

Validation and Estimation of Parameters for a General Probabilistic Model of the PCR Process

Nilanjan Saha, Layne T. Watson, Karen Kafadar,

Naren Ramakrishnan, Alexey Onufriev,

Shrinivas Rao Mane, and Cecilia Vasquez-Robinet

September 28, 2006

Abstract

Earlier work by Saha et al. rigorously derived a general probabilistic model for the PCR process that 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 function 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. Furthermore, 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. This paper extends the earlier Saha model with a physics based model of the dependence of the reaction rates on temperature, and estimates parameters in this new model by nonlinear regression. The calibrated model is validated using RT-PCR data.

Key words: Levenberg-Marquardt algorithm, multicycle PCR, nonlinear regression, polymerase chain reaction (PCR), probabilistic model, yield.

1 INTRODUCTION

The polymerase chain reaction (PCR) is a powerful technique used for the amplification of specific segments of DNA or mRNA. The PCR process has become a technique of choice for bioinformatics researchers 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 whose effects on the containers (where the reaction takes place) are difficult to model. 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 assessed by determining the amounts of co-amplified products from replicated proportions of the target with the dilution series of the competitor. A normalization based on co-amplification of a heterologous sequence, however, does not optimally address the difference in yield due to different template efficiencies in the amplification. It is quite difficult to rigorously quantify these differences.

Another popular version of PCR, called real time PCR (RT-PCR), is used widely as an industry standard 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 technique for biologists, with applications in research (e.g., clinical/food/veterinary microbiology), as well as clinical medicine (e.g., oncology, disease identification, chromosomal translocations.) One of the most important uses is in measuring gene expressions. Here, accuracy of quantification is extremely important as slight variations in estimating initial mRNA (or cDNA) concentrations can lead to false conclusions. Due to PCR's exponential growth in product, the process of estimating this initial concentration involves errors of many types and is,

mathematically, ill-conditioned. Sensitivities to these errors increase with cycle number and the yield (ratio of actual product to exponentially increasing product) 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 relied mostly on experimental inferences rather than on 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 extends and validates such a model developed by Saha *et al.* (2004). Before summarizing 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 (MgCl_2). 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 concentrations of the DNA polymerase, dNTPs, magnesium salt (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-to-tube variation may sometimes affect the outcome of the reaction. The

physical basis of such variation is not yet known. Some researchers [8,19] 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 co-amplification does not optimally characterize the variation in yield due to differences 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

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

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, 1997a), who used the law of mass action to derive the kinetic equations for PCR. Stochastic models for PCR have also been developed (Mullis and Faloona, 1987; Mullis *et al.*, 1994; Saiki *et al.*, 1988; Stolovitzky and Cecchi, 1996; Wang *et al.*, 1998; Weiss and Von Haeseler, 1995). 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 an accurate solution because the efficiency is 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, the model itself involved two unrealistic assumptions. First, the model assumed that the reaction rates of all nucleotides were identical. In reality, the chemical kinetics of nucleotides binding to the template strand depends strongly on the specific nucleotide (Goodman, 1995). Second, the model assumes that the initial concentrations of the four nucleotides are the same. In fact, the 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 (Saha *et al.*, 2004). Saha *et al.* (2004) modified the master equation developed by Velikanov and Kapral (1999) 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. The next section summarizes this general model derived rigorously in Saha *et al.* (2004).

2 PROBABILISTIC MODELING OF POLYMERASE CHAIN REACTION

Let L denote the length of the template strand and ℓ denote the length of the growing strand at time t ; ℓ_0 denotes the length at $t = 0$. A reasonable assumption is that, at the molecular level, the rate of change of the probability 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 rate of change of the probability 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 probability rate of change is denoted as $w(\ell, t)$. 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,

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

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 experiment is executed with target value for this number in mind, but, in practice, $m_{0\hat{\ell}}$ is known only to within some degree of error, which can be estimated from the data; see section IIID). 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

$$n_{\hat{\ell}} = m_{0\hat{\ell}} - \ell X_{\hat{\ell}}. \quad (2)$$

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

given by

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

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

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

Since $\eta(\ell, t)$ is strictly monotone in t , a new function can be defined by $\tilde{P}(\ell, \eta) = P(\ell, t)$. Using the chain rule,

$$\frac{\partial}{\partial t}P(\ell, t) = \left[\frac{\partial}{\partial \eta} \tilde{P}(\ell, \eta) \right] \left[\frac{\partial}{\partial t} \eta(\ell, t) \right], \quad (5)$$

the master equation (3) becomes

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

where

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

After some calculus, the recursive solution to (6) is

$$\tilde{P}(\ell, \eta) = \frac{\int_0^\eta e^{n_\ell u} \left(\tilde{k}(\ell, u) n_{\ell-1} \tilde{P}(\ell - 1, u) \right) du + B_\ell}{e^{n_\ell \eta}}. \quad (8)$$

Under the three assumptions

$$\tilde{k}(\ell, \eta) \approx 1, \text{ all } m_{0\ell} \text{ are the same, and } X_{\ell} \text{ is constant,} \quad (9)$$

$\tilde{P}(\ell, \eta)$ in (8) has a closed form, rather than recursive, solution. 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.,

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

$P(\ell_0, 0) = \tilde{P}(\ell_0, 0) = B_{\ell_0}$ 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 $t = \eta = 0$,

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

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 function. This number can be chosen empirically based on the reaction parameters.

