ML ESTIMATION OF DNA INITIAL COPY NUMBER IN POLYMERASE CHAIN REACTION (PCR) PROCESSES

H. Vikalo[*], *B. Hassibi*[*], *and A. Hassibi*[‡]

[*]Electrical Engineering Department, California Institute of Technology, Pasadena, CA
[‡]Electrical and Computer Engineering Department, University of Texas, Austin, TX

ABSTRACT

Estimation of DNA copy number in a given biological sample is an extremely important problem in genomics. This problem is especially challenging when the number of the DNA strands is minuscule, which is often the case in applications such as pathogen and genetic mutation detection. A recently developed technique, real-time polymerase chain reaction (PCR), amplifies the number of initial target molecules by replicating them through a series of thermal cycles. Ideally, the number of target molecules doubles at the end of each cycle. However, in practice, due to biochemical noise the efficiency of the PCR reaction, defined as the fraction of target molecules which are successfully copied during a cycle, is always less than 1. In this paper, we formulate the problem of joint maximum-likelihood estimation of the PCR efficiency and the initial DNA copy number. As indicated by simulation studies, the performance of the proposed estimator is superior with respect to competing statistical approaches. Moreover, we compute the Cramer-Rao lower bound on the mean-square estimation error.

Index terms – biomedical signal processing, estimation, error analysis

1. INTRODUCTION

Amplification and quantification of DNA strands with polymerase chain reaction (PCR) process [1] is an essential part of many biotechnological procedures. Applications of PCR [2] include genotyping, detection of infectious and hereditary diseases, and genetic fingertyping, to name a few. Typically, a given biological sample contains only a small amount of a target DNA. PCR amplifies the target *in vitro*, relying on an enzymatic replication process in each of its temperature-regulated cycles (typically, $30-40$ of them). A PCR cycle, illustrated in Figure 1, consists of three distinct phases: denaturing, annealing, and extension. During denaturing, the sample is heated (typically, to temperatures above 90°C) to break the hydrogen bonds between strands of the target DNA fragments, creating twice as many single-stranded fragments. Each of the single-stranded DNA fragments serves as a template during the second phase, annealing, when the sample is cooled down to temperatures typically between $40-72^\circ$C. At such temperatures, primers – short sequences of nucleotides, designed to be exact complements to specific regions on the templates – hybridize to the templates. There are two types of primers, one for each of the two types of templates. The primers serve as initiation sites for a DNA polymerase enzyme activated in the last phase of a cycle, extension. The sample is heated to 72°C maximizing the rate of extension while ensuring that the partially extended primers remain attached to the templates. Ideally, at the end of the extension phase, there are twice as many double-stranded target DNA strands as there were at the beginning of the cycle. This implies an exponential growth of the number of the target DNA. However, practical issues affect the replication process adversely and the efficiency of PCR – defined as the probability of generating a replica of each template molecule – is smaller than 1. The random nature of the underlying biochemical process leads to variations in the PCR yield. Moreover, the creation of non-specific byproducts in the replication process further diminishes the purity of the PCR product. The probabilistic nature of the replication process is addressed in [6]-[9], where various stochastic models have been proposed. In [10], the mutations-related effects that plague the efficiency of PCR have been studied.

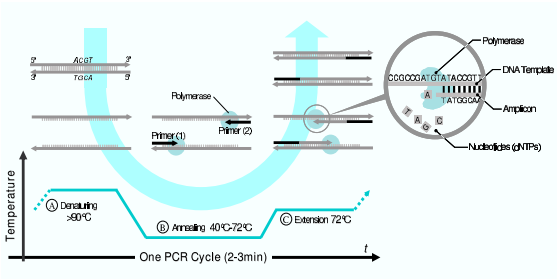

Fig. 1. *Typical PCR cycle*

Quantification of the created amplicons (DNA molecules obtained by the replication of the initial DNA strand) is based on measuring the light intensity originating from the fluorescent reporter molecules incorporated into the

This work was supported in part by the National Science Foundation under grant no. CCR-0133818, by the Office of Naval Research under grant no. N00014-02-1-0578, and by Caltech's Lee Center for Advanced Networking.

double-stranded DNA (dsDNA). One such reporter is SYBR Green I, a dye that binds to dsDNA after which its fluorescence activity increases significantly. Other reporters include hybridization and TaqMan probes (see, e.g., [3]).

