A General Probabilistic Model of the PCR Process

Nilanjan Saha
Department of Computer Science
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
Layne T. Watson
Departments of Computer Science and Mathematics
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
Karen Kafadar
Department of Mathematics
University of Colorado at Denver, Denver, CO 80217
Alexey Onufriev, Naren Ramakrishnan
Department of Computer Science
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061
Cecilia Vasquez-Robinet, Jonathan Watkinson
Department of Plant Pathology, Physiology, and Weed Science
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

Abstract

This paper rigorously derives a general probabilistic model for the PCR process; this model includes as a special case the Velikanov-Kapral model where all nucleotide reaction rates are the same. In this model the probability of binding of deoxy-nucleoside triphosphate (dNTP) molecules with template strands is derived from the microscopic chemical kinetics. A recursive solution for the probability distribution of binding of dNTPs is developed for a single cycle and is used to calculate expected yield for a multicycle PCR. The model is able to reproduce important features of the PCR amplification process quantitatively. With a set of favorable reaction conditions, the amplification of the target sequence is fast enough to rapidly outnumber all side products. Furthemore, the final yield of the target sequence in a multicycle PCR run always approaches an asymptotic limit that is less than one. The amplification process itself is highly sensitive to initial concentrations and the reaction rates of addition to the template strand of each type of dNTP in the solution.

1. Introduction

The polymerase chain reaction (PCR) is a powerful technique used for the amplification of specific segments of DNA or mRNA. Over the past decade the PCR has rapidly become a technique of choice for researchers in the field of bioinformatics due to its capabilities of detecting and amplifying low copy segments. However, in practice, the PCR process does not always have a consistent relation between the initial target amount and the absolute amount of the synthesized product. This is due to the PCR's high sensitivity to several variables that includes intractable variations between the containers where the reaction is carried out. Therefore, comparisons of the amount of product to that of an external control standard do not always lead to accurate quantifications. This problem, however, is addressed in quantitative competitive PCR (QC-PCR).

In the QC-PCR, a competitive mRNA or DNA, namely an allelic variant of the target template, is used as an internal standard to provide an internal control in the amplification process. Quantification is managed by determining the amounts of coamplified products from replicated proportions of the target with the dilution series of the competitor. A normalization based on coamplification of a heterologous sequence, however, does not optimally address the difference of yield due to different template efficiencies in the amplification. It is quite difficult to rigorously quantify these differences.

Another popular version of PCR is called real time PCR (RT-PCR) and is widely used as an industry standard. It is sometime used to validate gene expression data obtained from microarray experiments. Determining yield by following the real time kinetics of PCR eliminates the need for a competitor to be co-amplified with the target for the internal standard. Quantitation can be performed by the more basic method of preparing a standard curve and determining an unknown amount by comparison to the standard curve. Real time PCR quantitation eliminates post-PCR processing of PCR products (which is necessary in QC-PCR). This helps to increase throughput, reduce the chances of carryover contamination, and remove post-PCR processing as a potential source of error.

PCR is an extremely important process from a biologist's point of view, with application in clinical microbiology, food microbiology, and veterinary microbiology. It is also a very important technique in the areas of clinical oncology, minimal residual disease, and chromosomal translocations, but perhaps one of the most important uses of PCR is in determining gene expression data. In this case accuracy of quantification is extremely important as a slight variation in estimating initial concentration of mRNA (or cDNA) can lead to false conclusions. Due to the exponential nature of the PCR, estimating this initial concentration is an error prone process. One needs to be more careful because these sensitivities increase with cycle number and the process becomes nonlinear beyond a certain cycle number.

Therefore, in spite of the popularity of the PCR, theoretical considerations to reliably describe its different applications have been mostly limited to experimental inferences rather than mathematical derivations from biophysical principles; consequently, the currently used expressions lack consistency since their foundations have not been clearly established, frequently leading to empirical fitting procedures of experimental data that result in poor quantifications. It is clear that a physics based model will predict the yield of the PCR amplification in a much better fashion; however, developing any type of generalized model is not an easy task. The present study attempts to develop such a model. Before going into details of the formulation for this model some background on the PCR amplification is given.

The PCR amplification process in general is conducted in vitro. The three primary ingredients for this process are the three nucleic acid segments: a double-stranded DNA containing the sequence to be amplified and two single-stranded primers. They react in an environment containing a DNA polymerase enzyme, deoxy-nucleoside triphosphates (dNTPs), a buffer, and a magnesium salt. Through cycles of combined denaturing, annealing (a vast number of primers is added to ensure complete annealing), and DNA synthesis, the primers hybridize to opposite strands of the target sequence such that the synthesis stage proceeds across the region between the primers, thus doubling the DNA amount. Therefore, the products formed in successive cycles should result in geometric accumulation and the target amplification after n cycles can be approximated by

$$N_n = 2^n N_0,$$

where N_0 is the initial amount of DNA segment to be amplified.

The quantitative reliability of the PCR, however, is limited by the amplification process itself. Due to its geometric nature, small differences in any of the control variables will dramatically affect the reaction yield. The variables that influence the yield of the PCR process are the concentration of the DNA polymerase, dNTPs, $MgCl_2$, DNA, and primers; the denaturing, annealing and synthesis temperature; the length and the number of cycles; ramping times, and the presence of contaminating DNA and inhibitors in the sample. Even if extreme care is taken to strictly control these parameters the tube-tube variation may sometimes affect the outcome of the reaction. The physical basis of such variation is not yet known. Some researchers (Innis and Gelfand, 1990; Roux, 1995) indicated that this variation might be due to small temperature differences along the thermal cylinder block during the first few cycles. According to Wang *et al.* (1989) and Gilliland *et al.* (1990), normalization based on coamplification does not optimally address the variation of yield due to the difference in template efficiencies. In reality it is a well-observed fact that the reaction efficiency is never 100 percent and does not remain constant during the cycles. Hence, the accumulation trend is better represented as

$$N_n = \Big[\prod_{i=1}^{n}(1+\epsilon_i)\Big]N_0,$$

where ϵ_i is the cycle efficiency and is estimated empirically from the experimental data.

A different deterministic and more physics based approach was proposed by Schnell and Mendoza (1997), who used the law of mass action to derive the kinetic equations for PCR. Stochastic models for PCR have also been developed (Nedeman *et al.*, 1992; Sun, 1995; Peccoud and Jakob, 1996; Weiss and von Haeseler, 1995, 1997; Schierwater *et al.*, 1996). Finally, a combined deterministic and stochastic approach was proposed by Stolovitzky and Cecchi (1996). They used a deterministic mass action equation to compute the amplification efficiency and estimate the number of PCR cycles. Although these models lead to a better quantification for the phenomenon, they still do not provide a complete solution because the efficiency was assumed to be approximately constant during all cycles.

