

The Polymerase Chain Reaction and Branching Processes

Fengzhu Sun

Department of Mathematics, DRB-155
University of Southern California
Los Angeles, CA 90089-1113

Abstract

We construct a mathematical model for the polymerase chain reaction and its mutations using the theory of branching processes. Under this model we study the number of mutations in a randomly chosen sequence after n PCR cycles. A method for estimating the mutation rate is proposed and the variance of this estimator is studied. We also study the distribution of the Hamming distance between two randomly chosen sequences and a method for estimating the mutation rate based on pairwise differences is proposed.

1 Introduction

The polymerase chain reaction or PCR is a method that uses test tubes in laboratory, (i.e. it is an *in vitro* method), for producing large amount of identical copies of a specific gene from small amount of complex molecules (Saiki et al. 1985 1988, Mullis and Faloona 1987). The specificity, sensitivity, speed, and versatility of PCR are having a profound impact on molecular biological approaches to problems in human genetics, forensic science, infectious disease diagnosis, cancer research, evolutionary and developmental biology. For a survey of applications of PCR see Arnheim et al. (1990), White et al. (1989), Erlich and Arnheim (1992).

The principle of PCR can be outlined as follows. First a region of interest is chosen. This region is called a *target*. The nucleotide sequence of the target DNA may be unknown, but sequences of short stretches of DNA on either side of the target must be known. These sequences are used to design two oligonucleotide primers which are single-stranded sequences of DNA each usually 20 nucleotides long. The double-stranded DNA molecules are heated to near boiling temperature so that the double-stranded DNA molecules are separated completely into two single-stranded sequences (Figure 1 (b)). This process is called *denaturing*. The single-stranded sequences generated by denaturing are used as templates for the primers and the DNA polymerase. Then the temperature is lowered such that the primers anneal to the templates. Because DNA sequences can only grow from 5' to 3', the primers are oriented so that the 3' end of each primer directs toward the target sequence (Figure 1 (c)). This process is called *annealing*. The temperature is raised again to the temperature that is optimum for the polymerase to react. Because DNA polymerase can make phosphodiester bonds between nucleotides to make a long chain, The DNA polymerases use the single-stranded sequences as templates to extend the primers that have been annealed to the templates. Because of the specific base pairing, the newly synthesized strand is complementary to the original strand. The extension products of each primer are long enough so that they include the sequences complementary to the other primer. Therefore primer binding sites are generated on each newly synthesized DNA strand (Figure 1 (d)). This process is called *polymerase extension*.

—————-Figure 1 is around here—————

The three steps, DNA denaturing, primer annealing, and polymerase extension, form a PCR cycle. After the first cycle of PCR the number of DNA sequences that contain the target is doubled. If one cycle is followed by another one, the newly synthesized strands are separated from the original strands and all these single-stranded sequences can be used as templates for the primers and DNA polymerase. Thus each cycle essentially doubles the number of molecules containing the target sequence. After n PCR cycles, we can get a theoretical maximum of 2^n fold amplification. Unfortunately, in the experiment, not all

cycles are perfect, that is not every template can make a complete copy. Sometimes primers do not anneal to the templates or even if primers anneal to the templates, the primers can't be extended beyond the position of the the primer on the opposite strand. In that case the templates do not make complete copies. We can suppose a fraction λ of molecules make a complete copy. λ is called the *efficiency* of PCR. Strictly speaking, λ depends on the number of molecules in the experiment. It has been observed that if the number of molecules in the experiment is moderate, the efficiency is approximately a constant until the number of molecules reaches a very high level. During this period, the number of molecules increases exponentially. Thus this period is referred as the exponential region. After this high level, the efficiency begins to decrease. Finally the increase in the number of molecules becomes linear (Saiki et al. 1988). The reason for this phenomena is not clear. Presumably when the number of molecules is too high, the amount of enzymes present is not able to extend all the primer-template complexes in the allotted time. The efficiency and the maximum level of molecules depend on the enzymes used and the target sequence to be amplified. For practical purposes we concentrate on the exponential growth region.

Like all biochemical processes, PCR is not a perfect process and occasionally DNA polymerase substitutes, adds or deletes an nucleotide to the growing DNA chain. A mutation occurs in this case. Mathematical and statistical models are needed to study the distribution of the number of mutations in PCR and to estimate the mutation rate (probability of mutation per base per cycle). Several experimental and mathematical results have been presented in the literature. By directly cloning and sequencing the PCR products, Scharf et al. (1987) estimated the mutation rate for *E.coli* polymerase and Saiki et al. (1988) estimated the mutation rate for *Taq* polymerase using the formula $\mu = 2f/n$ where μ denotes the mutation rate, f is the observed error frequency per base in PCR products and n is the number of PCR cycles. By using denaturing gradient gel electrophoresis method—the wild type molecules and mutant molecules are hybridized to form heteroduplexes—to detect mutations, Keohavong and Thilly (1989) estimated the mutation rates of different DNA polymerases—T4, modified T7, *E. coli* and *Taq* polymerases. Eckert and Kunkel (1990) studied the change of mutation rates for *Taq* polymerase with the change of experimental conditions. All the above results are experimental. By using the theory of Galton—Watson processes, Krawczak et al. (1989) constructed a mathematical model for PCR mutations and obtained the proportion of PCR products with no mutations after n PCR cycles. They assumed that the efficiency of PCR is 1 and that a single-stranded sequence is falsely amplified with probability p_μ in each PCR cycle. Let S_0 be the number of original single-stranded sequences containing the target. Then the proportion C_n of PCR products with no mutations after n PCR cycles is:

$$EC_n = \left(1 - \frac{p_\mu}{2}\right)^n,$$

$$Var(C_n) = \frac{p_\mu \left(1 - \frac{p_\mu}{2}\right)^{2n-1}}{2S_0}.$$

Hayashi (1990) considered a similar problem by assuming that the efficiency is λ and that the number of mutations per single-stranded sequence per cycle of amplification is given by

a Poisson random variable with mean μG where G is the length of the target. Then the expected fraction of PCR products having no mutations is

$$\left(\frac{1}{1 + \lambda} + \frac{\lambda}{1 + \lambda} \exp(-\mu G) \right)^n .$$

This formula is applicable when the initial number of molecules is large. The above two papers only considered the fraction of correct copies and did not give any information about the number of errors occurred during the PCR reactions. By assuming that mutations occur randomly at a constant rate throughout PCR and efficiency is 1, Maruyama (1990) showed the total number M_n of mutations in n PCR cycles is

$$M_n = Gn\mu 2^n S_0.$$

Therefore the number of mutations for a randomly chosen sequence of length G is $Gn\mu/2$.

We can use different experimental methods to analyze mutations in DNA molecules. Denaturing gradient gel electrophoresis can in principle detect all point mutations. Without heteroduplex formation (the hybridization of wild and mutant sequences), we can not determine which base in a double-stranded molecule is changed. Therefore the probability of detecting the mutation is reduced to .50. Some other detection methods also have a 50% detection rate (Myers et al. 1988). Generally we can suppose a certain method can detect a mutation with probability c . Under the above considerations, Reiss et al. (1990) studied the proportion of PCR product with no detectable mutations and obtained a formula for the probability distribution of the number of detectable mutations. Because they did not consider the dependence structure of PCR products, the formula is wrong.

In this paper we construct a mathematical model for the polymerase chain reaction and its mutations using the theory of branching processes. Under this model we want to study the following questions

1. What is the distribution of a randomly chosen sequence after n PCR cycles?
2. Suppose we sample s sequences after n PCR cycles and obtain the number of mutations of these s sequences. How do we estimate the mutation rate?
3. What is the distribution of the Hamming distance between two randomly chosen sequences?

We learned the problem of pairwise Hamming distance from A. von Haeseler when he visited USC in 1993. The organization of this paper is as follows. In section 2 we construct a mathematical model for PCR and its mutations. The main results are presented without proof so that biologists can use these results without involving any mathematical details. In section 3 we give proofs for the results.

2 A mathematical model and results

Because of the complementary nature of the DNA sequences, we can study just one strand of the DNA. We consider a single-stranded model described below. Suppose initially we have S_0 identical copies of single-stranded sequences. They serve as templates for DNA replication. During each cycle, every template generates a new sequence with probability λ and itself always remains in the products. The original sequences and other sequences generated from them serve as templates for the next PCR cycle. This process is repeated for n cycles. Let S_n be the number of single-stranded sequences containing the target and the two primers after n PCR cycles. $S_0, S_1, S_2, \dots, S_n, \dots$ form a branching process. We also assume that the following two conditions are true.

1. Given S_n , the probability law governing $S_k, k \geq n+1$ depends only on S_n and not on the information before the n -th PCR cycle. That is the Markov property. The transition probabilities for the chain S_n do not depend on time because λ is a constant.
2. The behavior of each template does not depend on the behavior of other templates. That is the event that a template generates a complete copy is independent of the events that other templates generate a complete copy or not.

Then $S_0, S_1, S_2, \dots, S_n, \dots$ is a Galton—Watson process. We can use the general theory of branching processes to study S_n .