3 PHYSICS BASED REACTION RATE MODELLING

Numerical results in Saha *et al.* (2004) clearly showed that the effect of the assumptions (9) is profound; for the practical problem of estimating initial concentrations from yield curves, falsely assuming (9), when in fact these three assumptions do not hold, can result in over 200% error in predicting final yield (compared to the general model without these assumptions). There is thus no doubt that the general Saha model is a better starting point for a realistic model of the PCR process than the Velikanov-Kapral model that assumes (9). Equation (8) can be solved numerically to generate estimates of the yield of the reaction; see Saha *et al.* (2004). These results indicate that the model (8) 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. Furthermore, 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 (8) is more realistic it can be expected that the yield estimated by this model would be quantitatively more accurate than the yield predicted by the Velikanov-Kapral model (Velikanov and Kapral, 1999). This hypothesis will be rigorously tested by calibrating the model (8) with PCR data, and then comparing the model's quantitative predictions with experimental data.

It should be noted here that the magnitude of the probability $P(\ell, t)$ depends strongly on the value chosen for the

initial condition (B_{ℓ_0}). This parameter gives flexibility in the proposed model to accommodate effects due to inherent experimental variations. The goal here is to model some of these variations to improve upon the basic model (8). In the spirit of the physics based Velikanov-Kapral (Velikanov and Kapral, 1999) and Saha et al. (Saha *et al.*, 2004) models, the present study attempts to model other sources of variation in the PCR process.

The quantitative reliability of the PCR process 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 include

1. the concentration of the DNA polymerase,
2. the initial concentration of dNTPs,
3. the concentration of $MgCl_2$,
4. initial concentration of the DNA strand,
5. the concentration of primers,
6. the denaturing, annealing, and synthesis temperature,
7. the length and the number of cycles,
8. ramping times,
9. temperature,
10. the presence of contaminating DNA and inhibitors in the sample, and
11. the tube-tube variation.

The physical basis of the effects from all these variations is not yet known. Velikanov and Kapral (1999) modeled the stochastic kinetics of the PCR process, focusing on the microscopic nature of the amplification process. Saha *et al.* (2004) extended this model to accommodate the fact that the reaction rates of addition to the template strand of each type (A, C, T, G) of nucleotide are substantially different from each other and showed that these differences

matter. They also have shown that the model is highly sensitive to initial concentration of these nucleotides present in the buffer solution at the beginning of each cycle.

Parameter estimation for this basic Saha model is done first with experimental data obtained from RT-PCR experiments conducted by Grene (2004) using nonlinear regression, specifically the Levenberg-Marquardt algorithm in MINPACK (More *et al.*, 1980). Subsequently, a series of improved models are considered.

3.1 METHODOLOGY

DNase treatment, cDNA synthesis, primer design and SYBR Green I RT-PCR were carried out as described by Vandesompele *et al.* (2002). In brief, 2.25g of each total RNA sample was treated with the RNase-free rDNase I according to the manufacturer's instructions (Ambion). Treated RNA samples were purified before cDNA synthesis using Quia columns (Qiagen). First-strand cDNA was synthesized using Oligo dT (18 mer) and SuperscriptII reverse transcriptase according to the manufacturer's instructions (Invitrogen), and subsequently diluted with nuclease-free water (Sigma) to 12.5ng cDNA. RT-PCR amplification mixtures (25) contained 25ng template cDNA, 2× SYBR Green I Master Mix buffer (12.5) (Applied Biosystems) and 300nM forward and reverse primer. Primer sequences for GAPDH (forward: 5'-CGTGA TCTAA GGAGA GCAAG AG-3'; reverse: 5'-TTCCT TTGAG GTTAG GGAGC-3') and GR2 (forward: 5'-TG TTC TTGCT TTGTC GCTTC-3'; reverse: 5'-CGCCA CCTTA TCAAT CTCAC C-3') were synthesized commercially (Integrated DNA Technologies). For GR2, primers were designed near the 5' region of the gene (74 bp–135 bp) to avoid amplification of the antisense gene (which corresponds to 780–1827 bp region on GR2 cDNA). The reactions were run on an ABI 7300 real time PCR system (Applied Biosystems). The cycling conditions comprised 10 minutes of polymerase activation at 95°C, and 35 cycles at 95°C for 15 sec, 56°C for 30 sec, and 72°C for 30 sec. Each assay included: a standard curve of 7 serial dilution points of GAPDH or GR2 (≈ 600 bp) PCR fragments, a no-template control, and 25ng of each test cDNA. Each assay was performed in triplicate. All PCR efficiencies were above 95%. Sample identification was done using Sequence Detection Software (V 1.2.3, Applied Biosystems) (Grene, 2004).

3.2 MODEL FOR REACTION RATES

Assume that the reaction rates $k(\ell, t)$ depend on the temperature $T(t)$ as a function of time t , and on the length ℓ of the growing strand according to the Arrhenius equation

$$k(\ell, t) = K_0 e^{-\frac{E_0}{RT(t)}}. \quad (10)$$

The dependence of $k(\ell, t)$ on the length ℓ of the growing strand is plausible, since the energetics of a molecule are related in a coarse way to its size. Dependence on ℓ (i.e., the type of the nucleotide to be added and the current length of the growing strand) can be incorporated in K_0 , or in E_0 (activation energy). First assume E_0 , the Arrhenius energy, to be a single constant and $K_0 = K_0(\ell)$, where six different forms of $K_0(\ell)$ are considered:

$$\begin{aligned} f_1(\ell) &= \alpha_0, \\ f_2(\ell) &= \alpha_0 + \alpha_1 \ell, \\ f_3(\ell) &= \alpha_0 + \alpha_1 \ell + \alpha_2 \ell^2, \\ f_4(\ell) &= \alpha_0 e^{-\alpha_1 \ell}, \\ f_5(\ell) &= \frac{\alpha_0 + \alpha_1 \ell + \alpha_2 \ell^2}{1 + \beta \ell}, \\ f_6(\ell) &= \sqrt{\frac{\alpha_0 + \alpha_1 \ell}{1 + \alpha_2 \ell}}. \end{aligned}$$

