MAXIMUM A POSTERIORI (MAP) ESTIMATOR FOR POLYMERASE CHAIN REACTION (PCR) PROCESSES

Hossein Kakavand, Deirdre O'Brien, Arjang Hassibi and Thomas. H. Lee

Department of Electrical Engineering, Stanford University, CA, USA

ABSTRACT

A *maximum a posteriori* (MAP) estimator for polymerase chain reaction (PCR) assays is presented. The estimation relies on the replication efficiencies of the PCR assay and the observed end-point concentration of the PCR product at an arbitrary cycle. This derivation is carried out in view of the stochastic progression of the amplicons, and the assumption that the end-point concentrations for any particular initial value have a jointly Gaussian distribution. In addition, we provide an extension for the estimator which can be applied to various quantitative PCR assays.

1. INTRODUCTION

Amplification and quantification of specific sequences of DNA molecules has become an essential part of many molecular biology procedures and experiments. The original number of target nucleic acid molecules present in a typical sample is usually minuscule, making it practically undetectable via any direct means. Thus nucleic acid amplification is necessary in different areas of research, diagnostics and forensic sciences. Currently various techniques for nucleic acid amplification or quantification are available; yet, an enzymatic amplification methodology which uses temperature cycling, known as polymerase chain reaction (PCR) [1], is by far the most common. Each PCR cycle in theory, should double the number of target DNA strands; yet in practice, PCR shows lower efficiencies (defined as the probability of successful replication of individual strands).

There have been some attempts to model the stochastic process of PCR [2, 3] – all models assign a replication probability to individual DNA strands, which can be considered constant within early and mid cycles of PCR (typically the first 20-25 cycles). Hence, estimation of the initial values of the DNA strands, based on the observed quantities, calls for insight into the probability distribution of the PCR amplicons (i.e. amplified strands). In practice, competitive PCR methods [4] try to address this challenge by amplifying additional samples with a known quantity of analogous nucleic acid molecules (control samples) along with the target sample. In these techniques quantification by comparison is carried out, assuming that the control and the target have matching efficiencies. In addition, existing methods of PCR detection use different reporter molecules (e.g. fluorogenic probes [5]); thus in principle, each method detects certain aspects of amplification.

Typically in quantitative PCR assays, we observe the quantities of the amplified strands or a combination of them, from which we need to *estimate* the initial nucleic acid quantities. In this paper, we model the quantities at each cycle as random variables with statistics depending on the values of the previous cycle and the amplification efficiencies. Based on this model, we attempt to derive the best estimate of the initial DNA concentrations in view of the observable amplicons. In Section 2, we present the basic model and in section 3 we give the solution to the estimation problem, using only the approximation that the observed amplicons have a Gaussian distribution given the initial concentrations. We present the solution for the case where a linear combination of the amplicons is observed in section 4 which is applicable to most PCR detection methods.

2. AMPLIFICATION MODEL

Let A_0 and B_0 be the initial concentrations of the complementary strands of the target DNA, and let A_i and B_i be the corresponding concentrations following cycle i of PCR. Note that during cycle i, each single strand of type B produces a single strand of type A with probability P^i_{AB} (a complete analysis of the probabilities P^i_{AB}, P^i_{BA} is provided in [3]). In cycle i, the concentration of type A strands increases by U^i_{AB}, similarly the concentration of type B strands increases by U^i_{BA}, thus,

$$\begin{aligned} A_i &= A_{i-1} + U^i_{AB} \\ B_i &= B_{i-1} + U^i_{BA}, \end{aligned} \tag{1}$$

where,

$$\begin{aligned} U^i_{AB} &\sim \text{Binom}(P^i_{AB}, B_{i-1}) \\ U^i_{BA} &\sim \text{Binom}(P^i_{BA}, A_{i-1}). \end{aligned} \tag{2}$$

This work was partially funded by NSF (Grant number:CCR-0309701)

U^i_{AB} has a binomial distributions since it is the sum of B_{i-1} independent Bernoulli random trials each with success probability P^i_{AB} [7]. A similar argument follows for U^i_{BA}.

