Adv. Appl. Prob. **29**, 629–668 (1997)
Printed in N. Ireland

WHOLE GENOME AMPLIFICATION AND BRANCHING PROCESSES

FENGZHU SUN,* *Emory University*
MICHAEL S. WATERMAN,** *University of Southern California*

Abstract

Whole genome amplification is important for multipoint mapping by sperm or oocyte typing and genetic disease diagnosis. Polymerase chain reaction is not suitable for amplifying long DNA sequences. This paper studies a new technique, designated PEP-primer-extension-preamplification, for amplifying long DNA sequences using the theory of branching processes. A mathematical model for PEP is constructed and a closed formula for the expected target yield is obtained. A central limit theorem and a strong law of large numbers for the number of kth generation target sequences are proved.

BRANCHING PROCESSES; PEP; CENTRAL LIMIT THEOREM; STRONG LAW OF LARGE NUMBERS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 92A10
SECONDARY 60J85; 60K80

1. Introduction

The polymerase chain reaction or PCR is a method that allows biologists in the laboratory to produce a large number of identical copies of a specific DNA molecule from as few as one molecule (Saiki *et al.* 1985, 1988; Mullis and Faloona 1987). PCR is widely used in almost all branches of biological studies including human genetics, forensic science and cancer research. A key to understanding the importance of PCR is that most experimental procedures require 10^7 to 10^8 identical molecules. For a survey of applications of PCR, see Arnheim *et al.* (1990a, b), White *et al.* (1989), Erlich and Arnheim (1992).

PCR uses certain features of DNA replication. Thus we begin with some basic knowledge about DNA and its replication mechanisms. DNA is a double-stranded sequence formed by two purines (adenine(A) and guanine(G)) and two pyrimidines (thymine(T) and cytosine(C)) that are called *bases*. Each strand is composed of a linear sequence of these four bases which are connected by chemical bonds (called phosphodiester bonds). Every base in one strand pairs with another base in the other strand according to the following rules: adenine(A) can only pair with

Received 13 September 1994; revision received 4 April 1996.

* Postal address: Department of Genetics, Emory University School of Medicine, Atlanta, GA 30329, USA.

** Postal address: Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113, USA.

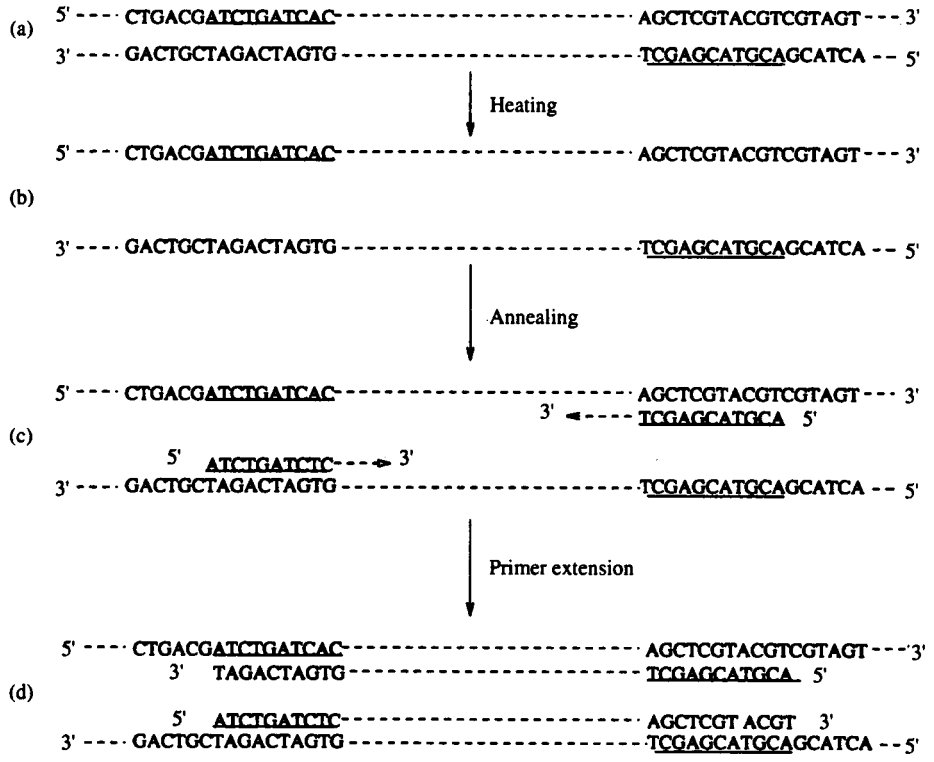

Figure 1. Principle of PCR. (a) The double-stranded DNA molecule. The primers are underlined. (b) DNA is separated into two strands by denaturing. (c) The primers anneal to the single-stranded sequences. (d) DNA polymerase synthesizes the primers that have annealed to the templates, generating double-stranded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

thymine(T), and guanine(G) can only pair with cytosine(C). Because of this specific base pairing, if the sequence of one chain is known, that of its partner is also known. The opposing strand is referred to as *complementary*. Each strand of DNA has an orientation; one end is the 5′ end and the other is the 3′ end. The complementary strand has a reciprocal orientation. Figure 1(a) gives an example of a double-stranded DNA molecule with its orientations. Although the bond between each pair is quite weak, each DNA molecule contains so many base pairs that two strands bind strongly. The two strands never separate spontaneously under normal conditions. If, however, DNA is exposed to high temperatures, so many base pairs fall apart that the two strands separate into two single-stranded sequences. This process is called *denaturation*. This is the basis of the polymerase chain reaction. Figure 1(b) shows two single-stranded DNAs denatured from the double-stranded DNA.

DNA can replicate or reproduce itself. In DNA replication, the double-stranded DNA first separates into two single-stranded DNAs. Enzymes called DNA polymerase use these single-stranded sequences as templates to synthesize two new double-stranded molecules. Because of the specific base pairing, the newly synthesized

strand is complementary to the original strand. This is another feature that is used in the polymerase chain reaction.

The principle of PCR can be outlined as follows. First a region of interest is chosen. This region is called a *target* (Figure 1). 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. Knowledge of these sequences is used to design two oligonucleotide primers which are single-stranded sequences of DNA (each usually 20 nucleotides long) that are complementary to these short stretches read from the 5′ ends. The double-stranded DNA molecules are heated to high temperatures so that the double-stranded DNA molecules are separated completely into two single-stranded sequences, i.e. they are denatured. The single-stranded sequences generated by denaturing are used as templates for the primers and the DNA polymerase. Then the temperature is lowered so 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. 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 form a long chain, the DNA polymerases use the single-stranded sequences as templates to extend the primers that have been annealed to the templates. The extension products of the primers 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. This process is called *polymerase extension.*

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, 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, i.e. 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 might not be extended beyond the position of the primer on the opposite strand. In that case the templates do not make complete copies. We can suppose a fraction E of molecules make a complete copy. E is called the *efficiency* of PCR. Under the above assumptions, the number of PCR products forms a branching process. A branching process model has been used by Krawczak *et al.* (1989), Weiss and von Haeseler (1995) and Sun (1995) to study the mutations in PCR.

Two problems make it hard to use PCR to amplify very long DNA molecules such as entire chromosomes, which in humans are about 10^8 bases in length. All the DNA encoding an organism is called the genome. For humans, the collection of chromosomes is the genome. For efficient language, we simply refer to amplifying

genomes in this paper. One problem is that the *Taq* polymerase usually used in PCR cannot extend the primers long enough to cover the whole genome. The length of the *Taq* polymerase extension is around 1000–2000 bases, far shorter than 10^8 bases. Another problem is that we must know the exact nucleotide sequences flanking the target we want to amplify so that we can use this information to design the primers used in PCR. In order to overcome these difficulties, a new method called *p*rimer-*e*xtension-*p*reamplification (PEP) has been proposed to amplify the whole genome (Zhang *et al.* 1992). The goal of this method is to amplify the whole genome simultaneously. The idea of this new technology is roughly the same as that of PCR, except that we use random primers instead of a fixed pair of primers. PEP can be contrasted with PCR in that the aim of the former is to amplify all DNA sequences in a sample whereas in the latter only one specific genomic sequence is the target. Thus PEP is a whole genome amplification method and it can be used to select those genomic sequences that bind specific proteins (Kinzler and Vogelstein 1989), to prepare DNA probes for FISH (Ludecke *et al.* 1989, Telenius *et al.* 1992), and to permit multiple PCR analysis on very small samples such as single cells (Zhang *et al.* 1991, Kristjansson *et al.* 1994) or molecules (Dear and Cook 1993). The mechanism and the realization of this technique is now described in more detail.

The whole genome to be amplified is sampled and random primers of a certain length are prepared. In the reported experiment, Zhang *et al.* (1991) chose all segments of 15 bases long as primers. Theoretically the random primers are composed of all the primers of 15 bases long with 4^{15} possibilities. The whole genome and the random primers together with the *Taq* polymerase are put into the test tube. The mixture is heated to near boiling so that the whole genome is separated completely into two strands. This is called *denaturing.* Then the temperature is lowered to a certain level so that random primers randomly anneal to the single-stranded DNA. This is called *annealing.* The third step is to raise the temperature again to the level that is optimum for the *Taq* polymerase to extend the primers that have annealed to the single-stranded DNA. This is called *polymerase extension.* For example we add the 4-base primer 5′-CTCA-3′ to the sequence below to obtain

```
5′-CGTATGACTGATGCTGAGTAGTCTGA-3′
                •••••
              3′-ACTC-5′.
```

Polymerase can extend this primer in the 3′ direction and we obtain

```
5′-CGTATGACTGATGCTGAGTAGTCTGA-3′
   •••••••••••••••••••
3′-GCATACTGACTACGACTC-5′.
```

Because we have a random number of primers annealing to the template, it may

happen that the extension product of one primer meets another primer or primer extension in front of it. In this case, the extension product replaces the sequence it meets leaving a nick in the sugar–phosphate backbone between them from the processivity of the *Taq* polymerase. Mathematically we can think of a nick as a gap between two sequences. For example two primers 5′-CTCA-3′ and 5′-AGAC-3′ anneal to the following sequence and are extended by 18 bases. The extension product of 5′-AGAC-3′ is 5′-AGACTACTCAGCATCAGT-3′ which replaces the primer 5-CTCA-3′ and its extension. Only 5′-CATACG-3′ is left for the extension of primer 5′-CTCA-3′ and there is a nick or gap (the '|' in the complementary strand) between the two extension products:

5′-CGTATG ACTGATGCTGAGTAGTCTGA-3′
• • • • • • • • • • • • • • • • • • • •
3′-GCATAC | TGACTACGACTCATCAGA-5′.

Denaturing, primer annealing and polymerase extension form an experimental *cycle*. In the second cycle of PEP, first we denature again. This time we get the original sequence and a random number of shorter sequences. In the annealing step, primers can anneal to the original sequence and to all the new shorter sequences. At the last step we use polymerase to extend the primers. This process is repeated for n cycles. Finally, we obtain a large number of short sequences from the original genome. Suppose there is a region of interest, and we want to know how many intact copies have been made of this specific region. We are also interested in what proportion of the genome has been amplified at least a certain number of times. Whole genome amplification has two goals. The first is to increase the total amount of DNA significantly (yield). The second is to insure that the amplification is not biased. Ideally all of the sequences of the original molecule should be amplified to the same extent.

In the rest of this section we construct a mathematical model for PEP. We only consider the single-stranded model. That is, we think of the genome as single-stranded. The results for expected numbers can be doubled if we consider the double-stranded case. Because the genome is usually very long, we model it as the real line. Primers are quite short compared to the genome so we model the primers as points. Primers anneal to the single-stranded sequence according to a Poisson process with parameter λ. For each annealed primer, the polymerase extends the primer by length L directed from 5′ to 3′. If the extension of one primer meets another primer or its extension, it replaces the original one leaving a nick or gap between them. Figure 2 shows the mechanism of PEP using these assumptions and notation.