The parameters α_0 , α_1 , α_2 , β , and E_0 are estimated from the data using nonlinear regression. Precisely, $k(\ell, t)$ from (10), with a particular form for $K_0(\ell)$, is used for $k(\ell, t)$ in (4), which yields $k(\ell, \eta)$ in (7); $\tilde{P}(\ell, \eta)$ is calculated from (8), converted to the probability $P(\ell, t)$, from which a yield $\hat{y}(t)$ at time t is ultimately calculated (by integrating the probability $P(L - 1, t)$ with respect to time t over all the cycles of the PCR process, adjusting the initial concentrations at the beginning of each cycle according to the probable yield from the previous cycle). The details of the calculation of the yield $\hat{y}(t)$ from the probability function $\tilde{P}(\ell, \eta)$ are given in an appendix I. The nonlinear least squares problem is then to minimize, over whatever parameters are involved, the sum of squared errors $\sum_i (\hat{y}(t_i) - y_i)^2$, where PCR yields y_i are measured at times t_i . The temperature ($^{\circ}K$) for each cycle is modeled as a

piecewise linear function of time t (sec) as

$$T(t) = \begin{cases} 327 + 0.6t, & 0 \leq t \leq 30, \\ 345, & 30 \leq t \leq 60, \\ 345 + 0.7(t + 60), & 60 \leq t \leq 90, \\ 366, & 90 \leq t \leq 150, \\ 366 - 1.3(t - 150), & 150 \leq t \leq 180, \\ 327, & 180 \leq t \leq 210. \end{cases}$$

The parameters for the function $K_0(\ell)$, shown in Table 1, were estimated using nonlinear regression (the Levenberg-Marquardt algorithm in MINPACK (More *et al.*, 1980)) with RT-PCR data. Three sets of RT-PCR data were obtained from experiments conducted under the same set of conditions. The first set was used to estimate the parameters and the other two sets were used to validate these parameter estimates. Yields were plotted against cycle numbers for each of these six models (f_1, \dots, f_6) for $K_0(\ell)$ and are shown in Figs. 1 through 6; the experimental measurements shown in those figures (with error bars showing the accepted experimental uncertainty for such data) are the averages of the results in the second and third data sets, while the predictions are based solely on the first data set. One can observe that the model is unable to estimate the final yield with any reasonable accuracy. It should be noted here that the maximum flexibility that can be achieved by the Velikanov-Kapral model using the similar model for the reaction rate constants is given by Fig. 5. Clearly this reveals the inadequacy of that (Velikanov-Kapral) model. The final predicted yield differs from the experimental value by more than 24 percent for the best case. Clearly modeling more than the scale factor K_0 in $k(\ell, t) = K_0 e^{-E_0/(RT(t))}$ and a single Arrhenius energy E_0 is needed.

3.3 MODEL ACTIVATION ENERGY

A more realistic model for activation energy allows different values for each nucleotide, say $E_{0\hat{\ell}}$. This is a reasonable assumption as the activation energy can be expected to vary depending on the specific nucleotide. It has been shown experimentally that the reaction rate constants for nucleotide addition in a polymerase reaction can vary by order of magnitude (Boosalis *et al.*, 1987; Mendelman *et al.*, 1989). Now using four values $E_{0\hat{\ell}}$ for E_0 and the (2, 1) rational

function that fit best for $K_0(\ell)$, namely

$$f_5(\ell) = \frac{\alpha_0 + \alpha_1 \ell + \alpha_2 \ell^2}{1 + \beta \ell}, \quad (11)$$

the results are plotted in Fig. 7. It shows considerable improvement in estimating the final yield, but the prediction is still not within the estimated experimental error of 10 percent.

3.4 PERTURBATION OF INITIAL CONDITIONS

As noted earlier, the exponential growth in PCR makes the model sensitive to initial concentrations of the nucleotides in the solution. Hence, assume

$$m_{0\hat{\ell}} = \bar{m}_{0\hat{\ell}} (1 + \delta m_{0\hat{\ell}}), \quad (12)$$

where $\bar{m}_{0\hat{\ell}}$ is the nominal measured experimental initial concentration and the perturbation $\delta m_{0\hat{\ell}}$ is a parameter to be estimated with the nonlinear regression concurrent with all the other parameters. If the magnitudes $|\delta m_{0\hat{\ell}}|$ of the estimated perturbations and the predicted yields fall within experimental error, then it may be concluded that the model is accurate within experimental error. Indeed (see Fig. 8), the least squares estimated perturbations in the initial concentrations were less than three percent, well within RT-PCR experimental error. Table 2 shows the estimated parameters for the model whose predictions are shown in Fig. 8, which plots the model's predicted yields against the data used to generate the model. Note that for this model with estimated $\delta m_{0\hat{\ell}}$ it does not make sense to compare the model predictions with data from an experiment *different* from that used to estimate the model parameters, since each experimental data set would have different perturbations $\delta m_{0\hat{\ell}}$ of the initial concentrations. It does make sense, however, to compare the other model parameters, and this is done in Table 3 for the three data sets.

In summary, the most complete model consists of the probability $\tilde{P}(\ell, \eta)$ defined by (8), the Arrhenius equation (10), the model $K_0(\ell) = f_5(\ell)$ from (11), the activation energies $E_{0\hat{\ell}}$ (determined by nonlinear regression), and initial concentration perturbations $\delta m_{0\hat{\ell}}$ defined by (12). Nonlinear least squares estimation (using, e.g., the Levenberg-Marquardt algorithm) is done to estimate the model parameters $\alpha_0, \alpha_1, \alpha_2, \beta, E_{0A}, E_{0C}, E_{0G}, E_{0T}, \delta m_{0A}, \delta m_{0C}, \delta m_{0G}, \delta m_{0T}$.

