A General Probabilistic Model of the PCR Process

Nilanjan Saha[1], Layne T. Watson[2], Karen Kafadar[3], Alexey Onufriev[1], Naren Ramakrishnan[1], Cecilia Vasquez-Robinet[4], Jonathan Watkinson[4]

[1] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA

[2] Departments of Computer Science and Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA

[3] Department of Mathematics, University of Colorado at Denver, Denver, CO

[4] Department of Plant Pathology, Physiology, and Weed Science, Virginia Polytechnic Institute and State University, Blacksburg, VA

Abstract— **This paper describes 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. This model also suggests that 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.**

Keywords— PCR, modeling, probability

I. Introduction

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. 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.

To address this problem, Velikanov and Kapral [1] proposed a probabilistic approach to the kinetics of the PCR. Though they were able to capture qualitative features, their model was based on the assumption that all nucleotides were identical. In reality, the chemical kinetics of nucleotides binding to the template strand is highly dependent on the type of the nucleotide and its initial concentration. In this paper, the master equation developed by Velikanov and Kapral [1] is modified to accommodate these variations.

II. Methodology

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 $w(\ell, t)$ 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 [1].

For a 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. So, in this notation, $w(\ell, t) = k(\ell, t) n_{\hat{\ell}}$, where $k(\ell, t)$ is the reaction rate coefficient that also depends on temperature. The evolution of the probability distribution is governed by a master equation [1] and is given by

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

where $P(\ell,t)$ is the probability of a reaction at time t when the growing strand length is ℓ. It can be shown (for details see Saha *et al.* [2]) that

$$\frac{\partial}{\partial\eta}\tilde{P}(\ell,\eta) = \tilde{k}(\ell,\eta)\left(m_{0\widehat{\ell-1}} - (\ell-1)X_{\widehat{\ell-1}}\right)\cdot \tilde{P}(\ell-1,\eta) - \left(m_{0\hat{\ell}} - \ell X_{\hat{\ell}}\right)\tilde{P}(\ell,\eta), \quad (2)$$

where $\eta(\ell,t)$ is a strictly monotonic function given by $\frac{\partial}{\partial t}\eta(\ell,t) = k(\ell,t)$, $\eta(\ell,0) = 0$, $\tilde{P}(\ell,\eta) = P(\ell,t)$, $\tilde{k}(\ell,\eta) = \frac{k(\ell-1,t)}{k(\ell,t)}$, $m_{0\hat{\ell}}$ denotes the initial number of nucleotides of type $\hat{\ell}$ in the system, and $X_{\hat{\ell}}$ 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.

A general recursive solution to this equation is possible [2] and is given by

$$\tilde{P}(\ell,\eta) = \frac{\int_0^\eta e^{n_\ell u}\left(\tilde{k}(\ell,u)n_{\widehat{l-1}}\tilde{P}(\ell-1,u)\right)du + B_\ell}{e^{n_\ell\eta}}. \quad (3)$$

It can be safely argued that $B_{\ell_0} = P(\ell_0,0) = \tilde{P}(\ell_0,0)$ is the initial probability of the primer growth and $\tilde{P}(\ell,0) = B_\ell = 0$ for $\ell > \ell_0$ at time $t = \eta = 0$.

III. Results and discussions

Equation (3) was solved numerically and was plotted in Figure 1. 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.

The concept is then extended to a multicycle PCR [1] 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 $\tilde{k}(\ell,\eta)$ were needed to estimate [3] a realistic yield. Figure 2 compares the yield predicted by the Velikanov-Kapral model (stars) with that of the model from equation (3) here (diamonds). 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 well known amplification plateau effect captured 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 3 highlights another interesting aspect, for the same sequence TCCCCCCCCCCCCCCCCCCC and parameters as in Figure 2. At every cycle, initial concentrations were estimated for the model (3) 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 3. 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 TCCCCCCCCCCCCCCCCCCCC 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 4, 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 4 (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 (3) 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.

III. Conclusions

Since the model (3) is more realistic than the Velikanov-Kapral model, it can be expected that the yield estimated by this model would be quantitatively more accurate. This hypothesis will be rigorously tested by calibrating the model (3) 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.

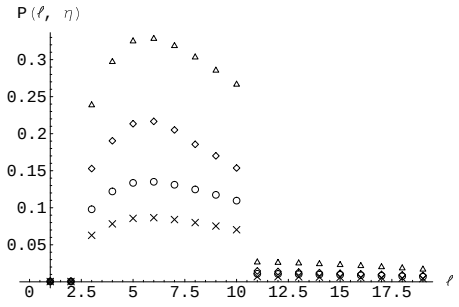

Figure 1: $P(\ell, \eta)$ vs. ℓ for different values of η (triangles for $\eta = 0.03$, diamonds for $\eta = 0.024$, circles for $\eta = 0.018$, crosses for $\eta = 0.012$) for the sequence TTTTTTTTTTCCCCCCCCCC, $m_{0G}/m_{0A} = 1.5$, $K_A/K_G = 10$.

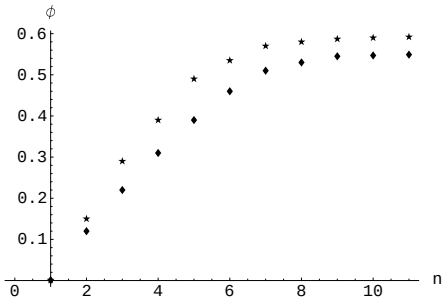

Figure 2: Comparison of yield ϕ estimated using the Velikanov-Kapral model (stars) and the present more general model (diamonds) for the sequence TCCCCCCCCCCCCCCCCCCC, all initial concentrations assumed equal, $K_A/K_G = 1.5$.

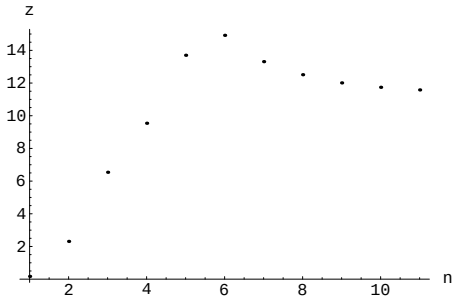

Figure 3: Required percentage change in initial concentration z for model (3) to match the yield estimated using the Velikanov-Kapral model at the end of each cycle for the sequence TCCCCCCCCCC-CCCCCCCCCC, all initial concentrations assumed equal, $K_A/K_G = 1.5$.

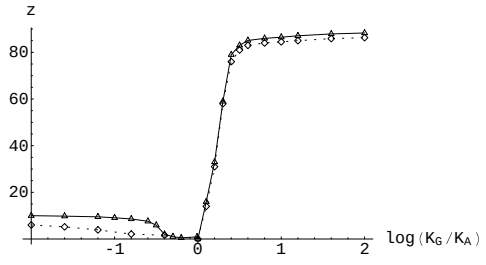

Figure 4: Required percentage change in initial concentration z for model (3) to match the yield estimated using the Velikanov-Kapral model with respect to variation in reaction rate $\log(K_G/K_A)$, for fixed duration $\eta = 0.03$, 12 cycles, and sequence TCCCCCCCCCCCCCCCCCCC (dotted line); maximum required percentage change in initial concentration z for model (3) 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).

Acknowledgement

This work was supported in part by NSF Grants EIA–0103660, IBN-0219322, and AFRL Grant F30 602-01-2-0572. The authors also gratefully acknowledge the generous assistance of Dr. Ruth Grene and Dr. Gregory Gonye.

References

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

[2] 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, Computer Science, Virginia Tech (http://eprints.cs.vt.edu:8000/archive/00000738).

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

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

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