Suppose we have a special region of interest called the *target*. This target can be modeled as a fixed interval on the real line. We use T to denote both the target and the target length and the meaning should be clear in context. The DNAs that contain the target T are called T-DNAs. Figure 3 shows the mechanism of

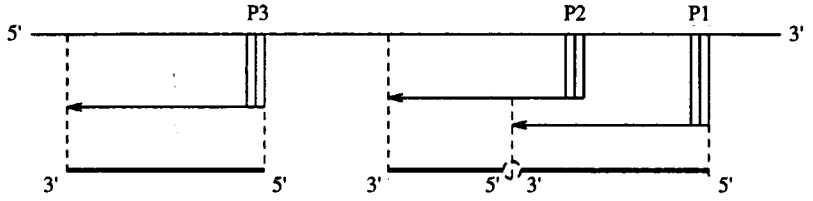

Figure 2. Mechanism of PEP. P1, P2, and P3 anneal and are extended by length L

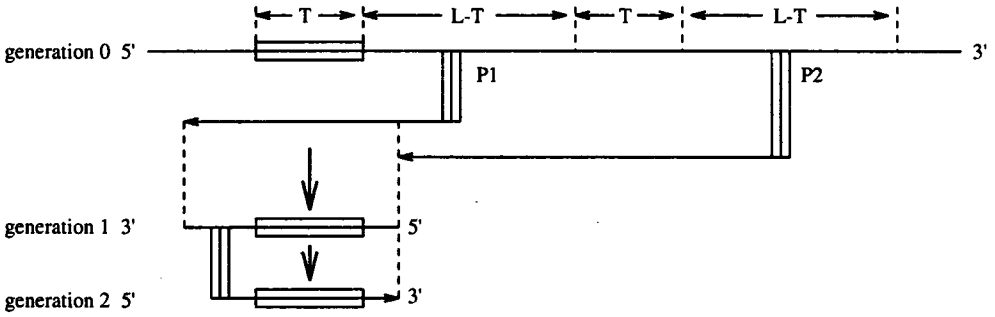

Figure 3. PEP with target length T and *Taq* product length L. Intact T-DNAs are generated

generating intact T-DNAs. We call the original sequence the 0th *generation* T-DNA. In order that we get a T-DNA from the original one there must be a primer (P1) in an interval of length $L - T$ at the 3′ end of the target so that its extension contains the target. There should be no primers in another interval of length T, otherwise their extension products replace the product of primer P1, destroying the target, and no T-DNAs are made. There might be primers in another interval of length $L - T$ since their products shorten the product of P1 but do not destroy the target. Under the above conditions a T-DNA is generated. We call it the *first generation* T-DNA. The first generation product as shown in Figure 3 can, in another cycle, generate another T-DNA called the *second generation* T-DNA. Inductively a kth *generation* T-DNA can generate a $(k+1)$th *generation* product as shown in Figure 4. Let Y_k^3 and Y_k^5 be the lengths of a kth generation T-DNA at the 3′ and 5′ ends beyond the target. Notice that $Y_k^5 = Y_{k+1}^3$ while usually $Y_k^3 \geqq Y_{k+1}^5$. This mechanism is a

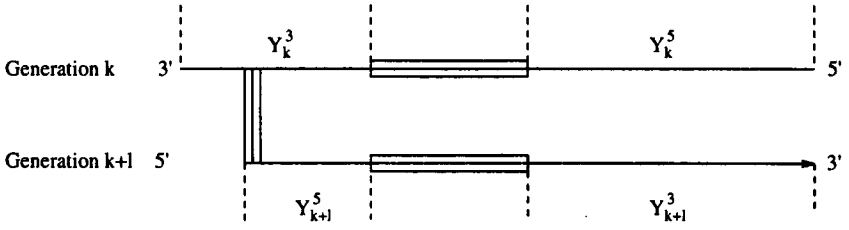

Figure 4. The formation of a $(k+1)$th generation T-DNA from a kth generation T-DNA

multitype branching process (Harris 1963). The type of a sequence is (Y^5, Y^3)—the lengths of the sequence at 5′ and 3′ ends beyond the target. In all, we have the following parameters:

T = length of the target;
L = length of the primer extension;
λ = annealing rate of the primers;
n = number of PEP cycles.

Let X_k^n and T_n be the number of kth generation T-DNAs and the total number of T-DNAs respectively. We are interested in the following problems.

(a) What is ET_n? What is the growth rate of ET_n with respect to n?
(b) What is the characteristic function and variance of X_k^n?
(c) Can we prove a central limit theorem and law of large numbers for X_k^n?

In this paper, we concentrate on the mathematical aspects of the analysis. The biological applications of our results are reported in a separate paper (Sun *et al.* 1995).

The organization of this paper is as follows. In Section 2 we study the distribution of (Y_k^5, Y_k^3), EX_k^n, ET_n, and their limiting behavior. In Section 3 we study the characteristic function and variance of X_k^n. We also prove that a central limit theorem and a strong law of large numbers hold for X_k^n. Section 4 extends the results to the case that L is random. Section 5 handles a related experimental approach known as tagged PCR (T-PCR).

2. Expected number of products

In this section we study the expected and limiting behavior of the number of T-DNAs after n PEP cycles. Recall that the whole genome is modeled as the real line, primers as points, and primers anneal to the genome according to a Poisson process with parameter λ. The target T is an interval on the real line. Any segments that contain the target T are called T-DNAs. The original genome is called 0th generation T-DNA and the T-DNAs directly generated from 0th generation T-DNA are called first generation T-DNAs and so on. (Y_k^5, Y_k^3) are the lengths of the kth generation T-DNAs at the 5′ and 3′ ends respectively beyond the target T. For different kth generation T-DNAs, they will have different values for (Y_k^5, Y_k^3). They have the same distribution and we use this generic notation. Let $X_k^n(l)$ and $T_n(l)$ be the number of kth generation T-DNAs and the total number of T-DNAs with length greater than $T + l$, $0 \leqq l \leqq L - T$ after n PEP cycles. Under the above notation we have the following results. (In the following, we denote $A(x) = \lambda(L - T - x)$. Theorem 1 gives the joint density of (Y_k^5, Y_k^3) and the marginal densities.)

Theorem 1. (*i*) *The joint density function of* (Y_k^5, Y_k^3) *is*

$$f_k(x, y) = \frac{\lambda^2 e^{-\lambda(T+x+y)}}{(k-2)!}\left(A^{k-2}e^{-2A} + \int_0^A z^{k-2}e^{-2z}\,dz\right), \qquad x > 0,\ y > 0,\ x + y < L - T,$$

where $k = 2, 3, 4, \cdots$, $A = A(x + y)$, *in the sense that, for any subset* $B \subset \{(x, y): x > 0, y > 0, x + y < L - T\}$,

$$P\{(Y_k^5, Y_k^3) \in B\} = \iint_B f_k(x, y)\, dx\, dy.$$

(ii) Y_k^5 *and* Y_k^3 *have the same marginal density function. If we denote their common marginal density function by* $f_k(x)$, *we have*

$$f_1(x) = \lambda e^{-\lambda(T+x)},$$

$$f_k(x) = \frac{\lambda e^{-\lambda(T+x)}}{(k-2)!} \int_0^{A(x)} z^{k-2} e^{-2z}\, dz,$$

where $0 < x < L - T$, $k = 2, 3, 4, \cdots$ *in the sense that*

$$P\{Y_k^5 > x\} = P\{Y_k^3 > x\} = \int_x^{L-T} f_k(s)\, ds.$$

This theorem asserts that statistically the 3′ and 5′ end lengths are indistinguishable; their distributions are the same. Theorem 2 gives an explicit formula for the probability that there exists a kth generation T-DNA of length at least $T + l$ at the kth cycle. This probability plays an important role in Theorem 3. Part (ii) also gives the expected length of a kth generation T-DNA.

Theorem 2. *(i) Let* $P_k(l)$ *be the probability that there exists a* kth *generation T-DNA at the* kth *cycle with length greater than* $T + l$. *Then*

$$P_1(l) = (1 + \lambda l)e^{-\lambda(T+l)} - e^{-\lambda L},$$

$$P_k(l) = \frac{e^{-\lambda(T+l)}}{(k-2)!} \int_0^{A(l)} z^{k-2} e^{-z}(e^{-z}(1 + \lambda l) - e^{-A(l)})\, dz, \qquad k \geqq 2.$$

(ii) Let Y_k *be the length of a* kth *generation T-DNA at* 3′ *or* 5′ *end beyond the target T. Then, given that the* kth *generation T-DNAs exist, the expectation of* Y_k *is*

$$E(Y_1) = \frac{e^{-\lambda L}}{P_1(0)\lambda}(e^{\lambda(L-T)} - 1 - \lambda(L - T)),$$

$$E(Y_k) = \frac{e^{-\lambda L}}{P_k(0)\lambda(k-2)!} \int_0^{A(0)} z^{k-2} e^{-z}(e^{A(0)-z} - 1 - A(0) + z)\, dz, \qquad k \geqq 2.$$

The following theorem gives an explicit formula for the expected number of kth generation T-DNAs after n PEP cycles and the limit behavior of the expected total

number of T-DNAs as n tends to infinity. It is important to note that $T_n(l)$ increases neither linearly nor exponentially with respect to n. The rate of increasing is approximately $\exp(2\sqrt{A(l)n})$.

Theorem 3. (*i*) *Let* $X_k^n(l)$ *be the total number of* kth *generation T-DNAs with length greater than* $T + l$. *Then*

$$EX_k^n(l) = \binom{n}{k} P_k(l).$$

(*ii*) (*Asymptotic result.*) *For fixed* n *and* l, *the maximum point* K_n *of* $EX_k^n(l)$ *with respect to* k *satisfies*

$$K_n \leqq [\sqrt{nA(l) + \tfrac{1}{4}(A(l)+1)^2} - \tfrac{1}{2} - \tfrac{1}{2}A(l)] + 1,$$

where $A(l) = \lambda(L - T - l)$.

Let $T_n(l)$ *be the total number of T-DNAs with length greater than* $T + l$ *after* n *cycles. Then*

$$\lim_{n\to\infty} \frac{\log T_n(l)}{\sqrt{n}} = 2\sqrt{A(l)}.$$

The annealing rate of primers certainly has a major effect on the final products. If the annealing rate is low, few primers anneal to the single-stranded genome and thus few PEP products are made. On the other hand if the annealing rate is too high, too many primers anneal to the genome and their extension products shorten each other. Many short strands are generated and few complete targets are made. The optimal annealing rate is important in the design of experiments. In the following we give some illustrative figures about the expected number of T-DNAs. In all these figures we use $L = 1000$, $T = 250$. Figure 5 shows the expected number of second generation T-DNAs after 20, 30, 40, and 50 cycles. Figures 6 and 7 show the total number of T-DNAs after 20 and 40 cycles respectively. From these figures we see the phenomena described above. Also we can find the optimal annealing rate under

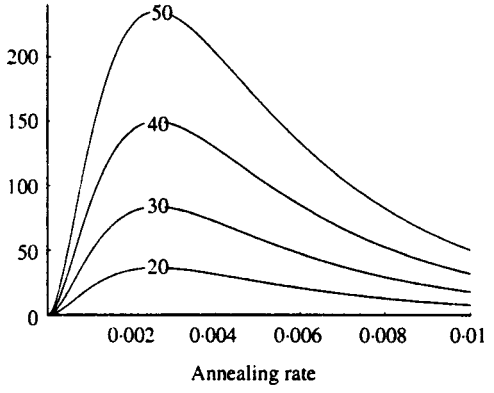

Figure 5. The expected number of second generation T-DNAs after 20, 30, 40, and 50 cycles

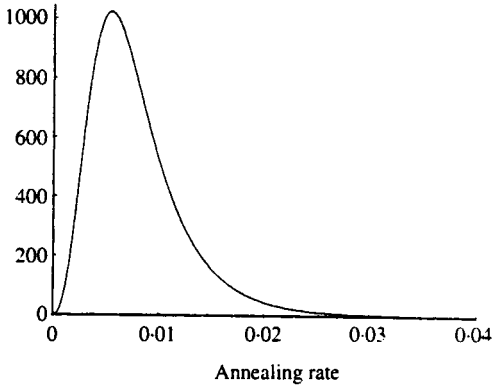

Figure 6. The total expected number of T-DNAs after 20 cycles

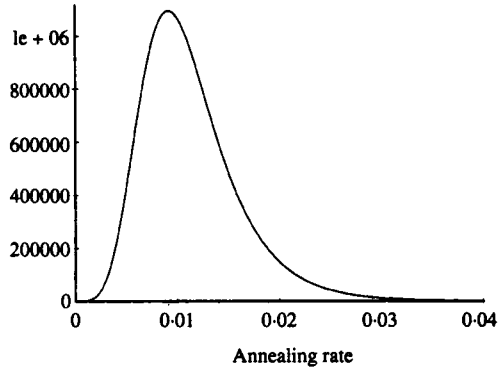

Figure 7. The total expected number of T-DNAs after 40 cycles

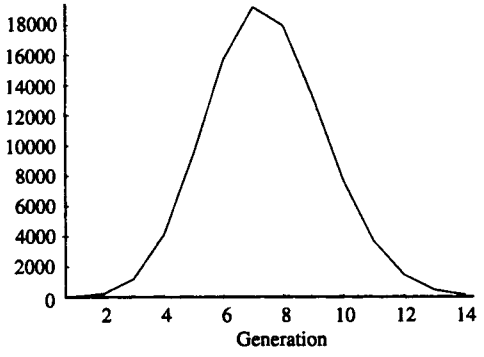

Figure 8. The expected number of kth generation T-DNAs as a function of generation after 50 cycles with $\lambda = 0.002$

the above conditions. Figure 8 shows the expected number of kth generation T-DNAs as a function of generation number k with $\lambda = 0.002$. From this figure we see EX_k^n first increases and then decreases with respect to k. The maximum is at $k = 7$. The upper bound given in Theorem 3(ii) is 8. The upper bound in the theorem gives an accurate estimate of the maximum generation number. A remarkable fact about PEP is that the expected number of target DNAs is neither polynomial nor exponential. The growth rate is about $\exp(e^2\sqrt{n\lambda(L-T)})$.