In the replication of new sequences, mutations can occur. During the synthesis of a new sequence, we suppose the number of mutations is a Poisson process with parameter μ . The probability that there are k mutations in a newly generated sequence of length G is $\exp(-\mu G)(\mu G)^k/k!$. We only consider substitutions and assume mutations occur in different places whenever a mutation occurs, so that there are no back mutations. Generally we have the following parameters.

S_0 = Initial number of sequences;
 G = Number of bases in the DNA segment to be amplified;
 n = Number of PCR cycles;
 μ = mutation rate per base per PCR cycle;
 λ = efficiency of PCR.

We suppose the original sequences are identical and do not have any mutations. If a sequence is generated directly from an original sequence, the number of mutations in the newly generated sequence is Poisson(μG) and if a sequence is generated from an original one through two replications, its number of mutations is Poisson($2\mu G$) and so on. This leads us to give the following definition.

Definition 1 *We call the original sequences the 0-th generation sequences. The sequences generated directly from the original sequences are called the first generation sequences; ... ; Inductively the sequences generated directly from the k -th generation sequences are called the $k + 1$ -st generation sequences.*

Figure 2 shows the mechanism of PCR. Sequence 0 generates 11 with probability λ and itself always remains among the products, and so on. After 2 PCR cycles we get at most 4 sequences. The first index in the notation denotes the generation number of the sequence and the second index counts the number of sequences of a given generation.

The first question we want to answer is the number of k -th generation sequences after n cycles. Let X_k^n be the number of k -th generation sequences after n cycles. We prove in section 3 that the expectation of X_k^n is $S_0 \binom{n}{k} \lambda^k$, $k \geq 0$, $n \geq 1$. From the theory of branching processes (Harris 1963), we can easily see $ES_n = S_0(1 + \lambda)^n$. After n PCR cycles, we choose a sequence. The probability that we get a k -th generation sequence is $E(X_k^n/S_n)$. When S_0 is sufficiently large, we can approximate this quantity by $\binom{n}{k} \lambda^k / (1 + \lambda)^n$. The exact formulation of this approximation is given in section 3. Simulations showed that if $\lambda > .85$, this approximation is good for any S_0 . Thus we will make the following assumption and refer it to assumption **A1**.

Assumption (A1). The distribution of the generation number K of a random chosen sequence after n PCR cycles is *Binomial*($n, \lambda/(1 + \lambda)$).

Then we have the following result.

Theorem 1 *Let M be the number of mutations of a randomly chosen sequence and assume (A1) holds. Then*

i). For any $m \geq 0$

$$P\{M = m\} = \frac{(\mu G)^m (1 + \lambda e^{-\mu G})^n}{m!(1 + \lambda)^n} E \left(\text{Bin} \left(n, \frac{\lambda e^{-\mu G}}{\lambda e^{-\mu G} + 1} \right) \right)^m.$$

ii). The probability generating function of M is

$$\frac{(1 + \lambda \exp(\mu G(s - 1)))^n}{(1 + \lambda)^n},$$

and

$$EM = \frac{n\lambda\mu G}{1 + \lambda}, \quad \text{Var}(M) = \frac{n(\lambda\mu G)}{(1 + \lambda)^2} (\mu G + 1 + \lambda).$$

iii). For any $x \in R$,

$$\lim_{n \rightarrow \infty} P \left\{ \frac{(1 + \lambda)M - n\lambda\mu G}{\sqrt{n\lambda\mu G(\mu G + 1 + \lambda)}} \leq x \right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds.$$

*iv). Suppose μ and G change with n , denoted by μ_n and G_n , such that $\lim_{n \rightarrow \infty} n\mu_n G_n = \nu$. Then M is approximately *Poisson*($\lambda\nu/(1 + \lambda)$) as n tends to infinity.*

Note 1: Part *iv*) is stated as a Poisson limit theorem and can be used in the following way. When μG is relatively small and $n\mu G$ is neither too small nor too large, then the number of mutations in a randomly chosen sequence can be approximated by $Poisson(\lambda\nu/(1 + \lambda))$.

Note 2: Just as in the approximation to binomial distribution, if $n\mu G$ is small, the Poisson approximation is better than normal approximation and if $n\mu G$ is large, normal approximation is better than Poisson approximation.

Note 3: From Theorem 1 we see the proportion of sequences without any replication errors is

$$\frac{(1 + \lambda \exp(-\mu G))^n}{(1 + \lambda)^n}.$$

which gives the result of Hayashi (1990). If $\lambda = 1$ this gives the result of Krawczak et al. (1989) and the expectation of M gives the result of Maruyama (1990). Because we have considered the dependence among the PCR products, the distribution of M is different from that given by Reiss et al. (1990).

—————Table 1 is around here—————

Table 1 shows the distribution of the number of mutations in a randomly chosen sequence after a) 20 and b) 50 cycles with $\mu G = 1/40$ and $\lambda = .9$ together with the Poisson and Normal approximations. In both cases, Poisson approximation almost coincides with the actual distribution. Normal approximation is much worse than Poisson approximation. For practical purposes, when the number of cycles is 20 to 50 and μG is less than .1, the Poisson approximation is usually better than Normal approximation.

—————Table 2 is around here—————

Table 2 gives the distribution of the number of mutations for different values of $\lambda = .8, .9, 1.0$ with a) 20 and b) 50 cycles and $\mu G = 1/40$. In the following we refer the fraction of sequences that are identical to the target sequence as purity of the PCR products. We see the purity of the final products does not change with the efficiency very much. On the other hand we see the number of mutations is stochastically decreasing as efficiency increases. This is intuitively true since less replications make less errors.

—————Table 3 is around here—————

Table 3 shows the change of the distribution of the number of mutations with target length and number of cycles when $\mu = 10^{-4}$, $\lambda = .9$. From this table we see that the target length and number of PCR cycles greatly affect the purity. When $n = 20$ and $G = 250$, 97.5 % of the product has at most one error and the number of errors ranges from 0 to 3. Whereas when $n = 50$ and $G = 1000$, the number of errors in a randomly chosen sequence ranges from 0 to 5.

Next let us study the estimation of the mutation rate. Suppose we sample s sequences from the PCR products with replacement. The case without replacement can be approximated by the analysis with replacement because the number of PCR products is large and the sample is small. Let M_1, M_2, \dots, M_s be the number of mutations of the sampled sequences. Then M_1, M_2, \dots, M_s have the same distribution but they are not independent. Under assumption **(A1)**, from Theorem 1 we know

$$E\left(\sum_{i=1}^s M_i\right) = sEM_1 = \frac{n\lambda\mu Gs}{1 + \lambda}.$$

Therefore

$$\hat{\mu} = \frac{(1 + \lambda) \sum_{i=1}^s M_i}{n\lambda Gs}$$

is an unbiased estimate of μ . For $\lambda = 1$, this estimate corresponds to the formula used by Saiki et al. (1988) and Scharf et al. (1988). The next results gives the asymptotic behavior of the variance of $\hat{\mu}$ as S_0 tends to infinity. We have the following result which does not depend on assumption **(A1)**.

Theorem 2 *Let S_0 be the initial number of sequences. Then for $\lambda = 1$,*

$$Var\left(\sum_{i=1}^s M_i\right) = \frac{sn\mu G}{4}(\mu G + 2) + \binom{s}{2} \frac{\mu G}{S_0}(1 - 2^{-n}).$$

For $0 < \lambda < 1$,

$$\lim_{S_0 \rightarrow \infty} S_0 \left\{ Var\left(\sum_{i=1}^s M_i\right) - \frac{sn\lambda\mu G}{(1 + \lambda)^2}(\mu G + 1 + \lambda) \right\} = sA + 2\binom{s}{2}B,$$

where

$$A = -\frac{(1 - \lambda)\mu G}{(1 + \lambda)^2} \left(1 + \frac{(1 - \lambda)\mu G}{1 + \lambda}\right) (1 - (1 + \lambda)^{-n}),$$

$$B = \frac{n\lambda^2\mu G}{(1 + \lambda)^{n+2}} + \frac{\mu G}{(1 + \lambda)^2} \left(2 - \frac{n\lambda + 2}{(1 + \lambda)^n}\right) \left(1 + \frac{(1 - \lambda)\mu G}{1 + \lambda}\right).$$

It is important to note that A and B are bounded with respect to n . From this theorem we see when S_0 is large, we can approximate $Var(\sum_{i=1}^s M_i)$ by $sn\lambda(\mu G + 1 + \lambda)/(1 + \lambda)^2$. In the following we use the above result to the data obtained in Saiki et al. (1988). In that

paper, after 30 cycles of PCR, they chose 28 separate clones each with 239 bps long from genomic DNA. 17 misincorporated bases were identified. In that paper, the efficiency λ of PCR is estimated at .85. Using our estimation we get an estimate of the mutation rate at

$$\hat{\mu} = \frac{(1 + \lambda) \sum_{i=1}^s M_i}{n\lambda Gs} = \frac{(1 + .85) \times 17}{28 \times 239 \times 30 \times .85} = 1.85 \times 10^{-4}.$$

The standard deviation of this estimate is approximately

$$\sqrt{\frac{\mu}{n\lambda Gs}(\mu G + 1 + \lambda)} = 4.5 \times 10^{-5}.$$

In the amplification of unknown DNA sequences, often we do not know the exact nucleotide sequence of the target. Thus it is impossible to get the number of mutations in a randomly chosen sequence. Von Haeseler proposed to use pairwise differences among a sample of PCR products to estimate the mutation rate (private communications). In the following we study the distribution of the pairwise differences between two randomly chosen sequences. From our model we see any two sequences are correlated through a branching process. Using the ideas in population genetics, we first give a definition of most recent common ancestor (Tavaré 1993).