In real-time PCR, the fluorescent signal is measured at the end of each cycle. The measured light intensities comprise a reaction profile, usually plotted against the number of cycles. A typical reaction profile has three distinct phases: a background phase, an exponential growth phase, and a saturation phase. During the first phase, the background noise originating from unbound probes dominates the useful signal emanating from the probes attached to the templates. Although the fluorescent level of the unbound probes is much lower than the fluorescent level of the probes bounded to the double-stranded target DNA, the former significantly outnumber the latter during the first $15-20$ cycles. The second phase starts when the signal from the PCR products rises sufficiently above the background noise. Typically, measurements collected during the exponential growth phase (also referred to as the log phase) are the only ones used to infer information about the original number of the DNA targets in the biological sample. The reason for imposing such a restriction is that the efficiency of PCR can be assumed constant in the first two phases, which makes the estimation tractable. In the third phase of PCR, however, the efficiency decreases rapidly as the reaction enters a plateau.

The ultimate goal of PCR is the estimation of the initial number of target molecules. In practice, this is commonly done by comparing a PCR reaction profile with the reaction profile of a so-called standard, where the latter are recorded for several initial concentrations of a target which has the same efficiency as the DNA target of interest. In recent work [7], [8], the reaction profile is used directly (i.e., without use of a standard) to estimate the efficiency, which in turn is then used to find an estimate of the initial number of the DNA target molecules.

In this paper, we find the joint maximum-likelihood estimate of the PCR efficiency and the number of initial target molecules. Furthermore, we find the Cramer-Rao lower bound on the minimum mean-square error of the estimated parameters and illustrate by simulations that the proposed estimator can achieve it.

2. THE MODEL AND JOINT ML ESTIMATION

Let x_0 denote the initial copy number of target DNA molecules which we want to estimate. We assume that the efficiency of replication during both the background phase and the exponential phase is constant, and denote it by p. (During the saturation phase, the efficiency drops as the reaction approaches a plateau. For the sake of simplicity of the estimation procedure, we use only the measurements taken at the end of the cycles wherein the efficiency is constant.) Furthermore, denote the number of target molecules at the end of the n^{th} cycle by x_n, and note that

$$x_n = x_{n-1} + a_n,$$

where a_n is the number of amplicons that have been created in the n^{th} cycle. Since the probability that each of the x_{n-1} available amplicons extends in the n^{th} cycle is p, it is easy to see that a_n is a binomial random variable with mean px_{n-1} and variance $p(1-p)x_{n-1}$. We may therefore write

$$x_n = (1+p)x_{n-1} + \tilde{x}_n, \tag{1}$$

where $\tilde{x}_n$ is a random variable with zero mean and variance $p(1-p)x_{n-1}$. Recursion (1) describes a branching process, often used to model replication in biological systems [5]. It is not too difficult to show (see, e.g., [6]) that the mean of x_n in (1) is given by

$$E\{x_n\} = (1+p)^n x_0. \tag{2}$$

Furthermore, its variance can be found as ([6])

$$\sigma_n^2 = \frac{1-p}{1+p}\left[(1+p)^{2n} - (1+p)^n\right] x_0. \tag{3}$$

Imperfect instrumentation and other biochemistry independent sources create a noise which corrupts the measurements of x_n. We assume that the noise is additive Gaussian $\mathcal{N}(0, \sigma_{\mathbf{w}}^2)$, and denote it by w_n. Hence, the quantity measured is given by

$$z_n = x_n + w_n.$$

Let us denote the number of temperature cycles in the background phase of PCR by k. Therefore, the first measurement taken beyond the background noise level is z_{k+1}. Furthermore, denote the number of temperature cycles in the exponential phase by l. Hence, the last measurement taken before the efficiency starts to rapidly deteriorate is z_{k+l}. Introduce a new variable, $\mathbf{y}$, defined as

$$\mathbf{y} = \begin{bmatrix} \frac{z_{k+1} - (1+p)^{k+1} x_0}{\sigma_{k+1}} \\ \frac{z_{k+2} - (1+p)^{k+2} x_0}{\sigma_{k+2}} \\ \vdots \\ \frac{z_{k+l} - (1+p)^{k+l} x_0}{\sigma_{k+l}} \end{bmatrix}.$$