To support the contention that the model is not merely a phenomenological fitting of the data, Fig. 9 shows the results of using the first half of the data to predict the second half of the yield curve, and conversely using the second

half of the data to predict the first half of the yield curve, for three 35-cycle data sets (for the same sequence) different from the data sets used for the earlier figures. Note that the second half of the data predicts the first half extremely well, but that the first half (where the yield curve is all essentially linear) does not predict well the second half (where the yield asymptotically levels off). Figs. 8 and 9 together support a significant anticipated use of the present model, namely to use an entire RT-PCR data curve to accurately estimate initial concentrations.

Another plausible usage of this model would be to estimate the parameters of the PCR process under a certain set of conditions such as the type of the DNA polymerase, nucleotides and primer concentrations, and temperature cycles from the RT-PCR experiment, and then use these estimated parameters in future PCR experiments. Since RT-PCR is an expensive process and is not always affordable, one could conduct simple PCR and use the final measured yield (one data point) after all cycles are completed, in conjunction with estimated parameters from the previous RT-PCR experiment, to estimate the initial concentration of the DNA product in the sample. Fig. 10 shows estimated yield at the end of each cycle calculated by using just one data point (to obtain the initial concentration of the DNA sequence to be amplified, as just described) against experimentally measured RT-PCR yield. Unfortunately, as demonstrated by Fig. 10, the model cannot be used with simple PCR data to reliably predict RT-PCR output.

4 CONCLUSION

This final version of the model (with model parameters $\alpha_0, \alpha_1, \alpha_2, \beta, E_{0A}, E_{0C}, E_{0G}, E_{0T}, \delta m_{0A}, \delta m_{0C}, \delta m_{0G}, \delta m_{0T}$) appears to successfully capture the dynamics of the PCR process, both qualitatively and quantitatively, and can be used to predict the yield of the process from given initial concentrations, or estimate initial concentrations from a given yield curve. As a by-product the nonlinear regression provides estimates of the Arrhenius energies for a given PCR process, and it will be an interesting future research task to design a PCR experiment to experimentally determine these values and compare with the regression estimated values.

Although the final model's estimates of yield appear to be within the limits of experimental error, note that in reality every cycle may potentially produce a significant number of incomplete strands. As the cycle number progresses these strands may as well be amplified and at the end of the process may contribute to the measured final yield.

In the present model incomplete strand contributions with length less than some threshold are ignored to reduce the computational complexity. However, a close look at Equation (8) suggests that the equation can be written as $P(\ell) = F(P(\ell - 1))$. Therefore, the entire probability distribution with respect to the length of the growing strands can be computed and contributions of each of these strands towards final yield can be estimated. However, this would be a computationally expensive process. For instance, for a 35 cycle PCR simulation for a strand that is 400 nucleotides long, the computation would need to be carried out for strands of all 401 different lengths (0 through 400) at the end of each of the 35 cycles instead of for just one length. This incomplete strand modeling issue will be investigated further in future work.

APPENDIX 1: YIELD CALCULATION

The PCR process involves multiple consecutive cycles. During the first cycle the complementary counterpart of the original template strand grows as the free dNTPs attach one by one to the template. By the end of the cycle a double stranded DNA molecule is created. The denaturation and primer annealing phases separate these two strands and at the beginning of the next cycle two different types (one is complementary to the other) of template strands exist in the solution. These two types are arbitrarily labeled as “+” and “-”. The primers that are attached to these two different templates during the next cycle are different in length, and they also bind at different locations from the 3’ end on the corresponding templates. This necessitates different conditions of extension for these two different types of primers. Therefore, two different types of probability distributions (denoted by “+” and “-”) are required in order to accurately represent the extension of these primers during each cycle.

The initial condition for the probability distributions for the second cycle can be written as

$$P_2^+(\ell^+, 0) = \omega_2^+ P_1^+(\ell^+, 0) \equiv \omega_2^+ P(\ell^+, 0)$$

and

$$P_2^-(\ell^-, 0) = \omega_2^- P_1^-(\ell^-, 0) \equiv \omega_2^- P(\ell^-, 0),$$

where ℓ^+ and ℓ^- are the lengths from the 3’ end of the original and complementary strands, ω_2^+ and ω_2^- represent the weights of the component distribution, and the subscript denotes the cycle number. The weights are the fractions,

out of the entire DNA matter accumulated, of DNA molecules that are to be amplified. A simple but reasonable assumption will be that the primer binds if the length of overlap between the primer and the template strand is higher than a threshold (ℓ_t^+ and ℓ_t^- in each case). Thus, interpreting the ω s as the probabilities of primer attachment,

$$\omega_2^+ = \sum_{j=\ell^+-\ell_t^+}^{\ell^+} P(j, \eta)$$

and

$$\omega_2^- = \sum_{j=\ell^--\ell_t^-}^{\ell^-} P(j, \eta).$$

The probability of extension $P_2^+(\ell^+, 0)$ is the probability of attachment (ω_2^+) times the conditional probability of extension given primer attachment ($P(\ell^+, 0)$) from the master equation. For each cycle $k > 1$, $P_k^+(\ell^+, 0)$ and $P_k^-(\ell^-, 0)$ are initialized similar to P_2^+ , P_2^- , and then $P_k^+(\ell^+, \eta)$, $P_k^-(\ell^-, \eta)$ are defined from (8) (with initial conditions analogous to that of second cycle) for $\eta > 0$. Following Velikanov and Kapral (1999), assume that $\ell_t^+ = \ell^+/2$, and $\ell_t^- = \ell^-/2$.