Definition 2 For any sequence α , if there exist $\alpha = \alpha_t, \alpha_{t-1}, \dots, \alpha_0 = 0$ such that α_i is generated from $\alpha_{i-1}, 1 \leq i \leq t$, then any $\alpha_i, 0 \leq i \leq t-1$ is called an ancestor of α . If γ is a common ancestor of two sequences α and β and there are no other common ancestors before, then γ is called the most recent common ancestor (MRCA) of α and β . For purposes of MRCA only we identify all 0-th generation sequences.

Example. In Figure 2, the MRCA of any two sequences after 2 PCR cycles are respectively $MRCA(21,11) = 11$ and $MRCA(\alpha, \beta) = 0$ for $\alpha \neq 21, \beta \neq 11$.

It is easy to see the pairwise difference should depend on μ and their MRCA. Thus we give the following definition.

Definition 3 For any pair of sequences α and β , let γ be their MRCA and $g(\cdot)$ be the generation number of that sequence. Then we define the distance between α and β by

$$d(\alpha, \beta) = (g(\alpha) - g(\gamma)) + (g(\beta) - g(\gamma)).$$

Note: $d(\alpha, \beta)$ counts the number of PCR replications that occurred between sequences α and β .

Example. For the sequences in Figure 2 we have

$$\begin{aligned} d(21, 11) &= 1, & d(21, 12) &= 3, & d(21, 0) &= 2, \\ d(11, 12) &= 2, & d(11, 0) &= 1, & d(12, 0) &= 1. \end{aligned}$$

In section 3 we prove that the expected number of pairs with distance k after n PCR cycles is $S_0 P_n(k) + \binom{S_0}{2} \lambda^k$, where $P_n(k)$ satisfies the following recursive equation

$$P_n(1) = (1 + \lambda)^n - 1,$$

$$P_{n+1}(k) = P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2), \quad k = 2, 3, \dots, 2n+1,$$

where $P_n(k) = 0$ if $k = 0$ or $k \geq 2n$. The generating function of $\{P_n(k), k \geq 1\}$ is

$$\varphi_{P_n}(s) = \lambda s \frac{(1 + \lambda s)^{2n} - (1 + \lambda)^n}{(1 + \lambda s)^2 - (1 + \lambda)}.$$

The total expected number of pairs is

$$E\binom{S_n}{2} = S_0(1 + \lambda)^{n-1}((1 + \lambda)^n - 1) + \binom{S_0}{2}(1 + \lambda)^{2n}.$$

When S_0 is sufficiently large, we approximate the distribution of pairwise distance D between two randomly chosen sequences by

$$P\{D = k\} = \frac{S_0 P_n(k) + \binom{S_0}{2} \binom{2n}{k} \lambda^k}{E\binom{S_n}{2}}.$$

Thus we make the following assumption.

Assumption (A2). The distribution of the pairwise distance between two randomly chosen sequences is given by the above equation.

Under this assumption we have the following result.

Theorem 3 *Under Assumption (A2), the probability generating function $\varphi_D(s)$ of D is*

$$\begin{aligned} \varphi_D(s) &= \frac{1}{E\binom{S_n}{2}} (S_0 \varphi_{P_n}(s) + \binom{S_0}{2} (1 + \lambda s)^{2n}). \\ ED &= \frac{2n\lambda}{1 + \lambda} - \frac{2}{(1 + \lambda)S_0 + 1 - \lambda} + O\left(\frac{1}{S_0(1 + \lambda)^n}\right). \\ \text{Var}(D) &= \frac{2n\lambda}{(1 + \lambda)^2} - \frac{2(3 + \lambda)}{((1 + \lambda)S_0 + (1 - \lambda))(1 + \lambda)} \\ &\quad - \frac{2}{((1 + \lambda)S_0 + 1 - \lambda)^2} + O\left(\frac{1}{S_0(1 + \lambda)^n}\right). \end{aligned}$$

In the following, we study the pairwise Hamming distance H —the number of different bases between two randomly chosen sequences. By the model described above we have

$$H = \sum_{i=1}^D X_i,$$

where $X_i, i = 1, 2, \dots$ are i.i.d. $Poisson(\mu G)$. From the above equation we have

Theorem 4 Under assumption **(A2)**, we have

i). The probability generating function $\varphi_H(s)$ of H is

$$\varphi_H(s) = \varphi_D(\exp(\mu G(s - 1))).$$

ii). The expectation and variance of H are

$$EH = (\mu G)ED, \quad \text{Var}(H) = (\mu G)ED + (\mu G)^2 \text{Var}(D).$$

iii). For $0 < \lambda \leq 1$, $\frac{(1+\lambda)H - 2\lambda n\mu G}{\sqrt{2\lambda n\mu G(1+\lambda+\mu G)}}$ is asymptotically normal $N(0,1)$ as $n \rightarrow \infty$.

iv). If μ and G change with n , denoted by μ_n and G_n , such that $\lim_{n \rightarrow \infty} n\mu_n G_n = \nu$, then H is approximately Poisson($\frac{2\lambda}{1+\lambda}\nu$).

As at the end of Theorem 1, we can compare the Poisson and normal approximation and study the effect of efficiency, mutation rate, number of PCR cycles, and the length of target DNA on the pairwise differences. Since the idea is the same, we omit the details here. We can also use the moment estimation method to get an estimation of the mutation rate based on a sample of pairwise Hamming distances by

$$\tilde{\mu} = \frac{\sum_{i \neq j, i, j=1}^s H_{i,j}}{\binom{s}{2} ED \times G},$$

where $H_{i,j}$ is the pairwise Hamming distance between sequences i and j . But the variance of this estimator is hard to study.

Weiss and von Haeseler (1995) study in detail about the pairwise distance between two randomly chosen sequences. They perform simulation studies on the applicability of assumption **(A2)** and propose a χ^2 -method to estimate the mutation rate.

3 Mathematical proofs

In this section we prove the results listed in section 2. We divide this section into three subsections based on the three questions in section 1.

3.1 Distribution of the Number of Mutations in a Sequence

In this subsection, we elucidate assumption **A1** and prove Theorem 1. First let us study the expected number of k -th generation sequences.

Lemma 1 Let X_k^n be the number of k -th generation sequences after n PCR cycles. If $X_0^0 = 1$, then

$$EX_k^n = \binom{n}{k} \lambda^k, \quad k \geq 0, n \geq 1.$$

Proof. We prove this by induction.

- (i). The lemma is obviously true for $k = 0$.
- (ii). Suppose the lemma is true for $k - 1$ and any n . Then by the mechanism of PCR

$$X_k^{n+1} = X_k^n + \sum_{j=1}^{X_{k-1}^n} I_j.$$

where $I_1, I_2, \dots, I_j, \dots$ are i.i.d random indicators with

$$P\{I_j = 1\} = \lambda, \quad P\{I_j = 0\} = 1 - \lambda.$$

Thus

$$\begin{aligned} EX_k^{n+1} &= EX_k^n + \lambda EX_{k-1}^n \\ &= EX_k^n + \binom{n}{k-1} \lambda^k \\ &= \left(\binom{k-1}{k-1} + \dots + \binom{n}{k-1} \right) \lambda^k \\ &= \binom{n+1}{k} \lambda^k \end{aligned}$$

By induction the lemma is true. \square

Now suppose we choose a sequence randomly from the PCR products and let K be the generation number of the chosen sequence. Conditioning on $\mathbf{X}^n = (X_0^n, X_1^n, \dots, X_n^n)$ we have

$$P\{K = k | (X_0^n, X_1^n, \dots, X_n^n)\} = \frac{X_k^n}{S_n}. \quad (1)$$

Averaging over the sample values \mathbf{X}^n we obtain

$$P\{K = k\} = E(P\{K = k | (X_0^n, X_1^n, \dots, X_n^n)\}) = E \frac{X_k^n}{S_n}, \quad (2)$$

where $S_n = \sum_{k=0}^n X_k^n$ is the total number of sequences after n PCR cycles. It is hard to calculate $P\{K = k\}$ explicitly from equation (2).