Velikanov and Kapral (1999) proposed a probabilistic approach to the kinetics of the PCR, which focused on the microscopic nature of the amplification process. Their results indicated that the model was able to reproduce the main qualitative features of PCR kinetics, namely sensitivity to reaction conditions and leveling-off of the yield with increasing number of cycles (the plateau effect). Though they were able to obtain a closed form solution for their model, it had a severe flaw. This model was based on the assumption that all nucleotides were identical, which is never the case in reality. The chemical kinetics of nucleotides binding to the template strand is highly dependent on the type of the nucleotide. Moreover, the initial number of each type of nucleotide at the beginning of each cycle may not be the same and may influence the dynamics of the reaction in subsequent cycles. In this paper, the master equation developed by Velikanov and Kapral (1999) is modified to accommodate the fact that the initial template strand consists of the four different types of nucleotides, namely A, C, T, and G, and that the initial numbers of these nucleotides present in the buffer solution at the beginning of each cycle are independent of each other.

2. Polymerase Chain Reaction: A Probabilistic Approach

To begin with, let the length of the template strand be L and the length of the growing strand be ℓ at a given time t, ℓ_0 being the length at $t = 0$. A reasonable assumption is that at the molecular level, the probability rate of a reaction event is proportionate to the number of ways in which the molecules of the reactants available in the system can be combined for the reaction to take place. It can be further assumed that the template strand consists of all of the four different nucleotides ($A, C, T,$ and G) in an arbitrary but given order.

For this given template strand, the probability rate of a single nucleotide to be added depends on the rate of reaction of the particular nucleotide $\hat{\ell} \in \{A, C, T, G\}$ that is complementary to the $(\ell + 1)$st nucleotide on the template strand, and the number $n_{\hat{\ell}}$ of such nucleotides present in the system. This, the probability rate, is denoted as $w(l, t)$. It is important to note here that $\hat{\ell}$ is the type of nucleotide that is complementary to the next nucleotide on the template strand when the length of the growing strand is ℓ. So, in this notation,

$$w(\ell, t) = k(\ell, t) n_{\hat{\ell}} \ , \tag{1}$$

where $k(\ell, t)$ is the reaction rate coefficient that also depends on temperature. Two more parameters are necessary. The first one is $m_{0\hat{\ell}}$, which denotes the initial number of nucleotides of type $\hat{\ell}$ in the system. The other one is $X_{\hat{\ell}}$, which indicates the ratio of the number of nucleotides of type $\hat{\ell}$ to the total number of nucleotides of all types in the growing strand when the length of the growing strand is ℓ. It is reasonable to assume that the total number of nucleotides of each type remains constant, so

$$n_{\hat{l}} = m_{0\hat{\ell}} - \ell X_{\hat{\ell}} \ . \tag{2}$$

The evolution of the probability distribution is governed by a master equation (Gardiner, 1985; Cox and Miller, 1965). The master equation for the primer extension process was further developed by Velikanov and Kapral (1999) and is given by

$$\frac{\partial}{\partial t} P(\ell, t) = w(\ell - 1, t) P(\ell - 1, t) - w(\ell, t) P(\ell, t), \tag{3}$$

where $P(\ell, t)$ is the probability of a reaction at time t when the growing strand length is ℓ. Now a strictly monotonic function $\eta(\ell, t)$ can be introduced by

$$\frac{\partial}{\partial t} \eta(\ell, t) = k(\ell, t), \qquad \eta(\ell, 0) = 0. \tag{4}$$

Since $\eta(\ell, t)$ is strictly monotone in t, a new function can be defined by $\tilde{P}(\ell, \eta) = P(\ell, t)$. (The analysis in Velikanov and Kapral is mathematically incorrect since they write $P(\ell, \eta) = P(\ell, t)$.) Using the chain rule,

$$\frac{\partial}{\partial t} P(\ell, t) = \Big[\frac{\partial}{\partial \eta} \tilde{P}(\ell, \eta)\Big]\Big[\frac{\partial}{\partial t} \eta(\ell, t)\Big], \tag{5}$$

the master equation becomes

$$\begin{aligned} \frac{\partial}{\partial \eta} \tilde{P}(\ell, \eta) &= \frac{\frac{\partial}{\partial t} P(\ell, t)}{k(\ell, t)} \\ &= \frac{1}{k(\ell, t)} \left(k(\ell - 1, t) \left(m_{0\widehat{\ell-1}} - (\ell - 1) X_{\widehat{\ell-1}} \right) \tilde{P}(\ell - 1, \eta) - k(\ell, t) \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}} \right) \tilde{P}(\ell, \eta) \right) \\ &= \frac{k(\ell - 1, t)}{k(\ell, t)} \left(m_{0\widehat{\ell-1}} - (\ell - 1) X_{\widehat{\ell-1}} \right) \tilde{P}(\ell - 1, \eta) - \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}} \right) \tilde{P}(\ell, \eta) \\ &= \tilde{k}(\ell, \eta) \left(m_{0\widehat{\ell-1}} - (\ell - 1) X_{\widehat{\ell-1}} \right) \tilde{P}(\ell - 1, \eta) - \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}} \right) \tilde{P}(\ell, \eta), \end{aligned} \tag{6}$$

where

$$\tilde{k}(\ell,\eta) = \frac{k(\ell-1,t)}{k(\ell,t)}. \tag{7}$$

Following Gardiner (1985), a generating function $G(s,\eta)$ is introduced that is related to the probability distribution $\tilde{P}(\ell,\eta)$ by

$$G(s,\eta) = \sum_{\ell=0}^{\infty} s^{\ell} \tilde{P}(\ell,\eta). \tag{8}$$

Assuming $\tilde{k}(\ell,\eta)$ is uniformly bounded in η, the derived series from (8) converges uniformly in η, so an alternate form of the master equation is

$$\frac{\partial}{\partial\eta} G(s,\eta) = \sum_{\ell=0}^{\infty} s^{\ell} \left(\tilde{k}(\ell,\eta) \left(m_{0\widehat{\ell-1}} - (\ell-1) X_{\widehat{\ell-1}}(\ell-1) \right) \tilde{P}(\ell-1,\eta) - \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}}(\ell) \right) \tilde{P}(\ell,\eta) \right). \tag{9}$$

Three assumptions are necessary to derive a tractable closed form solution:

$$\tilde{k}(\ell,\eta) \approx 1, \text{ all } m_{0\hat{\ell}} \text{ are the same, and } X_{\hat{\ell}} \text{ is constant.} \tag{10}$$

