescence units.

Note that model (1) proposed by Jagers and Klebaner (2003) is an efficiency model for which saturation occurs at the beginning of the reaction, $S = N_0$ whereas in model (5), there exists an exponential phase if $S > N_0$.

We estimate the parameters of the efficiency model (5) thanks to the conditional least squares method using $n-h+1$ consecutive observations of the process, starting from the observation at cycle h. The conditional least squares estimator of (K, S, C), denoted by $(\widehat{K}_n, \widehat{S}_n, \widehat{C}_n)$, minimizes the sum of squared differences between the process and its conditional expectation, each squared difference being adequately weighted by a positive quantity of the order of the variance of the process. This sum of squared differences, also called contrast in the statistical literature, will be denoted by $SS_n(K, S, C)$. See the appendix for more detail.
In practice, the starting cycle h taken into account in the contrast will be set large enough so that we do not consider the first noisy observations from the exponen-

tial phase. It is well-known that the early observations are below the background noise and therefore useless for quantitative purposes, and that they become relatively less noisy as more and more DNA molecules accumulate (Bustin, 2003). From a theoretical perspective, one can describe the asymptotic properties of $\widehat{K}_n$ as $n \to \infty$ (Lalam et al., 2004). We denote by n_s the first cycle of the saturation phase: if $n < n_s$ then N_n belongs to the exponential phase ($N_n < S$), and if $n \geq n_s$ then N_n belongs to the saturation phase ($N_n \geq S$). Then let $\Phi_n^{-1}(n_s) = \sqrt{\sum_{k=h+1}^{n_s} (1+p)^{k-1} + n - n_s}$ be the rate of convergence of the estimator. Then we have

Proposition. *(a) Strong consistency:* $\lim_{n\to\infty} \widehat{K}_n \stackrel{a.s.}{=} K$,
(b) Asymptotic distribution:

$$\lim_{n\to\infty} \Phi_n^{-1}(n_s) \frac{\widehat{K}_n - K}{\sqrt{2K}} \stackrel{D}{=} N(0,1). \tag{6}$$

Proof. *See Proposition 5.1 of Lalam et al. (2004).*

This proposition means that on each trajectory, $\{\widehat{K}_n\}_n$ tends to the true parameter value K with a rate of convergence $\Phi_n^{-1}(n_s)$, and that the asymptotic distribution of $\Phi_n^{-1}(n_s)(\widehat{K}_n - K)/\sqrt{2K}$ is the standard Gaussian distribution. This entails that, if the saturation threshold S is known, one can construct a confidence interval for the efficiency of the exponential phase $p = K/(K+S)$. In practice, since the saturation threshold S and the cycle n_s of the end of the non-saturated phase are unknown, one would construct an approximate confidence interval for which the conditional least squares estimator $\widehat{S}_n$ is plugged-in instead of S and the value of n_s is estimated by the cycle $\widehat{n}_s$ such that, from this cycle on, the process is larger than $\widehat{S}_n$.

Although the results of the proposition are asymptotic, one may also obtain accurate estimators at finite n when using a single PCR amplification trajectory. The efficiency model (5) was validated with two data sets obtained on an ABI PRISM^{TM} measuring device (Applied Biosystems, Foster City, CA, USA) when using observations in the exponential phase above the background noise and in the early saturation phase (Lalam et al., 2004). We proceed as follows when analyzing data expressed in fluorescence units. Recall that F_k represents the measured fluorescence at replication cycle k, and is assumed proportional to the number of DNA molecules N_k. Let $\overline{SS}_{h,n}(K,S,C) = SS_n(K,S,C)/(n-h)$ be the normalized contrast. In theory, one should normalize $SS_n(\cdot)$ with the quantity $\Phi_n^{-1}(n_s)$ but, since this rate of convergence contains unknown parameters, we use the normalization $n-h$ instead of $\Phi_n^{-1}(n_s)$. In order to derive the reaction efficiency estimator using the normalized contrast, we consider a window of observations $[h_0, n_0]$ such that the observations belonging to this interval are reliable, that is above the background noise. Cycle h_0 belongs to the exponential phase, and cy-

