A STOCHASTIC MODEL AND SIMULATION ALGORITHM FOR POLYMERASE CHAIN REACTION (PCR) SYSTEMS

Arjang Hassibi, Hossein Kakavand and Thomas H. Lee

Department of Electrical Engineering
Stanford University, CA 94305-4070, USA

ABSTRACT

A new stochastic approach to model Polymerase Chain Reaction (PCR) kinetic is presented, in which primer and template DNA sequences, enzyme concentration, temperature profile, PCR duration, hybridization kinetics and enzymatic rates are all incorporated. By studying the underlying biochemical processes of PCR, we show that under certain conditions, the extension length of DNA strands during a given time interval can be modeled as a Poisson random variable. We further use this fact to derive the distribution of the number of replicated strands, along with the probability of DNA replication (i.e. efficiency) in each PCR cycle. This in turn enables one to follow the stochastic behavior of the biochemical process as the PCR cycles progress. A simulation algorithm with preliminary results is also included, which demonstrates the feasibility and applicability of this modeling technique for a wide range of PCR applications.

1. INTRODUCTION

Today amplification and quantification of specific sequences of deoxyribonucleic acid (DNA) molecules using Polymerase Chain Reaction (PCR) has turned into a requisite part of Genomics as well as biotechnology in general [1,2]. PCR in principle and practice, mimics the basic processes of the natural DNA replication within the cell, except in a reaction tube. PCR relies on the use of temperature cycles which initiate and subsequently end an enzymatic process, supposedly doubling the number of target DNA molecules after each cycle. While in theory, one would expect an exponential growth for the target as a function of PCR cycles (i.e. 2^n times the original DNA copy number, after n cycles), in practice, replication processes measured by different real-time PCR systems show varying yields, suggesting a biochemical random process. In addition to variable *gains* and inconsistent amplification levels within a PCR process, there is also the likelihood of creating non-specific byproducts (i.e. DNA strands different than the target) as well as inserting mutations into the product, which further degrades the *quality* of the PCR product.

Because of the widespread use of PCR today, and the significance of its reliable performance and predictability, theoretical discussions for its accurate modeling are of great importance. Most of the existing probabilistic PCR models are based on the assumption that efficiency is constant within all relevant cycles [3,4], which clearly makes the model inapplicable to the plateau region [5] or PCR procedures with non-uniform temperature cycling schemes. In addition to kinetics modeling, theoretical aspects of mutation and branching processes in PCR have independently been studied [6]; yet a comprehensive model which includes both phenomena along with non-specific hybridization, enzyme degradation and product inhibition, has not been demonstrated.

In this paper, we introduce a new stochastic model for the underlying biochemical processes governing PCR kinetics. In Section 2, we show that by extracting the parameters of the random process from the assay conditions of PCR (e.g. primer sequences, enzyme concentration, and etc.), we can derive the probability of generated all DNA strands (both specific and non-specific products). Based on this derivation, we present the closed-form approximation of the replication efficiency for a three-step PCR in Section 3. Finally in Section 4, we provide simulation and numerical results verifying the applicability of the model.

2. PCR PROCESS

2.1. Assay procedure

Generally the goal of a PCR assay is to replicate a double-stranded DNA fragment (i.e., complementary DNA strands s_1 and s_2) using two primers, p_1 and p_2, that act as initiation sites for DNA polymerase enzyme. Each PCR cycle is typically carried out in three distinct steps governed by temperature. The three step are, denaturing, annealing and extension. In the first step, denaturing, the reaction mixture is heated (typically to above 90°C) to break the hydrogen bonds between base pairs of the complementary strands, ensuring all DNA fragments are single-stranded; hence, almost no double stranded DNA fragment is present at the end of this step (e.g. no $s_1 * s_2$ complex). In the second step of the cycle,

annealing, the reaction mix is cooled to a temperature (typically between 40-72°C) where the primers can efficiently hybridize to possible templates (e.g. p_1 to s_1 and p_2 to s_2). Both the primer and the template binding site sequence, which are usually designed to be exactly complementary, have a direct effect on the hybridization kinetics. During the annealing step, the thermo-stable DNA polymerase q initiates primer extension (incorporating deoxynucleotides – dNTPs onto the strand) subsequent to a successful hybridization. Finally during the extension phase, the temperature is increased to 72°C (optimum temperature for enzyme activity) creating a condition which maximizes the rate of extension, while keeping the partially extended primers attached. Based on this procedure, at the end of each cycle, s_1 and s_2 have a likelihood of creating s_2 and s_1 respectively, by primer extension, which can potentially result in an exponential growth of s_1 and s_2 as the cycles progress.

