C. R. Acad. Sci. Paris, t. 3??, Série I, p. 1–4, 2005 – PXMA????.TEX –
Rubrique/*Mathematics*
(Sous-rubrique/*Statistics*)

Estimation of the PCR efficiency based on a size-dependent modelling of the amplification process

Nadia LALAM [a], Christine JACOB [b], Peter JAGERS [c]

[a] **EURANDOM, P.O. Box 513, 5600 MB Eindhoven, Pays-Bas**
E-mail: lalam@eurandom.tue.nl, tel:+31 40 247 81 12, fax:+31 40 247 81 90

[b] **Unity of Applied Mathematics and Informatics, INRA, 78352 Jouy-en-Josas Cedex, France**
E-mail: christine.jacob@jouy.inra.fr, tel: +331 34 65 22 25, fax: +331 34 65 22 17

[c] **Chalmers University of Technology S-412 96 Göteborg, Suède**
E-mail: jagers@math.chalmers.se, tel: +46 31 772 35 20, fax: +46 31 772 35 08

(Reçu le jour mois année?, accepté le jour mois année?)

Abstract. We propose a stochastic modelling of the PCR amplification process by a size-dependent branching process starting as a supercritical Bienaymé-Galton-Watson transient phase and then having a saturation near-critical size-dependent phase. This model based on the concept of saturation allows one to estimate the probability of replication of a DNA molecule at each cycle of a single PCR trajectory with a very good accuracy. © 2005 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Estimation de l'efficacité de la PCR basée sur une modélisation taille-dépendante du processus d'amplification

Résumé. *Nous proposons une modélisation stochastique du processus d'amplification par PCR s'appuyant sur un processus de branchement taille-dépendant qui débute par une phase transitoire de type Bienaymé-Galton-Watson supercritique et qui présente ensuite une phase de saturation taille-dépendante presque-critique. Cette modélisation basée sur le concept de saturation permet d'estimer la probabilité de réplication d'une molécule d'ADN à chaque cycle d'amplification à partir de l'observation d'une unique trajectoire d'amplification par PCR avec une très bonne précision.* © 2005 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Note présentée par ? ?

S0764-4442(00)0????-?/FLA
© 2005 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS. Tous droits réservés.

1. Introduction

Polymerase Chain Reaction (PCR [9]) is a largely used enzymatic technique of molecular biology allowing to amplify the number of DNA molecules called the target through successive replication cycles. The beginning of the reaction is characterized by an exponential accumulation of target population. Then, because of a depletion of reaction components and a loss of activity of the polymerase enzyme, the probability that a molecule will be duplicated after one replication cycle, called the efficiency, decreases leading to a saturation phase ([15]).
We propose to model the whole PCR amplification process by a size-dependent branching process in order to estimate the efficiency of the reaction from a single amplification trajectory whereas current quantification methods need many amplification trajectories and use only one observation in the exponential phase per trajectory ([1]). The results presented here are detailed in [6] and [7].

2. Stochastic modelling of the amplification process

It is natural to model the amplification process by a branching process. In the literature, several authors used the classical Bienaymé-Galton-Watson model for the exponential phase of the reaction either for estimating the mutation rate (Krawczak et al. [4], Sun [17], Weiss and Von Haeseler [19], Wang et al. [18], Piau [12], [13], [14]), for simulation purposes (Weiss and Von Haeseler [20]) or for estimating some other characteristics of this process (Jacob and Peccoud [2], Peccoud and Jacob [10], [11]). Jagers and Klebaner [3] used a size-dependent branching process relying on the enzymological study of the reaction of Schnell and Mendoza [16]. They proposed the following efficiency

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

where N_n is the molecules number present at cycle n and K is the Michaelis-Menten reaction constant. We model the sequence of DNA molecules number at each replication cycle by a branching process $\{N_n\}_n$. We assume that each molecule can give birth at the next cycle to two identical molecules if the replication succeeds or remains unchanged otherwise. The number of molecules at cycle $n+1$ is given by the formula

$$N_{n+1} = \sum_{i=1}^{N_n} Y_{n+1,i}, \tag{2}$$

where $Y_{n+1,i}$ is the number of descendants at cycle $n+1$ of the ith molecule belonging to cycle n: $Y_{n+1,i} = 2$ when the replication has succeeded, $Y_{n+1,i} = 1$ otherwise. We assume that the $\{Y_{n+1,i}\}_i$ are independent and identically distributed (i.i.d.) random variables conditionally to $\mathcal{F}_n$, the σ-algebra generated by $N_0, \ldots, N_n$, and that the replication depends only on the initial conditions and on the amount of molecules already synthesized. Then, the process may be considered as a size-dependent process and denoting $m(N_n) = E(Y_{n+1,i}|\mathcal{F}_n)$ and $\sigma^2(N_n) = \text{var}(Y_{n+1,i}|\mathcal{F}_n)$, we have

$$P(Y_{n+1,i} = 2|\mathcal{F}_n) = p(N_n), P(Y_{n+1,i} = 1|\mathcal{F}_n) = 1 - p(N_n),$$
$$m(N_n) = 1 + p(N_n), \sigma^2(N_n) = p(N_n)(1 - p(N_n)),$$