In order to prove the theorems we first prove a lemma that plays a crucial role in the following proofs.

Lemma 1. *Let Y_k^5 and Y_k^3 be the lengths of the* kth *generation T-DNA at the* 5′ *and* 3′ *ends beyond the target T. Then (Y_k^5, Y_k^3) has a density function $f_k(x, y)$ in the sense of Theorem* 1 *and $f_k(x, y)$ depends only on $x + y$ for any $k \geq 2$ and $x > 0$, $y > 0$, $0 < x + y < L - T$.*

Proof. We prove this lemma by induction.

(a) To prove this lemma, it is enough to study the marginal distributions of Y_1^5 and Y_1^3. For later use, we derive the joint distribution for (Y_1^5, Y_1^3) here. We refer to Figure 9. For $k = 1$, in order that we can get a first generation T-DNA with $\{Y_1^5 \geqq x$,

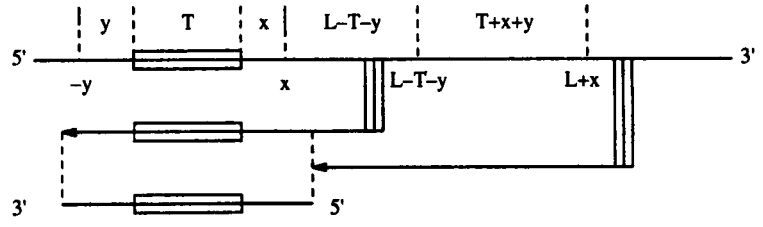

Figure 9. The mechanism by which a first generation T-DNA with $\{Y_1^5 \geqq x, Y_1^3 \geqq y\}$ is generated

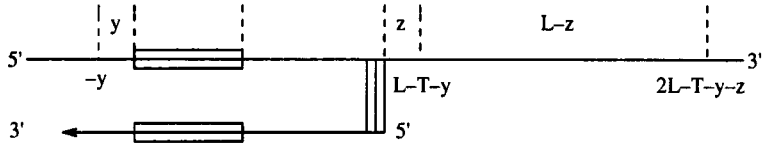

Figure 10. The mechanism by which a first generation T-DNA with $\{Y_1^5 + Y_1^3 = L - T,\ Y_1^3 \geqq y\}$ is generated

$Y_1^3 \geqq y\}$ where $x > 0$, $y > 0$, $x + y < L - T$, there must be a primer P1 in the interval $(x, L - T - y]$ so that its extension product covers both x and $-y$. There must be no primers in the interval $(L - T - y, L + x]$ otherwise their extension products shorten the product of P1 through x and no products with $\{Y_1^5 \geqq x, Y_1^3 \geqq y\}$ are made. Therefore

$$P\{Y_1^5 \geqq x,\ Y_1^3 \geqq y\} = P\{\geqq 1 \text{ primer in } (x, L - T - y) \text{ and}$$

$$0 \text{ primers in } (L - T - y, L + x)\}$$

$$= (1 - e^{-\lambda(L-T-y-x)})e^{-\lambda(T+x+y)}$$

$$= e^{-\lambda(T+x+y)} - e^{-\lambda L}, \qquad x + y < L - T.$$

Next we calculate $P\{Y_1^5 + Y_1^3 = L - T,\ Y_1^3 \geqq y\}$. Look at Figure 10. In order that we can get a first generation T-DNA with length L and $Y_1^3 \geqq y$, there must be a primer in $(0, L - T - y]$ and to the right of this primer of length L there are no primers so that the product of the primer is of length L. Let Z be the distance from $L - T - y$ to the first primer to the left. Then Z is exponentially distributed with parameter λ. Given $Z = z$, $\{Y_1^5 + Y_1^3 = L - T,\ Y_1^3 \geqq y\}$ happens if and only if there are no primers in $(L - T - y, 2L - T - y - z]$ with probability $e^{-\lambda(L-z)}$. Thus

$$P\{Y_1^5 + Y_1^3 = L - T,\ Y_1^3 \geqq y\} = E(P\{Y_1^5 + Y_1^3 = L - T,\ Y_1^3 \geqq y \mid Z\})$$

$$= \int_0^{L-T-y} \lambda e^{-\lambda z} e^{-\lambda(L-z)}\, dz$$

$$= \lambda e^{-\lambda L}(L - T - y), \qquad 0 < y < L - T.$$

Therefore,

$$P\{Y_1^5 \geqq x\} = e^{-\lambda(T+x)} - e^{-\lambda L}, \qquad 0 \leqq x \leqq L - T,$$

and

$$P\{Y_1^3 \geqq y\} = e^{-\lambda(T+y)} - e^{-\lambda L}, \qquad 0 \leqq y \leqq L - T.$$

We see that Y_1^5 and Y_1^3 have the same density function

$$f_1(x) = \lambda e^{-\lambda(T+x)}, \qquad 0 \leqq x \leqq L - T.$$

(b) Now we prove that the lemma is true for $k = 2$. Note that the first equality comes from the experimental mechanism.

$$\begin{aligned} P\{Y_2^5 \geqq x, Y_2^3 \geqq y\} &= P\{Y_2^5 \geqq x, Y_1^5 \geqq y\} = E(P(Y_2^5 \geqq x \mid Y_1^5 \geqq y, Y_1^3)) \\ &= E(1 - e^{-\lambda(Y_1^3 - x)})I_{\{Y_1^5 \geqq y, Y_1^3 \geqq x\}} \\ &= P\{Y_1^5 \geqq y, Y_1^3 \geqq x\} - \int_x^{L-T-y} e^{-\lambda(x'-x)} \lambda e^{-\lambda(T+x'+y)}\, dx' \\ &= e^{-\lambda(T+x+y)} - e^{-\lambda L} - \tfrac{1}{2} e^{-\lambda(T+y-x)}(e^{-2\lambda x} - e^{-2\lambda(L-T-y)}) \\ &= \tfrac{1}{2}(e^{-\lambda(T+x+y)} + e^{-\lambda(2L-T-x-y)}) - e^{-\lambda L}, \qquad x + y \leqq L - T. \end{aligned}$$

Therefore the density function of (Y_2^5, Y_2^3) is

$$(1) \qquad f_2(x, y) = \tfrac{1}{2}\lambda^2(e^{-\lambda(T+x+y)} + e^{-\lambda(2L-T-x-y)}), \qquad x + y \leqq L - T,$$

and the lemma is true for $k = 2$.

(c) Suppose the lemma is true for k, i.e. $f_k(x + c, y - c) = f_k(x, y)$ for any c such that $0 \leqq x + c \leqq L - T$, $0 \leqq y - c \leqq L - T$. Then by the mechanism of PEP (Figure 4) we have

$$(2) \qquad f_{k+1}(x, y) = \int_x^{L-T-y} \lambda e^{-\lambda(x'-x)} f_k(y, x')\, dx', \qquad x + y \leqq L - T.$$

Thus for any c as above we have

$$\begin{aligned} f_{k+1}(x + c, y - c) &= \int_{x+c}^{L-T-(y-c)} \lambda e^{-\lambda(x'-(x+c))} f_k(y - c, x')\, dx' \\ &= \int_x^{L-T-y} \lambda e^{-\lambda(z-x)} f_k(y - c, z + c)\, dz \\ &= \int_x^{L-T-y} \lambda e^{-\lambda(z-x)} f_k(y, z)\, dz \qquad \text{(by induction)} \\ &= f_{k+1}(x, y). \end{aligned}$$

Therefore $f_k(x, y)$ depends only on $x + y$ for any $k \geqq 2$.

Proof of Theorem 1. From Lemma 1 we can define $H_k(x+y) = f_k(x, y)$, where $x > 0$, $y > 0$, $x + y < L - T$. From (2) we have

$$\begin{aligned} H_{k+1}(x+y) &= f_{k+1}(x, y) \\ &= \int_x^{L-T-y} \lambda e^{-\lambda(x'-x)} f_k(y, x')\, dx' \\ &= \int_{x+y}^{L-T} \lambda e^{\lambda(x+y-z)} f_k(y, z-y)\, dz \\ &= \lambda e^{\lambda(x+y)} \int_{x+y}^{L-T} e^{-\lambda z} H_k(z)\, dz. \end{aligned}$$

Let $g_k(r) = e^{-\lambda r} H_k(r)$. Then

$$\begin{aligned} g_{k+1}(r) &= \lambda \int_r^{L-T} g_k(s)\, ds \\ &= \lambda^2 \int_r^{L-T} \int_s^{L-T} g_{k-1}(t)\, dt\, ds \\ &= \lambda^2 \int_r^{L-T} (t-r) g_{k-1}(t)\, dt \\ &\vdots \qquad\qquad \vdots \\ &= \frac{\lambda^{k-1}}{(k-2)!} \int_r^{L-T} (t-r)^{k-2} g_2(t)\, dt \\ &= \frac{\lambda^{k+1}}{2(k-2)!} \left(\int_r^{L-T} (t-r)^{k-2} e^{-2\lambda t}\, dt - e^{-\lambda(2L-T)} \frac{(L-T-r)^{k-1}}{k-1} \right) \\ &= \frac{\lambda^2 e^{-\lambda(T+2r)}}{(k-1)!} \left(A(r)^{k-1} e^{-2A(r)} + \int_0^{A(r)} t^{k-1} e^{-2t}\, dt \right), \end{aligned}$$

and

$$\begin{aligned} f_{k+1}(x, y) &= e^{\lambda(x+y)} g_{k+1}(x+y) \\ &= \frac{\lambda^2 e^{-\lambda(T+x+y)}}{(k-1)!} \left(A^{k-1} e^{-2A} + \int_0^A z^{k-1} e^{-2z}\, dz \right). \end{aligned}$$

Now (i) of Theorem 1 is proved.

From $f_k(x) = \int_0^{L-T-x} f_k(x, y)\, dy$, we can easily prove (ii) of Theorem 1 by careful calculations, and Theorem 1 is proved.

Proof of Theorem 2. First we calculate

$$Q_k(l) = \iint_{0<x+y<l} f_k(x, y)\, dx\, dy.$$

According to Theorem 1 we know

$$\begin{aligned}\int_0^{l-x} f_k(x, y)\, dy &= \frac{\lambda e^{-\lambda(T+x)}}{(k-2)!}\left(\int_0^{l-x} \lambda e^{-\lambda y-2A} A^{k-2}\, dy - \int_0^{l-x}\int_0^{A} z^{k-2}e^{-2z}\, dz\, de^{-\lambda y}\right)\\ &= \frac{\lambda e^{-\lambda(T+x)}}{(k-2)!}\left(\int_0^{A(x)} z^{k-2}e^{-2z}\, dz - e^{-\lambda(l-x)}\int_0^{A(l)} z^{k-2}e^{-2z}\, dz\right)\\ &= f_k(x) - f_k(l).\end{aligned}$$

Clearly

$$Q_k(l) = \int_0^l f_k(x)\, dx - l f_k(l),$$

and

$$\begin{aligned}P_k(l) &= \int_0^{L-T} f_k(x)\, dx - Q_k(l)\\ &= \int_l^{L-T} f_k(x)\, dx + l f_k(l).\end{aligned}$$

From Theorem 1 it is easy to see that

$$\begin{aligned}\int_l^{L-T} f_k(x)\, dx &= \frac{\lambda^k e^{-\lambda T}}{(k-2)!}\int_l^{L-T} e^{-\lambda x}\, dx \int_0^{L-T-x} y^{k-2}e^{-2\lambda y}\, dy\\ &= \frac{\lambda^k e^{-\lambda T}}{(k-2)!}\int_0^{L-T-l} y^{k-2}e^{-2\lambda y}\, dy \int_l^{L-T-y} e^{-\lambda x}\, dx\\ &= \frac{\lambda^{k-1} e^{-\lambda T}}{(k-2)!}\int_0^{L-T-l} y^{k-2}e^{-2\lambda y}(e^{-\lambda l} - e^{-A(y)})\, dy\\ &= \frac{e^{-\lambda(T+l)}}{(k-2)!}\int_0^{A(l)} z^{k-2}e^{-z}(e^{-z} - e^{-A(l)})\, dz.\end{aligned}$$

From the above equations and Theorem 1 it is easy to see that (i) of Theorem 2 holds. By almost the same method we can prove (ii) of Theorem 2.