cle n_0 belongs to the linear part of the saturation for which model (5) is valid (Lalam, 2003; Lalam et al., 2004). Once this window $[h_0, n_0]$ is selected, we consider several windows $[h', n']$ inside $[h_0, n_0]$, with h' from the exponential phase and n' from the early saturation phase, and we search the best window of observations $[h, n]$ included in $[h_0, n_0]$ which leads to the best fit. This means that the observations from cycles h to n are such that $\overline{SS}_{h,n}(\widehat{K}_n, \widehat{S}_n, \widehat{C}_n)$ minimizes the set

$$(\overline{SS}_{h',n'}(\widehat{K}_{n'}, \widehat{S}_{n'}, \widehat{C}_{n'}))_{h_0 \leq h' < \widehat{n}_s^{obs,graph} < n' \leq n_0}. \tag{7}$$

Cycle $\widehat{n}_s^{obs,graph}$ is a graphical estimation of the end of the exponential phase defined as the first cycle of the decrease of 10 consecutive values of the simple estimator of the amplification rate $\{F_k/F_{k-1}\}_k$ (Peccoud and Jacob, 1996). The constraint $h' < \widehat{n}_s^{obs,graph} < n'$ aims at ensuring that cycle h' belongs to the exponential phase, and cycle n' to the saturation phase. We set $h_0 = \sup(k : F_{k-1} \leq 0)$, that is $F_k > 0$ for all $k \geq h_0$, since the values of the measurements of the emitted fluorescence have a meaning only when they are positive. By trial-and-error, we set $n_0 = \widehat{n}_s^{obs,graph} + 7$. In order to compute the estimates more efficiently, the preliminary interval $[h_0, n_0]$ may be given by an experienced experimenter who should select h_0 from the exponential phase and above the background noise, and n_0 from the linear part of the saturation phase.

We present the results obtained for a simulated PCR trajectory and real-time PCR data. The simulation is done as follows. Recall that the offspring $\{Y_{n+1,i}\}_i$ are i.i.d. conditionally to N_n with $Y_{n+1,i} = 2$ when the replication has succeeded, and $Y_{n+1,i} = 1$ otherwise. Since we have proposed that the probability of replication at cycle n is modelled by $p(N_n)$ defined in (5), then $P(Y_{n+1,i} = 2|N_n) = p(N_n)$. Therefore, in view of (3), the process $\{N_n\}_n$ is recursively defined by

$$N_{n+1} = N_n + \text{Bin}(N_n, p(N_n)),\ n \geq 0, \tag{8}$$

where $\text{Bin}(N, q)$ is a random variable having a binomial distribution with parameters N and q. Consequently, in view of (5), once N_0, K, S and C are given, one can simulate $\{N_n\}_n$ using (8). In order to reproduce fluorescence data that are very noisy in the early exponential phase and relatively more accurate after a threshold has been crossed, we add a Gaussian noise to $\{N_n\}_n$. The noise was tuned such that its influence decreases as n increases and disappears after some threshold cycle. Figure 1 shows the estimators obtained with a PCR simulation trajectory for which the true reaction efficiency in the exponential phase is $p = 0.8$. Figure 2 shows the results obtained with a real-time PCR trajectory obtained on ABI PRISM^{TM} 7700 (Applied Biosystems, Foster City, CA, USA) and provided by the Laboratory of Phytopathology and Methodology of Detection, INRA, France. The plot in dotted line is the plot of $F_k/F_{k-1} - 1$ versus