The actual volume of the original and complementary strands (denoted by $E_n^+(\ell^+, \eta)$ and $E_n^-(\ell^-, \eta)$, $n > 0$ being the cycle number) can be estimated recursively as

$$E_n^+(\ell^+, \eta) = E_{n-1}^+(\ell^+, \bar{\eta}) + E_{n-1}^-(\ell^-, \bar{\eta})P_n^-(\ell^-, \eta)$$

and

$$E_n^-(\ell^-, \eta) = E_{n-1}^-(\ell^-, \bar{\eta}) + E_{n-1}^+(\ell^+, \bar{\eta})P_n^+(\ell^+, \eta),$$

where $\bar{\eta}$ is the scaled final time for that cycle, and E_0^+ , E_0^- are the initial volumes of the original and complementary strands, respectively, to be amplified. For double stranded DNA amplification, $E_0^+ = E_0^-$. For single stranded mRNA, $E_0^- = 0$.

The final yield is defined as the fold change of the volume of the initial strands to be amplified, normalized by the initial volume of the respective original strand, and scaled with the fold change for maximum theoretical amplification (2^N for N cycles):

$$\Phi_n^+ \equiv \frac{\log\left(\frac{E_n^+(\ell^+, \bar{\eta})}{E_0^+}\right)}{\log(2^N)} = \frac{\log(E_n^+(\ell^+, \bar{\eta})) - \log(E_0^+)}{N \log(2)}.$$

APPENDIX 2: SEQUENCE AMPLIFIED

CTTTACCTCT AGCAGACGCA GCAACTCTAC ACTCGGTACG GGATAATCCA GCCGAATTTCG GAAAGATTTCG
ATTGATGGCA TCAGTTGGTT GGGGATTTCGC TATGTTTCATT ATGGGAATAG CACTCGATTA TTCGGATACA
AAAAATCATT CGAGGTGGAG CTCTTGGAAT GCTCAAAGAG ACACTTGTTG GTGGATCAAC TTCCGAAATC
ATCCATGTCC CTACCGTCAA CCAGCAGTTG AACAAAGCTTT CATGCTTCTT CTCATTCTTA TCTGCCTTTG
TGAATTCTTC TCAGCTCCTG TGCTGAAAAT ACAACAGAGA AAAATTATAC GTTATGTTTC GTAATGTGTG
TCATATTTAT GTTGGCTGCC ATGGGATTGG CCAGTA

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

- Boosalis, M. S., Petruska, J., and Goodman, M. F. 1987. DNA polymerase insertion fidelity. *J. Biol. Chem.* 262, 14689–14696.
- Chelly, J., Kaplan, J. C., Maire, P., Gautron, S. and Kahn, A. 1988. Transcription of the dystrophin gene in human muscle and non-muscle tissue. *Nature* 333, 858–860.
- 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.

- Gilliland G., Perrin S., Blanchard, K., 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.* 87(7), 2725–2729.
- Goodman, M. F. 1995. PCR strategies, in *DNA Polymerase Fidelity: Misinsertions and Mismatched Extensions*, M. A. Innis (ed.), pp. 17–31, Academic Press, San Diego.
- Greene, R. 2004. *Unpublished RT-PCR data*, PPWS, Blacksburg, VA 24061.
- 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.* 4, 219–226.
- Mendelman, L. V., Boosalis, M. S., Petruska, J., and Goodman, M. F. 1989. Nearest neighbor influences on DNA polymerase insertion fidelity. *J. Biol. Chem.* 264, 14415–14423.
- More, J. J., Garbow, B. S., and Hillstrom, K. E. 1980. *User Guide for MINPACK-1*, Argonne National Laboratory Report ANL-80-74, Argonne, Ill.
- 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. 1988. Comparison between DNA melting thermodynamics and DNA polymerase fidelity. *Proc. Natl. Acad. Sci. U.S.A.* 85(17), 6252–6256.

- Raeymaekers, L. 1993. Quantitative PCR: theoretical considerations with practical implications. *Anal. Biochem.* 214, 582–585.
- Rashtchian, A. 1994. Amplification of RNA. *PCR Meth. Applic.* 4(2), S83–S91.
- Roux, K. H. 1995. Optimization and troubleshooting in PCR. *PCR Meth. Applic.* 4, S185–S194.
- Saha, N., Watson, L. T., Kafadar, K., Onufriev, A., Ramakrishnan, N., Vasquez-Robinet, C., and Watkinson, J. 2006. A general probabilistic model of the PCR process. *Appl. Math. Comput.*, to appear.
- Saha, N., Watson, L. T., Kafadar, K., Onufriev, A., Ramakrishnan, N., Vasquez-Robinet, C., and Watkinson, J. 2004. *A general probabilistic model of the PCR process*, Technical Report TR-04-06, Dept. Computer Science, Virginia Polytechnic Institute & State Univ., Blacksburg, VA 24061.
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. and Arnheim, N. 1985. Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230, 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.
- Santagati, S., Bettini, E., Asdente, 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.* 46(10), 1797–1803.
- Schierwater, B., Metzler, D., Kruger, K., 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(2), 235–251.
- Schnell, S. and Mendoza, C. 1997. 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. 1997. 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.* 4, 234–238.
- Segel, I. H. 1993. *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) 12947–12952.
- Sun, F. 1995. The polymerase chain reaction and branching processes. *J. Comp. Biol.* 2(1), 63–86..
- Vandesompele, J., De Paepe, A., Speleman, F. 2002. Elimination of primer-dimer artifacts and genomic coamplification using a two-step SYBR Green I real-time RT-PCR. *Anal. Biochem* 303, 95–98.
- Velikanov, M. V. and Kapral, R. 1999. Polymerase chain reaction: a Markov process approach. *J. Theor. Biol.* 201(4), 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.* 86(24) 9717–9721.
- Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J., and Block, S. M. 1998. Force and velocity measured for single molecules for RNA polymerase. *Science* 2825390, 902–907.
- Weiss, G. and Von Haeseler, A. 1995. Modeling the polymerase chain reaction. *J. Comp. Biol.* 2(1), 49–61.
- Weiss, G. and Von Haeseler, A. 1997. A coalescent approach to the polymerase chain reaction. *Nucleic Acids Res.* 25, 3082–3087.