The goal of a typical quantitative PCR assay is to *estimate* A_0 and B_0 given A_n and B_n. From the experimental setup, we know that either $A_0 = B_0$, or $B_0 = 0$, and so in both cases it suffices to estimate A_0. From estimation theory we know that the best estimator of A_0 is the *maximum a posteriori* (MAP) estimator [8], given by $\hat{A}_0$;

$$\hat{A}_0 \quad = \quad \arg\max_{a_0} P(A_0 = a_0 | A_n, B_n)$$

Usually we have no prior knowledge about A_0 except that it lies within a certain range, and so we assume a uniform prior on A_0 over its range. Although not presented here, it is straightforward to extend this analysis to the case where the prior distribution is not uniform but is known. Using Bayes' rule,

$$\begin{aligned} P(A_0|A_n, B_n) &= \frac{P(A_n, B_n|A_0)P(A_0)}{P(A_n, B_n)} \\ \Rightarrow \hat{A}_0 &= \arg\max_{a_0} P(A_n, B_n|A_0 = a_0) \end{aligned}$$

A_n and B_n are strictly integers, however even for small n ($n \geq 5$), A_n and B_n are generally quite large (greater than 100) and the distribution of $A_N, B_N|A_0$ can be well approximated by a joint Gaussian distribution, $f(A_N, B_N|A_0)$. This approximation is supported by our simulation results. Defining the random vector $X_n = [A_n, B_n]^T$ and letting x_n be the observed instance of X_n yields,

$$f(X_n|A_0) = \frac{1}{\sqrt{(2\pi)^2|\Sigma_n|}} e^{(-\frac{1}{2}(X_n-\mu_n)^T\Sigma_n^{-1}(X_n-\mu_n))} \qquad (3)$$

where μ_n and Σ_n are given by,

$$\begin{aligned} \mu_n &= \begin{bmatrix} \mu_{A,n} \\ \mu_{B,n} \end{bmatrix} = \begin{bmatrix} E[A_n|A_0] \\ E[B_n|A_0] \end{bmatrix} \\ \Sigma_n &= \begin{bmatrix} \sigma^2_{A,n} & \sigma_{AB,n} \\ \sigma_{AB,n} & \sigma^2_{B,n} \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(A_n|A_0) & \text{cov}(A_n, B_n|A_0) \\ \text{cov}(A_n, B_n|A_0) & \text{var}(B_n|A_0) \end{bmatrix} \end{aligned}$$

From (1) and (2) it follows that,

$$\begin{aligned} \begin{bmatrix} \mu_{A,n} \\ \mu_{B,n} \end{bmatrix} &= \begin{bmatrix} 1 & P^n_{AB} \\ P^n_{BA} & 1 \end{bmatrix} \begin{bmatrix} \mu_{A,n-1} \\ \mu_{B,n-1} \end{bmatrix} \\ &= \prod_{i=1}^{n} \begin{bmatrix} 1 & P^i_{AB} \\ P^i_{BA} & 1 \end{bmatrix} \begin{bmatrix} \mu_{A,0} \\ \mu_{B,0} \end{bmatrix} \end{aligned}$$

Similar derivations were carried out for the covariance matrix, details can be found in [9]. Combining the results gives,

$$\begin{bmatrix} \mu_{A,n} \\ \mu_{B,n} \\ \sigma^2_{A,n} \\ \sigma^2_{B,n} \\ \sigma_{AB,n} \end{bmatrix} = \prod_{i=1}^{n} G_i \begin{bmatrix} \mu_{A,0} \\ \mu_{B,0} \\ \sigma^2_{A,0} \\ \sigma^2_{B,0} \\ \sigma_{AB,0} \end{bmatrix} = H_n \begin{bmatrix} \mu_{A,0} \\ \mu_{B,0} \end{bmatrix} \qquad (4)$$

where each G_i is given by,

$$\begin{bmatrix} 1 & P^i_{AB} & 0 & 0 & 0 \\ P^i_{BA} & 1 & 0 & 0 & 0 \\ 0 & P^i_{AB}\bar{P}^i_{AB} & 1 & (P^i_{AB})^2 & 2P^i_{AB} \\ P^i_{BA}\bar{P}^i_{BA} & 0 & (P^i_{BA})^2 & 1 & 2P^i_{BA} \\ 0 & 0 & P^i_{BA} & P^i_{AB} & (1 + P^i_{AB}P^i_{BA}) \end{bmatrix}$$