the replication cycle k. The quantity $F_k/F_{k-1} - 1$ is an estimate of the observed efficiency (Peccoud and Jacob, 1996, 1998). The plot in solid line represents the estimator of the efficiency $p(F_{k-1})$ defined in (5) for which (K, S, C) is replaced by $(\widehat{K}_n, \widehat{S}_n, \widehat{C}_n)$ which is the conditional least squares estimator computed for the best window $[h, n]$ chosen by applying criterion (7). Since $p = K/(K + S)$, the efficiency of the exponential phase is estimated by $\widehat{p}_n = \widehat{K}_n/(\widehat{K}_n + \widehat{S}_n)$. The fitted reaction efficiency equals the estimate $\widehat{p}_n$ as long as the replication cycle k is less than the estimated end of the non-saturated phase $\widehat{n}_s = \sup(l : F_{l-1} < \widehat{S}_n)$. For $k \geq n_s$, the fitted efficiency according to model (5) is the following decreasing function of F_k:

$$(\frac{\widehat{K}_n}{\widehat{K}_n + F_k})(\frac{1 + \exp(-\widehat{C}_n(F_k/\widehat{S}_n - 1))}{2}).$$

In Figures 1 and 2, for the early cycles, the ratio F_k/F_{k-1} (dotted lines) behaves very erratically as a result of the fact that the fluorescence values are below the background noise. In the reliable part of the exponential phase, this ratio stabilizes, and then decreases in the saturation part. The obtained fit in Figures 1 and 2 are quite accurate at the end of the exponential phase and in the saturation phase. As concerning Figure 2 related to real-time PCR data, the fit at the end of the saturation phase is relatively less accurate: the solid line is above the dotted line. This is due to the fact that the process undergoes the linear part of the saturation for which model (5) is no more valid.

In order to make a distinction between the parameters related to numbers of molecules and those related to fluorescence data, we denote by K^F (respectively S^F and C^F) the counterpart of the parameter K (respectively S and C) when the quantity of DNA molecules is measured through the fluorescence emitted by the molecules.

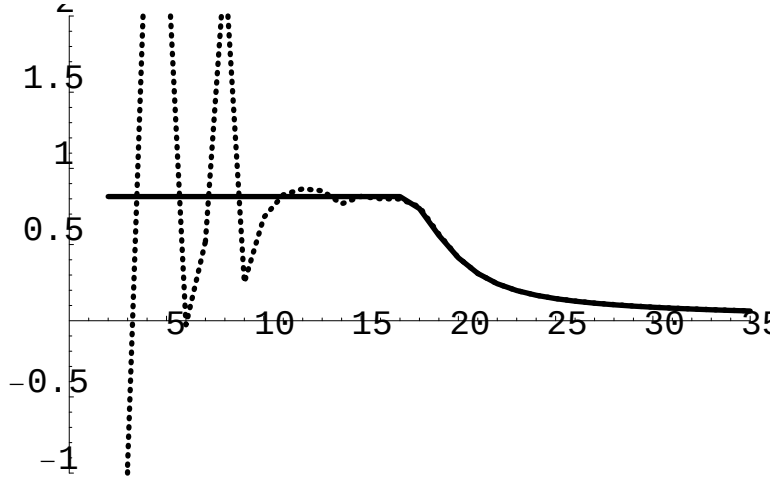

Figure 1: Simulation. y-coordinate: reaction efficiency; x-coordinate: replication cycle; dashed line: observed reaction efficiency; solid line: estimated reaction efficiency. The window of observations selected via criterion (7) is $[h, n] = [14, 24]$. For this window, the estimate of the reaction efficiency of the exponential phase is $\widehat{p}_n = 0.815$, the estimate of the end of the non-saturated phase is $\widehat{n}_s = 17$, and $\widehat{K}_n = 1.965\ 10^6$, $\widehat{S}_n = 4.463\ 10^5$, $\widehat{C}_n = 0.25$.