The latter is a reasonable assumption for long sequences. The effect of the former is the motivation for this paper. With assumption (10) and the convention $\tilde{P}(-1,\eta) \equiv 0$,

$$\begin{aligned}
\frac{\partial}{\partial\eta} G(s,\eta) &= \sum_{\ell=0}^{\infty} s^{\ell} \left(\left(m_{0\widehat{\ell-1}} - (\ell-1) X_{\widehat{\ell-1}} \right) \tilde{P}(\ell-1,\eta) - \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}}(\ell) \right) \tilde{P}(\ell,\eta) \right) \\
&= m_{0\hat{j}} s \sum_{j=-1}^{\infty} s^{j} \tilde{P}(j,\eta) - m_{0\hat{\ell}} \sum_{\ell=0}^{\infty} s^{\ell} \tilde{P}(\ell,\eta) \\
&\quad + \sum_{\ell=0}^{\infty} s^{\ell} \ell X_{\hat{\ell}} \tilde{P}(\ell,\eta) - \sum_{j=-1}^{\infty} s^{j+1} j X_{\hat{j}} \tilde{P}(j,\eta) \\
&= m_{0\hat{\ell}} s G(s,\eta) - m_{0\hat{\ell}} G(s,\eta) + \sum_{\ell=0}^{\infty} s^{\ell} \ell X_{\hat{\ell}} \tilde{P}(\ell,\eta) - \sum_{\ell=0}^{\infty} s^{\ell+1} \ell X_{\hat{\ell}} \tilde{P}(\ell,\eta) \\
&= m_{0\hat{\ell}} s G(s,\eta) - m_{0\hat{\ell}} G(s,\eta) + X_{\hat{\ell}} \sum_{\ell=0}^{\infty} s^{\ell} \ell \tilde{P}(\ell,\eta) - X_{\hat{\ell}} \sum_{\ell=0}^{\infty} s^{\ell+1} \ell \tilde{P}(\ell,\eta) \\
&= m_{0\hat{\ell}} (s-1) G(s,\eta) + X_{\hat{\ell}} \left(s \frac{\partial}{\partial s} G(s,\eta) - s^2 \frac{\partial}{\partial s} G(s,\eta) \right) \\
&= (s-1) \left(m_{0\hat{\ell}} G(s,\eta) - X_{\hat{\ell}} s \frac{\partial}{\partial s} G(s,\eta) \right) \\
&= (s-1) X_{\hat{\ell}} \left(\frac{m_{0\hat{\ell}}}{X_{\hat{\ell}}} G(s,\eta) - s \frac{\partial}{\partial s} G(s,\eta) \right).
\end{aligned} \tag{11}$$

Let $\dfrac{m_{0\hat{\ell}}}{X_{\hat{\ell}}} = M_{0\hat{\ell}}$ and $F(s,\eta) = s^{-M_{0\hat{\ell}}} G(s,\eta)$. Then

$$\begin{aligned}
\frac{\partial}{\partial \eta} F(s,\eta) &= s^{-M_{0\hat{\ell}}} \frac{\partial}{\partial \eta} G(s,\eta) \\
&= s^{-M_{0\hat{\ell}}}(s-1) X_{\hat{\ell}} \left(M_{0\hat{\ell}} G(s,\eta) - s \frac{\partial}{\partial s} G(s,\eta) \right) \\
&= (1-s) s X_{\hat{\ell}} \left(-M_{0\hat{\ell}} s^{-(M_{0\hat{\ell}}+1)} G(s,\eta) + s^{-M_{0\hat{\ell}}} \frac{\partial}{\partial s} G(s,\eta) \right) \\
&= s(1-s) X_{\hat{\ell}} \frac{\partial}{\partial s} F(s,\eta).
\end{aligned} \tag{12}$$

Since the series in (8) converges absolutely for $|s| < 1$, it makes sense to define the strictly monotone function $\sigma = \dfrac{\ln\left(\frac{s}{1-s}\right)}{X_{\hat{\ell}}}$ and define $\tilde{F}(\sigma,\eta) = F(s,\eta)$. Then

$$\begin{aligned}
\frac{\partial}{\partial s} F(s,\eta) &= \frac{\partial}{\partial \sigma} \tilde{F}(\sigma,\eta) \frac{d\sigma}{ds} \\
&= \frac{\frac{\partial}{\partial \sigma} \tilde{F}(\sigma,\eta)}{X_{\hat{\ell}} s(1-s)},
\end{aligned} \tag{13}$$

or

$$X_{\hat{\ell}} s(1-s) \frac{\partial}{\partial s} F(s,\eta) = \frac{\partial}{\partial \sigma} \tilde{F}(\sigma,\eta). \tag{14}$$

Then, combining equations (12) and (14),

$$\frac{\partial}{\partial \eta} \tilde{F}(\sigma,\eta) = \frac{\partial}{\partial \sigma} \tilde{F}(\sigma,\eta). \tag{15}$$

The solution $\tilde{F}$ to this hyperbolic partial differential equation is constant along the lines $\eta + \sigma = \mathrm{c}$ (characteristics). The initial condition will determine the analytical solution. The initial probability distribution is $P(\ell, 0) = \delta_{\ell,\ell_0}$, where δ_{ℓ,ℓ_0} is the Kronecker delta, so with $\tilde{G}(\sigma,\eta) = G(s,\eta)$, The initial condition is

$$\tilde{G}(\sigma, 0) = (1 + e^{-X_{\hat{\ell}}\sigma})^{-\ell_0}. \tag{16}$$

Now $G(s,\eta) = s^{M_{0\hat{\ell}}} F(s,\eta)$ gives

$$\tilde{G}(\sigma,\eta) = (1 + e^{-X_{\hat{\ell}}\sigma})^{-M_{0\hat{\ell}}} \tilde{F}(\sigma,\eta). \tag{17}$$

Equation (16) and (17), together with the fact that $\tilde{F}(\sigma,\eta)$ is constant along the characteristics, yield

$$\tilde{G}(\sigma,\eta) = (1 + e^{-X_{\hat{\ell}}\sigma})^{-M_{0\hat{\ell}}} \left(1 + e^{-X_{\hat{\ell}}(\eta+\sigma)}\right)^{M_{0\hat{\ell}} - \ell_0}. \tag{18}$$

Now substituting for σ, equation (18) becomes

$$
\begin{aligned}
G(s,\eta) &= s^{M_{0\hat{\ell}}}\left(1-\frac{(1-s)}{s}e^{-X_{\hat{\ell}}\eta}\right)^{M_{0\hat{\ell}}-\ell_0} \\
&= s^{M_{0\hat{\ell}}}\left(1-e^{-X_{\hat{\ell}}\eta}+\frac{e^{-X_{\hat{\ell}}\eta}}{s}\right)^{M_{0\hat{\ell}}-\ell_0}.
\end{aligned}
\tag{19}
$$