Address correspondence to:

Naren Ramakrishnan

Department of Computer Science

660 McBryde Hall, MC 0106

Virginia Polytechnic Institute and State University

Blacksburg, VA 24061

E-mail: naren@cs.vt.edu

Table 1: Values of Parameters Estimated for Each Model of $K_0 = K_0(\ell)$.

α_0	α_1	α_2	β	$E_0(\text{J/Mol})$	RMSE
$f_1(\ell) = \alpha_0$					
2.01(+02)	—	—	—	1.14397(+05)	1.14(+01)
$f_2(\ell) = \alpha_0 + \alpha_1\ell$					
6.79(+02)	-1.47(+01)	—	—	1.16352(+05)	8.33(+00)
$f_3(\ell) = \alpha_0 + \alpha_1\ell + \alpha_2\ell^2$					
7.63(+02)	-3.86(+01)	5.61(-01)	—	1.1482(+05)	6.42(+00)
$f_4(\ell) = \alpha_0 e^{-\alpha_1\ell}$					
5.89(+02)	1.21(-01)	—	—	1.10323(+05)	7.03(+00)
$f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$					
1.68(+03)	4.50(+01)	-5.72(-01)	1.29(+00)	1.11158(+05)	4.10(+00)
$f_6(\ell) = \sqrt{(\alpha_0 + \alpha_1\ell) / (1 + \alpha_2\ell)}$					
1.63(+06)	-4.02(+04)	2.47(+00)	—	1.1572(+05)	5.37(+00)

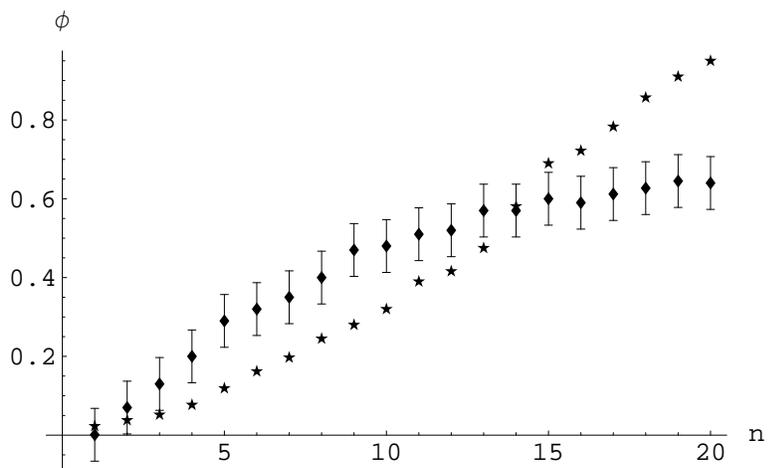


Figure 1: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_1(\ell) = \alpha_0$ (stars).

Table 2: Estimated Parameters and Initial Concentrations $\bar{m}_{0\hat{\ell}}$ for the Model Shown in Fig. 8.

α_0	1.73(+03)
α_1	4.63(+01)
α_2	5.63(-01)
β	1.32(+01)
E_{0A} (kJ/Mole)	1.042(+05)
E_{0C} (kJ/Mole)	1.052(+05)
E_{0G} (kJ/Mole)	1.057(+05)
E_{0T} (kJ/Mole)	1.046(+05)
\bar{m}_{0A} (ng/l)	2.5(+03)
\bar{m}_{0C} (ng/l)	2.5(+03)
\bar{m}_{0G} (ng/l)	2.5(+03)
\bar{m}_{0T} (ng/l)	2.5(+03)
δm_{0A} (percent)	2.72(+00)
δm_{0C} (percent)	1.87(+00)
δm_{0G} (percent)	-1.12(+00)
δm_{0T} (percent)	1.37(+00)

Table 3: Model Parameters for Three 20-cycle Data Sets for the Same Sequence.

Parameters	Data Set 1	Data Set 2	Data Set 3
α_0	1.733(+03)	2.668(+03)	1.012(+03)
α_1	4.634(+01)	5.724(+01)	3.827(+01)
α_2	5.632(-01)	6.117(-01)	4.321(-01)
β	1.322(+01)	1.101(+01)	2.054(+01)
E_{0A} (kJ/Mole)	1.042(+05)	1.026(+05)	1.054(+05)
E_{0C} (kJ/Mole)	1.052(+05)	1.037(+05)	1.069(+05)
E_{0G} (kJ/Mole)	1.057(+05)	1.039(+05)	1.065(+05)
E_{0T} (kJ/Mole)	1.046(+05)	1.032(+05)	1.043(+05)