Note that $\mathbf{y}$ is zero-mean. Finding the exact probability density function (pdf) of $\mathbf{y}$ appears to be difficult. On the other hand, we can express $\mathbf{y}$ as a sum of x_0 identically distributed random variables,

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \cdots + \mathbf{y}_{x_0},$$

where each summand is given by

$$\mathbf{y}_i = \begin{bmatrix} \frac{x_{i,k+1} + w_{k+1}/x_0 - (1+p)^{k+1}}{\sigma_{k+1}} \\ \frac{x_{i,k+2} + w_{k+2}/x_0 - (1+p)^{k+2}}{\sigma_{k+2}} \\ \vdots \\ \frac{x_{i,k+l} + w_{k+l}/x_0 - (1+p)^{k+l}}{\sigma_{k+l}} \end{bmatrix},$$

where $x_{i,j}$ denotes the number of the amplicons at the end of the j^{th} cycle which have originated from the i^{th} target molecule in the original sample. Moreover, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{x_0}$ are independent since they are comprised of the measurements taken in the first two phases of PCR, during which targets do not inhibit each other's replication. Therefore, $\mathbf{y}$ can be represented as a sum of a large number of zero-mean independent, identically distributed (iid) random variables. We thus invoke the central limit theorem to argue that the distribution of $\mathbf{y}$ may be approximated by a multivariate Gaussian distribution.

Note that the (i,j)-entry of the $l \times l$ covariance matrix of $\mathbf{y}$, R, is given by

$$\begin{aligned} R(i,j) &= E\{y_i y_j\} \\ &= \frac{E\{z_{k+i} z_{k+j}\} - (1+p)^{2k+i+j} x_0^2}{\sigma_{k+i}\sigma_{k+j}}, \end{aligned} \quad (4)$$

where y_i and y_j denote the i^{th} and the j^{th} component of $\mathbf{y}$, respectively. We can show that

$$R(i,j) = (1+p)^{j-i}\frac{\sigma_{k+i}}{\sigma_{k+j}} + \frac{\sigma_{\mathbf{w}}^2}{\sigma_{k+i}\sigma_{k+j}}\delta_{i-j},$$

where

$$\delta_{i-j} = \begin{cases} 1, \text{ if } i = j, \\ 0, \text{ otherwise,} \end{cases}$$

Now that we computed the covariance matrix R, the probability density function of $\mathbf{y}$ can be approximated by the multi-variate Gaussian distribution

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{(2\pi)^{l/2}(\det R)^{1/2}} e^{-\frac{1}{2}\mathbf{y}^T R^{-1}\mathbf{y}}. \quad (5)$$

Note that $f_{\mathbf{y}}(\mathbf{y})$ depends on x_0 and p through both $\mathbf{y}$ and R.

2.1. Optimal estimation of x_0 and p

The joint maximum-likelihood estimate of x_0 and p can be found by solving the maximization problem

$$\max_{x_0,p} f_{\mathbf{y}}(\mathbf{y}),$$

or, equivalently, by solving the minimization

$$\min_{x_0,p}\{\mathbf{y}^T R^{-1}\mathbf{y} + \log\det R\}. \quad (6)$$

On the other hand, the traditional approach to the estimation of the initial population in a branching process (see, e.g., [11]) first focuses on finding the maximum-likelihood estimator of p,

$$\hat{p} = \frac{z_{k+1} + \cdots + z_{k+l}}{z_k + \cdots + z_{k+l-1}} - 1. \quad (7)$$

Then, the above estimate $\hat{p}$ is used to estimate x_0 as

$$\hat{x}_0 = \frac{z_{k+l}}{(1+\hat{p})^{k+l}}. \quad (8)$$

Note that for the reliability of the estimate $\hat{p}$ in (7), we only use measurements taken in the exponential phase of the PCR.