The probability distribution $\tilde{P}(\ell,\eta)$ can be obtained by expanding this equation in powers of s and reading off the coefficients of the series (Velikanov and Kapral, 1999):

$$
\tilde{P}(\ell,\eta) = \binom{M_{0\hat{\ell}}-\ell_0}{M_{0\hat{\ell}}-\ell}(1-e^{-X_{\hat{\ell}}\eta})^{\ell-\ell_0}e^{-X_{\hat{\ell}}(M_{0\hat{\ell}}-\ell)\eta}. \tag{20}
$$

Equation (20) indirectly gives the solution for $P(\ell,t)$ in (3), after integrating with respect to t, under three simplifying assumptions in (10). As is often the case with analytic solutions, a more general recursive computational solution is possible (assumptions in (10) are not necessary and the solution is more computationally efficient than the closed form solution). From (3) and (6) the master equation is

$$
\frac{\partial}{\partial\eta}\tilde{P}(\ell,\eta) = \tilde{k}(\ell,\eta)n_{\widehat{l-1}}\tilde{P}(\ell-1,\eta) - n_{\hat{\ell}}\tilde{P}(\ell,\eta), \tag{21}
$$

or

$$
\frac{\partial}{\partial\eta}\tilde{P}(\ell,\eta) + n_{\hat{\ell}}\tilde{P}(\ell,\eta) = \tilde{k}(\ell,\eta)n_{\widehat{l-1}}\tilde{P}(\ell-1,\eta). \tag{22}
$$

For fixed ℓ, assume the solution $\tilde{P}(\ell-1,\eta)$ is known and let

$$
y(\eta) = \tilde{P}(\ell,\eta). \tag{23}
$$

Then, equation (22) becomes

$$
y'(\eta) + ay(\eta) = g(\eta), \tag{24}
$$

where $g(\eta) = \tilde{k}(\ell,\eta)n_{\widehat{l-1}}\tilde{P}(\ell-1,\eta)$, and $a = n_{\hat{\ell}}$ is constant with respect to η.

Using an integrating factor, the solution to equation (24) is

$$
y(\eta) = \frac{\int_0^\eta e^{au}g(u)du + B}{e^{a\eta}}. \tag{25}
$$

$$
\tilde{P}(\ell,\eta) = y(\eta) = \frac{\int_0^\eta e^{n_{\hat{\ell}}u}\left(\tilde{k}(\ell,u)n_{\widehat{l-1}}\tilde{P}(\ell-1,u)\right)du + B_\ell}{e^{n_{\hat{\ell}}\eta}}. \tag{26}
$$

The constant B_ℓ can be determined from the initial conditions, $\ell = \ell_0$ and $\eta = 0$. It can be safely assumed that $\tilde{P}(\ell_0-1,\eta) = 0$, and argued that the probability of the length of the growing strand that is more than ℓ_0 is zero at $\eta = 0$, i.e.,

$$
\tilde{P}(\ell,0) = \delta_{\ell\ell_0}B_{\ell_0}; \qquad \ell = \ell_0, \ell_0+1, \ell_0+2, \ldots L.
$$

$P(\ell_0,0) = \tilde{P}(\ell_0,0) = B_{\ell_0}$ is the initial probability of the primer growth, which must be estimated empirically. Since the primer length is ℓ_0 and there should be no growing strand of length $\ell > \ell_0$ at time $t = \eta = 0$,

$$
\tilde{P}(\ell,0) = B_\ell = 0 \qquad \text{for } \ell > \ell_0.
$$

It is important to note here that the initial condition B_{ℓ_0} appears just as a multiplication factor in the final solution for the probability distribution. This number can be chosen empirically based on the reaction parameters.

3. Results and discussions

Equation (26) was solved numerically and the probability distributions were plotted in Figures 1–4. These figures represent the probability distribution $\tilde{P}(\ell, \eta)$ with respect to ℓ and η for different template strands of equal length. The strands in Figure 1 and Figure 2 consist of type T and type C nucleotides with a single transition in the base pair sequence at the 11th base pair. For Figure 1, it is assumed that the initial number (i.e., concentration) of type A (complement of type T) nucleotides is one and a half times that of type G (complement of type C) nucleotides. It is also assumed that the reaction rate of type A nucleotide is the same as that of type G nucleotide ($K_A/K_G = 1$), where K_A and K_G are reaction rates of addition of a single type A nucleotide and type G nucleotide, respectively. Consider a growing strand at length ℓ. Now let the ℓth nucleotide on the template strand be T and the $(\ell + 1)$st nucleotide be C. Then $\tilde{k}(\ell, \eta) = K_A/K_G = 1$ in equation (7). The goal was to verify that the model could capture the variations in yield of PCR due to differences in concentration, reaction rate, and type of nucleotides by varying one parameter at a time. In Figure 2 the initial concentrations for different type nucleotides were assumed to be the same but the reaction rate ratio was changed to $K_A/K_G = 1.5$. In both cases there is a drop in probability distribution when the template sequence changes from T to C. In Figure 1 this drop can be attributed to lower initial concentration of type G nucleotide. In Figure 2 this drop is due to a lower reaction rate. Careful observation reveals the drop in Figure 1 is more prominent. This suggests that when the concentrations and reaction rates are varied separately the effect on the efficiency of the process due to initial concentration of nucleotides is dominant over the rate of reaction. However, this is an unrealistic situation as the reaction rate is not the same for different types of nucleotides.

Figure 3 and Figure 4 show the effect on the probability distribution due to transitions in the nucleotide sequence when both initial concentration and reaction rate ratio are varied simultaneously (a more realistic situation). The template strand in Figure 3 consists of only type T nucleotides and the template strand in Figure 4 is a combination of type T and type C nucleotides. The initial concentration and reaction rate ratio were varied opposing each other in terms of their effect on the probability distribution. In accordance with the nature of the extension process for fixed η, the behavior of the distribution is similar to a wave passing through the interval between ℓ_0 and L. As the length of the growing strand increases, the concentration of that particular type of nucleotide decreases in the solution, and as a result, the probability of the addition of subsequent nucleotides of the same type starts to fall (Figure 3). So it can be concluded that this model correctly predicts that the PCR process is highly dependent on the initial concentration of nucleotides. In reality it is expected that the rates of reaction for nucleotide addition are different for different types of nucleotides. In the present model the reaction rate also plays a significant role toward the yield of the process only when there is a change in the sequence of the type of nucleotide. Figure 4 clearly captures this jump in the probability distribution in the case where the nucleotide to be added to the template strand is different from the previous nucleotide. Interestingly the rate of reaction appears to have a more dominant effect on this change than initial concentration when both of them are varied together and in opposition. Both rises and falls in the probability distribution, in the direction dictated by the reaction rate ratio, can be observed when the template sequence switches back and forth between T and C, as shown in Figure 4.