where $p(N_n)$ is the efficiency at cycle $n+1$. The efficiency $p(N_n)$ is then modelled according to the following assumption: there exists a saturation threshold, denoted by S with $S \geq N_0$, such that, when $N_n < S$, the underlying branching process may be considered as a supercritical Bienaymé-Galton-Watson process, whereas as soon as $N_n \geq S$, the efficiency decreases. Therefore we assume that the efficiency is a decreasing function of $\mathcal{S}(N)$ that we denote by $p(N)$, where $\mathcal{S}(N) = S1_{\{N<S\}} + N1_{\{N \geq S\}}$ and $1_{\{\}}$ is the indicator function. We will use the following efficiency model generalizing (1) and deduced from a real-time PCR data analysis performed in [8]:

$$p(N) = \frac{K}{K + \mathcal{S}(N)} \frac{1 + \exp\left(-C(\mathcal{S}(N)S^{-1} - 1)\right)}{2}. \tag{3}$$

The efficiency parameter is either (K, S, C) with $C > 0$ or (K, S) when taking $C = 0$. When $S = N_0$ and $C = 0$, (3) is equivalent to (1) and in this case, the parameter to be estimated is K.
Let $n_s = \sup\{n : N_{n-1} < S\}$ be the last cycle of the non-saturated phase. Then, in the exponential phase, $p(N) = K/(K + S) = p$, if $N_n < S$ which is equivalent to $n \le n_s - 1$.

Remark 1. – *The real-time PCR amplification trajectories are observed through the fluorescence emitted by the DNA molecules. We assume that the observed fluorescence F is proportional to the DNA molecules number N which is a classical condition in the PCR setting and we do not take into account the measurement errors. When considering real-time PCR amplification trajectories, the efficiency $p(F)$ expressed in fluorescence units may be written according to (3) where K, S and $\mathcal{S}(N)$ are replaced by their equivalents in fluorescence units K_F, S_F and $\mathcal{S}_F(F) = S_F 1_{\{F<S_F\}} + F 1_{\{F \ge S_F\}}$. K_F and S_F are proportional to K and S, and $\mathcal{S}_F(F)S_F^{-1} = \mathcal{S}(N)S^{-1}$.*

3. Asymptotic behavior of the process

Let $a_0 = 1$ and for all $n \ge 1$, $a_n = a_{n-1}m(a_{n-1})$. Set $K_C = K(1 + 1_{\{C=0\}})/2$.

THEOREM 3.1. – *(i)* $\lim_{n\to\infty} N_n a_n^{-1} \stackrel{a.s.}{=} 1$.
(ii) $\lim_{n\to\infty} a_n n^{-1} = K_C$.

4. Estimation of the efficiency

We estimate the unknown parameters K_C, S and C from a single PCR trajectory by using the nonlinear autoregressive model deduced from (2):

$$N_k = m(N_{k-1})N_{k-1} + \eta_k, \quad m(N_{k-1}) = 1 + p(N_{k-1}),$$
$$E(\eta_k|\mathcal{F}_{k-1}) = 0, \quad \sigma^2(\eta_k|\mathcal{F}_{k-1}) = N_{k-1}\sigma^2(N_{k-1}).$$

Since

$$p(N) = \frac{K_C}{\mathcal{S}(N)} + O_{K,S,C}((\mathcal{S}(N))^{-2}), \tag{4}$$

then $\sigma^2(\eta_k|\mathcal{F}_{k-1}) = O_{K,S,C}(N_{k-1}(\mathcal{S}(N_{k-1}))^{-1})$. So, relying on [5], we define the estimators $(\widehat{K_C}_{h,n}, \widehat{S}_{h,n}, \widehat{C}_{h,n}) = \text{argmin}_{K_C,S,C} SS_{h,n}(K_C, S, C)$, where

$$SS_{h,n}(K_C, S, C) = \sum_{k=h+1}^{n} (N_k - (1 + p(N_{k-1}))N_{k-1})^2 N_{k-1}^{-1}\mathcal{S}(N_{k-1}),$$

where, for $k \le n_s$, the weight $\mathcal{S}(N_{k-1}) = S$ is estimated by N_{n_s}. According to (4), K_C is asymptotically identifiable at the rate N whereas $O_{K,S,C}((\mathcal{S}(N))^{-2})N$ is asymptotically negligible ([5]). Consequently, relying on [21], we will study the properties of $\{\widehat{K_C}_{h,n}\}_n$ and we will consider $\{(\widehat{S}_{h,n}, \widehat{C}_{h,n})\}_n$ as a nuisance parameter. The existence and uniqueness of these estimators are guaranteed asymptotically. Notice that the properties of $\{\widehat{K_C}_{h,n}\}_n$ do not depend on the values of $\{(\widehat{S}_{h,n}, \widehat{C}_{h,n})\}_n$.