Proof of Theorem 3. We prove this theorem by induction on k.

(a) $k = 1$. At each cycle, the original DNA strand generates a first generation

T-DNA of length greater than $T+l$ with probability $P_1(l)$. Let I_i, $i=1,2,\cdots,n$, be the number of first generation T–DNAs of length greater than $T+l$ generated by the original DNA strand. Then $X_1^n = I_1 + I_2 + \cdots + I_n$, where $I_1, I_2, \cdots, I_n$ are i.i.d. with distribution $P\{I_1=1\} = P_1(l)$, and $P\{I_1=0\} = 1 - P_1(l)$. Therefore

$$EX_1^n = EI_1 + EI_2 + \cdots + EI_n = nP_1(l).$$

(b) Suppose the assertion is true for $k-1$ and for any n. Then for k, we have $X_k^n = X_k^{n-1} + Z_{k-1}^{n-1}$, where X_k^{n-1} is the number of kth generation T–DNAs with length greater than $T+l$ generated by the original DNA after $n-1$ cycles, and Z_{k-1}^{n-1} is defined by

$$Z_{k-1}^{n-1} = \begin{cases} 0, & \text{if } \alpha_0 \text{ does not generate } \alpha_1, \\ \text{number of } k\text{th generation T-DNAs generated by } \alpha_1, \end{cases}$$

where α_0 denotes the original DNA and α_1 denotes the first generation T-DNA generated directly from α_0 after the first cycle. Then

$$\begin{aligned} EZ_{k-1}^{n-1} &= E(E(Z_{k-1}^{n-1} \mid Y_1^3, Y_1^5)) \\ &= \binom{n-1}{k-1} EP_{k-1}(Y_1^3, Y_1^5)(l) = \binom{n-1}{k-1} P_k(l), \end{aligned}$$

where $P_k(x,y)(l)$ denotes the probability that there exists a kth generation T-DNA with length greater than $T+l$ at the kth cycle when $Y_0^5 = x$, $Y_0^3 = y$. Finally

$$\begin{aligned} EX_k^n &= EX_k^{n-1} + \binom{n-1}{k-1} P_k(l) \\ &= \left(\binom{k-1}{k-1} + \binom{k+1}{k-1} + \cdots + \binom{n-1}{k-1}\right) P_k(l) \\ &= \binom{n}{k} P_k(l). \end{aligned}$$

By induction, (i) of Theorem 3 is true.

From Theorem 1 (i) and integration by parts it is not difficult to verify that

$$P_k(l) = \mathrm{e}^{-\lambda(2L-T-l)}\left(\sum_{i=0}^{\infty} \frac{(2^i(1+\lambda l)-1)A^{k+i-1}(l)}{(k+i-1)!}\right). \tag{3}$$

It is easy to prove that

$$\frac{\binom{n}{k+1}A^k(l)}{k!} \Big/ \frac{\binom{n}{k}A^{k-1}(l)}{(k-1)!} = \frac{(n-k)A(l)}{k(k+1)} \leqq 1,$$

iff

$$k \geqq [\sqrt{nA(l) + \tfrac{1}{4}(A(l)+1)^2} - \tfrac{1}{2}(1+A(l))] + 1 \stackrel{\text{def}}{=} M_n, \tag{4}$$

where '$\stackrel{\text{def}}{=}$' means 'define'. From (3) for any $k \geqq M_n$, we have $\binom{n}{k+1}P_{k+1}(l) \leqq \binom{n}{k}P_k(l)$. Thus the maximum point does not exceed M_n. From (3) it is easy to see

that there exist constants c and C which do not depend on k such that

$$c\frac{A^{k-1}(l)}{(k-1)!} \leqq P_k(l) \leqq C\frac{A^{k-1}(l)}{(k-1)!}.$$

By $T_n(l) = \sum_{k=1}^{n} \binom{n}{k} P_k(l)$ we have

$$c\frac{\binom{n}{M_n}A^{M_n-1}(l)}{(M_n-1)!} \leqq T_n(l) \leqq Cn\frac{\binom{n}{M_n}A^{M_n-1}(l)}{(M_n-1)!}.$$

From Stirling's formula and (4) we get

$$\frac{\binom{n}{M_n}A^{M_n-1}(l)}{(M_n-1)!} \sim C' n^{\frac{1}{2}} e^{2\sqrt{A(l)n}}.$$

Therefore

$$\lim_{n\to\infty} \frac{\log T_n(l)}{\sqrt{n}} = 2\sqrt{A(l)},$$

and (ii) of Theorem 3 is proved.

3. The characteristic function, variance and limit theorems

In Section 2 we studied the expectation of X_k^n and T_n. This study gives the average number of kth generation sequences in a large number of experiments. It does not tell us the distribution of X_k^n and T_n. The distribution of X_k^n and T_n is important in answering questions such as the probability that the target is replicated at least M_k times by kth generation T-DNAs and so on. Because of the complicated mechanism of PEP, it is hard to get the explicit distribution of X_k^n and T_n. In this section, we study the characteristic function of X_k^n. Of course the characteristic function gives us all the information about the distribution. By using the characteristic function, we obtain a recursive formula for calculating the variance of X_k^n and prove that a central limit theorem and strong law of large numbers hold for X_k^n. The limit theorems can be used to get an approximate probability that the target is replicated at least M_k times by kth generation T-DNAs. Before we state the theorems, we need some notation. Let $X_k^n(x, y)$ denote the number of kth generation T-DNAs after n PEP cycles when the original DNA satisfies $Y_0^5 = x$, $Y_0^3 = y$, $P_k(x, y)$ be the probability that we get a kth generation T-DNA after k cycles when $Y_0^5 = x$, $Y_0^3 = y$, and $g_k^n(t; x, y)$ be the characteristic function of $X_k^n(x, y)$. Throughout this section, we assume that $x + y < L - T$. Then we have the following theorems.

Theorem 4. (*i*) *The characteristic function* $g_k^n(t; x, y)$ *of* $X_k^n(x, y)$ *satisfies the following recursive equation*:

$$g_1^n(t; x, y) = (e^{-\lambda y} + (1 - e^{-\lambda y})e^{it})^n,$$

$$g_k^n(t; x, y) = \prod_{j=k}^{n} \left(e^{-\lambda y} + \int_0^y g_{k-1}^{j-1}(t; x_1, x)\lambda e^{-\lambda(y-x_1)}\, dx_1 \right), \qquad n, k = 2, 3, \cdots.$$

(*ii*) *The variance of* $X_k^n(x, y)$ *satisfies the following recursive equation:*

$$\mathrm{Var}\,(X_1^n(x, y)) = ne^{-\lambda y}(1 - e^{-\lambda y}),$$

$$\mathrm{Var}\,(X_k^n(x, y)) = \sum_{j=k}^{n} \int_0^y \mathrm{Var}\,(X_{k-1}^{j-1}(x_1, x))\lambda e^{-\lambda(y-x_1)}\, dx_1$$

$$+ \sum_{j=k}^{n} \binom{j-1}{k-1}^2 \left(\int_0^y P_{k-1}^2(x_1, x)\lambda e^{-\lambda(y-x_1)}\, dx_1 - P_k^2(x, y) \right),$$

$$n, k = 2, 3, \cdots$$

where

$$P_{2k}(x, y) = \Gamma_k(\lambda x)\Gamma_k(\lambda y), \qquad P_{2k+1}(x, y) = \Gamma_k(\lambda x)\Gamma_{k+1}(\lambda y),$$

and

$$\Gamma_k(x) = \frac{1}{(k-1)!} \int_0^x s^{k-1} e^{-s}\, dx, \qquad k = 1, 2, \cdots.$$

Theorem 5. (Central limit theorem.) *Under the conditions of Theorem* 4, *for any fixed* $k \geqq 1$,

$$Y_k^n(x, y) = \frac{X_k^n(x, y) - \binom{n}{k} P_k(x, y)}{\sqrt{\mathrm{Var}\,(X_k^n(x, y))}}$$

is asymptotically normal as $n \to \infty$.

Theorem 6. (Strong law of large numbers.) *Under the conditions of Theorem* 4, *for any fixed* $k \geqq 1$,

$$\lim_{n\to\infty} \frac{X_k^n(x, y)}{\binom{n}{k} P_k(x, y)} = 1$$

almost surely.

Because of the dependence among the PEP products, it is hard to get the exact distribution of X_k^n. From the recursive formula in Theorem 4 we can get the variance of X_k^n. Then from the central limit theorem we can get the approximate distribution of X_k^n.

Proof of Theorem 4. Suppose now the original DNA satisfies the condition

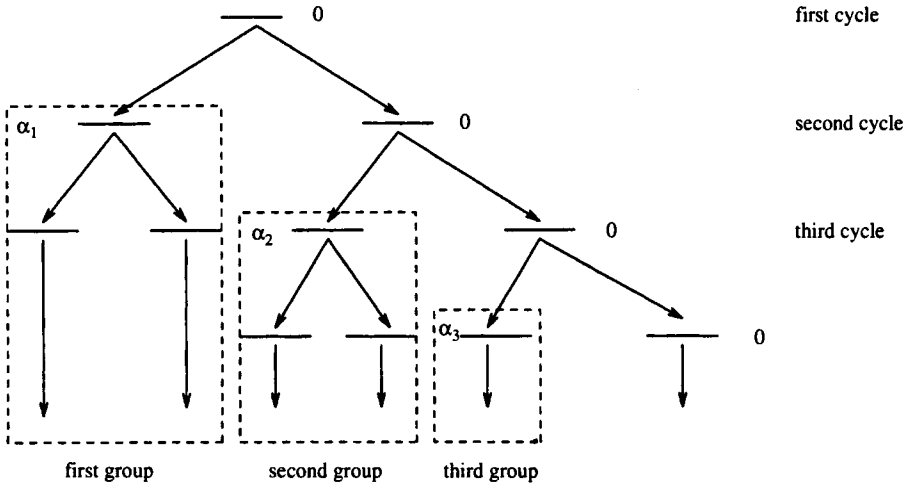

Figure 11. The final PEP products are divided into n groups

$Y_0^5 = x$, $Y_0^3 = y$. The final products of T-DNAs after n PEP cycles can be divided into n groups (Figure 11). The jth group is composed of all the T-DNAs generated from α_j where α_j is the first generation T-DNA generated in the jth cycle, $1 \leqq j \leqq n$. If α_j does not exist, the jth group is empty.

Let $N_j(k)$ be the number of kth generation T-DNAs in the jth group. Then we have

$$X_k^n(x, y) = N_1(k) + N_2(k) + \cdots + N_{n-k+1}(k)$$

and $N_1(k)$, $N_2(k)$, $\cdots$, $N_{n-k+1}(k)$ are independent. The characteristic function of $N_j(k)$ can be calculated in the following way. If α_j does not exist, then $N_j(k) = 0$. Conditioning on the lengths of both sides of α_j and noting that given $(Y_0^5, Y_0^3) = (x, y)$, then $Y_1^3 = x$ and Y_1^5 has density function $\lambda e^{-\lambda(y-x_1)}$. Thus

$$\begin{aligned} Ee^{itN_j(k)} &= E(E(e^{itN_j(k)} \mid Y_1^5 = x_1, Y_1^3 = x)) \\ &= E(e^{-\lambda y} + Ee^{itX_{k-1}^{n-j}(x_1, x)}) \\ &= e^{-\lambda y} + \int_0^y g_{k-1}^{n-j}(t; x_1, x) e^{-\lambda(y-x_1)}\, dx_1. \end{aligned}$$

By independence, we see that the characteristic function of $X_k^n(x, y)$ is

$$\begin{aligned} &\prod_{j=1}^{n-k+1} \left(e^{-\lambda y} + \int_0^y g_{k-1}^{n-j}(t; x_1, x)\lambda e^{-\lambda(y-x_1)}\, dx_1\right) \\ &= \prod_{j=k}^{n} \left(e^{-\lambda y} + \int_0^y g_{k-1}^{j-1}(t; x_1, x)\lambda e^{-\lambda(y-x_1)}\, dx_1\right), \end{aligned} \tag{5}$$

and (i) is proved.