Note that G_i depends only on the efficiencies at cycle i, i.e. P^i_{AB}, P^i_{BA}. The second equation 4 follows since given A_0, $\sigma^2_{A,0} = \sigma^2_{B,0} = \sigma_{AB,0} = 0$, and so H_n is the matrix formed by taking the first two columns of $\prod_{i=1}^{n} G_i$. Also given A_0, $\mu_{A,0} = A_0$ and $\mu_{B,0}$ is either A_0 or 0. This implies,

$$\begin{bmatrix} \mu_{A,n} & \mu_{B,n} & \sigma^2_{A,n} & \sigma^2_{B,n} & \sigma_{AB,n} \end{bmatrix}^T \quad = \quad \tilde{H}_n A_0,$$

where $\tilde{H}_n = [\tilde{h}_1\ \tilde{h}_2\ \tilde{h}_3\ \tilde{h}_4\ \tilde{h}_5]^T$ is a vector, equal to the sum of the first two columns of H_n if $A_0 = B_0$ and equal to the first column of H_n if $B_0 = 0$. Thus,

$$\mu_n \quad = \quad \begin{bmatrix} \tilde{h}_1 \\ \tilde{h}_2 \end{bmatrix} A_0 = m_n A_0 \qquad (5)$$

$$\Sigma_n \quad = \quad \begin{bmatrix} \tilde{h}_3 & \tilde{h}_5 \\ \tilde{h}_5 & \tilde{h}_4 \end{bmatrix} A_0 = K_n A_0, \qquad (6)$$

where m_n and K_n depend only on the efficiencies. Since both the mean and variance of A_n and B_n are linear in A_0, the variance relative to the mean remains constant with varying A_0. Thus for large values of A_0, the range of values of A_n, B_n relative to the size of the mean ($\mu_{A,n}, \mu_{B,n}$) decreases. This is emphasized by considering the logarithm of the values as in Fig. 1.

3. ESTIMATION

In this section we shall derive the MAP estimator of A_0 given A_n and B_n. Based on the model in section 2 the estimator is given by

$$\begin{aligned} \hat{A}_0 &\approx \arg\max_{a_0} f(X_n|A_0 = a_0) \\ &= \arg\min_{a_0} \left(-\ln f(X_n|A_0 = a_0)\right) \end{aligned}$$

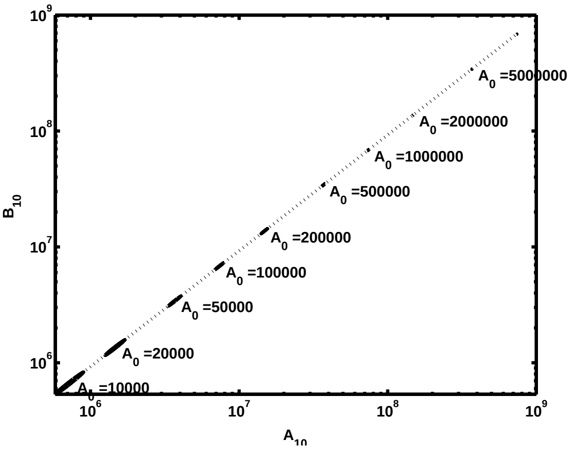

Fig. 1. Distribution of A_{10} and B_{10} given A_0, with $B_0 = 0$, $P^i_{AB} = 0.7$, $P^i_{BA} = 0.6$. $E[A_{10}, B_{10}|A_0]$ for different values of A_0 lie along the dotted line. The ellipses around the expectations show the positions where $f(A_{10}, B_{10}|A_0) > e^{-1000}$. It is reasonable to assume that all data generated for a particular A_0 lies well inside its ellipse.

Substituting (5) and (6) into (3) yields,

$$\begin{aligned}
&-\ln f(x_n|A_0)\\
&= \frac{1}{2}\ln((2\pi)^2|K_n|) + \ln(A_0) +\\
&\quad \frac{1}{2A_0}(x_n - m_nA_0)^T K_n^{-1}(x_n - m_nA_0)\\
&= \frac{1}{2}\ln((2\pi)^2|K_n|) + \ln(A_0) + \frac{x_n^T K_n^{-1} x_n}{2A_0}\\
&\quad -x_n^T K_n^{-1} m_n + \frac{A_0}{2} m_n^T K_n^{-1} m_n
\end{aligned}$$

Since the first and fourth terms do not depend on A_0,

$$\hat{A}_0 = \arg\min_{A_0}\left[\ln(A_0) + \frac{x_n^T K_n^{-1} x_n}{2A_0} + \frac{A_0}{2} m_n^T K_n^{-1} m_n\right]$$