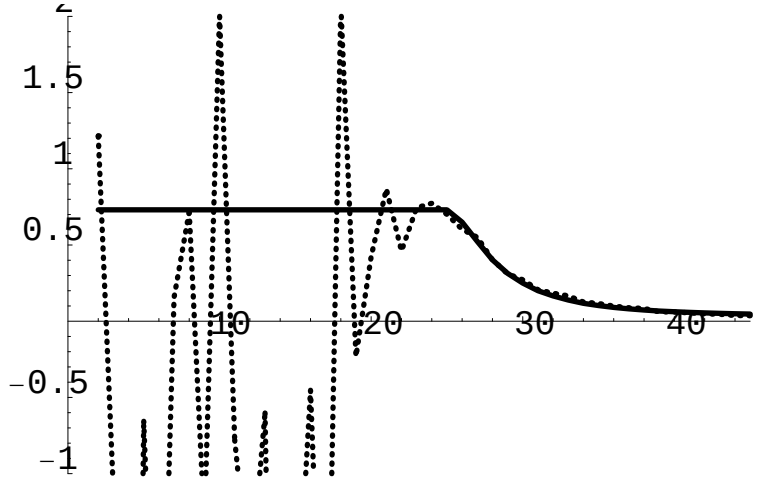

Figure 2: Real-time PCR data. y-coordinate: reaction efficiency; x-coordinate: replication cycle; dashed line: observed reaction efficiency; solid line: estimated reaction efficiency. The window of observations selected via criterion (7) is $[h, n] = [22, 29]$. For this window, the estimate of the reaction efficiency of the exponential phase is $\widehat{p}_n = 0.731$, the estimate of the end of the non-saturated phase is $\widehat{n}_s = 25$, and $\widehat{K}_n^F = 0.254$, $\widehat{S}_n^F = 0.094$, $\widehat{C}_n^F = 0.07$.

4 Discussion

Q-PCR is widespread in molecular biology and has various applications spanning from medical diagnosis to forensic science. The ability to collect data as PCR proceeds thanks to fluorescence-based methods had a tremendous impact on quantitative analyses of PCR. Many quantification procedures are available for Q-PCR (Bustin, 2003), and a growing number of studies addresses PCR efficiency calculations (Larionov et al., 2005). The method we propose allows us to construct an estimator of the reaction efficiency from the observation of consecutive measurements of a single trajectory as opposed to the threshold-based procedure that needs many amplification trajectories for the generation of standard curves. Estimation is based here on a stochastic modelling of the PCR amplification process relying on the size-dependent branching process theory, and generalizing the model of Jagers and Klebaner (2003). This stochastic modelling arises naturally when considering the size evolution of in vitro populations for which the offspring distribution depends on the existing population size, and possibly on saturation phenomena. Based on this modelling, we have provided a novel method to determine the reaction efficiency thanks to a conditional least squares procedure using reliable observations from both phases of a single PCR trajectory, that is the exponential and the early saturation part. Preliminary results on two data sets obtained on an ABI PRISM^{TM} platform (Applied Biosystems, Foster City, CA, USA) have led to satisfying fits (Lalam et al., 2004).

Our method is based on a parametric modelling of the reaction efficiency in both PCR phases as a function of the quantity of present DNA molecules, and this reaction efficiency model is used to define the branching process accounting for the stochastic accumulation of the DNA molecules. However, other quantification methods relying on the kinetics of a single amplification trajectory are available. For example, some authors parameterized directly the fluorescence process F_k versus the replication cycle k. Rutledge (2004) proposed to fit a sigmoid function to the PCR amplification trajectory. Goll et al. (2006) performed nonlinear regressions of fluorescence data F_k versus cycle k by using sigmoid type functions possibly corrected with a linear term to model a baseline drift. They considered fluorescence data belonging to both phases of PCR, and they either defined a specific weight function or they $\log_{10}$ transformed the data to account for the late plateau phase. Tichopad et al. (2003a, b) inferred the reaction efficiency from a single amplification trajectory by using successive observations of the early exponential phase, the other observations being discarded from the estimation via adequate algorithms. The advantage of our method is that it uses both phases of PCR and it accounts for the stochastic variability inherent to the amplification trajectory.

An interesting line of research would be to propose an automated method to

select the preliminary window of observations $[h_0, n_0]$ appearing for selecting the observations from cycles $[h, n]$ in (7) upon which the reaction efficiency is estimated. Future investigation consists in implementing the proposed estimation procedure with other data sets.