PCR is a homogenous (closed-tube) assay, in which primers, dNTPs, polymerase and DNA strands are always present in the reaction mix. Consequently, there is always the possibility of hybridization incidents and enzymatic extensions influenced by temperature. In the following subsections, we look into their probabilistic model independent of the temperature or step.

2.2. Hybridization and enzyme binding

During a PCR cycle, various DNA-DNA hybridization incidents are possible. For instance for the hybridization incident between two DNA strands s_i and s_j resulting in the $s_i * s_j$ complex, we have

$$s_i + s_j \xleftrightarrow{K_{ij}^h} s_i * s_j , \tag{1}$$

where K_{ij}^h is the reaction equilibrium constant qhich is a function of temperature. Consequently, p_{ij}^h, the hybridized portion of s_j is

$$p_{ij}^h = [s_i * s_j]/[s_j^0], \tag{2}$$

where $[s_j^0]$ is the initial concentration of s_j ([ ] indicates concentration). p_{ij}^h can always be derived using the hybridization rate, and K_{ij}^h. Additionally in thermal equilibrium, and for large number of $s_i * s_j$ and s_i^0 molecules, we can use the p_{ij}^h value as an approximation for the probability of s_i being hybridized to s_j.

The enzyme q can subsequently attach to any hybridized pair $s_i * s_j$ with some probability; however, before or after any dNTP incorporation, there is the likelihood of its detachment from the complex. Thus, we can define p_{ij}^E, the portion of $s_i * s_j$ complex attached to an enzyme molecule with equilibrium binding constant K_{ij}^E, using the following

$$p_{ij}^E = [s_i * s_j * q]/[s_i * s_j]. \tag{4}$$

Similar to p_{ij}^h, we can also approximate p_{ij}^E as the probability of attachment of an enzyme to the primer DNA-strand pair, given the primer and the DNA strand are attached.

2.3. Polymerization stochastic model

Consider two general DNA strands s_i and s_j. They randomly attach and detach from each other as time goes on. The attachment (binding) time instances of the $s_i * s_j$ pairs form a point process. It can be shown that this process is indeed a Poisson point process [7]. Additionally, if we replace each attachment event with a Delta function, we can represent the binding process as $S_h(t)$, the sum of delta functions, such that

$$S_h(t) = \sum_i \delta(t - t_i), \tag{3}$$

where t_i's are the time instances of each attachment event.

Next, we consider the enzyme attachment to the $s_i * s_j$ complex. As a function of time this attachment can be viewed as a random variable E, which randomly deletes some of the points in the Poisson point process associated with the binding times. These deletion sites are actually the points where the enzyme is not attached to the complex and hence the pair cannot extend. Basically the random variable E, deletes each point in the point process with probability p_{ij}^E. The remaining time instances $\{t_k\}$ form a Poisson point process; hence we can define the following process:

$$R(t) = \sum_k \delta(t - t_k) \tag{4}$$

where $\{t_k\}$ is a subset of $\{t_i\}$. $R(t)$ represents the time instances where the enzyme is attached to the $s_i * s_j$ complex and is derived from the Poisson point process.

During polymerization (primer extension), whenever the enzyme is attached to the $s_i * s_j$ complex, s_i (e.g. primer) extends with an average rate of ν (typically between 30-80s^{-1} in 72ºC). This rate is a function of temperature and dNTP concentration. Note that we can model polymerization as a Poisson random process, if the extension rate is sequence independent. Under this assumption, we define the extension length of the $s_i * s_j$ complex, N_{ij}, as the number of base increases during time epoch Δt. N_{ij} is actually the integration of $R(t)$ over the interval Δt; thus, the distribution of N_{ij} is

Assay Condition	Biochemical Process	Model Parameter
Temperature	Molecular bindings reaction rates	$K_{ij}^{h}, K_{ij}^{E}, \nu$
DNA Sequence	Hybridization	K_{ij}^{h}
DNA concentration	Molecular bindings	p_{ij}^{h}, p_{ij}^{E}
Salt concentration	Hybridization	K_{ij}^{h}
Enzyme concentration	Enzyme binding	p_{ij}^{E}
Polymerase Type	Enzyme rate and stability	ν, τ_E
dNTP concentration	Polymerization rate	ν