If the initial number S_0 of sequences is large enough, we can use the following approximation. Let $S_n(i)$ be the total number of sequences and $X_k^n(i)$ be the number of k -th generation sequences generated from 0-th generation sequence i after n cycles. Then by strong law of large numbers (SLLN) we have

$$\lim_{S_0 \rightarrow \infty} \frac{\sum_{i=1}^{S_0} S_n(i)}{S_0} = ES_n(1) = (1 + \lambda)^n,$$

and

$$\lim_{S_0 \rightarrow \infty} \frac{\sum_{i=1}^{S_0} X_k^n(i)}{S_0} = EX_k^n(1) = \binom{n}{k} \lambda^k.$$

Therefore

$$\lim_{S_0 \rightarrow \infty} \frac{X_k^n}{S_n} = \lim_{S_0 \rightarrow \infty} \frac{\sum_{i=1}^{S_0} X_k^n(i)}{\sum_{i=1}^{S_0} S_n(i)} = \frac{\binom{n}{k} \lambda^k}{(1 + \lambda)^n}. \quad (3)$$

Since the conditional probabilities are given by equation (1), we can ask how reasonable our approximation in equation (3) is. We apply the SLLN to $\lim_{S_0 \rightarrow \infty} S_0^{-1} \sum_{i=1}^{S_0} S_n(i)$. Each $S_n(i)$ has mean $(1 + \lambda)^n$ and variance $(1 - \lambda)(1 + \lambda)^{n-1}((1 + \lambda)^n - 1)$ by application of standard branching process theory (Harris 1963). Therefore the sum $S_0^{-1} \sum_{i=1}^{S_0} S_n(i)$ has mean $(1 + \lambda)^n$ and variance $c(1 + \lambda)^{2n}/S_0$ for a constant c . To make $S_0^{-1} \sum_{i=1}^{S_0} S_n(i)$ close to $(1 + \lambda)^n$ requires $\sqrt{S_0}$ to be large relative to $(1 + \lambda)^n$. This is unlikely to obtain for small initial number of target molecules and should be a topic for future investigations.

Proof of Theorem 1.

i). For any $m \geq 0$,

$$\begin{aligned} P\{M = m\} &= \sum_{k=0}^n P(M = m | K = k) P(K = k) \\ &= \sum_{k=0}^n \exp(-k\mu G) \frac{(k\mu G)^m}{m!} \frac{\binom{n}{k} \lambda^k}{(1 + \lambda)^n} \\ &= \frac{(\mu G)^m}{m!(1 + \lambda)^n} \sum_{k=0}^n \exp(-k\mu G) \binom{n}{k} k^m \lambda^k \\ &= \frac{(\mu G)^m (1 + \lambda e^{-\mu G})^n}{m!(1 + \lambda)^n} E\left(\text{Bin}(n, \frac{\lambda e^{-\mu G}}{\lambda e^{-\mu G} + 1})\right)^m. \end{aligned}$$

ii). By the model described above we have

$$M = \sum_{i=1}^K X_i.$$

where $X_i, i = 1, 2, \dots$ are i.i.d Poisson(μG).

Let $\varphi_M(s), \varphi_K(s)$ and $\varphi_X(s)$ are the corresponding probability generating functions of M, K and X . Then we have

$$\varphi_M(s) = \varphi_K(\varphi_X(s)).$$

Because K is *Binomial*($n, \lambda/(1 + \lambda)$) and X is Poisson(μG) we have

$$\varphi_K(s) = \frac{(1 + \lambda s)^n}{(1 + \lambda)^n},$$

and

$$\varphi_X(s) = \exp(\mu G(s - 1)).$$

It follows that

$$\varphi_M(s) = \frac{(1 + \lambda \exp(\mu G(s - 1)))^n}{(1 + \lambda)^n}.$$

From the last equation we know that M can be represented as a sum of i.i.d random variables. Let Y_1, Y_2, \dots be i.i.d random variables with probability generating function $\frac{1+\lambda \exp(\mu G(s-1))}{1+\lambda}$. Then M has the same distribution as $\sum_{i=1}^n Y_i$. From the generating function of Y_i , it is easy to see

$$EY_1 = \frac{\lambda\mu G}{1+\lambda},$$

and

$$Var(Y_1) = \frac{\lambda\mu G(\mu G + 1 + \lambda)}{(1+\lambda)^2}.$$

Thus

$$EM = \frac{n\lambda\mu G}{1+\lambda},$$

and

$$Var(M) = \frac{n(\lambda\mu G)(\mu G + 1 + \lambda)}{(1+\lambda)^2}.$$

iii). Because M is a sum of i.i.d random variables, the central limit theorem holds. iii) is proved.

iv). From the condition $\lim_{n \rightarrow \infty} n\mu_n G_n = \nu$ it follows

$$\lim_{n \rightarrow \infty} n(\exp(\mu_n G_n(s-1)) - 1) = \nu(s-1).$$

Thus

$$\begin{aligned} \lim_{n \rightarrow \infty} \varphi_M(s) &= \lim_{n \rightarrow \infty} \left(\frac{1 + \lambda \exp(\mu_n G_n(s-1))}{1 + \lambda} \right)^n \\ &= \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda(1 - \exp(\mu_n G_n(s-1)))}{1 + \lambda} \right)^n \\ &= \exp\left(\frac{\lambda\nu(s-1)}{1 + \lambda}\right), \end{aligned}$$

which is the generating function of $Poisson(\lambda\nu/(1+\lambda))$. \square

3.2 Estimation of the Mutation Rate

In this subsection we mainly prove Theorem 2. We separate the proof of Theorem 2 into several lemmas. First note

$$\begin{aligned} Var\left(\sum_{i=1}^s M_i\right) &= \sum_{i=1}^s Var(M_i) + 2 \sum_{1 \leq i < j \leq s} Cov(M_i, M_j) \\ &= s Var(M_1) + 2\binom{s}{2} Cov(M_1, M_2). \end{aligned}$$

In the following we study $Var(M_1)$ and $Cov(M_1, M_2)$ separately. First we study $Cov(M_1, M_2)$. Let α and β be a pair of randomly chosen sequences. The following lemma relates $Cov(M(\alpha), M(\beta))$ to $Eg(\gamma)$ and $Cov(g(\alpha), g(\beta))$ where γ is the MRCA of α and β , $g(\cdot)$ and $M(\cdot)$ are the generation number and number of mutations of the corresponding sequences.

Lemma 2 For any pair of sequences α and β after n PCR cycles, let γ be their most recent common ancestor, $g(\cdot)$ be the generation number, and $M(\cdot)$ be the number of mutations. Then

$$Cov(M(\alpha), M(\beta)) = (\mu G)Eg(\gamma) + (\mu G)^2Cov(g(\alpha), g(\beta)).$$

Proof. Let α and β be the two chosen sequences and γ be their most recent common ancestor (MRCA). Let $g(\cdot), M(\cdot)$ be the corresponding generation number and number of mutations of the sequence, $M(\gamma\alpha)$ and $M(\gamma\beta)$ be the number of mutations from γ to α and β respectively. By the mechanism of PCR and the mutation process, $M(\gamma), M(\gamma\alpha)$ and $M(\gamma\beta)$ are independent given α, β and γ and

$$E(M(\gamma)|\alpha, \beta, \gamma) = g(\gamma)\mu G,$$

$$E(M(\gamma\alpha)|\alpha, \beta, \gamma) = (g(\alpha) - g(\gamma))\mu G,$$

$$E(M(\gamma\beta)|\alpha, \beta, \gamma) = (g(\beta) - g(\gamma))\mu G.$$

In these formulas the topology is assumed and the sequences $\alpha, \beta,$ and γ are random. Therefore

$$\begin{aligned} & EM(\alpha)M(\beta) \\ &= E(M(\gamma) + M(\gamma\alpha))(M(\gamma) + M(\gamma\beta)) \\ &= E\{E[(M(\gamma) + M(\gamma\alpha))(M(\gamma) + M(\gamma\beta))|\alpha, \beta, \gamma]\} \\ &= E\{E(M^2(\gamma)|\alpha, \beta, \gamma) + E(M(\gamma)M(\gamma\beta)|\alpha, \beta, \gamma) \\ &\quad + E(M(\gamma)M(\gamma\alpha)|\alpha, \beta, \gamma) + E(M(\gamma\alpha)M(\gamma\beta)|\alpha, \beta, \gamma)\} \\ &= E(M^2(\gamma)) + E\{E(M(\gamma)|\alpha, \beta, \gamma)E(M(\gamma\beta)|\alpha, \beta, \gamma) \\ &\quad + E(M(\gamma)|\alpha, \beta, \gamma)E(M(\gamma\alpha)|\alpha, \beta, \gamma) + E(M(\gamma\alpha)|\alpha, \beta, \gamma)E(M(\gamma\beta)|\alpha, \beta, \gamma)\} \\ &= E(M^2(\gamma)) + (\mu G)^2\{Eg(\gamma)(g(\beta) - g(\gamma)) \\ &\quad + Eg(\gamma)(g(\alpha) - g(\gamma)) + E(g(\alpha) - g(\gamma))(g(\beta) - g(\gamma))\} \\ &= E(M^2(\gamma)) + (\mu G)^2E(g(\alpha)g(\beta) - g^2(\gamma)) \\ &= Var(M(\gamma)) + (EM(\gamma))^2 + (\mu G)^2E(g(\alpha)g(\beta) - g^2(\gamma)) \\ &= Var(M(\gamma)) + (\mu G)^2(Eg(\gamma))^2 + (\mu G)^2E(g(\alpha)g(\beta) - g^2(\gamma)) \end{aligned}$$

$$\begin{aligned}
&= \text{Var}(M(\gamma)) - (\mu G)^2 \text{Var}(g(\gamma)) + (\mu G)^2 E(g(\alpha)g(\beta)) \\
&= \mu G E g(\gamma) + (\mu G)^2 E g(\alpha)g(\beta).
\end{aligned}$$

The last equation holds because

$$M(\gamma) = \sum_{i=1}^{g(\gamma)} X_i,$$

where X_i are i.i.d. $Poisson(\mu G)$. Thus

$$\begin{aligned}
\text{Var}(M(\gamma)) &= (E X_1)^2 \text{Var}(g(\gamma)) + \text{Var}(X_1) E g(\gamma) \\
&= (\mu G)^2 \text{Var}(g(\gamma)) + \mu G E g(\gamma).
\end{aligned}$$

Therefore

$$\text{Cov}(M(\alpha), M(\beta)) = \mu G E g(\gamma) + (\mu G)^2 \text{Cov}(g(\alpha), g(\beta)).$$

Lemma 2 is proved. \square

From Lemma 2, we see if we want to study the covariance of $M(\alpha)$ and $M(\beta)$, we only need to study $E g(\gamma)$ and $\text{Cov}(g(\alpha), g(\beta))$. Now let us first study $E g(\gamma)$.