The objective function of the optimization (6) is not convex. To solve it, one can use, e.g., a gradient search initialized by $\hat{x}_0$ and $\hat{p}$ obtained from (8) and (7), respectively.

3. LIMITS OF PERFORMANCE OF PCR: THE CRAMER-RAO BOUND

The minimum mean-square error of any estimation procedure is lower bounded by the Cramer-Rao bound [12]. We will derive the Cramer-Rao lower bound (CRLB) and use it to quantify the limits of achievable performance of DNA estimation in PCR.

Collect the parameters that need to be estimated into a vector, $\mathbf{c}^T = [x_0 \quad p]^T$. The Fisher information matrix, F, is given by the negative of the expected value of the Hessian matrix of $\log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})$, i.e.,

$$F = -E_{\mathbf{y}}\left\{\nabla_{\mathbf{c}}\nabla_{\mathbf{c}}^T \log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})\right\}.$$

Therefore, the entries of the 2×2 matrix F are given by

$$F_{ij} = -E_{\mathbf{y}}\left\{\frac{\partial^2}{\partial c_i \partial c_j}\log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})\right\},$$

where, for compactness of the notation, c_i and c_j denote the entries of $\mathbf{c}$ (i.e., $c_1 = x_0$, $c_2 = p$). Assuming an unbiased estimator, the CRLB on the minimum mean-square error of estimating x_0 is given by

$$E\left\{(\hat{x}_0 - x_0)^2\right\} \geq [F^{-1}]_{11}, \quad (9)$$

where $[F^{-1}]_{11}$ denotes the $(1,1)$-entry of F^{-1}. Similarly, the CRLB on the minimum mean-square error of estimating p is

$$E\left\{(\hat{p} - p)^2\right\} \geq [F^{-1}]_{22}, \quad (10)$$

where $[F^{-1}]_{22}$ denotes the $(2,2)$-entry of F^{-1}.

Let us denote $L_1(\mathbf{c}) = \log\det R$ and $L_2(\mathbf{c}) = \mathbf{y}^T R^{-1}\mathbf{y}$, so that we can write

$$L(\mathbf{c}) = \log(p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})) = -\log(2\pi) - \frac{1}{2}L_1(\mathbf{c}) - \frac{1}{2}L_2(\mathbf{c}).$$

Therefore, the Fisher information matrix can be written as

$$F = \frac{1}{2}E_{\mathbf{y}}\left\{\nabla_{\mathbf{c}}\nabla_{\mathbf{c}}^T L_1(\mathbf{c})\right\} + \frac{1}{2}E_{\mathbf{y}}\left\{\nabla_{\mathbf{c}}\nabla_{\mathbf{c}}^T L_2(\mathbf{c})\right\}.$$

It is easy to find the components of $\nabla_{\mathbf{c}}\nabla_{\mathbf{c}}^T L_1(\mathbf{c})$,

$$\frac{\partial^2}{\partial c_i \partial c_j} L_1(\mathbf{c}) = \mathrm{Tr}\left\{ -R^{-1}\frac{\partial R}{\partial c_j} R^{-1}\frac{\partial R}{\partial c_i} + R^{-1}\frac{\partial^2 R}{\partial c_j \partial c_i} \right\}, \tag{11}$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace operation over its argument. Finding the components of $\nabla_{\mathbf{c}}\nabla_{\mathbf{c}}^T L_2(\mathbf{c})$ is somewhat more involved. Due to space limitation, we omit the derivation – although straightforward, the final expressions for F_{ij} is fairly cumbersome to write (for details, we refer the interested reader to [13]). It suffices to say that one may use a symbolic math manipulation package (e.g., *Mathematica*, *Maple*) to efficiently compute F_{ij} for a given set of parameters $(\sigma_{\mathbf{w}}^2, k, l, x_0, p)$. This, in fact, is how we proceed: we use *Mathematica* to evaluate the CRLBs in (9) and (10) for any given set of the RT-PCR experiment parameters.