For the values of interest, the expression in square brackets is a convex function in A_0. The maximum likelihood estimate of A_0 is the value $\hat{A}_0$, which minimizes this expression. $\hat{A}_0$ is given by,

$$\hat{A}_0 = \frac{-1 + \sqrt{1 + (m_n^T K_n^{-1} m_n)(x_n^T K_n^{-1} x_n)}}{m_n^T K_n^{-1} m_n},$$

where (7) follows since $\hat{A}_0$ is strictly positive. Thus given the observed values of $X_n = [A_n, B_n]^T$, the MAP estimate of A_0 is obtained by evaluating the expression given in (7). A graphic of the estimator is shown in Fig. 2. From Fig. 1 it is evident that the area of greatest interest lies close to the mean line, $\hat{A}_0$ is approximately linear in the distance along this line from the origin. To validate the accuracy of the estimator, we illustrate its relative error $|A_0 - \hat{A}_0|/A_0$ in Fig. 3. The relative error decreases with A_0 (as would be expected from inspection of Fig. 1).

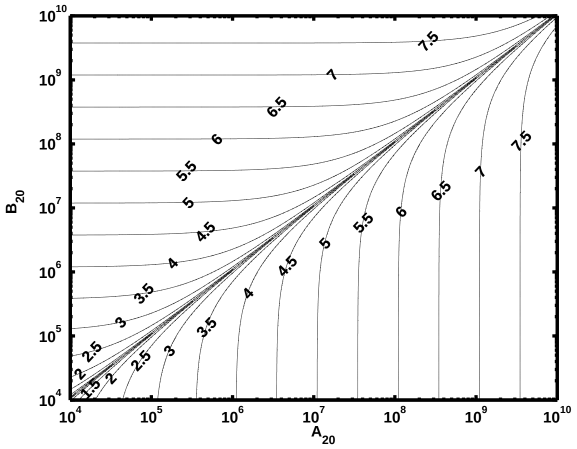

Fig. 2. Contour map of $\log_{10}(\hat{A}_0)$ against A_{20}, B_{20} with $P^i_{AB} = 0.7$ and $P^i_{BA} = 0.6$ for $i = \{1, 2, \ldots 20\}$.

To get some intuition about this estimation, assume that $(m_n^T K_n^{-1} m_n)(x_n^T K_n^{-1} x_n) \gg 1$, then,

$$\begin{aligned}
\hat{A}_0 &\approx \sqrt{\frac{x_n^T K_n^{-1} x_n}{m_n^T K_n^{-1} m_n}}\\
&= A_0\sqrt{\frac{x_n^T \Sigma_n^{-1} x_n}{\mu_n^T \Sigma_n^{-1} \mu_n}}
\end{aligned}$$

Note that $\sqrt{x_n^T \Sigma_n^{-1} x_n}$ is a measure of the length of x_n scaled appropriately to account for the covariance, similarly $\sqrt{\mu_n^T \Sigma_n^{-1} \mu_n}$ is a measure of the length of μ_n ($E[X_n|A_0]$). For an accurate approximation ($\hat{A}_0 \approx A_0$) we want

$$\frac{x_N^T \Sigma_N^{-1} x_N}{\mu_N^T \Sigma_N^{-1} \mu_N} \approx 1. \qquad (7)$$

The operation of the estimator can be thought of as finding $\hat{A}_0$ such that the ratio in (7) is as close to unity as possible.

4. LINEAR COMBINATIONS OF AMPLICONS

In many practical cases, A_n and B_n are not both observable, rather some linear combination of them can be observed. Examples of this combination include the sum of A_n and B_n (intercalator dyes [5]) or just one of the two (fluorogenic reporters[6]).

In general, consider the random variable C_n which is the linear combination of A_n, B_n given by,

$$C_n = \alpha A_n + \beta B_n,$$

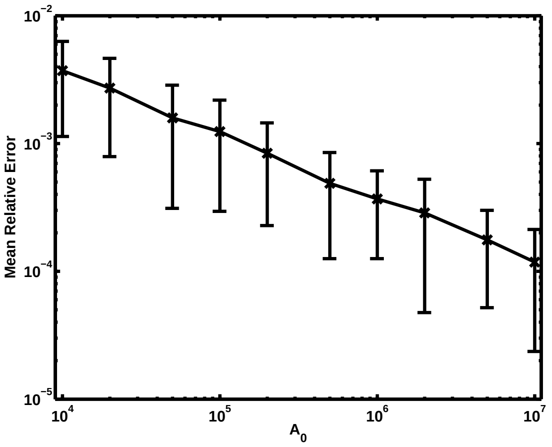

Fig. 3. Mean (with 1 standard deviation error-bars) of the absolute relative error ($|(A_0 - \hat{A}_0)/A_0|$) using our estimation method and the observed concentrations after cycle 20. Equations (1) and (2) were used with $P^i_{AB} = 0.7$ and $P^i_{BA} = 0.6$ for each i.