Figure 5 shows the probability distribution for different cycle durations (η). The template strand used in this case was TTTTTTTTTTCCCCCCCCCC to emphasize the importance of change in reaction rate constant. For this plot it is assumed that $K_A/K_G = 10$, which is a reasonable assumption. It is quite evident from this plot that whenever there is a change in type of nucleotide in the template strand there is a very significant jump in the probability distribution for the addition of the next nucleotide. This jump in the probability distribution is present for different cycle durations (corresponding to different values of η in this case). However, this jump is not independent of the duration of the cycle.

This concept is then extended to a multicycle PCR (Velikanov and Kapral, 1999) and the probability distribution is converted to yield (expectation). The yield is defined as the ratio of estimated number of DNA strands generated at each cycle using this model over the number of DNA strands (2^n) at the end of all cycles in the ideal case. Four independent values for $\tilde{k}(\ell, \eta)$ (Equation 7) were needed to estimate a realistic yield. Goodman (1995) estimated these values experimentally. Figure 6 compares the yield predicted by the Velikanov-Kapral model (stars) with that of the model from equation (26) here (diamonds). Figure 6 is the plot of target strand yield ϕ against cycle number. One can see that in both cases the yield approaches an asymptotic value that is less than one as the cycle number increases. This is the manifestation of the amplification plateau effect, which is well known in PCR experiments, and the model was able to capture this effect successfully.

It can be easily explained why the asymptotic yield is less than one. Only a fraction of the total number of strands synthesized in each cycle is utilized as template strands in subsequent cycles. Furthermore, with the concentration of each type of dNTP decreasing from cycle to cycle and the cycle duration remaining constant, there is a decrease in the fraction of synthesized strands that can serve as templates. This results in a decrease in template efficiency as the run progresses through the cycles.

The Velikanov-Kapral model predicts 7 percent more final yield at the end of the 12th cycle. In order to compute the yield by the Velikanov-Kapral model it was assumed that all nucleotides were the same. Clearly there is a significance difference in yield between these two models, showing that using a different reaction rate of addition for each type of nucleotide matters. Figure 7 highlights another interesting aspect, for the same sequence TCCCCCCCCCCCCCCCCCCC and parameters as in Figure 6. At every cycle, initial concentrations were estimated for the model (26) in order to match the yield predicted by the Velikanov-Kapral model at that cycle. The percentage change (z) in this initial concentration is plotted in Figure 7. It is quite evident that there is a significant difference in the computed initial concentration.

In order to probe sensitivity to the reaction conditions quantitatively, a set of numerical experiments was conducted with the duration (η) of the extension phase being the same for each cycle in each run. The sequence used for this experiment was TCCCCCCCCCCCCCCCCCCC and the reaction rate ratio (K_G/K_A) was varied from 0.01 to 100. The percentage change in initial concentration (z) required in order to match the yield predicted by the Velikanov-Kapral model is plotted against reaction rate ratio (Figure 8, dotted line). This percentage change required is zero at $K_G/K_A = 1$, which essentially means that the template sequence consists of the same type of nucleotide. However, as the reaction rate ratio increases this value z changes rapidly with small changes in the reaction rate ratio. The skewness of the plot can be explained by the asymmetric composition of the template strand in terms of type of nucleotides. The most important feature of this plot is perhaps the very high relative difference in initial concentration (on the order of 80 percent) for a relatively small change in reaction rate ratio. One needs to keep in mind that this variation is highly dependent on the composition and length of the template strand and the concentration of each type of free nucleotide available in the solution. However, the present study suggests that the computed difference in initial concentration between two template strands is higher when the difference in sequence is at the beginning of the sequence.

Figure 8 (solid line) also shows the largest percentage difference in initial concentration for all η in this experiment ($0.012 \leq \eta \leq 0.03$) between the Velikanov-Kapral model and that for model (26) with respect to variation in reaction rate $\log(K_G/K_A)$, all other parameters being the same. This plot suggests that the Velikanov-Kapral model may not be adequate to capture the variation in yield (which can be quite significant in some cases) due to the change in composition of the template strand.

In order to investigate the effect of the length of the sequence to be replicated on the initial concentration an arbitary sequence $(\mathrm{TCGGCGCGGGTTTAATAATT})^x$ was chosen, where x is the number of repetitions of the base sequence of length 20. Based on the data in Goodman (1995), the six reaction ratios were chosen as $K_A/K_G = 100$, $K_T/K_G = 100$, $K_C/K_G = 1$, $K_A/K_T = 1$, $K_C/K_T = 1.5$, and $K_A/K_C = 0.667$. The percentage difference in initial concentration (z) is plotted (Figure 9) against L, where $L = 20x$ for $x = 1, 2, 3, \ldots, 15$. It was found that for this sequence $z = -78.91 + 33.34L^{0.364}$. From this result it is clear that for a longer (i.e., more realistic) sequence the difference in reaction rates is even more significant.

These results indicate that the model (26) is able to reproduce important features of the PCR amplification process quantitatively. The model implies that with a set of favorable reaction conditions, the amplification of the target sequence is fast enough to rapidly outnumber all side products. Furthemore, the final yield of the target sequence in a multicycle PCR run always tends toward an asymptotic limit that is less than one. The amplification process itself is highly sensitive to initial concentrations and the reaction rates of addition to the template strand of each type of dNTP in the solution.

In practice the initial concentration of mRNA is estimated from the yield after a certain number of cycles. This value determines whether the corresponding gene is up expressed or down expressed, and is sensitive to the measured yield. Since the model (26) is more realistic it can be expected that the yield estimated by this model would be quantitatively more accurate than the yield predicted by Velikanov-Kapral model. This hypothesis will be rigorously tested by calibrating the model (26) with PCR data, and then comparing quantitative prediction with experimental data; such a thorough calibration and validation will be topic of a future paper.

It should be noted here that the magnitude of the probability distribution is highly dependent on the value chosen for the initial condition (B_{ℓ_0}). This parameter gives the much needed flexibility in the proposed model to accommodate effects due to the variation in type and concentration of the DNA polymerase; the concentration of dNTPs, $MgCl_2$, DNA, and primers; the denaturing, annealing, and synthesis temperature; the length and the number of cycles; ramping times; and the presence of contaminating DNA and inhibitors in the sample. Some of these parameters, depending upon their degree of importance, will be included in a future model.