Table1: List of assay variables and the immediate path through which they influence the PCR biochemical process and the model parameters.

$$P(N_{ij} = k) = \frac{(\lambda_{ij}\Delta t)^k}{k!}\exp(-\lambda_{ij}\Delta t), \qquad (5)$$

where $\lambda_{ij} = \nu\ p_{ij}^{R}$. Note that p_{ij}^{R} is the probability of an enzyme being attached to the $s_i * s_j$ complex, which equals to $p_{ij}^{R} = p_{ij}^{h} p_{ij}^{E}$.

Having derived the distribution of the extension lengths starting from s_i as a function of hybridization kinetics and enzymatic processes, we can calculate the probability of generating all relevant DNA strands in time epoch Δt. Note that, (5) is valid as long as p_{ij}^{h}, p_{ij}^{E} and temperature remain relatively constant within that certain time epoch.

In Table.1, we have listed the assay conditions and the manner in which they directly affect the biochemical process in addition to model parameters. The stability parameter τ_E in Table.1 represents the decay half-life of polymerase as a function of temperature (e.g. τ_E for Taq polymerase is approximately 40min in 95°C).

3. THREE-STEP PCR APPROXIMATION

So far, we have modeled the basic underlying processes of PCR within an arbitrary time interval. While the general time increment approach can be implemented using simulation, it is mathematically intractable to find a closed form solution to the probability of successful DNA replication. Nonetheless, if we consider the three step PCR process, we can decouple the phases of each cycle and derive a closed-form approximation for its amplification levels (successful replication). As mentioned in Section 2.1, a three step PCR process has constant temperature during each of the denaturing, annealing and extension phases. And so the time epoch Δt can be set equal to duration of each of the three phases.

We can derive the probability of successful replication for each cycle by making the following assumptions. First, we assume that the binding probabilities do not change during the annealing phase of duration t_A, which is a rather true assumption for the early and mid-cycles. During this time interval polymerase has also a fixed extension rate, ν_A. As the primer extends over the template the overall hydrogen bonds between the pair become stronger, thus its breaking becomes less probable; however, it is again mathematically intractable to follow every single strand and calculate all the probabilities associated with each strand. The second simplifying assumption is to have a minimum extension length, m, such that any primer with a smaller extension length at the end of the annealing phase will get detached in the extension phase and any primer with a longer extension length than m will remain attached during the extension phase. The duration of the extension phase is t_E, and its enzyme extension rate is ν_E (note that $\nu_E >> \nu_A$).

A successful strand replication is one in which a primer completely extends and reaches the end of the template-DNA which is of length l. Based on the assumptions mentioned above the probability of a successful strand replication is equal to the probability of N_{ij}, reaching l, given that it reaches m in the annealing phase. It can be shown that

$$P(N_{ij} \geq l) = \sum_{k=l}^{\infty}\sum_{k_1=m}^{k}\frac{(\lambda_{ij}^{A})^{k-k_1}}{(k-k_1)!}\cdot\exp(-\lambda_{ij}^{A})\cdot\frac{(\lambda_{ij}^{E})^{k_1}}{(k_1)!}\cdot\exp(-\lambda_{ij}^{E}), \qquad (6)$$

where $\lambda_{ij}^{A} = p_{ij}^{R}\nu_A t_A$ and $\lambda_{ij}^{E} = \nu_E t_E$. This probability is, in fact, the probability of the extension length extending by k_1 bases, $(k_1 \geq m)$ during the annealing phase of duration t_A, times the probability of the extension length extending by at least another $l - k_1$ bases during the extension phase of duration t_E. Note that this probability is generally known as efficiency η_{ij}. It can be shown that under typical PCR conditions, the primer extension length can be approximated by a Poisson random variable $N_{ij} \sim Pos(p_{ij}^{R}\nu_A t_A + \nu_E t_E)$. As a result we can approximate the efficiency (probability of a complete combined extension) as

$$P(N_{ij} \geq l) \approx \sum_{k=l}^{\infty}\frac{(p_{ij}^{R}\nu_A t_A + \nu_E t_E)^k}{k!}\exp(-p_{ij}^{R}\nu_A t_A - \nu_E t_E). \qquad (7)$$

It is important to realize that all assay parameters and physical constants are all included in (7), except for the sequence-dependency of the enzyme activity (see

Table.1). While (7) is an approximation for the efficiency of a PCR cycle, it is not limited to the mid-cycles or to the plateau region, since the fundamental processes which affect the efficiency are already built in the proposed model.