Lemma 3 *Suppose $S_0 = 1$. Let $C_n(k)$ be the expected number of pairs with k -th generation MRCA. (The two sequences of the pair are different and order is not considered here.) Then*

$$C_{n+1}(k) = (1 + \lambda)^2 C_n(k) + \binom{n}{k} \lambda^{k+1}, \quad (4)$$

where $C_n(k) = 0$ for $k \geq n$.

The generating function of $C_n(k)$, $k = 0, 1, 2, \dots$ defined by $\varphi_{C_n}(s) = \sum_{k=0}^{n-1} C_n(k) s^k$, is

$$\varphi_{C_n}(s) = \lambda \frac{(1 + \lambda s)^n - (1 + \lambda)^{2n}}{(1 + \lambda s) - (1 + \lambda)^2}. \quad (5)$$

—————Figure 3 is around here—————

Proof. The recursive formula (4) can be proved as follows. Choose a pair of sequences with k -th generation MRCA after $n + 1$ PCR cycles. There are two cases. In the first case the two sequences of the pair are the result of a PCR amplification of a k -th generation

sequence during cycle $n+1$. There are X_k^n sequences that can be amplified to produce one of these pairs. The random number of pairs is therefore

$$\sum_{i=1}^{X_k^n} I_i,$$

with

$$P\{I_i = 1\} = 1 - P\{I_i = 0\} = \lambda.$$

The expectation of the number of these pairs is $EX_k^n EI_1 = \binom{n}{k} \lambda^{k+1}$. In the other case, the two sequences of the pair have different ancestors at n -th cycle (Figure 3) denoted by α and β . There are four possibilities for their choice corresponding to the amplification/nonamplification of the sequences.

1. The pair is (α, β) with $C_n(k)$ possibilities;
2. The pair is (α', β) or (α, β') with $2\lambda C_n(k)$ possibilities;
3. The pair is (α', β') with $\lambda^2 C_n(k)$ possibilities;

Summing over all these cases we have equation (4).

Let $\varphi_{C_n}(s)$ be the generating function of $C_n(k)$, $k = 0, 1, 2, \dots, n-1$. Then from equation (4) we have

$$\varphi_{C_{n+1}}(s) = (1 + \lambda)^2 \varphi_{C_n}(s) + \lambda(1 + \lambda s)^n.$$

By induction we can prove

$$\varphi_{C_n}(s) = \lambda \frac{(1 + \lambda s)^n - (1 + \lambda)^{2n}}{(1 + \lambda s) - (1 + \lambda)^2}.$$

Equation (5) holds and the lemma is proved. \square

The next lemma shows the behavior of expected generation number of the MRCA of a randomly chosen pair as $S_0 \rightarrow \infty$. It follows from the lemma that as S_0 tends to infinity the MRCA of two random PCR products is from generation 0.

Lemma 4 *Let A_n be the generation number of the MRCA of a randomly chosen pair with replacement from the products after n PCR cycles. Then*

$$\lim_{S_0 \rightarrow \infty} S_0 EA_n = \frac{2}{(1 + \lambda)^2} - \frac{2 + n\lambda(1 - \lambda)}{(1 + \lambda)^{n+2}}.$$

Proof. First suppose $S_0 = 1$. If the two sequences are chosen with replacement and order is considered, from Lemma 3 and taking into account the pairs with the same sequences whose expectation is $\binom{n}{k} \lambda^k$, the expected number $C_n^*(k)$ of pairs with k -th generation MRCA is

$$C_n^*(k) = 2C_n(k) + \binom{n}{k} \lambda^k. \tag{6}$$

Next suppose initially we have S_0 sequences. Let $Y_k^n(i)$ be the number of pairs with k -th generation MRCA and both of the sequences of the pairs are generated from 0-th generation sequence i . Let $S_n(i)$ be the total number of sequences generated from 0-th generation sequence i . Then the probability that chosen a pair, the pair is of k -th generation MRCA is

$$P\{A_n = k\} = E \frac{\sum_{i=1}^{S_0} Y_k^n(i)}{(\sum_{i=1}^{S_0} S_n(i))^2}, \quad 1 \leq k \leq n,$$

where A_n denote the generation number of the MRCA of a randomly chosen pair. From strong law of large numbers and Lemma 3 we have

$$\begin{aligned} \lim_{S_0 \rightarrow \infty} S_0 P\{A_n = k\} &= \lim_{S_0 \rightarrow \infty} S_0 E \frac{\sum_{i=1}^{S_0} Y_k^n(i)}{(\sum_{i=1}^{S_0} S_n(i))^2} \\ &= \lim_{S_0 \rightarrow \infty} E \frac{(\sum_{i=1}^{S_0} Y_k^n(i))/S_0}{(\sum_{i=1}^{S_0} S_n(i))^2/S_0^2} \\ &= \frac{E Y_k^n(1)}{(E S_n(1))^2} \\ &= \frac{C_n^*(k)}{(1 + \lambda)^{2n}}. \end{aligned}$$

Therefore from Lemma 3 we have

$$\begin{aligned} \lim_{S_0 \rightarrow \infty} S_0 E A_n &= \lim_{S_0 \rightarrow \infty} S_0 \sum_{k=1}^n k P\{A_n = k\} \\ &= \frac{1}{(1 + \lambda)^{2n}} \sum_{k=1}^n k (2C_n(k) + \binom{n}{k} \lambda^k) \\ &= \frac{1}{(1 + \lambda)^{2n}} (2\varphi'_{C_n}(1) + n\lambda(1 + \lambda)^{n-1}) \\ &= \frac{2}{(1 + \lambda)^2} - \frac{2 + n\lambda(1 - \lambda)}{(1 + \lambda)^{n+2}}. \end{aligned}$$

Lemma 4 is proved. \square

Next we study $Cov(g(\alpha), g(\beta))$. As in section 3.1, let X_k^n be the number of k -th generation sequences after n PCR cycles. Then we have

$$P\{g(\alpha) = k, g(\beta) = l\} = E \frac{X_k^n X_l^n}{S_n^2}.$$

Therefore

$$\begin{aligned} Cov(g(\alpha), g(\beta)) &= E \sum_{k,l} \frac{kl X_k^n X_l^n}{S_n^2} - \left\{ E \sum_k \frac{k X_k^n}{S_n} \right\}^2 \\ &= E \left\{ \sum_k \frac{k X_k^n}{S_n} \right\}^2 - \left\{ E \sum_k \frac{k X_k^n}{S_n} \right\}^2 \\ &= Var\left(\sum_k \frac{k X_k^n}{S_n}\right). \end{aligned}$$

Let $T_n = \sum_{k=0}^n kX_k^n$ and $T_n(i)$ be the corresponding quantity generated by 0-th generation sequence i . Then we have

$$T_n = \sum_{i=1}^{S_0} T_n(i),$$

and

$$\text{Cov}(g(\alpha), g(\beta)) = \text{Var}\left(\frac{T_n}{S_n}\right) = \text{Var}\left(\frac{\bar{T}_n}{\bar{S}_n}\right), \quad (7)$$

where

$$\bar{T}_n = \frac{\sum_{i=1}^{S_0} T_n(i)}{S_0},$$

and

$$\bar{S}_n = \frac{\sum_{i=1}^{S_0} S_n(i)}{S_0}.$$

In order to get the limit behavior of $\text{Var}(\bar{T}_n / \bar{S}_n)$, we give a lemma which gives the limit behavior of the expectation and variance of the ratio between two sample means.