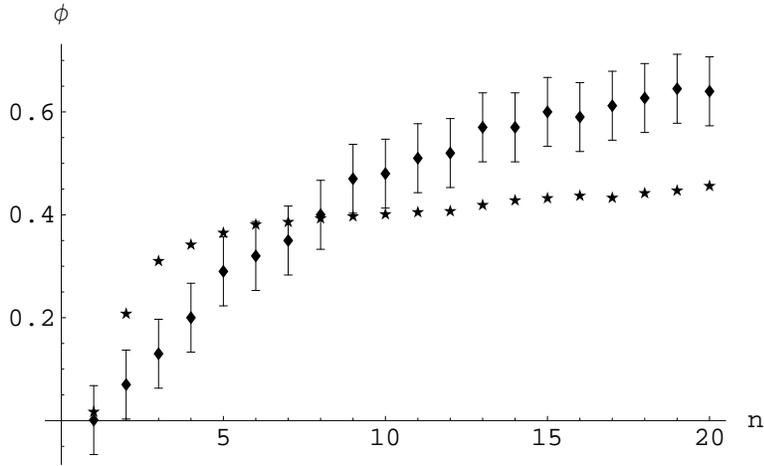


Figure 2: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with

$$K_0 = f_2(\ell) = \alpha_0 + \alpha_1 \ell \text{ (stars).}$$

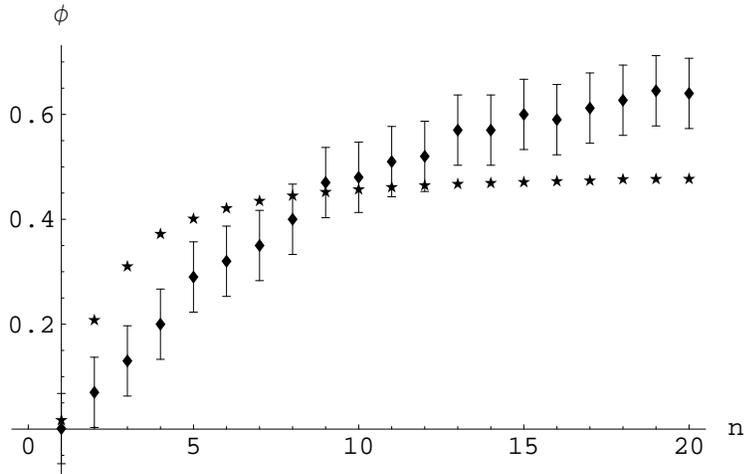


Figure 3: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_4(\ell) = \alpha_0 (1 - e^{-\alpha_1 \ell})$ (stars).

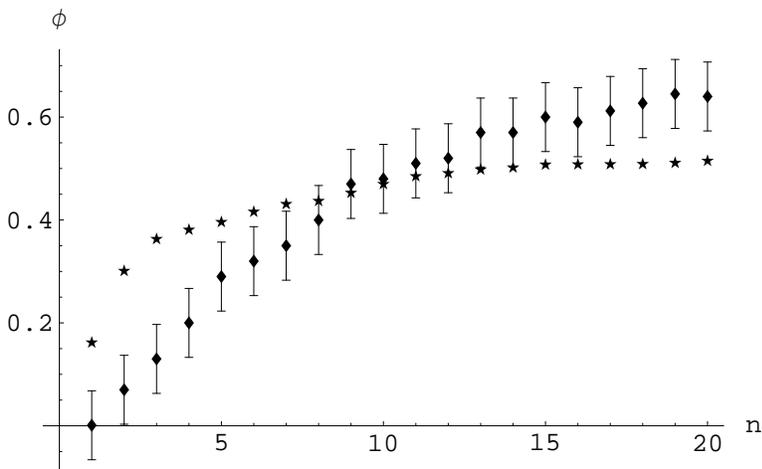


Figure 4: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_3(\ell) = \alpha_0 + \alpha_1 \ell + \alpha_2 \ell^2$ (stars).

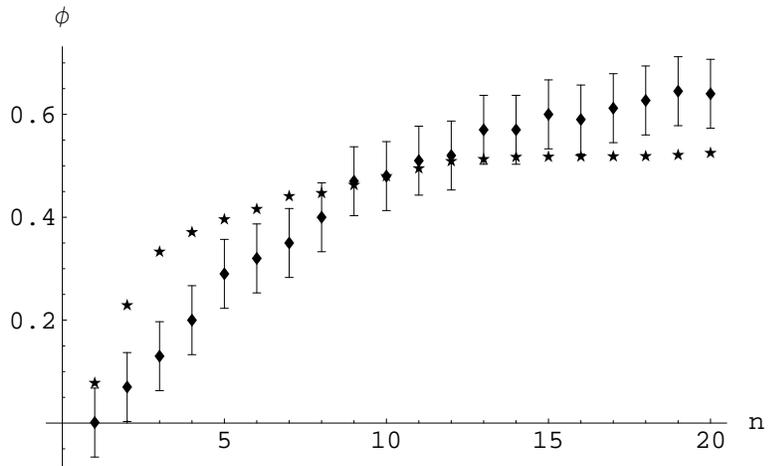


Figure 5: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$ (stars).

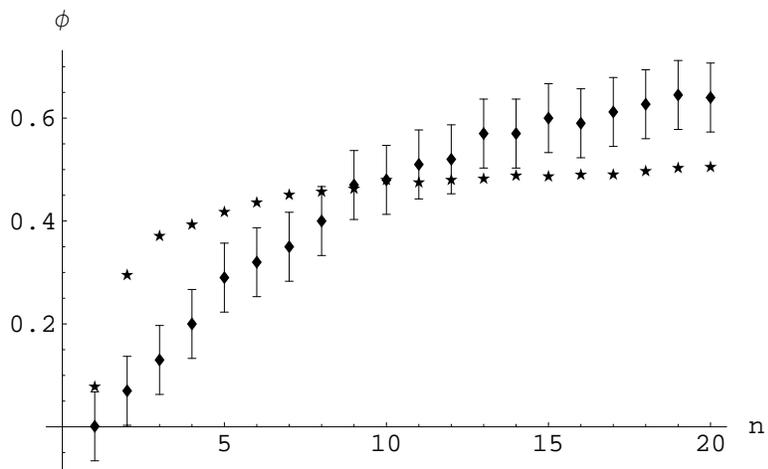


Figure 6: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_6(\ell) = \sqrt{(\alpha_0 + \alpha_1\ell) / (1 + \alpha_2\ell)}$ (stars).