where $\alpha \geq 0$ and $\beta \geq 0$. Note that conditioned on A_0, A_n and B_n are assumed jointly Gaussian and so any linear combination of them (C_n) is a Gaussian random variable, with,

$$\begin{aligned}\mu_{C,n} &= \alpha\mu_{A,n} + \beta\mu_{B,n} = \gamma_n A_0 \\ \sigma^2_{C,n} &= \alpha^2\sigma^2_{A,n} + \beta^2\sigma^2_{B,n} + 2\alpha\beta\sigma_{AB,n} = \tau^2_n A_0,\end{aligned}$$

where γ_n and τ_n can be calculated directly from the amplification efficiencies. Following the same procedure as in section 3 with observation c_n as a sample of C_n, yields,

$$\hat{A}_0 = \frac{-\tau^2_n + \sqrt{\tau^4_n + 4\gamma^2_n c^2_n}}{2\gamma^2_n}.$$

Again, to gain some intuition about this estimator, assume that $\frac{c_n\gamma_n}{\tau^2_n} \gg 1$ (this approximation is supported by simulation results) then

$$\hat{A}_0 \approx \frac{c_n}{\gamma_n} = A_0\frac{c_n}{\mu_{C,n}}$$

The operation of the estimator can be thought of as finding $\hat{A}_0$ such that the ratio of c_n to $\mu_{C,n}$ is as close to unity as possible.

5. CONCLUSION

A *maximum a posteriori* (MAP) estimator for polymerase chain reaction (PCR) assays is presented. The estimator relies on the replication efficiencies of the PCR assay and the observed end-point concentration of the PCR product. This derivation is based on the stochastic progression of the amplicons. We assume that the end-point concentrations are well approximated by a joint Gaussian distribution. The estimator finds the closest mean in the sense of the ratio in equation (7) rather than finding the closest point on the 'mean line' (the dotted line in Fig. 1) in a least squares sense. The estimator, effectively tries to choose an estimate that makes the 'length' of the observed vector as close to the length of the average vector as possible.

In many practical cases, only a linear combination of the end-point concentrations can be observed. Based on this fact we also provide an extension for the estimator which can be applied to various quantitative PCR assays and detection methods. This feature makes the estimation method applicable to a variety of PCR detection techniques, such as fluorogenic probes [6], or intercalator dye [5], The methods presented in this paper can also be implemented in the design of accurate PCR assays, such as medical diagnostic applications, where estimation performance is significant.

6. REFERENCES

[1] M. J. McPherson, S. G. Mller, "PCR: The Basics from Background to Bench," BIOS Scientific Publishers, 2000.

[2] G. Stolovitzky, G. Cecchi, "Efficiency of DNA replication in PCR," National Academy of Sciences, volume 93, pp. 12947-12952, 1996.

[3] A. Hassibi, H. Kakavand, T. H. Lee, "A Stochastic Model and Simulation Algorithm for Polymerase Chain Reaction (PCR) Systems of DNA," accepted for publication, GENSIPS, 2004.

[4] M. Becker-Andr, K. Hahlbrock, "Absolute mRNA quantification using PCR," Nucleic Acids Research", volume 17, pp. 9437-9446, 1989.

[5] V. Lyamichev, J. E. Dahlberg, "Structure-specific cleavage of nucleic acids by eubacterial DNA polymerases," Science, volume 260, pp. 778-783, 1993.

[6] R. Higuchi, G. Dollinger, and R. Griffith, "Simultaneous amplification and detection of specific DNA sequences," Biotech volume 10 pp.413-417, 1992.

[7] A. Papoulis, "Probability, random variables, and stochastic processes," McGraw-Hill, 3'rd Ed., 1991.

[8] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification," John Wiley & Sons, 2'nd Ed. 2001.

[9] D. B. O'Brien, H. Kakavand,"MAP estimation for PCR, E.E. technical report, Stanford Unversity, Apr. 2004.