5 Appendix

At cycle $n+1$, the conditional probability that $Y_{n+1,i} = 2$, given the number of molecules N_n at the previous cycle, equals $P(Y_{n+1,i} = 2|N_n) = p(N_n)$. Note that, since $Y_{n+1,i}$ equals either 1 or 2, this entails $P(Y_{n+1,i} = 1|N_n) = 1 - p(N_n)$. Furthermore the conditional expectation and variance of $Y_{n+1,i}$, given N_n, reads

$$E(Y_{n+1,i}|N_n) = 1 + p(N_n),\ \text{var}(Y_{n+1,i}|N_n) = p(N_n)(1 - p(N_n)).$$

We denote here $p(N_n)$ by $p_{K,S,C}(N_n)$ to indicate in the notation that the model efficiency (5) is parametric with unknown parameter (K, S, C). The offspring mean model $E(Y_{n+1,i}|N_n) = m(N_n) = 1 + p_{K,S,C}(N_n)$ reads

$$m(N_n) = \begin{cases} 1 + p \text{ if } N_n < S \\ 1 + \frac{K}{2N_n} + r(N_n) \text{ if } N_n \geq S, \end{cases} \tag{9}$$

with the remainder term $r(N_n)$ satisfying

$$r(N_n) = \frac{K}{N_n(K + N_n)}(-\frac{K}{2} + N_n \frac{\exp\left(-C(N_n/S - 1)\right)}{2}) = 0(N_n^{-2}).$$

The notation $a_n = 0(b_n)$ means that $\lim_{n\to\infty} a_n/b_n < \infty$.
In view of (9), it was proved in Lalam et al. (2004) that the conditional least squares estimator of K is strongly consistent (part (a) of the proposition), and that its asymptotic distribution, under appropriate normalization, is Gaussian (part (b)). More precisely, we consider the contrast $SS_n(\cdot)$ defined by the following sum of conditional weighted squares:

$$\begin{aligned} SS_n(K, S, C) &= \sum_{k=h+1}^{n_s} (N_k - (1 + p_{K,S,C}(S))N_{k-1})^2 N_{k-1}^{-1} N_{n_s} \\ &\quad + \sum_{k=n_s+1}^{n} (N_k - (1 + p_{K,S,C}(N_{k-1}))N_{k-1})^2. \end{aligned} \tag{10}$$

The conditional least squares estimator $(\widehat{K}_n, \widehat{S}_n, \widehat{C}_n)$ minimizes $SS_n(K, S, C)$ with respect to K, S, and C.

Acknowledgements

The author is grateful to the referees for their suggestions that improved the presentation of the paper. This study is part of the author's PhD thesis supervised by Dr Christine Jacob, INRA, France. The author thanks Professor Peter Jagers, Chalmers University of Technology, Sweden, and Professor Nicolas Yanev, Academy of Sciences of Sofia, Bulgaria, for having reviewed her PhD thesis. The author also thanks Dr Yves Bertheau, Laboratory of Phytopathology and Methodology of Detection, INRA, France, for providing real-time PCR data.

References

[1] Alvarez, M. J., Depino, A. M., Podhajcer, O. L., Pitossi, F. J., 2000. Bias in estimations of DNA content by competitive Polymerase Chain Reaction. Anal. Biochem., 287, 87–94.

[2] Bustin, S. A., 2003. A-Z of quantitative PCR. IUL Biotechnology series 5.

[3] Demidov, V. V., Broude, N. E., 2004. DNA amplification: current technologies and applications. Horizon Bioscience, Norfolk, United Kingdom.

[4] Ginzinger, D. G., 2002. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp. Hematol., 30, 503–512.

[5] 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.

[6] Haccou, P., Jagers, P., Vatutin, V. A., 2005. Branching processes: variation, growth, and extinction of populations. Cambridge University Press.

[7] Higuchi, R., Dollinger, G., Walsh, P. S., Griffith, R., 1992. Simultaneous amplification and detection of specific DNA sequences. BioTechnology, 10, 413–417.

[8] Jagers, P., 1975. Branching processes with biological applications. Wiley, New York.