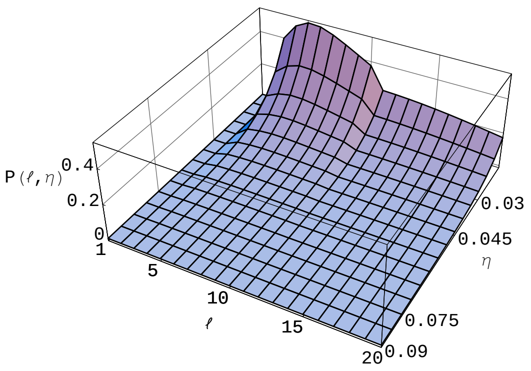

Figure 1: $P(\ell, \eta)$ vs. ℓ and η
($m_{0A}/m_{0G} = 1.5$, $K_A/K_G = 1$)
(Sequence: TTTTTTTTTTCCCCCCCCCC)

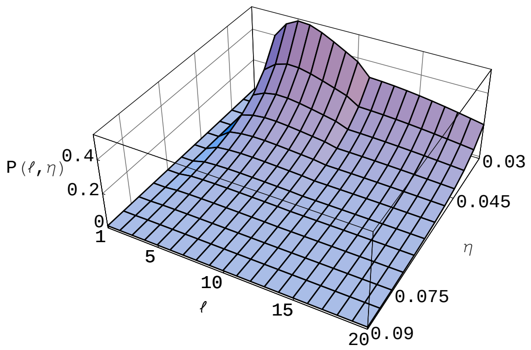

Figure 2: $P(\ell, \eta)$ vs. ℓ and η
($m_{0A}/m_{0G} = 1$, $K_A/K_G = 1.5$)
(Sequence: TTTTTTTTTTCCCCCCCCCC))

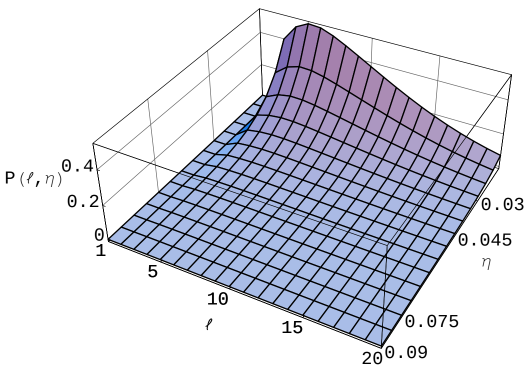

Figure 3: $P(\ell, \eta)$ vs. ℓ and η
($m_{0G}/m_{0A} = 1.5$, $K_A/K_G = 1.5$)
(Sequence: TTTTTTTTTTTTTTTTTTTTTTTT)

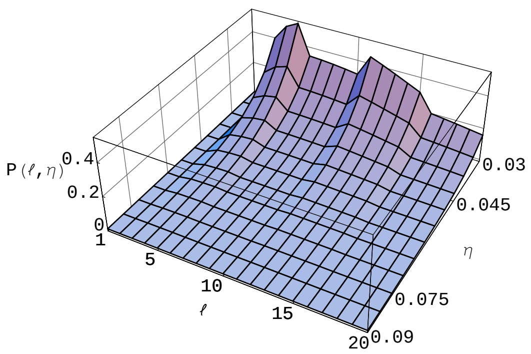

Figure 4: $P(\ell, \eta)$ vs. ℓ and η
($m_{0G}/m_{0A} = 1.5$, $K_A/K_G = 1.5$)
(Sequence: TTTTTCCCCCTTTTTCCCCC)

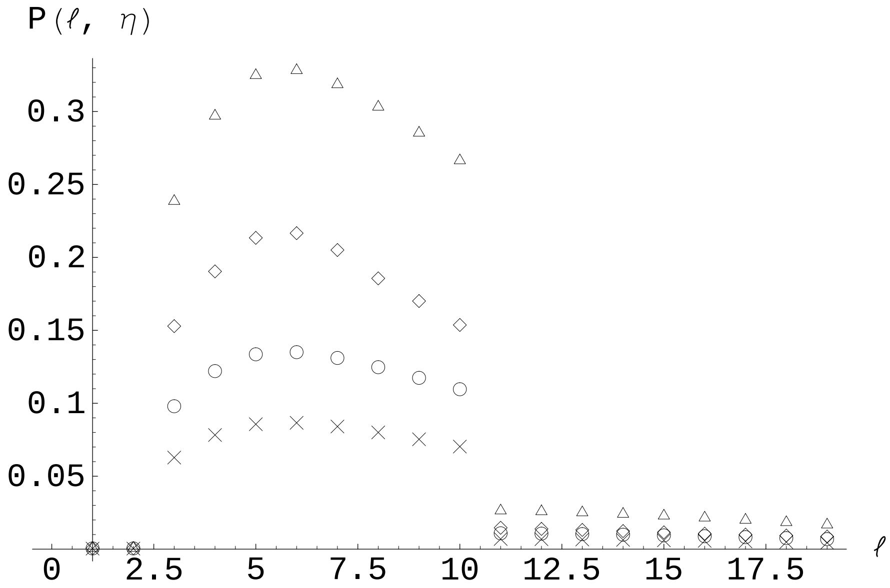

Figure 5: $P(\ell, \eta)$ vs. ℓ for different values of η (triangles for $\eta = 0.03$, diamonds for $\eta = 0.024$, circles for $\eta = 0.018$, crosses for $\eta = 0.012$) for the sequence TTTTTTTTTTC-CCCCCCCCC, $m_{0G}/m_{0A} = 1.5$, $K_A/K_G = 10$.

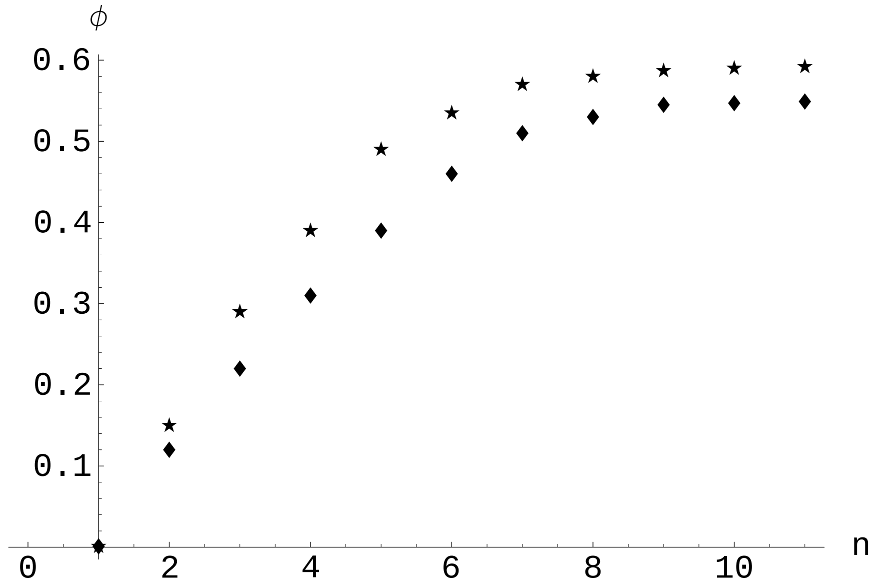

Figure 6: Comparison of yield ϕ estimated using the Velikanov-Kapral model (stars) and the present more general model (diamonds) for the sequence TCCCCCCCCCCCCCCC-CCCCC, all initial concentrations assumed equal, $K_A/K_G = 1.5$.