4. SIMULATION RESULTS AND CONCLUSION

In Figure 2, we compare the mean-square error of the estimate of x_0 computed by (6) and that of (8), and compare them with the Cramer-Rao lower bound. The PCR is simulated as a branching process with $x_0 = 1000$, while the variance of the noise in the exponential phase is assumed to be $1/100$ of the measured signal intensity.

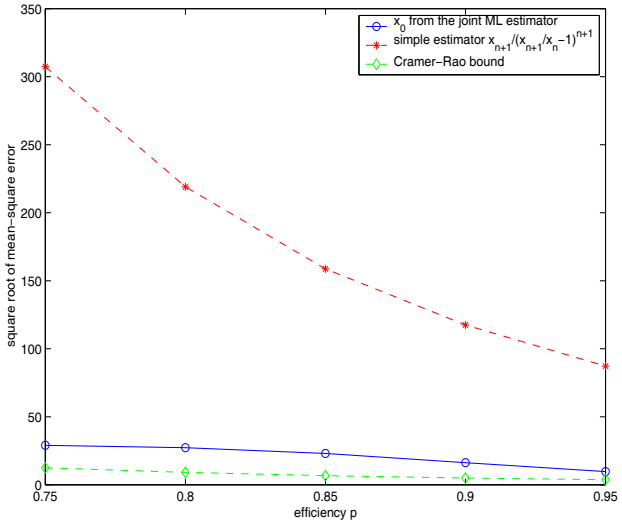

Fig. 2. *Comparison of the estimation mean-square errors* $E(\hat{x}_0 - x_0)^2$

From Figure 2, we see that the proposed joint ML estimator (6) significantly outperforms the estimator (8) for all considered values of p. Furthermore, the mean-square error of the joint ML estimator is almost achieving the Cramer-Rao lower bound. The slight discrepancy could be caused by the approximation of the distribution of $\mathbf{y}$ by a Gaussian.

It is of interest to extend the results presented here to the case where the efficiency is not constant, but changes according to a known model. Furthermore, a study of the PCR in the saturation phase is also of interest.

5. REFERENCES

[1] K. Mullis and F. Faloona, "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction," *Methods Enzymol* 155:335-350, 1987.

[2] M. A. Innis, D. H. Gelfand, and J. J. Sninsky, *PCR Applications: Protocols for Functional Genomics*, Academic Press, 1999.

[3] C.T. Wittwer, M. G. Herrmann, A.A. Moss, and R.P. Rasmussen, "Continuous fluorescence monitoring of rapid cycle DNA amplification," *BioTechniques*, 22:130-138, 1997.

[4] P. M. Holland, R. D. Abramson, R. Watson, and D. H. Gelfand, " Detection of Specific Polymerase Chain Reaction Product by Utilizing the 5" to 3" Exonuclease Activity of Thermus aquaticius DNA polymerase," *PNAS*, 88, 7276-80, 1991.

[5] M. Kimmel and D. E. Axelrod, *Branching Processes in Biology*, Springer-Verlag, New York, 2002.

[6] G. Stolovitzky and G. Cecchi, "Efficiency of DNA replication in the polymerase chain reaction," *PNAS*, vol. 93, pp. 12947-12952, November 1996.

[7] C. Jacob and J. Peccoud, "Estimation of the parameters of a branching process from migrating binomial observation," *Adv. in Appl. Prob.,* 30,948-967, 1998.

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

[9] A. Hassibi, H. Kakavand, and T. H. Lee, "A stochastic model and simulation algorithm for polymerase chain reaction (PCR) systems," GENSIPS 2004.

[10] D. Wang, C. Zhao, R. Cheng, F. Z. Sun, "Estimating the mutation rate during error-prone polymerase chain reaction," *J. of Computational Biology*, 7, 143-158, 2000.

[11] J.-P. Dion, "Estimation of the mean and the initial probabilities of a branching process," *J. of Applied Probability*, 11, 687-694, 1974.

[12] H. Cramer, *Mathematical Models of Statistics*, Princeton University Press, Princeton, NJ 1946.

[13] H. Vikalo, B. Hassibi, and A. Hassibi, "ML estimation of DNA initial copy number in polymerase chain reaction (PCR) process," http://www.its.caltech.edu/˜hvikalo/PCRest.pdf.