[9] Jagers, P., Klebaner, F., 2003. Random variation and concentration effects in PCR. J. Theoret. Biol., 224, 299–304.

[10] Kang, J. J., Watson, R. M., Fisher, M. E., Higuchi, R., Gelfand, D. H., Holland, M. J., 2000. Transcript quantitation in total yeast cellular RNA using kinetic PCR. Nucleic Acids. Res., 28, e2.

[11] 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.

[12] Lalam, N., 2003. Estimation in the setting of size-dependent branching processes. Application to quantitative PCR. PhD thesis in Applied Mathematics, University Paris XI, France.

[13] Lalam, N., Jacob, C., 2004. Estimation of the offspring mean in a supercritical or near-critical size-dependent branching process. Adv. Appl. Prob., 36, 582–601.

[14] Lalam, N., Jacob, C., Jagers, P., 2004. Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. Adv. Appl. Prob., 36, 602-615.

[15] Larionov, A., Krause, A., Miller, W., 2005. A standard curve based method for relative real time PCR data processing. BMC Bioinformatics 2005, 6:62, 16 pp.

[16] Liu, W., Saint, D. A., 2002. Validation of a quantitative method for real time PCR kinetics. Biochemical and Biophysical Research Communications, 294, 347–353.

[17] Mullis, K. B., Faloona, F., 1987. Specific synthesis of DNA in vitro via a polymerase-catalysed chain reaction. Methods Enzymol., 155, 335–350.

[18] Mullis, K. B., Ferré, F., Gibbs, R. A., 1994. The Polymerase Chain Reaction. Birkhauser, New York.

[19] Palmer, S., Wiegand, A. P., Maldarelli, F., Bazmi, H., Mican, J. M., Polis, M., Dewar, R. L., Planta, A., 2003. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. J. Clin. Microbiol., 41, 4531–4536.

[20] Peccoud, J., Jacob, C., 1996. Theoretical uncertainty of measurements using quantitative Polymerase Chain reaction. Biophysical journal, 71, 101–108.

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

[22] Pfaffl, M. W., Tichopad, A., Prgomet, C., Neuvians, T. P., 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper-Excel-based tool using pair-wise correlations. Biotechnol. Lett., 26, 509–515.

[23] Piau, D., 2001. Processus de branchement et champ moyen. Adv. Appl. Prob., 33, 391–403.

[24] Raeymakers, L., 1995. A commentary on the practical applications of competitive PCR. Genome Res., 5, 91–94.

[25] Ramakers, C., Ruijter, J. M., Lekanne Deprez, R. H., Moorman, A. F. M., 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neuroscience Lett., 339, 62–66.

[26] Rutledge, R. G., Côté, C., 2003. Mathematics of quantitative kinetic PCR and the application of standard curves. Nucleic Acids Res., 31, e96.

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

[28] Schnell, S., Mendoza, C., 1997. Enzymological considerations for a theoretical description of the Quantitative Competitive Polymerase Chain Reaction. J. Theoret. Biol., 184, 433–440.

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

[30] Sun, F., 1995. The PCR and branching processes. J. of Computational Biology, 2, 63–86.

[31] Tichopad, A., Dilger, M., Schwarz, G., Pfaffl, M. W., 2003a. Standardized determination of real-time PCR efficiency from a single reaction set-up. Nucleix Acids Res., 31(20), e122.

[32] Tichopad, A., Dilger, M., Schwarz, G., Pfaffl, M. W., 2003b. Erratum: Standardized determination of real-time PCR efficiency from a single reaction set-up. Nucleix Acids Res., 31(22), e122.

[33] Van, T.-L., Paquet, N., Calvo, E., Cumps, J., 2005. Improved real-time RT-PCR method for high-throughput measurements using second derivatives calculations and double correction. Biotechniques, 38, 287–293.

[34] Weiss, G., Von Haeseler, A., 1995. Modeling the PCR. J. of Computational Biology, 2, 49–61.

[35] Wong, M. L., Medrano, J. F., 2005. Real-time PCR for mRNA quantitation. BioTechniques, 39, 11 pp.