Differentiating (5) twice, letting $t = 0$ and noting that

$$(g_{k-1}^{j-1}(t; x, y))'_t\big|_{t=0} = iEX_{k-1}^{j-1}(x, y) = i\binom{j-1}{k-1}P_{k-1}(x, y),$$

we get

$$E(X_k^n(x, y))^2 = \sum_{j=k}^{n}\binom{j-1}{k-1}\Bigg(\sum_{\substack{l=k\\ l\neq j}}^{n}\binom{l-1}{k-1}\Bigg)P_k^2(x, y) + \sum_{j=k}^{n}\int_0^y E(X_{k-1}^{j-1}(x_1, x))^2\lambda e^{-\lambda(y-x_1)}\,dx_1$$

$$= \binom{n}{k}^2 P_k^2(x, y) - \sum_{j=k}^{n}\binom{j-1}{k-1}^2 P_k^2(x, y)$$

$$+ \sum_{j=k}^{n}\int_0^y \binom{j-1}{k-1}^2 P_{k-1}^2(x_1, x)\lambda e^{-\lambda(y-x_1)}\,dx_1.$$

Thus

$$\operatorname{Var}(X_k^n(x, y)) = \sum_{j=k}^{n}\int_0^y \operatorname{Var}(X_{k-1}^{j-1}(x_1, x))\lambda e^{-\lambda(y-x_1)}\,dx_1$$

$$+ \sum_{j=k}^{n}\binom{j-1}{k-1}^2\Bigg(\int_0^y P_{k-1}^2(x_1, x)\lambda e^{-\lambda(y-x_1)}\,dx_1 - P_k^2(x, y)\Bigg).$$

In order to derive $P_{2k}(x, y)$ we note that, by the mechanism of PEP, the 5′ and 3′ ends can be thought of as being shortened independently k times according to a Poisson process with parameter λ. The density $h_k(s, y)$ of the 3′ end after the kth shortening satisfies the following recursive equation:

$$h_k(s, y) = \int_s^y \lambda e^{-\lambda(t-s)}h_{k-1}(t, y)\,dt, \qquad k = 2, 3, \cdots, \tag{6}$$

$$h_1(s, y) = \lambda e^{-\lambda(y-s)}, \qquad 0 < s < y.$$

By induction on (6), we can prove

$$h_k(s, y) = \frac{\lambda^k}{(k-1)!}(y-s)^{k-1}e^{-\lambda(y-s)}, \qquad 0 < s < y.$$

Thus

$$P_{2k}(x, y) = \int_0^x h_k(s, x)\,ds\int_0^y h_k(s, y)\,ds = \Gamma_k(\lambda x)\Gamma_k(\lambda y).$$

In exactly the same manner, we can prove

$$P_{2k+1}(x, y) = \Gamma_k(\lambda x)\Gamma_{k+1}(\lambda y), \qquad k = 1, 2, \cdots.$$

Note. Once we have Theorem 4, we can obtain the characteristic function and variance of X_k^n by the following formulas:

$$g_k^n(t) = \prod_{j=k}^{n} \left(1 - P_1 + \iint_{(x_1,y_1)} g_{k-1}^{j-1}(t; x_1, y_1)\, dF(x_1, y_1)\right),$$

$$\begin{aligned}\operatorname{Var}(X_k^n) = &\sum_{j=k}^{n} \iint_{(x_1,y_1)} \operatorname{Var}(X_{k-1}^{j-1}(x_1, y_1))\, dF(x_1, y_1) \\ &+ \sum_{j=k}^{n} \binom{j-1}{k-1}^2 \left(\iint_{(x_1,y_1)} P_{k-1}^2(x_1, y_1)\, dF(x_1, y_1) - P_k^2 \right),\end{aligned}$$

where $F(x_1, y_1)$ is the distribution of (Y_1^5, Y_1^3) and $P_k = P_k(0)$ as we obtain in Section 2. The proof of the above formulas is the same as that of the above theorem and we omit the proof here.

In order to prove Theorems 5 and 6, we first prove two lemmas. Lemma 2 gives the growth rate of $\operatorname{Var}(X_k^n)$ as n tends to infinity for any fixed k. In the following we denote $a_n = \Theta(b_n)$ if

$$0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leqq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty.$$

Lemma 2. *If* $0 < x + y < L - T$, *then for any fixed* $k \geqq 1$,

$$\operatorname{Var}(X_k^n(x, y)) = \Theta(n^{2k-1}).$$

Proof. We prove by induction. (i) For $k = 1$, the result is obviously true. (ii) Suppose the result is true for k, then from Theorem 4 we have

$$\operatorname{Var}(X_{k+1}^n) = \Theta\left(\sum_{j=k+1}^{n} (j-1)^{2k-1}\right) + \Theta\left(\sum_{j=k+1}^{n} \binom{j-1}{k}^2\right) = \Theta(n^{2k+1}).$$

By induction the lemma is true.

Lemma 3. *For any*

$$s \leqq \frac{t}{\sqrt{\operatorname{Var}(X_k^n(x, y))}} = \Theta\left(\frac{t}{n^{k-\frac{1}{2}}}\right)$$

we have

$$g_k^n(s; x, y) = \exp\left(\mathrm{i}EX_k^n(x, y)s - \tfrac{1}{2}s^2 \operatorname{Var}(X_k^n(x, y)) + n^{2k-1}o(s^2)\right).$$

Proof. Again the proof is by induction. (i) For $k = 1$,

$$\begin{aligned} g_1^n(s, x, y) &= (Q_1(x, y) + P_1(x, y)\mathrm{e}^{\mathrm{i}s})^n \\ &= \exp\left(n \log\left(Q_1(x, y) + P_1(x, y)\left(1 + \mathrm{i}t - \tfrac{1}{2}s^2 + o(s^2)\right)\right)\right) \\ &= \exp\left(n\left(\mathrm{i}P_1(x, y) - \tfrac{1}{2}s^2 P_1(x, y) - \tfrac{1}{2}P_1^2(x, y)s^2 + o(s^2)\right)\right) \\ &= \exp\left(\mathrm{i}EX_1^n(x, y) - \tfrac{1}{2}s^2 \operatorname{Var}(X_1^n(x, y)) + no(s^2)\right), \end{aligned}$$

where $Q_1(x, y) = 1 - P_1(x, y)$. Thus the lemma is true for $k = 1$.

(ii) Suppose the lemma is true for k, then for

$$s \leqq \frac{t}{\sqrt{\operatorname{Var}(X_{k+1}^n(x, y))}} = \Theta\left(\frac{1}{n^k\sqrt{n}}\right)$$

it is obvious that when n is sufficiently large,

$$s \leqq \frac{t}{\sqrt{\operatorname{Var}(X_k^j(x, y))}}, \qquad j = k+1, \cdots, n.$$

Therefore by the assumption and Lemma 2 we have

$$\begin{aligned}
g_{k+1}^n(s; x, y) &= \prod_{j=k+1}^{n} \Big(\exp(-\lambda y) + \int_0^y \exp\Big(\mathrm{i}\binom{j}{k}P_k(x_1, x)s - \tfrac{1}{2}s^2 \operatorname{Var}(X_k^{j-1}(x_1, x)) \\
&\quad + j^{2k-1}o(s^2)\Big)\lambda \exp(-\lambda(y - x_1))\, dx_1\Big) \\
&= \prod_{j=k}^{n-1} \Big(\exp(-\lambda y) + \int_0^y \Big(1 + \mathrm{i}\binom{j}{k}P_k(x_1, x)s - \tfrac{1}{2}s^2 \operatorname{Var}(X_k^j(x_1, x)) \\
&\quad - \tfrac{1}{2}\binom{j}{k}^2 P_k(x_1, x)s^2 + j^{2k}o(s^2)\Big)\lambda \exp(-\lambda(y - x_1))\, dx_1\Big) \\
&= \prod_{j=k}^{n-1} \Big(1 + \mathrm{i}\binom{j}{k}P_{k+1}(x, y) - \tfrac{1}{2}s^2\int_0^y \operatorname{Var}(X_k^j(x_1, x))\lambda \exp(-\lambda(y - x_1))\, dx_1 \\
&\quad - \tfrac{1}{2}s^2\binom{j}{k}^2\int_0^y P_k^2(x_1, x)\lambda \exp(-\lambda(y - x_1))\, dx_1 + j^{2k}o(s^2)\Big) \\
&= \prod_{j=k}^{n-1} \exp\Big(\mathrm{i}\binom{j}{k}P_{k+1}(x, y)s - \tfrac{1}{2}s^2\int_0^y \operatorname{Var}(X_k^j(x_1, x))\lambda \exp(-\lambda(y - x_1))\, dx_1 \\
&\quad - \tfrac{1}{2}s^2\binom{j}{k}^2\Big(\int_0^y P_k^2(x_1, x)\lambda \exp(-\lambda(y - x_1))\, dx_1 - P_{k+1}^2(x, y)\Big) + j^{2k}o(s^2)\Big) \\
&= \exp(\mathrm{i}EX_k^n(x, y)s - \tfrac{1}{2}s^2 \operatorname{Var}(X_k^n(x, y)) + n^{2k+1}o(s^2)).
\end{aligned}$$

By induction the lemma is true.

Proof of the central limit theorem. (*Theorem* 5.) Next we use the above lemmas to prove the central limit theorem. Let

$$Y_k^n(x, y) = \frac{X_k^n(x, y) - \binom{n}{k}P_k(x, y)}{\sqrt{\operatorname{Var}(X_k^n(x, y))}}.$$

Then

$$\begin{aligned} f_{Y_k^n(x,y)}(t) &= E\exp(itY_k^n(x,y)) \\ &= \exp\left(-it\frac{\binom{n}{k}P_k(x,y)}{\sqrt{\operatorname{Var}(X_k^n(x,y))}}\right)E\exp\left(it\frac{X_k^n(x,y)}{\sqrt{\operatorname{Var}(X_k^n(x,y))}}\right) \\ &= \exp\left(-it\frac{\binom{n}{k}P_k(x,y)}{\sqrt{\operatorname{Var}(X_k^n(x,y))}}\right)f_{X_k^n(x,y)}\left(\frac{t}{\sqrt{\operatorname{Var}(X_k^n(x,y))}}\right) \\ &= \exp\left(-\tfrac{1}{2}t^2 + n^{2k-1}o\left(\frac{t^2}{\operatorname{Var}(X_k^n(x,y))}\right)\right) \qquad \text{(Lemma 2)} \\ &= \exp(-\tfrac{1}{2}t^2 + o(1)). \end{aligned}$$

Therefore

$$\lim_{n\to\infty} f_{Y_k^n(x,y)}(t) = \exp(-\tfrac{1}{2}t^2)$$

and Theorem 5 is proved.

Proof of the strong law of large numbers. (*Theorem* 6.)

$$P\left\{\left|\frac{X_k^n(x,y)}{\binom{n}{k}P_k(x,y)} - 1\right| > \varepsilon\right\} = P\{|X_k^n(x,y) - \binom{n}{k}P_k(x,y)| > \varepsilon\binom{n}{k}P_k(x,y)\}$$

$$\leqq \frac{\operatorname{Var}(X_k^n(x,y))}{\varepsilon^2\binom{n}{k}^2P_k^2(x,y)}.$$

By Lemma 2, $\operatorname{Var}(X_k^n(x,y)) = \Theta(n^{2k-1})$. Therefore

$$\sum_{i=1}^{\infty}\frac{\operatorname{Var}(X_k^{i^2}(x,y))}{\binom{i^2}{k}^2P_k^2(x,y)} < \infty.$$

By the Borel–Cantelli lemma,

$$\lim_{i\to\infty}\frac{X_k^{i^2}(x,y)}{\binom{i^2}{k}P_k(x,y)} = 1, \tag{7}$$

almost surely. Because $X_k^n(x,y)$ is increasing with respect to n, we have

$$X_k^{[\sqrt{n}]^2}(x,y) \leqq X_k^n(x,y) \leqq X_k^{([\sqrt{n}]+1)^2}(x,y).$$

Thus

$$\frac{\binom{[\sqrt{n}]^2}{k}}{\binom{n}{k}}\frac{X^{[\sqrt{n}]^2}(x,y)}{\binom{[\sqrt{n}]^2}{k}} \leqq \frac{X_k^n(x,y)}{\binom{n}{k}} \leqq \frac{\binom{([\sqrt{n}]+1)^2}{k}}{\binom{n}{k}}\frac{X^{([\sqrt{n}]+1)^2}(x,y)}{\binom{([\sqrt{n}]+1)^2}{k}}.$$

Letting $n\to\infty$, using (7) and noting

$$\lim_{n\to\infty}\frac{\binom{[\sqrt{n}]^2}{k}}{\binom{n}{k}} = \lim_{n\to\infty}\frac{\binom{([\sqrt{n}]+1)^2}{k}}{\binom{n}{k}} = 1,$$

we have

$$\lim_{n\to\infty} \frac{X_k^n(x, y)}{\binom{n}{k}} = P_k(x, y),$$

almost surely and Theorem 6 is proved.