Lemma 5 *Suppose $(X, Y)^T$ is a random vector with expectation $(\mu, \nu)^T$ and covariance matrix $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, and $Y \geq c > 0$, $\nu \neq 0$. Further we assume X is bounded. Let $(X_1, Y_1)^T, (X_2, Y_2)^T \dots (X_m, Y_m)^T$ be a sample of size m . Then*

- (i). $\lim_{m \rightarrow \infty} m \left(E \frac{\bar{X}}{\bar{Y}} - \frac{\mu}{\nu} \right) = \frac{\mu}{\nu^3} \sigma_2^2 - \frac{1}{\nu^2} \rho \sigma_1 \sigma_2,$
- (ii). $\lim_{m \rightarrow \infty} m \text{Var}\left(\frac{\bar{X}}{\bar{Y}}\right) = \frac{\sigma_1^2}{\nu^2} - 2 \frac{\mu}{\nu^3} \rho \sigma_1 \sigma_2 + \sigma_2^2 \left(\frac{\mu}{\nu^2}\right)^2.$

This lemma can be proved by Taylor expansion of x/y at (μ, ν) . (The so called δ -method.)

From equation (7) and Lemma 5, in order to obtain the limit behavior of $\text{Cov}(g(\alpha), g(\beta))$, we only need to know $\text{Var}(T_n)$, $\text{Cov}(T_n, S_n)$, and $\text{Var}(S_n)$. The following lemma gives these quantities. For later use, we also include a formula for the covariance of $R_n = \sum_{k=0}^n k^2 X_k^n$ and S_n .

Lemma 6 *Suppose initially we have one sequence. Let $T_n = \sum_{k=0}^n kX_k^n$, $R_n = \sum_{k=0}^n k^2 X_k^n$, and S_n be the total number of sequences after n cycles. Then*

$$\begin{aligned} \text{Var}(S_n) &= (1 - \lambda)(1 + \lambda)^{n-1}((1 + \lambda)^n - 1), \\ \text{Cov}(T_n, S_n) &= (1 - \lambda)(1 + \lambda)^{n-2}(n\lambda + 1)((1 + \lambda)^n - 1), \\ \text{Var}(T_n) &= (1 - \lambda)(1 + \lambda)^{n-3} \left[(1 + \lambda)^n (n^2 \lambda^2 + 2n\lambda + 2) - n\lambda(n\lambda + 3) - 2 \right], \\ \text{Cov}(R_n, S_n) &= (1 - \lambda)(1 + \lambda)^{n-3} \left[(1 + \lambda)^n (n^2 \lambda^2 + 3n\lambda + 1 - \lambda) \right. \\ &\quad \left. - n\lambda(n\lambda + 3) + \lambda - 1 \right], \\ n &= 0, 1, 2, \dots \end{aligned}$$

Proof. Because $S_n, n = 1, 2, \dots$ is a standard Galton—Watson process, $Var(S_n)$ follows directly from the theory of branching processes (Harris 1963).

In order to prove the second formula, we first study $ET_{n+1}S_{n+1}$ and condition on the behavior of the 0-th generation sequence after the first cycle. We consider two cases.

Case I. The 0-th generation sequence does not generate a new copy after the first cycle with probability $1 - \lambda$. Then after the first cycle we have only the 0-th generation sequence. We think of the second cycle as the starting point and let T_n^* and S_n^* be the corresponding quantities generated from this sequence after n cycles. (T_n^*, S_n^*) has the same distribution as (T_n, S_n) .

Case II. The 0-th generation sequence generates a new copy with probability λ . Then after the first cycle we have one 0-th generation sequence and one first generation sequence. We next consider the 0-th generation sequence and first generation sequence separately. For the 0-th generation sequence, as in case I, we think of the second cycle as the starting point and let $(T_n^{(0)}, S_n^{(0)})$ be the corresponding quantities. For the first generation sequence, we think of this sequence as the original sequence and the second cycle as starting point again. Let \tilde{X}_k^n be the number of k -th generation sequences generated by this newly generated sequence under this way of thinking and $T_n^{(1)} = \sum_{k=0}^n k\tilde{X}_k^n$ and $S_n^{(1)} = \sum_{k=0}^n \tilde{X}_k^n$. Because in the later case the actual generation number equal to the imagined generation number plus 1, the actual corresponding quantity of T is

$$\tilde{T}_n^{(1)} = \sum_{k=0}^n (k+1)\tilde{X}_k^n = \sum_{k=0}^n k\tilde{X}_k^n + \sum_{k=0}^n \tilde{X}_k^n = T_n^{(1)} + S_n^{(1)}.$$

Therefore in case II, we have

$$\begin{aligned} T_{n+1} &= T_n^{(0)} + T_n^{(1)} + S_n^{(1)}, \\ S_n &= S_n^{(0)} + S_n^{(1)}. \end{aligned}$$

From the model described above, we see $(T_n^{(0)}, S_n^{(0)})$ and $(T_n^{(1)}, S_n^{(1)})$ are independent and have the same distribution as (T_n, S_n) .

Combining the above two cases and using law of total probability we have

$$\begin{aligned} &ET_{n+1}S_{n+1} \\ &= (1 - \lambda)ET_n^*S_n^* + \lambda E(T_n^{(0)} + T_n^{(1)} + S_n^{(1)})(S_n^{(0)} + S_n^{(1)}) \\ &= (1 + \lambda)ET_nS_n + \lambda(2ET_nES_n + (ES_n)^2 + E(S_n)^2). \end{aligned}$$

Noting the facts that

$$\begin{aligned} ES_n &= (1 + \lambda)^n, \quad ET_n = n\lambda(1 + \lambda)^{n-1}, \\ E(S_n)^2 &= Var(S_n) + (ES_n)^2 = (1 + \lambda)^{n-1} (2(1 + \lambda)^n - (1 - \lambda)). \end{aligned}$$

We have

$$ET_{n+1}S_{n+1} = (1 + \lambda)ET_nS_n + \lambda(1 + \lambda)^{n-1} \left(((2n + 1)\lambda + 3)(1 + \lambda)^n - 1 + \lambda \right).$$

Therefore

$$Cov(T_{n+1}, S_{n+1}) = (1 + \lambda)Cov(T_n, S_n) + (1 - \lambda)\lambda(1 + \lambda)^{n-1} \left(((n + 1)\lambda + 2)(1 + \lambda)^n - 1 \right).$$

Then by induction we can prove the formula for $Cov(T_n, S_n)$ is true.

Using the same idea we can prove

$$\begin{aligned} Var(T_{n+1}) &= (1 + \lambda)Var(T_n) + (1 - \lambda)\lambda(1 + \lambda)^{n-2} \\ &\times \left\{ \left[(n + 1)^2\lambda^2 + (4n + 3)\lambda + 4 \right] (1 + \lambda)^n - \left[(2n + 1)\lambda + 3 \right] \right\}. \end{aligned}$$

By induction we can prove the formula for $Var(T_n)$.

Using the same idea we can prove

$$\begin{aligned} Cov(R_{n+1}, S_{n+1}) &= (1 + \lambda)Cov(R_n, S_n) + (1 - \lambda)\lambda(1 + \lambda)^{n-2} \\ &\times \left\{ \left[(n + 1)^2\lambda^2 + (5n + 3)\lambda + 4 \right] (1 + \lambda)^n - \left[(2n + 1)\lambda + 3 \right] \right\}. \end{aligned}$$

By induction we can prove the formula for $Cov(R_n, S_n)$. The lemma is proved. \square

From Lemmas 5 and 6, equation (7) and the fact

$$ET_n(1) = \sum_{k=1}^n k \binom{n}{k} \lambda^k = n\lambda(1 + \lambda)^{n-1}, \quad ES_n(1) = (1 + \lambda)^n,$$

we can prove

Lemma 7 *Let $g(\alpha)$ and $g(\beta)$ be the generation numbers of a randomly chosen pair from the products after n PCR cycles with replacement. Then*

$$\lim_{S_0 \rightarrow \infty} S_0 Cov(g(\alpha), g(\beta)) = \frac{(1 - \lambda)}{(1 + \lambda)^3} \left(2 - \frac{n\lambda + 2}{(1 + \lambda)^n} \right).$$

From Lemmas 2, 4 and 7 we obtain the limit behavior of $Cov(M_1, M_2)$. Next we study the limit behavior of $Var(M)$. Since $M = \sum_{i=1}^K X_i$ where X_i are *i.i.d* $Poisson(\mu G)$ and independent of K , we have

$$Var(M) = EKVar(X_1) + Var(K)(EX_1)^2 = \mu G EK + (\mu G)^2 Var(K).$$

The next lemma gives the limit behavior of EK and $Var(K)$.

Lemma 8 *Let K be the generation number of a randomly chosen sequence after n PCR cycles. Then*

$$\begin{aligned} (i). \lim_{S_0 \rightarrow \infty} S_0 \left(EK - \frac{n\lambda}{1 + \lambda} \right) &= -\frac{1 - \lambda}{(1 + \lambda)^2} (1 - (1 + \lambda)^{-n}). \\ (ii). \lim_{S_0 \rightarrow \infty} S_0 \left(Var(K) - \frac{n\lambda}{(1 + \lambda)^2} \right) &= -\frac{(1 - \lambda)^2}{(1 + \lambda)^3} (1 - (1 + \lambda)^{-n}). \end{aligned}$$

Proof. From equation (2) it follows

$$EK = \sum_{k=0}^n k E \frac{X_n^k}{S_n} = E \frac{\sum_{k=0}^n k X_n^k}{S_n} = E \frac{T_n}{S_n} = E \frac{\bar{T}_n}{\bar{S}_n}.$$

Then using Lemma 5 (i) to (\bar{T}_n, \bar{S}_n) and the formulas for $Cov(T_n, S_n)$ and $Var(S_n)$ in Lemma 6, we can prove (i).