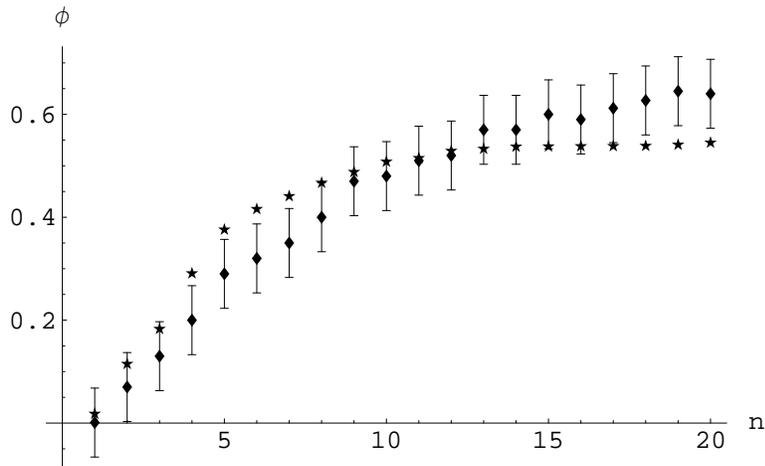


Figure 7: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$, and activation energy $E_0(\ell) = E_{\hat{\ell}}$ (stars).

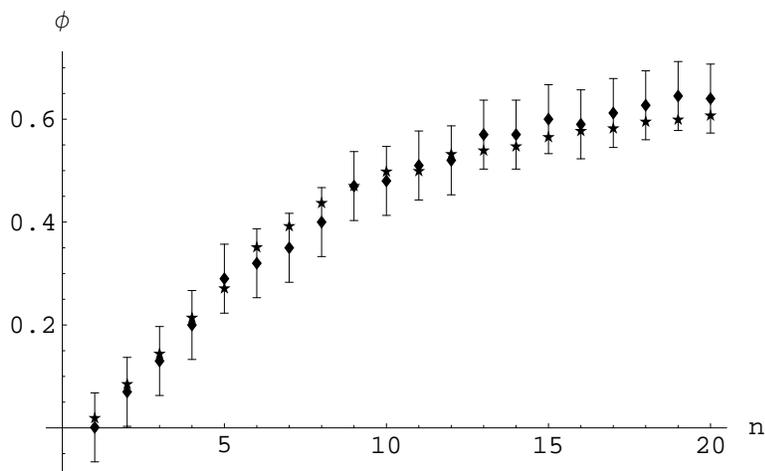


Figure 8: Comparison of yield ϕ measured experimentally by RT-PCR (diamonds) and estimated by model (10) with $K_0 = f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$, $E_0(\ell) = E_{\hat{\ell}}$, $m_{0\hat{\ell}} = \bar{m}_{0\hat{\ell}} + \delta m_{0\hat{\ell}}$ determined by regression (stars).

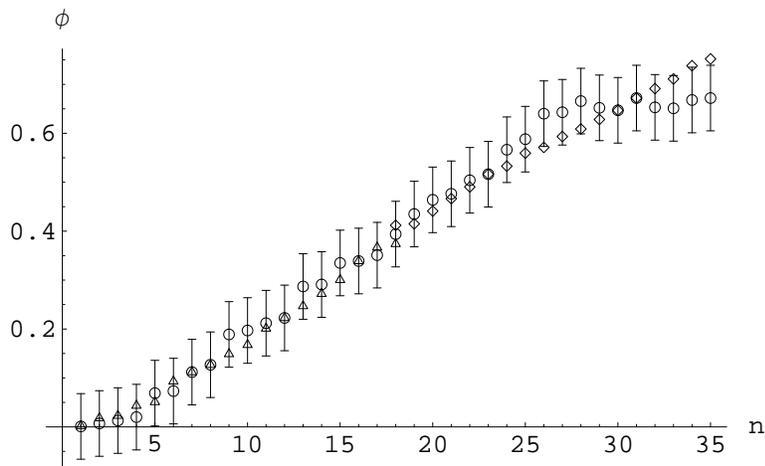


Figure 9: Comparison of yield ϕ measured experimentally by RT-PCR (circle) and estimated by model (10) with $K_0 = f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$, $E_0(\ell) = E_{\hat{\ell}}$, $m_{0\hat{\ell}} = \bar{m}_{0\hat{\ell}} + \delta m_{0\hat{\ell}}$ determined by regression for a 35 cycle PCR with first half of the yield estimated from 2nd half of the data (triangle) and vice versa (diamond) for three different data sets for the same sequence.

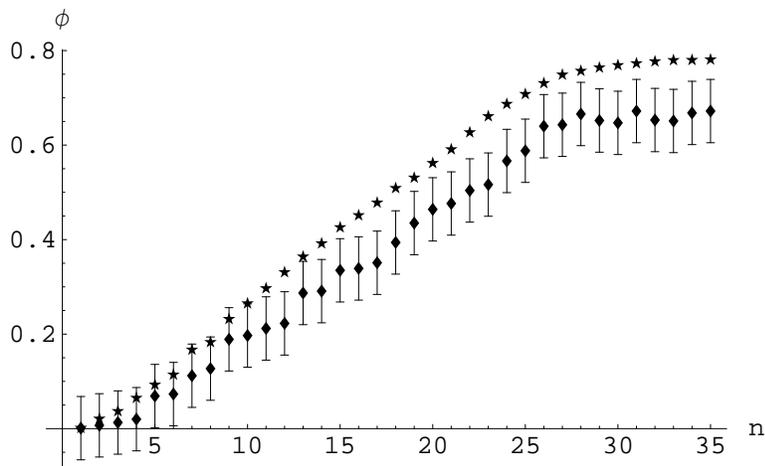


Figure 10: Comparison of yield ϕ measured experimentally by RT-PCR (circle) and estimated by model (10) with $K_0 = f_5(\ell) = (\alpha_0 + \alpha_1\ell + \alpha_2\ell^2) / (1 + \beta\ell)$, $E_0(\ell) = E_{\hat{\ell}}$, using just one (simple PCR final yield) data point to estimate the initial concentration of the DNA sequence to be amplified.