Next let us consider the total number of products of the first k generations. We have the following result.

Theorem 7. Let $S_k^n(x, y)$ be the total number of products of the first k generations after n PEP cycles when $Y_0^5 = x$, $Y_0^3 = y$, i.e $S_k^n(x, y) = \sum_{i=0}^k X_i^n(x, y)$. Then

$$\lim_{n\to\infty} \frac{S_k^n(x, y)}{EX_k^n(x, y)} = 1,$$

almost surely and

$$\frac{S_k^n(x, y) - ES_k^n(x, y)}{\sqrt{\text{Var}\,(X_k^n(x, y))}}$$

is asymptotically normal.

Proof. (i) By Theorem 6 for any $0 \leqq i < k$

$$\lim_{n\to\infty} \frac{X_i^n(x, y)}{EX_k^n(x, y)} = \lim_{n\to\infty} \frac{X_i^n(x, y)}{EX_i^n(x, y)} \lim_{n\to\infty} \frac{EX_i^n(x, y)}{EX_k^n(x, y)}$$

$$= \lim_{n\to\infty} \frac{\binom{n}{i} P_i(x, y)}{\binom{n}{k} P_k(x, y)} = 0.$$

Thus

$$\lim_{n\to\infty} \frac{S_k^n(x, y)}{EX_k^n(x, y)} = \lim_{n\to\infty} \frac{X_k^n(x, y)}{EX_k^n(x, y)} = 1,$$

almost surely.

(ii) By Lemma 2, for any $0 \leqq i < k$

$$\lim_{n\to\infty} \frac{EX_i^n(x, y)}{\sqrt{\text{Var}\,(X_k^n(x, y))}} = 0.$$

Thus

$$\lim_{n\to\infty} \frac{X_i^n(x, y) - EX_i^n(x, y)}{\text{Var}\,(X_k^n(x, y))} = \lim_{n\to\infty} \frac{X_i^n(x, y) - EX_i^n(x, y)}{E(X_i^n(x, y))} \lim_{n\to\infty} \frac{EX_i^n(x, y)}{\sqrt{\text{Var}\,(X_k^n(x, y))}} = 0.$$

By Slutsky's theorem (Durrett 1991) and Theorem 5 we have that

$$\frac{S_k^n(x, y) - ES_k^n(x, y)}{\sqrt{\text{Var}\,(X_k^n(x, y))}}$$

is asymptotically normal.

Note. Theorems 5, 6 and 7 still hold if we replace $X_k^n(x, y)$ by $X_k^n(l)$, the number of kth generation T-DNAs with length at least $T + l$. The proofs are almost the same but are more involved. We omit the proofs here.

Once we have the above theorems we can approximate the probability that a target of length T is covered by at least M_k kth generation T-DNAs in the following way. From Theorem 3 we can get EX_k^n. By the inductive formula in Theorem 4 we can calculate Var (X_k^n). Then from Theorem 5 we have

$$P\{X_k^n \geqq M_k\} = P\left\{\frac{X_k^n - EX_k^n}{\sqrt{\text{Var}\,(X_k^n)}} \geqq \frac{M_k - \binom{n}{k}P_k}{\sqrt{\text{Var}\,(X_k^n)}}\right\}$$

$$\approx 1 - \phi\left(\frac{M_k - \binom{n}{k}P_k}{\sqrt{\text{Var}\,(X_k^n)}}\right).$$

This approximation is good only when k is small. Simulation studies show that for 20 PEP cycles, this approximation is only good for $k = 1$ or 2. It is hard to get any limit distribution for T_n. Simulations show that the variance of T_n is very large compared to its expectation. So a central limit theorem cannot hold for T_n. For the probability that a target is replicated at least M times, we can resort to simulations. In our simulation, we use $L = 1000$, $T = 250$, $n = 20$, $M = 60$, and the annealing rate changes from 0 to 0.01. The simulations show that the above probability increases at first with the annealing rate until the annealing rate is 0.002. When the annealing rate is between 0.002 and 0.004, the above probability is around 94%. Then it begins to decrease as the annealing rate increases. At annealing rate 0.01, the above probability is only around 40%. This again shows the important role played by the annealing rate. From the above discussions, we see that in the design of PEP experiments, experimental conditions should be carefully arranged to obtain the maximum number of products and maximum coverage.

4. Random length case

In previous sections we assumed that both the length of the *Taq* extension and the annealing rate are constants. This may not be true in reality. The length of the *Taq* extension can vary from one reaction to another. In that case we can model L as a random variable. In the amplification of the whole genome, some regions may be easier to amplify than others. This is presumably caused by the fact that primers do not anneal with a constant rate. In that case we can model the primers as an inhomogeneous Poisson process. Because we lack the information about the DNA sequences we want to amplify, it is impossible to know the primer annealing rate. Thus we still assume primers anneal to the whole genome according to a

homogeneous Poisson process. In this section we relax the condition that the length of the *Taq* extension is a constant. Let now the length $\mathscr{L}$ of the *Taq* extension be a random variable having a continuous distribution function $F(x)$. Our main result of this section is Theorem 9, which shows that approximately $T_n(l)$ increases at most like $\exp(2\sqrt{\lambda n(E\mathscr{L} - \int_0^{T+l} \mathscr{F}(s)\,ds)})$. On the other hand, from the lower bound we see $T_n(l)$ increases at least like $\exp(2\exp(-\lambda E\mathscr{L}/2)\sqrt{2\lambda n(E\mathscr{L} - \int_0^{T+l} \mathscr{F}(s)\,ds)})$ which is much faster than linear. We still use the above notation. First we give a lemma which is similar to Lemma 1.

Lemma 4. *Suppose the length $\mathscr{L}$ of the Taq extension has continuous distribution function $F(x)$. Then $P\{Y_k^5 \geqq x,\ Y_k^3 \geqq y\}$ depends only on $x+y$ for $x>0$, $y>0$. Let $\mathscr{F}(x) = 1 - F(x)$,*

$$H(x, y) = \lambda \int_x^y \mathscr{F}(T+t) \exp\left(-\lambda \int_t^y \mathscr{F}(s-x)\,ds\right) dt,$$

and $F_k(x+y) = P\{Y_k^5 \geqq x,\ Y_k^3 \geqq y\}$. Then $F_k(x)$ satisfies the following recursive equation:

$$F_1(x) = \int_x^\infty \lambda \mathscr{F}(T+t) \exp\left(-\lambda \int_t^\infty \mathscr{F}(s-x)\,ds\right) dt = H(x, \infty),$$

$$F_{k+1}(x) = -\int_x^\infty H(x, y)\,dF_k(y), \qquad k \geqq 1,$$

and $F_k(x)$ is continuously differentiable.

Proof. We refer to Figure 12. In order that we have a first generation T-DNA with $\{Y_k^5 \geqq x,\ Y_k^3 \geqq y\}$, there must be a primer P1 to the right of A whose extension covers A and B, and to the right of P1 there are no primers whose extensions cover A. Let the position of P1 be $\mathscr{T}$. Note the number $N_1(s)$ of primers in $(x, s]$ whose extensions cover A and B is an inhomogeneous Poisson process with rate $\lambda\mathscr{F}(T+s+y)$ at s (Ross 1971). Therefore

$$P\{\mathscr{T} \leqq t\} = P\{N_1(t, \infty) = 0\} = \exp\left(-\lambda \int_t^\infty \mathscr{F}(T+s+y)\,ds\right).$$

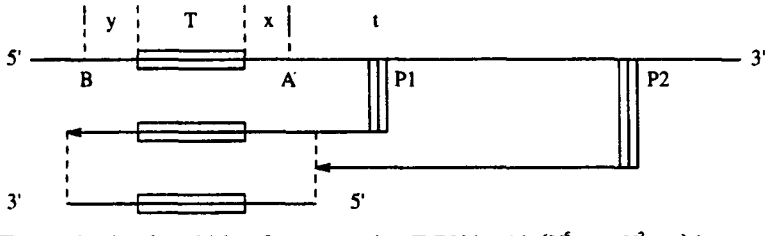

Figure 12. The mechanism by which a first generation T-DNA with $\{Y_1^5 \geqq x,\ Y_1^3 \geqq y\}$ is generated in the random length case

Thus the density function of $\mathcal{T}$ is

$$f_{\mathcal{T}}(t) = \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_t^\infty \mathcal{F}(T+s+y)\, ds\right). \tag{8}$$

Given $\{\mathcal{T} = t\}$, the number $N_2(s)$ of primers in $(t, s]$ whose extensions cover A but not B is an inhomogeneous Poisson process with rate $\lambda(\mathcal{F}(s-x) - \mathcal{F}(T+s+y))$ at s. Therefore

$$P\{N_2(t, \infty) = 0 \mid \mathcal{T} = t\} = \exp\left(-\lambda \int_t^\infty (\mathcal{F}(s-x) - \mathcal{F}(T+s+y))\, ds\right). \tag{9}$$

Using the law of total probability and (8) and (9) we have

$$\begin{aligned} P\{Y_1^5 \geqq x, Y_1^3 \geqq y\} &= \int_x^\infty P\{N_2(t, \infty) = 0 \mid \mathcal{T} = t\} f_{\mathcal{T}}(t)\, dt \\ &= \int_x^\infty \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_t^\infty \mathcal{F}(s-x)\, ds\right) dt \\ &= \int_x^\infty \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_{t-x}^\infty \mathcal{F}(s)\, ds\right) dt \\ &= \int_0^\infty \lambda \mathcal{F}(T+t+x+y) \exp\left(-\lambda \int_t^\infty \mathcal{F}(s)\, ds\right) dt. \end{aligned} \tag{10}$$

Therefore $g_1(x, y) = P\{Y_1^5 \geqq x, Y_1^3 \geqq y\}$ depends only on $x+y$ and $F_1(x) = P\{Y_k^5 \geqq x, Y_1^3 \geqq 0\}$ is continuously differentiable from (10). Suppose $g_k(x, y) = P\{Y_k^5 \geqq x, Y_k^3 \geqq y\}$ depends only on $x+y$ and is continuously differentiable. Then by a similar argument to that above we have

$$\begin{aligned} &P\{Y_{k+1}^5 \geqq x, Y_{k+1}^3 \geqq y\} \\ &= \iint_{y' \geqq x, x' \geqq y} \left[\int_x^{y'} \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_t^{y'} \mathcal{F}(s-x)\, ds\right) dt\right] dg_k(x', y') \\ &= \iint_{y' \geqq x, x' \geqq y} \left[\int_x^{y'} \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_{t-x}^{y'-x} \mathcal{F}(s)\, ds\right) dt\right] dg_k(x', y') \\ &= \iint_{y' \geqq x, x' \geqq y} \left[\int_0^{y'-x} \lambda \mathcal{F}(T+t+x+y) \exp\left(-\lambda \int_t^{y'-x} \mathcal{F}(s)\, ds\right) dt\right] dg_k(x', y') \\ &= \iint_{y' \geqq 0, x' \geqq y} \left[\int_0^{y'} \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_t^{y'} \mathcal{F}(s)\, ds\right) dt\right] dg_k(x', y'+x) \\ &= \iint_{y' \geqq 0, x' \geqq 0} \left[\int_0^{y'} \lambda \mathcal{F}(T+t+y) \exp\left(-\lambda \int_t^{y'} \mathcal{F}(s)\, ds\right) dt\right] dg_k(x'+y, y'+x). \end{aligned} \tag{11}$$

Because $g_k(x, y)$ depends only on $x + y$, we see that $g_{k+1}(x, y)$ also depends only on $x + y$.