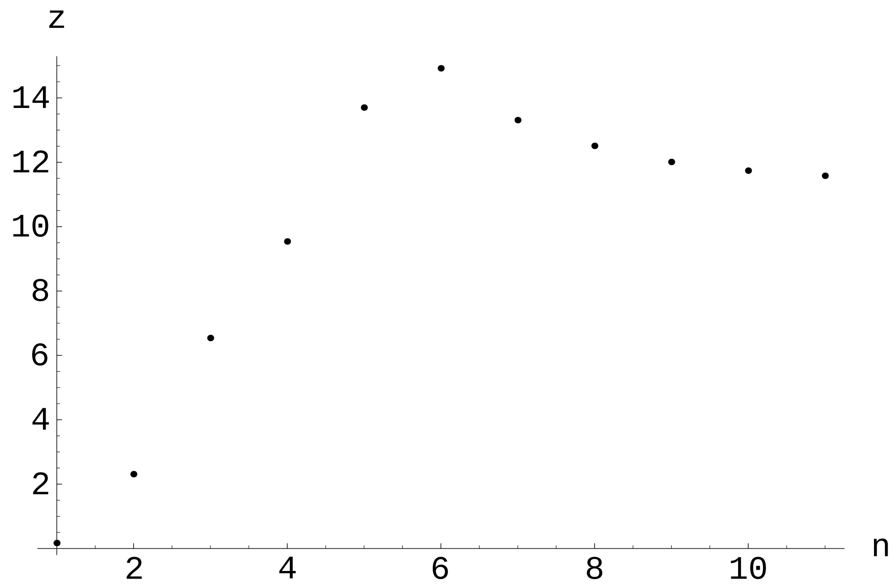

Figure 7: Required percentage change in initial concentration z for model (26) to match the yield estimated using the Velikanov-Kapral model at the end of each cycle for the sequence TCCCCCCCCCCCCCCCCCCC, all initial concentrations assumed equal, $K_A/K_G = 1.5$.

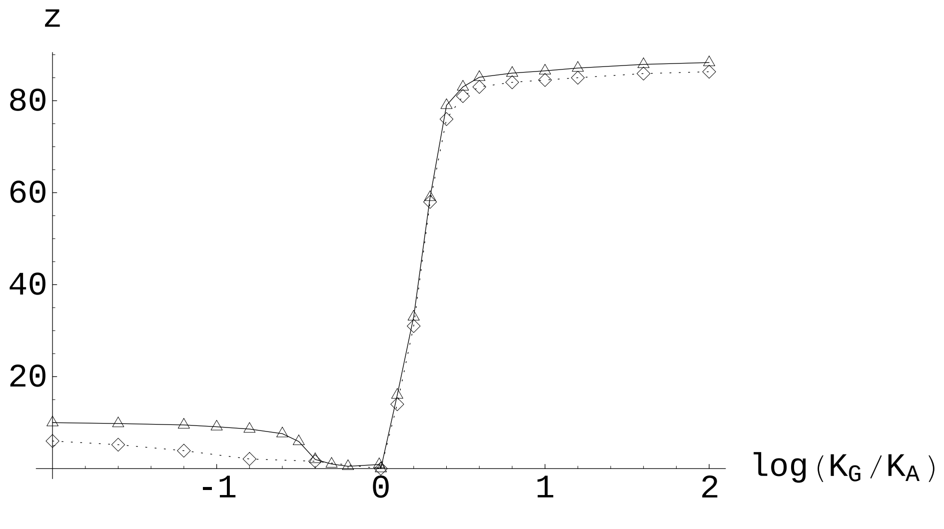

Figure 8: Required percentage change in initial concentration z for model (26) to match the yield estimated using the Velikanov-Kapral model with respect to variation in reaction rate $\log(K_G/K_A)$, for fixed duration $\eta = 0.03$, 12 cycles, and sequence TCCCCCCCCCCCCCCCCCCC (dotted line); maximum required percentage change in initial concentration z for model (26) to match the yield estimated using the Velikanov-Kapral model with the maximum taken over $0.012 \leq \eta \leq 0.03$, all other parameters being the same (solid line).

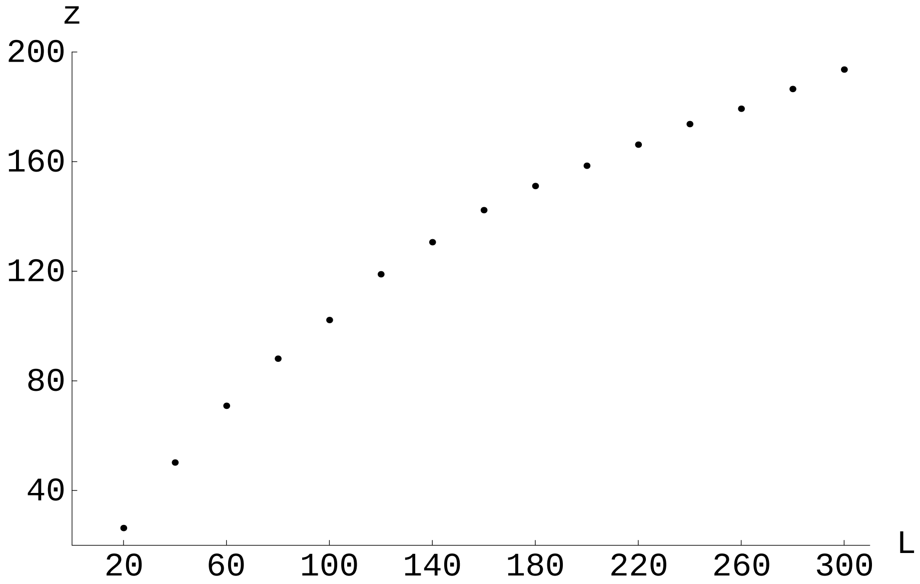

Figure 9: Required percentage change in initial concentration z for model (26) to match the yield estimated using the Velikanov-Kapral model with respect to length of the sequence L for fixed duration $\eta = 0.03$, 12 cycles. The sequence is (TCGGCGCGCCTTT-AATAATT)^x, $L = 20x$, $x = 1, 2, 3, \ldots, 15$, $K_A/K_G = 100$, $K_T/K_G = 100$, $K_C/K_G = 1$, $K_A/K_T = 1$, $K_C/K_T = 1.5$, $K_A/K_C = 0.667$.