Similar as above we have

$$EK^2 = E \frac{\bar{R}_n}{\bar{S}_n}.$$

Then using Lemma 5 (i) to (\bar{R}_n, \bar{S}_n) and the formulas for $Cov(R_n, S_n)$ and $Var(S_n)$ in Lemma 6, we can prove

$$\lim_{m \rightarrow \infty} S_0 \left(EK^2 - \frac{n\lambda(1+n\lambda)}{(1+\lambda)^2} \right) = -\frac{(1-\lambda)(2n\lambda+1-\lambda)}{(1+\lambda)^3} (1 - (1+\lambda)^{-n}).$$

Therefore

$$\begin{aligned} & \lim_{m \rightarrow \infty} S_0 \left(Var(K) - \frac{n\lambda}{(1+\lambda)^2} \right) \\ &= \lim_{m \rightarrow \infty} S_0 \left(EK^2 - \frac{n\lambda(1+n\lambda)}{(1+\lambda)^2} \right) - \lim_{m \rightarrow \infty} S_0 \left((EK)^2 - \left(\frac{n\lambda}{1+\lambda} \right)^2 \right) \\ &= -\frac{(1-\lambda)(2n\lambda+1-\lambda)}{(1+\lambda)^3} (1 - (1+\lambda)^{-n}) + \frac{1-\lambda}{(1+\lambda)^2} (1 - (1+\lambda)^{-n}) \times \frac{2n\lambda}{1+\lambda} \\ &= -\frac{(1-\lambda)^2}{(1+\lambda)^3} (1 - (1+\lambda)^{-n}). \end{aligned}$$

The lemma is proved. \square

Proof of Theorem 2. Once we have the above lemmas, it is easy to prove Theorem 2. From Lemmas 2, 4, 6, and 8 we see the second assertion of the Theorem holds .

When $\lambda = 1$, all X_k^n , T_n , and S_n are constants. From equation (7) we see $Cov(g(\alpha), g(\beta)) = 0$. Lemma 4 holds without taking limits. Thus the first assertion of the theorem holds.

Theorem 2 is proved. \square

3.3 Distribution of the Pairwise Differences

In this subsection we study the pairwise differences between two randomly chosen sequences and prove Theorems 3 and 4. First we give a lemma on the expected number of pairs with distance k .

Lemma 9 Assume $S_0 = 1$. Let $P_n(k)$ be the expected number of pairs with distance k after n PCR cycles. Then

$$P_n(1) = (1 + \lambda)^n - 1, \quad (8)$$

$$P_{n+1}(k) = P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2), \quad k = 2, 3, \dots, 2n+1, \quad (9)$$

where $P_n(k) = 0$ if $k = 0$ or $k \geq 2n$. The generating function of $\{P_n(k), k \geq 1\}$ is

$$\varphi_{P_n}(s) = \lambda s \frac{(1 + \lambda s)^{2n} - (1 + \lambda)^n}{(1 + \lambda s)^2 - (1 + \lambda)}. \quad (10)$$

Proof. In the following we use $N_n(k)$ to denote the number of pairs with distance k and S_n to denote the total number of sequences after n cycles. First we prove equation (8). It is obvious that $P_1(1) = \lambda$. After $n+1$ cycles the pairs of sequences that are distance 1 apart have two cases. In the first case, the pairs are already distance 1 apart after n cycles having $N_n(1)$ possibilities. In the second case, one of the sequences in the pair is generated from the other one at $(n+1)$ -st cycle. Because there are S_n sequences after n cycles, the number of pairs in the second case is

$$\sum_{i=1}^{S_n} I_i,$$

with $P\{I_i = 1\} = \lambda, P\{I_i = 0\} = 1 - \lambda$. Therefore

$$N_{n+1}(1) = N_n(1) + \sum_{i=1}^{S_n} I_i.$$

Taking expectation on both sides we have

$$EN_{n+1}(1) = EN_n(1) + ES_n EI_i = EN_n(1) + \lambda(1 + \lambda)^n,$$

and

$$P_{n+1}(1) = P_n(1) + \lambda(1 + \lambda)^n.$$

By induction we can get equation (8).

Equation (9) can be proved as follows. The pairs in the second case described above can only have distance 1. Therefore in order that they have distance $k > 1$, they must have different ancestors after n cycles. For any pair of sequences after $n+1$ cycles, suppose their ancestors at n -th cycle are α and β . There are four possibilities corresponding to the amplification/nonamplification of the sequences. The chosen pair has distance k (Figure 3)

1. The pair is (α, β) with $d(\alpha, \beta) = k$ after n cycles having $P_n(k)$ possibilities.
2. The pair is (α', β) or (α, β') with $d(\alpha, \beta) = k - 1$ after n cycles having $2\lambda P_n(k - 1)$ possibilities.
3. The pair is (α', β') with $d(\alpha, \beta) = k - 2$ after n cycles having $\lambda^2 P_n(k - 2)$ possibilities.

Summing over all the cases we have

$$P_{n+1}(k) = P_n(k) + 2\lambda P_n(k-1) + \lambda^2 P_n(k-2).$$

From equation (9) it follows

$$\begin{aligned} \varphi_{P_{n+1}}(x) &= ((1+\lambda)^{n+1} - 1)x + \sum_{k=2}^{2n+1} P_{n+1}(k)x^k \\ &= ((1+\lambda)^{n+1} - 1)x + \sum_{k=2}^{2n-1} P_n(k)x^k \\ &\quad + 2\lambda x \left(\sum_{k=1}^{2n-1} P_n(k)x^k \right) + (\lambda x)^2 \left(\sum_{k=1}^{2n-1} P_n(k)x^k \right) \\ &= (1+\lambda x)^2 \varphi_{P_n}(x) + \lambda(1+\lambda)^n x. \end{aligned}$$

By induction we can prove equation (10). \square

Next suppose initially we have S_0 sequences. Every 0-th generation sequence generates a set of sequences after n PCR cycles. The expected number of pairs with distance k where both of the sequences of the pair have the same 0-th generation ancestor is $S_0 P_n(k)$. The expected number of pairs with distance k and the two sequences of the pair have different 0-th generation ancestors is

$$\begin{aligned} E \left\{ \sum_{k_1 < k_2} \sum_{i+j=k} X_i^n(k_1) X_j^n(k_2) \right\} &= \sum_{k_1 < k_2} \sum_{i+j=k} E X_i^n(k_1) E X_j^n(k_2) \\ &= \binom{S_0}{2} \sum_{i+j=k} \binom{n}{i} \lambda^i \binom{n}{j} \lambda^j \\ &= \binom{S_0}{2} \binom{2n}{k} \lambda^k. \end{aligned}$$

Therefore the total expected number of pairs with distance k is

$$S_0 P_n(k) + \binom{S_0}{2} \binom{2n}{k} \lambda^k.$$

The total expected number of pairs is

$$\begin{aligned} E \binom{S_n}{2} &= \sum_{k=0}^{2n} \left\{ S_0 P_n(k) + \binom{S_0}{2} \binom{2n}{k} \lambda^k \right\} \\ &= S_0 (1+\lambda)^{n-1} ((1+\lambda)^n - 1) + \binom{S_0}{2} (1+\lambda)^{2n}. \end{aligned}$$

When S_0 is sufficiently large, we approximate the distribution of pairwise distance D between two randomly chosen sequences by

$$P\{D = k\} = \frac{S_0 P_n(k) + \binom{S_0}{2} \binom{2n}{k} \lambda^k}{E \binom{S_n}{2}}. \quad (11)$$

Letting $S_0 \rightarrow \infty$, we obtain

$$P\{D = k\} = \frac{\binom{2n}{k}\lambda^k}{(1 + \lambda)^{2n}}, \quad k = 0, 1, \dots.$$

The quantification we make regarding assumption **(A1)** which requires the size of S_0 to be large relative to $(1 + \lambda)^{2n}$ holds here also.

Proof of Theorem 3. From equation (11) we have

$$\begin{aligned} \varphi_D(s) &= \sum_{k=0}^{2n} P\{D = k\} s^k \\ &= \frac{1}{E\binom{S_n}{2}} \sum_{k=0}^{2n} (S_0 P_n(k) s^k + \binom{S_0}{2} \binom{2n}{k} \lambda^k s^k) \\ &= \frac{S_0 \varphi_{P_n}(s) + \binom{S_0}{2} (1 + \lambda s)^{2n}}{E\binom{S_n}{2}}. \end{aligned}$$

Once we get the generating function of D , it is a routine exercise to get the expectation and the variance of D . Because of the complicated form of $\varphi_D(x)$, the calculation is lengthy and we omit the calculation here. \square

Proof of Theorem 4.