From (11) we have

$$\begin{aligned} F_{k+1}(x) &= P\{Y_1^5 \geqq x,\ Y_1^3 \geqq 0\} \\ &= \iint_{y \geqq x,\, x' \geqq 0} H(x, y)\, dg_k(x', y) \\ &= -\int_x^\infty H(x, y)\, dP\{Y_1^5 \geqq 0,\ Y_1^3 \geqq y\} \qquad (12) \\ &= -\int_x^\infty H(x, y)\, dF_k(y) \\ &= -\int_x^\infty \left[\int_x^y \lambda \mathscr{F}(T + t) \exp\left(-\lambda \int_t^y \mathscr{F}(s - x)\, ds\right) dt\right] dF_k(y). \end{aligned}$$

Using induction and (12) we can prove that $F_k(x)$ is continuously differentiable. The lemma is proved.

Although we have a relatively simple recursive formula for $F_k(x)$, unlike the constant length case we do not have an explicit formula for $F_k(x)$. In the following we want to compare $F_k(x)$ when the length $\mathscr{L}$ of the *Taq* extension is random with $F_k^{(c)}(x) = P\{Y_k^5 \geqq x,\ Y_k^3 \geqq 0\}$ when the length of extension is a constant $E\mathscr{L}$. The following theorem gives a relationship between $F_k(x)$ and $F_k^{(c)}(x)$.

Theorem 8. *Suppose* $E\mathscr{L} < \infty$. *Let* $F_k(x)$ *be defined as above and* $F_k^{(c)}(x) = P\{Y_k^5 \geqq x,\ Y_k^3 \geqq 0\}$ *when the length of the Taq extension is a constant* $E\mathscr{L}$ *and the target length is* 0. *Then for any* $k \geqq 1$,

$$\exp\left(-(k-1)\lambda E\mathscr{L}\right) F_k^{(c)}\left(\int_0^{T+x} \mathscr{F}(t)\, dt\right) \leqq F_k(x) \leqq \exp\left(\lambda E\mathscr{L}\right) F_k^{(c)}\left(\int_0^{T+x} \mathscr{F}(t)\, dt\right).$$

Let $f_k(x) = -F_k'(x)$. *Then for any* $k \geqq 1$,

$$f_k(x) \leqq \frac{\lambda \mathscr{F}(T + x)}{(k-1)!} \left(\lambda \int_{T+x}^\infty \mathscr{F}(s)\, ds\right)^{k-1}.$$

Proof. We prove the theorem by induction. First we prove the bounds for $F_k(x)$. For the constant length case, from Section 2 we see

$$F_1^{(c)}(x) = \exp(-\lambda x) - \exp(-\lambda E\mathscr{L}), \qquad 0 < x < E\mathscr{L},$$

and

$$F_{k+1}^{(c)}(x) = \int_x^{E\mathscr{L}} F_k^{(c)}(y) \lambda \exp(-\lambda(y - x))\, dy, \qquad 0 < x < E\mathscr{L}. \qquad (13)$$

Now we prove the upper bound. Note that

$$H(x, y) = \int_x^y \lambda \mathscr{F}(T+t) \exp\left(-\lambda \int_t^y \mathscr{F}(s-x)\, ds\right) dt$$

$$\leqq \int_x^y \lambda \mathscr{F}(T+t) \exp\left(-\lambda \int_t^y \mathscr{F}(T+s)\, ds\right) dt$$

$$= 1 - \exp\left(-\lambda \int_x^y \mathscr{F}(T+s)\, ds\right).$$

Let $y \to \infty$ and from (10) we have

$$F_1(x) \leqq 1 - \exp\left(-\lambda \int_x^\infty \lambda \mathscr{F}(T+s)\, ds\right)$$

$$= \exp\left(\lambda \int_0^{T+x} \mathscr{F}(s)\, ds\right)\left(\exp\left(-\lambda \int_0^{T+x} \mathscr{F}(s)\, ds\right) - \exp(-\lambda E\mathscr{L})\right)$$

$$\leqq \exp(\lambda E\mathscr{L}) F_1^{(c)}\left(\int_0^{T+x} \mathscr{F}(t)\, dt\right).$$

Therefore the upper bound is true for $k = 1$. Suppose the upper bound is true for k. Then from (12) we have

$$F_{k+1}(x) = -\int_x^\infty H(x, y)\, dF_k(y)$$

$$\leqq -\int_x^\infty \left(1 - \exp\left(-\lambda \int_x^y \mathscr{F}(T+s)\, ds\right)\right) dF_k(y)$$

$$= \int_x^\infty F_k(y) \lambda \mathscr{F}(T+y) \exp\left(-\lambda \int_{T+x}^{T+y} \mathscr{F}(s)\, ds\right) dy$$

(induction)

$$\leqq \exp(\lambda E\mathscr{L}) \int_x^\infty \lambda \mathscr{F}(T+y) F_k^{(c)}\left(\int_0^{T+y} \mathscr{F}(s)\, ds\right) \exp\left(-\lambda \int_{T+x}^{T+y} \mathscr{F}(s)\, ds\right) dy$$

$(u = \int_0^{T+y} \mathscr{F}(t)\, dt)$

$$= \exp(\lambda E\mathscr{L}) \int_{\int_0^{T+x} \mathscr{F}(s)\, ds}^{E\mathscr{L}} \lambda F_k^{(c)}(u) \exp\left(-\lambda\left(u - \int_0^{T+x} \mathscr{F}(s)\, ds\right)\right) du$$

(equation (13))

$$= \exp(\lambda E\mathscr{L}) F_{k+1}^{(c)}\left(\int_0^{T+x} \mathscr{F}(t)\, dt\right).$$

By induction the upper bound is true for any k.

Now let us prove the lower bound. First note

$$H(x,y)=\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(s-x)\,ds\right)dt$$

$$=\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(T+s)\,ds\right)$$

$$\times\exp\left(-\lambda\int_t^y (\mathscr{F}(s-x)-\mathscr{F}(T+s))\,ds\right)dt$$

$$=\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(T+s)\,ds\right) \tag{14}$$

$$\times\exp\left(-\lambda\left(\int_{t-x}^{t+T}\mathscr{F}(s)\,ds-\int_{y-x}^{y+T}\mathscr{F}(s)\right)ds\right)dt$$

$$\geqq\exp\left(-\lambda\int_0^{T+x}\mathscr{F}(s)\,ds\right)\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(T+s)\,ds\right)dt$$

$$=\exp\left(-\lambda\int_0^{T+x}\mathscr{F}(s)\,ds\right)\left(1-\exp\left(-\lambda\int_{T+x}^{T+y}\mathscr{F}(s)\,ds\right)\right).$$

Letting $y\to\infty$ we have

$$F_1(x)\geqq F_1^{(c)}\left(\int_0^{T+x}\mathscr{F}(s)\,ds\right).$$

That is, the lower bound is true for $k=1$. Suppose the lower bound is true for k. Then

$$F_{k+1}(x)=-\int_x^\infty H(x,y)\,dF_k(y)$$

(equation (14))

$$\geqq-\exp(-\lambda E\mathscr{L})\int_x^\infty\left(1-\exp\left(-\lambda\int_{T+x}^{T+y}\mathscr{F}(s)\,ds\right)\right)dF_k(y)$$

$$=\exp(-\lambda E\mathscr{L})\int_x^\infty \lambda\mathscr{F}(T+y)F_k(y)\exp\left(-\lambda\int_{T+x}^{T+y}\mathscr{F}(s)\,ds\right)dy$$

(induction)

$$\geqq\exp(-k\lambda E\mathscr{L})\int_x^\infty \lambda\mathscr{F}(T+y)F_k^{(c)}\left(\int_0^{T+y}\mathscr{F}(s)\,ds\right)$$

$$\times\exp\left(-\lambda\int_{T+x}^{T+y}\mathscr{F}(s)\,ds\right)\Bigg)dy$$

$\left(u=\int_0^{T+y}\mathscr{F}(s)\,ds\right)$

$$\geqq\exp(-k\lambda E\mathscr{L})\int_{\int_0^{T+x}\mathscr{F}(s)ds}^{E\mathscr{L}}\lambda F_k^{(c)}(u)\exp\left(-\lambda\left(u-\int_0^{T+x}\mathscr{F}(s)\,ds\right)\right)dy$$

(equation (13))

$$\geqq \exp(-k\lambda E\mathscr{L})F_{k+1}^{(c)}\left(\int_0^{T+x}\mathscr{F}(s)\,ds\right).$$

By induction the lower bound is true.

Next we prove the upper bound for $f_k(x)$. Note that, for any fixed y and $x \leqq y$, we have

$$\begin{aligned}
-\frac{d}{dx}H(x,y) &= -\frac{d}{dx}\left[\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(s-x)\,ds\right)dt\right] \\
&= \lambda\mathscr{F}(T+x)\exp\left(-\lambda\int_x^y \mathscr{F}(s-x)\,ds\right) \\
&\quad + \lambda\int_x^y \lambda\mathscr{F}(T+t)\exp\left(-\lambda\int_t^y \mathscr{F}(s-x)\,ds\right)(\mathscr{F}(t-x)-\mathscr{F}(y-x))\,dt \\
&\leqq \lambda\mathscr{F}(T+x)\left(\exp\left(-\lambda\int_x^y \mathscr{F}(s-x)\,ds\right)\right. \\
&\quad \left. + \int_x^y \lambda\mathscr{F}(t-x)\exp\left(-\lambda\int_t^y \mathscr{F}(s-x)\,ds\right)dt\right) \\
&= \lambda\mathscr{F}(T+x).
\end{aligned}$$

Let $y \to \infty$ we have $f_1(x) \leqq \lambda\mathscr{F}(T+x)$. The upper bound is true for $k=1$. Suppose the upper bound is true for k, then

$$\begin{aligned}
f_{k+1}(x) &= -\int_x^\infty \frac{d}{dx}H(x,y)f_k(y)\,dy \\
&\leqq \lambda\mathscr{F}(T+x)\int_x^\infty f_k(y)\,dy
\end{aligned}$$

(induction)

$$\begin{aligned}
&\leqq \lambda\mathscr{F}(T+x)\int_x^\infty \frac{\lambda\mathscr{F}(T+y)}{(k-1)!}\left(\lambda\int_{T+y}^\infty \mathscr{F}(s)\,ds\right)^{k-1}dy \\
&= \frac{\lambda\mathscr{F}(T+x)}{k!}\left(\lambda\int_{T+x}^\infty \mathscr{F}(s)\,ds\right)^k.
\end{aligned}$$

By induction the inequality is true. Theorem 8 is proved.

Next we study the total number $T_n(l)$ of T-DNAs with length at least $T+l$. The following theorem contains our main result of this section, which gives lower and upper bounds for the growth rate of $T_n(l)$.

Theorem 9. (*i*) *Let* $P_k(l)$ *be the probability that there exist a* kth *generation T-DNA with length at least* $T+l$ *at the* kth *cycle and* $f_k(x) = -(d/dx)F_k(x)$. *Then* $P_k(l) = F_k(l) + lf_k(l)$.

(ii) The expected number of kth *generation T-DNAs with length at least* $T + l$ *after* n *PEP cycles is* $\binom{n}{k}P_k(l)$.

(iii) Let $T_n(l)$ *be the total number of T-DNAs with length at least* $T + l$ *after* n *cycles. Then*

$$2\exp\left(\tfrac{1}{2}\lambda E\mathscr{L}\right)\sqrt{\lambda\int_{T+l}^{\infty}\mathscr{F}(s)\,ds} \leqq \liminf_{n\to\infty}\frac{\log T_n(l)}{\sqrt{n}}$$

$$\leqq \limsup_{n\to\infty}\frac{\log T_n(l)}{\sqrt{n}}$$

$$\leqq 2\sqrt{\lambda\int_{T+l}^{\infty}\mathscr{F}(s)\,ds}.$$

Proof. (i) First note that

$$\begin{aligned}P_k(l) &= P\{Y_k^5 + Y_k^3 \geqq l\}\\ &= P\{Y_k^5 \geqq l,\ Y_k^3 \geqq 0\} + P\{0 < Y_k^5 < l,\ Y_k^5 + Y_k^3 \geqq l\}\\ &= F_k(l) + P\{0 < Y_k^5 < l,\ Y_k^5 + Y_k^3 \geqq l\}.\end{aligned}$$

Since $f_k(x) = -F_k'(x)$, we have

$$\frac{d^2}{dx\,dy}P_k\{Y_k^5 \geqq x,\ Y_k^3 \geqq y\} = F_k''(x+y) = -f_k'(x+y).$$

Therefore

$$\begin{aligned}P\{0 < Y_k^5 < l,\ Y_k^5 + Y_k^3 \geqq l\} &= -\iint_{0<x<l,\ x+y\geqq l} f_k'(x+y)\,dx\,dy\\ &= -\int_0^l\left[\int_{l-x}^{\infty} f_x'(x+y)\,dy\right]dx\\ &= -\int_0^l f_k(x+y)\Big|_{l-x}^{\infty}\,dx = lf_k(l).\end{aligned}$$

Thus

$$P_k(l) = F_k(l) + lf_k(l) \tag{15}$$

and (i) is proved.