Acknowledgement

This work was supported in part by NSF Grants EIA-0103660, IBN-0219322, and AFRL Grant F30602-01-2-0572. The authors also gratefully acknowledge the generous assistance of Dr. Ruth Grene and Dr. Gregory Gonye.

REFERENCES

Chelly, J., Kaplan, J. C., Maire, P., Gautron, S. and Kahn, A. (1988), "Transcription of the dystrophin gene in human muscle and non-muscle tissues", *Nature* 222, 747-759.

Cox, D.R. and Miller, H. D. (1965), *The Theory of Stochastic Processes*, London: Methuen.

Dimitrov, D.S. and Apostolova, M. A. (1996), "The limit of PCR amplification", *J. theor. Biol.* 178, 425-426..

Fersht, A. R. (1985), *Enzyme Structure and Mechanism*, New York: Freeman.

Gardiner, C. W. (1985), *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Berlin: Springer.

Gilliand, G., Perrin, S., Blanchard, K. and Bunn, H. F. (1990), "Analysis of cytokine mRNA and DNA: Detection and quantitation by competitive polymerase chain reaction. Proc.", *Natl. Acad. Sci. U.S.A.* 76, 1614-1618.

Goodman, M. F. (1995), "PCR Strategies. DNA Polymerase Fidelity: Misinsertions and Mismatched Extensions", *(Innis, M. A., Gelfand, D. H., Sninsky,J. J, eds)*, pp. 17-31, San Diego: Academic Press.

Hayward, A. L., Oefner, P. J., Sabatini, S., Kainer, D. B., Hinojos, C.A.and Doris, P. A. (1998), "Modeling and analysis of competitive RT-PCR", *Nucleic Acids Res. 26,* 2511-2518.

Innis, M.A. and Gelfand, D. H. (1990), "Optimization of PCRs. PCR Protocols: A Guide to Methods and Applications", *(Innis, M. A. and Gelfand, D. H., Sninsky, J. J. and White, T. J., eds)*, pp. 21-27, San Diego: Academic Press.

Kornberg, A. (1961), *Enzymatic Synthesis of DNA.*, New York: Wiley.

Lu, H. P., Xun, L. and Xie, X. S. (1998), "Single-molecule enzymatic dynamics", *Science* 282, 1877-1882.

Mcculloch, R. K., Choong, C. S. and Hurley, D. M. (1995), "An evaluation of competitor type and size for use in the determination of mRNA by competitive PCR", *PCR Meth. Applic.* 3, 108-115.

Mullis, K.B. and Faloona, F. A. (1987), "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction", *Meth. Enzymol* 155, 335-350.

Mullis, K. B., Ferred, F. and Gibbs, R. A. (1994), *The Polymerase Chain Reaction*, Cambridge, MA: BirkhaKuser.

Nedelman, J., Heagerty, P. and Lawrence, C. (1992), "Quantitative PCR: Procedures and Precision", *Bull. Math. Biol.* 54, 477-502.

Peccoud, J. and Jacob, C. (1996), "Theoretical uncertainty of measurements using quantitative polymerase chain reaction", *Biophys. J.* 71, 101-108.

Petruska, J., Goodman, M. F., Boosalis, M. S., Sowers, L. C., Cheong, C. and Tinoco, I. (1994), "Comparison between DNA melting thermodynamics and DNA polymerase fidelity", *Proc. Natl. Acad. Sci. U.S.A.* 74, 5141-5145.

Raeymaekers, L. (1993), "Quantitative PCR-theoretical considerations with practical implications", *Anal. Biochem.* 214, 582-585.

Rashtchian, A. (1984), "Amplification of RNA", *PCR Meth. Applic.* 3, S72-S80.

Roux, K. H. (1995), "Optimization and troubleshooting in PCR", *PCR Meth. Applic.* 3, S074-S083.

Roux, K. H. (1995), "Optimization and troubleshooting in PCR.", *PCR Meth. Applic.* 4, S185-S194.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A.and Arnheim, N. (0874), "Enzymatic amplification of b-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia", *Science* 129, 1350-1354.

Saiki, R. K., Gelfand, D. H. and Stoffel, S. (1988), "Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase", *Science* 239, 487-491.

Ssantagati, S., Bettini, E., Aadente, M., Muramatsu, M. and Maggi, A. (1993), "Theoretical considerations for the application of competitive polymerase chain reaction to the quantitation of a low abundance mRNA: Estrogen receptor", *Biochem. Pharmacol.* 35, 1797-1803.

Schierwater, B., Metzler, D., Kruger, K. and Streit, B. (1996), "The effects of nested primer binding sites on the reproducibility of PCR: mathematical modeling and computer simulation studies", *J. Comp. Biol.* 3, 235-251.

Schnell, S. and Mendoza, C. (1997a), "Enzymological considerations for a theoretical description of the quantitative competitive polymerase chain reaction (QC-PCR)", *J. Theor. Biol.* 184, 433-440.

Schnell, S. and Mendoza, C. (1997b), "Theoretical description of the polymerase chain reaction", *J. Theor. Biol.* 188, 313-318.

Schneeberger, C., Speiser, P., Kury, F. and Zeillinger, R. (1995), "Quantitative detection of reverse transcriptase-PCR products by means of a novel and sensitive DNA stain", *PCR Meth. Applic.* 3, 123-127.

Segel, I. H. (1975), *Enzyme Kinetics. Behavior and Analysis of Rapid Equilibrium and Steady State Enzyme Systems*, New York: Wiley.

Stolovitzky, G. and Cecchi, G. (1996), "Efficiency of DNA replication in the polymerase chain reaction", *Proc. Nat. Acad. Sci.; U.S.A.* 93, 12 947-12 952.

Sun, F. (1995), "The polymerase chain reaction and branching processes", *J. Comp. Biol.* 2, 63-86..

Velikanov, M. V. and Kapral, R. (1999), "Polymerase Chain Reaction: A Markov Process Approach", *J. Theor. Biol.* 201, 239-249.

Wang, A. M., Doyle, M. V. and Mark, D. F. (1989), "Quantitation of mRNA by polymerase chain reaction", *Proc. Natl. Acad. Sci.* U.S.A. 75, 8606-8610.

Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J. and Block, S. M. (1988), "Force and velocity measured for single molecules for RNA polymerase", *Science* 282, 902-907.

Weiss, G. and Von Haeseler, A. (1995), "Modeling the polymerase chain reaction", *J. Comp. Biol.* 2, 49-61.

Weiss, G. and Von Haeseler, A. (1997), "A coalescent approach to the polymerase chain reaction", *Nucleic Acids Res.* 25, 3082-3087.