4. SIMULATION

As shown in Fig.1, we have simulated a PCR assay with realistic assay conditions (mentioned in the figure footnotes), using the models presented in Section 2 and 3. In Fig.1a we see that the effects of primer concentration can considerably affect the annealing process. The saturation phenomenon in plateau region of this PCR assay simulation, based on the result demonstrated in Fig.2b, was due to limited concentration of polymerase, which resulted in decrease of p_{ij}^{E}. Our stochastic model also predicted the expected value and the standard deviation of the PCR product as the PCR cycles progress (Fig.1c).

5. CONCLUSION

In this paper, we have presented a stochastic model for Polymerase Chain Reaction (PCR) kinetics. We have shown that using the proposed methodology one can accurately simulate the biochemical process and calculate the expected value and the variation of the amplification. This modeling methodology can also be implemented to study the behavior of PCR assays with non-specific hybridization, product inhibition and non-uniform temperature cycling.

6. ACKNOWLEDGMENT

The authors want to thank Professor Abbas El Gamal for valuable discussions and Professor Douglas L. Brutlag for insightful technical feedback.

7. REFERENCES

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

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

[3] Stolovitzky, G., and G. Cecchi, "Efficiency of DNA replication in the polymerase chain reaction," *Proceedings of the National Academy of Sciences*, 93:12947–12952, 1996.

[4] Jagers, P., and F. Klebanar, "Random variation and concentration effects in PCR, " *J Theor Biol.* 224(3):299-304, 2003.

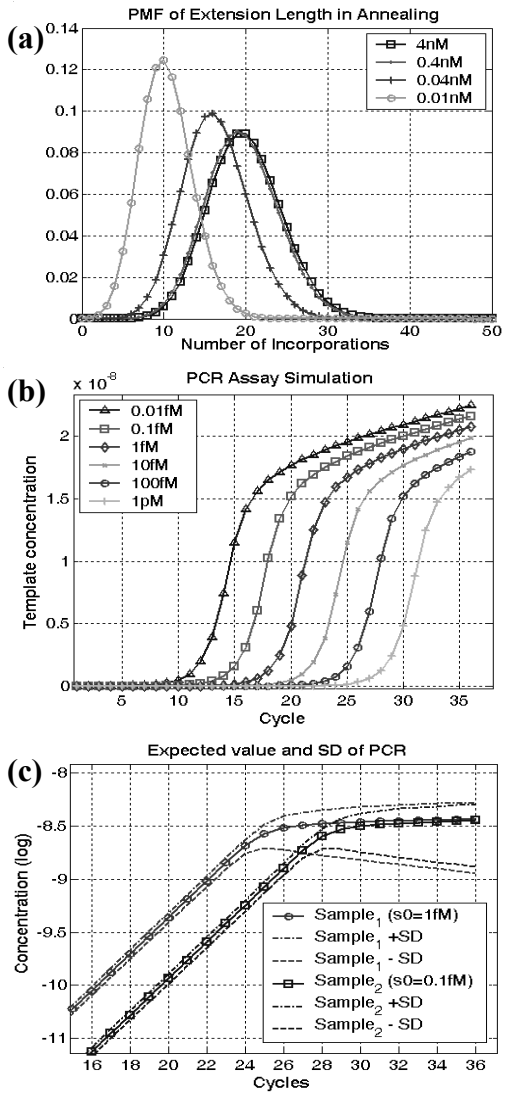

Figure1: Simulation results based on models presented in Section 2 and 3 of a PCR assay where primer concentration is 4nM, enzyme concentration is 1μM, and DNA initial concentration is 10fM, unless mentioned otherwise in (b) and (c). Annealing is carried out in the precise primer-template melting temperature, v_A and v_E are 2 and 30 respectively, and extension length l is 300 while minimum extension length m, is 5.

[5] Kainz, P., "The PCR plateau phase – toward an understanding of its limitations," *Biochimica et Biophysica Acta*, 1494:23-27, 2000.

[6] Piau, D., "Mutation-replication statistics of polymerase chain reactions," *J Comput Biol.,* 9(6):831-47, 2002.

[7] Papoulis, A., and U.S. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 2001.