THEOREM 4.1. – *Let h be fixed. Then* $\lim_{n\to\infty} \widehat{K_C}_{h,n} \stackrel{a.s.}{=} K_C$ *and*

$$\lim_{n\to\infty} \sqrt{n}(\widehat{K_C}_{h,n} - K_C)/\sqrt{K_C} \stackrel{d}{=} N(0, 1).$$

Remark 2. – *Relying on the exponential phase and on a Conditional Least Squares Estimate $\widehat{p}_n$ of the efficiency p during this phase, Jacob and Peccoud [2] build an asymptotic confidence interval of N_0, as $n \to \infty$. Since $\lim_{n\to\infty} N_n(1 + p)^{-n} \stackrel{a.s.}{=} W_{N_0,p}$, where $E(W_{N_0,p}) = N_0$ and $\sigma^2(W_{N_0,p}) = N_0[1 -$*

$p][1+p]^{-1}$, they considered $\widehat{N}_{0,\widehat{n}_s} = N_{\widehat{n}_s}/(1+\widehat{p}_{\widehat{n}_s})^{\widehat{n}_s}$, where the cycle $\widehat{n}_s$ is an estimator of the end of the exponential phase. One can extend their results to an estimator based on observations until a replication cycle n belonging to the early linear part of the saturation phase where model (3) is adequate. Such an extension is of interest since real-time PCR data are less noisy in the saturation phase.

Acknowledgements. The authors are thankful to anonymous referees who made insightful comments on a preliminary version of this note. We especially thank one referee who suggested some further investigations and who noticed the theoretical overestimation of the true value K when $C = 0$ and $n - h = 1$, but this was not detected on simulation studies.

References

[1] L. P. FERNANDO, P. J. KURIAN, D. J. FERNANDES, Quantitation of Gene-Specific DNA Damage by Competitive PCR, *Analytical Biochemistry*, 306 (2002) 212-221.

[2] C. JACOB, J. PECCOUD, Estimation of the parameters of a branching process from migrating binomial observations, *Adv. in Appl. Probab.*, 30 (1998) 948-967.

[3] P. JAGERS, F. C. KLEBANER, Random variation and concentration effects in PCR, *J. Theor. Biology*, 224/3 (2003) 299-304.

[4] M. KRAWCZAK, J. REISS, J. SCHMIDTKE, U. ROSLER , Polymerase chain reaction : replication errors and reliability of gene diagnosis, *Nucleic Acids Res.*, 17 (1989) 2197-2201.

[5] N. LALAM, C. JACOB, Estimation of the offspring mean in a supercritical or near-critical size-dependent branching process, *Rapport technique*, Unité de biométrie, INRA Jouy-en-Josas (2003).

[6] N. LALAM, C. JACOB, P. JAGERS, Modelling the PCR amplification process with size-dependent branching processes and estimation of the yield, *Rapport technique*, Unité de biométrie, INRA Jouy-en-Josas (2003).

[7] N. LALAM, C. JACOB, P. JAGERS, Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency, *Adv. in Appl. Probab.*, 36 (2004) 602-615.

[8] N. LALAM, Estimation dans le cadre de processus de branchement taille-dépendants. Application à la PCR quantitative, *Thèse de doctorat en Mathématiques appliquées*, Université Paris XI, (2003).

[9] K. B. MULLIS, F. FALOONA, Specific synthesis of DNA in vitro via a polymerase-catalysed chain reaction, *Methods Enzymol.*, 155 (1987) 335-350.

[10] J. PECCOUD, C. JACOB, Theoretical uncertainty of measurements using quantitative Polymerase Chain reaction, *Biophysical Journal*, 71 (1996) 101-108.

[11] J. PECCOUD, C. JACOB, Statistical estimations of PCR amplification rates, in: Ferré F. (Ed.), *Gene Quantification*, Birkhauser, New York, 1998, pp. 111-128.

[12] D. PIAU, Processus de branchement et champ moyen, *Adv. in Appl. Probab.*, 33 (2001) 391-403.

[13] D. PIAU, Immortal branching Markov processes: averaging properties and PCR applications, *Ann. Probab.*, 32 (2004) 337-364.

[14] D. PIAU, Confidence intervals for non homogeneous branching processes and polymerase chain reactions, *Ann. Probab.*, 33 (2005) 674-702.

[15] R. K. SAIKI, D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS, H. A. ERLICH, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science*, 239 (1988) 487-491.

[16] S. SCHNELL, C. MENDOZA, Enzymological considerations for a theoretical description of the Quantitative Competitive Polymerase Chain Reaction, *J. Theor. Biol.*, 184 (1997) 433-440.

[17] F. SUN, The PCR and branching processes, *Journal of Computational Biology*, 2 (1995) 63-86.

[18] D. WANG, C. ZHAO, R. CHENG, F. SUN, Estimation of the mutation rate during error-prone PCR, *Journal of Computational Biology*, 7 (2000) 143-158.

[19] G. WEISS, A. VON HAESELER, Modeling the PCR, *Journal of Computational Biology*, 2 (1995) 49-61.

[20] G. WEISS, A. VON HAESELER, A coalescent approach to the polymerase chain reaction, *Nucleic Acids Res.*, 25 (1997) 3082-3087.

[21] C. F. WU, Asymptotic theory of nonlinear least squares estimation, *Ann. Statist.*, 9 (1981) 501-513.