By our model we have

$$H = \sum_{i=0}^D X_i, \quad (12)$$

where X_i are i.i.d Poisson(μG) and independent of D . From equation (12) we have

$$\begin{aligned} \varphi_H(s) &= E s^H \\ &= E(E(s^H/D)) \\ &= E \exp(\mu G D (s - 1)) \\ &= \varphi_D(\exp(\mu G (s - 1))), \end{aligned}$$

and i) is proved.

ii). From equation (12) we have

$$EH = ED \times EX_1 = \mu G ED,$$

$$Var(H) = ED \times Var(X_1) + (EX_1)^2 \times Var(D) = (\mu G)ED + (\mu G)^2 Var(D).$$

iii). In order to prove iii), we separate the distribution of H into two parts—pairwise difference within the groups and pairwise difference between the groups. That is

$$P\{H = k\} = c_n P_w(k) + (1 - c_n) P_b(k),$$

where $c_n = \frac{S_0(1+\lambda)^{n-1}((1+\lambda)^n-1)}{E\binom{2n}{2}}$ and

$$P_w(k) = P\{H_w = k\} = \frac{P_n(k)}{(1+\lambda)^{n-1}((1+\lambda)^n-1)}, \quad 1 \leq k \leq 2n-1,$$

and

$$P_b(k) = P\{H_b = k\} = \frac{\binom{2n}{k}\lambda^k}{(1+\lambda)^{2n}}, \quad 0 \leq k \leq 2n.$$

The probability generating function of H_w and H_b are

$$\varphi_{H_w}(s) = \frac{\varphi_{P_n}(\exp(\mu G(s-1)))}{(1+\lambda)^{n-1}((1+\lambda)^n-1)}, \quad (13)$$

and

$$\varphi_{H_b}(s) = \left(\frac{1 + \lambda \exp(\mu G(s-1))}{1 + \lambda} \right)^{2n}.$$

Just as in the proof of Theorem 1, H_b can be represented as a sum of $2n$ i.i.d random variables with generating function $\frac{1+\lambda \exp(\mu G(s-1))}{1+\lambda}$. By central limit theorem $\frac{(1+\lambda)H_b - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}}$ is approximately normal $N(0, 1)$.

Let $g_n(t) = \varphi_{H_b}(\exp(it))$ be the characteristic function of H_b . Then

$$\begin{aligned} & \lim_{n \rightarrow \infty} E \exp \left(it \frac{(1+\lambda)H_b - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \\ &= \lim_{n \rightarrow \infty} g_n \left(\frac{(1+\lambda)t}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \exp \left(-it \frac{2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \\ &= \exp\left(-\frac{t^2}{2}\right). \end{aligned}$$

Thus from equations (10) and (13), after lengthy computations we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} E \exp \left(it \frac{(1+\lambda)H_w - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \\ &= \lim_{n \rightarrow \infty} g_n \left(\frac{(1+\lambda)t}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \exp \left(-it \frac{2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}} \right) \\ &= \exp\left(-\frac{t^2}{2}\right). \end{aligned}$$

Therefore $\frac{(1+\lambda)H_w - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1+\lambda+\mu G)}}$ is also approximately normal $N(0, 1)$.

Let $c_\infty = \lim_{n \rightarrow \infty} c_n = \frac{2}{1 - \lambda + s_0(1 + \lambda)}$, it follows

$$\begin{aligned}
& \lim_{n \rightarrow \infty} P \left\{ \frac{(1 + \lambda)H - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1 + \lambda + \mu G)}} \leq x \right\} \\
&= c_\infty \lim_{n \rightarrow \infty} P \left\{ \frac{(1 + \lambda)H_b - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1 + \lambda + \mu G)}} \leq x \right\} \\
&\quad + (1 - c_\infty) \lim_{n \rightarrow \infty} P \left\{ \frac{(1 + \lambda)H_w - 2\lambda n \mu G}{\sqrt{2\lambda n \mu G(1 + \lambda + \mu G)}} \leq x \right\} \\
&= c_\infty \phi(x) + (1 - c_\infty) \phi(x) \\
&= \phi(x),
\end{aligned}$$

where $\phi(x)$ is the distribution of $N(0, 1)$. iii) is proved.

iv). The proof of iv) is almost the same as that of iii). From $\lim_{n \rightarrow \infty} n \mu_n G_n = \nu$, it follows

$$\lim_{n \rightarrow \infty} \exp(\mu_n G_n(s - 1)) = 1,$$

and

$$\lim_{n \rightarrow \infty} n(1 - \exp(\mu_n G_n(s - 1))) = \nu(s - 1).$$

Thus

$$\begin{aligned}
\lim_{n \rightarrow \infty} \varphi_{H_b}(s) &= \lim_{n \rightarrow \infty} \left(\frac{1 + \lambda \exp(\mu_n G_n(s - 1))}{1 + \lambda} \right)^{2n} \\
&= \lim_{n \rightarrow \infty} \left\{ 1 - \frac{\lambda}{1 + \lambda} [1 - \exp(\mu_n G_n(s - 1))] \right\}^{2n} \\
&= \exp \left(\frac{2\lambda\nu}{1 + \lambda} (s - 1) \right).
\end{aligned}$$

Therefore H_b is approximately $Poisson(\frac{2\lambda\nu}{1 + \lambda})$. From the formula for $\varphi_{H_w}(s)$ we also have

$$\lim_{n \rightarrow \infty} \varphi_{H_w}(s) = \exp \left(\frac{2\lambda\nu}{1 + \lambda} (s - 1) \right),$$

which is the generating function of $Poisson(2\lambda\nu/(1 + \lambda))$. Thus H_b is also approximately $Poisson(2\lambda\nu/(1 + \lambda))$. By the same argument as in iii) we can prove H is approximately $Poisson(2\lambda\nu/(1 + \lambda))$. \square

Acknowledgments. This paper is part of my dissertation at USC. I thank my advisor, M. Waterman, for his help and encouragements during my studies at USC. I would also like

to thank Professor N. Arnheim for explaining PCR to me. I am grateful to Professors S. Tavaré and T. Harris for suggestions that improved the presentation of the paper. I learned the problem of pairwise differences from A. von Haeseler when he visited USC in 1993. Thanks are also due to Weiss and von Haeseler for making their paper available to me and for their valuable comments on the manuscript.

This work was supported by grants from the National Science Foundation (DMS-90-05833) and the National Institutes of Health (GM-36230). Some of this work was also done while I was visiting DIMACS and was supported by National Science Foundation (STC-91-19999 and BIR-9412594).

References

- [1] Arnheim, N., White, T., and Rainey, W. E. 1990. Application of PCR: organismal and population biology. *BioScience* 40: 174-82
- [2] Erlich, H. A. and Arnheim, N. 1992. Genetic analysis using the polymerase chain reaction. *Annu. Rev. Genet. Vol. 26: 479-506*
- [3] Harris, T. E. 1963. *The theory of branching processes*. Springer, Berlin
- [4] Hayashi, K. 1990. Mutations induced during the polymerase chain reaction. *Technique, Vol. 2: 216-217*
- [5] Eckert, K. A. and Kunkel, T. A. 1990. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Nucleic Acids Research, Vol. 18: 3739-44*
- [6] Keohavong, P. and Thilly, W. G. 1989. Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. USA, Vol. 86: 9253-57*
- [7] Krawczak, M., Reiss, J., Schmidtke, J., and Rosler, U. 1989. Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucleic Acids Research, Vol. 17: 2197-2201*
- [8] Maruyama, I. N. 1990. Estimation of errors in the polymerase chain reaction. *Technique, Vol.2: 302-303*
- [9] Mullis, K. B. and Faloona, F. A. 1987. Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. *Methods Enzymol. Vol. 155: 335-51*
- [10] Myers, R. M., Sheffield, V. C., and Cox, D. R. 1988. Detection of single base changes in DNA: ribonuclease cleavage and denaturing gradient gel electrophoresis. *Genome analysis—a practical approach*. Ed. Davis, K. E. IRL Press, Oxford
- [11] Reiss, J., Krawczak, M., Schloesser, M., Wagner, M., and Cooper, D. N. 1990. The effect of replication errors on the mismatch analysis of PCR-amplified DNA. *Nucleic Acids Research, Vol. 18: 973-78*

- [12] Saiki, R., Scharf, S., Faloona, F., Mullis, K., 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*. Vol. 230: 1350-54
- [13] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., et al. 1988. Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*. Vol. 239: 487-91
- [14] Scharf, S. J., Horn, and Erlich, H. A. 1986. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*. Vol. 223: 1076-78
- [15] Tavaré, S. 1993. *Lecture notes in population genetics*. University of Southern California
- [16] Weiss, G. and von Haeseler, A. 1995. Modelling the polymerase chain reaction. *J. Computational Biology*. This issue
- [17] White, T. J., Arnheim, N., and Erlich, H. A. 1989. The polymerase chain reaction. *Trends Genet*. Vol. 5: 185-89

Figure Legends

Figure 1. Principle of PCR. (a). The double stranded molecule. (b). DNA is separated into two strands by denaturing. (c). The primers anneal to the single-stranded sequences. (d). DNA polymerase synthesizes the primers that have annealed to the templates, generating double-stranded DNA.

Figure 2. The mechanism of PCR. 0 generates 11 with probability λ and itself always remains in the products, so on.

Figure 3. α and β generate pairs (α', α) and (β', β) respectively.