(ii) The proof of (ii) is the same as that of Theorem 3(i).

(iii) Similar to the proof of Theorem 3(ii) we can prove, for any $B \geqq 0$,

$$\lim_{n\to\infty}\frac{\log\left(\sum_{k=1}^{n}\binom{n}{k}B^k/k!\right)}{\sqrt{n}} = 2\sqrt{B}, \tag{16}$$

and there exist constants c and C which do not depend on k such that

$$c\frac{\lambda^k(E\mathscr{L}-l)^k}{k!} \leqq F_k^{(c)}(l) \leqq C\frac{\lambda^k(E\mathscr{L}-l)^k}{k!}.$$

From Theorem 8 and (15) we have

$$\begin{aligned} T_n(l) &= \sum_{k=1}^{n} \binom{n}{k} P_k(l) \geqq \sum_{k=1}^{n} \binom{n}{k} F_k(l) \\ &\geqq \sum_{k=1}^{n} \binom{n}{k} \exp\left(-(k-1)\lambda E\mathscr{L}\right) F_k^{(c)}\left(\int_0^{T+l} \mathscr{F}(s)\,ds\right) \\ &\geqq c \exp\left(\lambda E\mathscr{L}\right) \sum_{k=1}^{n} \binom{n}{k} \frac{\left(\lambda \exp\left(-\lambda\mathscr{L}\right)\left(E\mathscr{L} - \int_0^{T+l} \mathscr{F}(s)\,ds\right)\right)^k}{k!} \\ &= c \exp\left(\lambda E\mathscr{L}\right) \sum_{k=1}^{n} \binom{n}{k} \frac{\left(\lambda \exp\left(-\lambda E\mathscr{L}\right) \int_{T+l}^{\infty} \mathscr{F}(s)\,ds\right)^k}{k!}. \end{aligned}$$

From (16) we have

$$\liminf_{n\to\infty} \frac{\log T_n(l)}{\sqrt{n}} \geqq 2\sqrt{\lambda \exp\left(-\lambda E\mathscr{L}\right) \int_{T+l}^{\infty} \mathscr{F}(s)\,ds}.$$

Similarly, from Theorem 8 and (15) we have

$$\begin{aligned} T_n(l) &= \sum_{k=1}^{n} \binom{n}{k} \left(F_k(l) + lf_k(l)\right) \\ &\leqq \sum_{k=1}^{n} \binom{n}{k} \left(C\frac{\left(\lambda \int_{T+l}^{\infty} \mathscr{F}(s)\,ds\right)^k}{k!} + \lambda l \frac{\left(\lambda \int_{T+l}^{\infty} \mathscr{F}(s)\,ds\right)^{k-1}}{(k-1)!} \right) \\ &\leqq nC \sum_{k=1}^{n} \binom{n}{k} \frac{\left(\lambda \int_{T+l}^{\infty} \mathscr{F}(s)\,ds\right)^k}{k!}, \end{aligned}$$

where C is a constant but may be different each time it appears. From (16) we have

$$\limsup_{n\to\infty} \frac{\log T_n(l)}{\sqrt{n}} \leqq 2\sqrt{\lambda \int_{T+l}^{\infty} \mathscr{F}(s)\,ds}.$$

The theorem is proved.

As for the characteristic function, variance and limiting behavior of $X_k^n(x, y)$—the kth generation T-DNAs after n cycles when $Y_0^5 = x$, $Y_0^3 = y$, we have the following results.

Theorem 10. *Let $X_k^n(x, y)$ be the number of* kth *generation T-DNAs after n cycles when $Y_0^5 = x$, $Y_0^3 = y$. Then we have the following.*

(*i*) *The characteristic function $g_k^n(t; x, y)$ of $X_k^n(x, y)$ satisfies the following recursive equation:*

$$g_1^n(t; x, y) = (1 - P_1(x, y) + P_1(x, y)e^{it})^n,$$

$$g_k^n(t; x, y) = \prod_{j=k}^{n} \left(1 - P_1(x, y) + \iint_{(x_1, y_1)} g_{k-1}^{j-1}(t; x_1, y_1)\, dP(x_1, y_1 \mid x, y)\right), \qquad k \geqq 2,$$

where

$$P(x_1, y_1 \mid x, y) = P\{Y_1^5 \geqq x_1,\ Y_1^3 \geqq y_1 \mid Y_0^5 = x,\ Y_0^3 = y\}$$

$$= \int_{x_1}^{y} \lambda \mathcal{F}(T + t + y_1) \exp\left(-\lambda \int_t^y \mathcal{F}(s - x_1)\, ds\right) dt,$$

and

$$P_1(x, y) = P(0, 0 \mid x, y) = \int_0^y \lambda \mathcal{F}(T + t) \exp\left(-\lambda \int_t^y \mathcal{F}(s)\, ds\right) dt.$$

(*ii*) *The variance of $X_k^n(x, y)$ satisfies the following recursive equation:*

$$\mathrm{Var}\,(X_1^n(x, y)) = nP_1(x, y)(1 - P_1(x, y)),$$

$$\mathrm{Var}\,(X_k^n(x, y)) = \sum_{j=k}^{n} \iint_{(x_1, y_1)} \mathrm{Var}\,(X_{k-1}^{j-1}(x_1, y_1))\, dP(x_1, y_1 \mid x, y)$$

$$+ \sum_{j=k}^{n} \binom{j-1}{k-1}^2 \left(\iint_{(x_1, y_1)} P_{k-1}^2(x_1, y_1)\, dP(x_1, y_1 \mid x, y) - P_k^2(x, y)\right),$$

$$n, k = 2, 3, 4 \cdots,$$

where

$$P_k(x, y) = \iint_{(x_1, y_1)} P_{k-1}(x_1, y_1)\, dP(x_1, y_1 \mid x, y), \qquad k = 2, 3, \cdots.$$

(*iii*) *The central limit theorem and strong law of large numbers as in Theorems* 5 *and* 6 *still hold.*

The proof of this theorem is almost the same as in Section 3 and is omitted.

5. Whole genome amplification with T-PCR

In previous sections we studied PEP and proved that the expected total number of T-DNAs that contain a fixed target T increases as $e^{c\sqrt{n}}$, where c is a constant and n is the number of PEP cycles. Experiments show while PEP does amplify the DNA

from a single cell, the amount of amplification cannot be detected on ethidium bromide stained gels after 50 PEP cycles (Zhang *et al.* 1992). A new method was proposed by using tagged random primers (Grothues *et al.* 1993). The principle of this technique is described as follows. Tagged random primers (Jeffreys *et al.* 1991) with a random 3′ tail that can bind to arbitrary DNA sequences, and a constant 5′ head (the tag) for the subsequent detection of the incorporated primers, are synthesized first. In the first step, $n \geqq 2$ PEP cycles are done using these tagged random primers. In the first cycle, the 3′ tails of the tagged random primers anneal to the single-stranded sequences and the *Taq* polymerase extends the primers by a constant length L. In later cycle sequences tags at both ends are generated. After n PEP cycles, sequences without tags at either end are removed. In the second step, the usual PCR is applied using primers complementary to the constant region of the tagged random primers. During this step molecules containing random primers at both ends are amplified exponentially. This technique is referred to as T-PCR. The T-PCR technique combines the coverage properties of PEP and the exponential growth rate of amplification by PCR. Because in the second step only sequences with tags at both ends are amplified, we are only concerned about sequences with tags at both ends, which we refer to as *Tag sequences.* Note that the first generation sequences defined in Section 1 are not Tag sequences because, at most, only the 5′ end is tagged. A second generation sequence is a Tag sequence if and only if the 5′ end of its first generation ancestor is tagged. Because we suppose the length of extension is a constant, third or higher generation sequences are always Tag sequences (Figure 13). Here again we suppose there is a target T which we model as an interval on the real line. We want to answer the following questions.

(a) For a fixed target of length T, what is the expected number of Tag sequences with length at least $T + l$ covering the target completely?
(b) What is the probability that the target is covered by any Tag sequences of length at least $T + l$?

First let us fix a target of length T. Let Y_k^5 and Y_k^3 be the lengths of a kth generation Tag sequence at its 5′ and 3′ ends beyond the target. From the above discussions we see that Theorems 1, 2, and 3 still hold for $k \geqq 3$. For $k = 2$, the

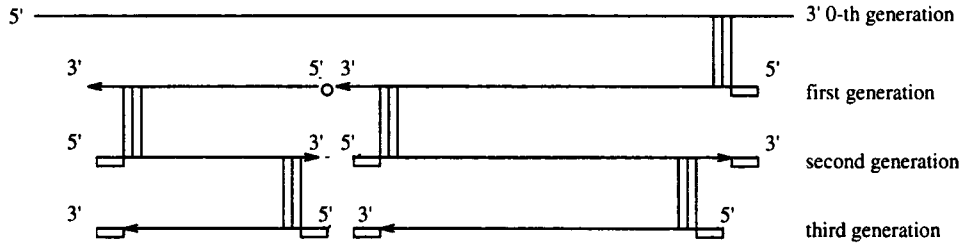

Figure 13. The mechanism of T-PCR. The sequences with boxes at both ends are Tag sequences. Third or higher generation sequences are always Tag sequences

results are changed because not all second generation sequences are Tag sequences. The following theorem gives modified versions of Theorems 1, 2, and 3 for $k = 2$.

Theorem 11. (*i*) *The joint density function of* Y_2^5 *and* Y_2^3 *is*

$$f_2(x, y) = \lambda^2 e^{-\lambda(2L-T-x-y)}, \qquad 0 < x + y < L - T,$$

in the sense that for any subset $B \subset \{(x, y): x > 0,\ y > 0,\ x + y < L - T\}$,

$$P\{(Y_2^5, Y_2^3) \in B\} = \iint_B f_2(x, y)\, dx\, dy.$$

(*ii*) Y_2^5 *and* Y_2^3 *have the same density function. If we denote their common density function by* $f_2(x)$, *we have*

$$f_2(x) = \lambda e^{-\lambda L}(1 - e^{-\lambda(L-T-x)}), \qquad 0 < x < L - T,$$

in the sense that for any $0 < x < L - T$,

$$P\{Y_2^5 > x\} = P\{Y_2^3 > x\} = \int_x^{L-T} f_2(s)\, ds.$$

(*iii*) *Let* $P_2(l)$ *be the probability that there exists a second generation Tag sequence at the second cycle with length greater than* $T + l$. *Then*

$$P_2(l) = (\lambda(L - T) - 1)e^{-\lambda L} - (\lambda l - 1)e^{-\lambda(2L-T-l)}.$$

(*iv*) *Given that second generation Tag sequences covering the target exist, the expected lengths of* Y_2^5 *and* Y_2^3 *are* $\int_0^{L-T} x f_2(x)\, dx / P_2(0)$.

(*v*) *The expected number of second generation Tag sequences with length at least* $T + l$ *containing the target* T *is* $\binom{n}{2} P_2(l)$.

The next theorem gives a recursive formula for the probability that a target of length T is covered by Tag sequences of length at least $T + l$.

Theorem 12. *Suppose initially we have one single-stranded sequence. The expected fraction* c_n *of the genome covered by Tag sequences of length at least* $T + l$ *after* n *PEP cycles satisfies the following recursive equation*: $1 - c_{n+1} = (1 - c_n)(1 - h_n)$, *where* $c_1 = 0$ *and* $\exp(\lambda T)h_n$ *is given by*

$$\iint_{l<x+y<L-T} \left\{1 - \prod_{i=1}^{n-1}\left[1 - \int_{(l-x)^+}^{y} \lambda e^{-\lambda(y-x_1)}(1 - e^{-\lambda i(x-(l-x_1)^+)})\, dx_1\right]\right\}$$
$$\times \lambda^2 e^{-\lambda(x+y)}\, dx\, dy + e^{-\lambda(L-T)}\left[\left(\lambda(L - T) - \frac{1}{n}\right) - \left(\lambda l - \frac{1}{n}\right)e^{-\lambda n(L-T-l)}\right].$$

In the following we illustrate the coverage result with an example. As in Section 2,

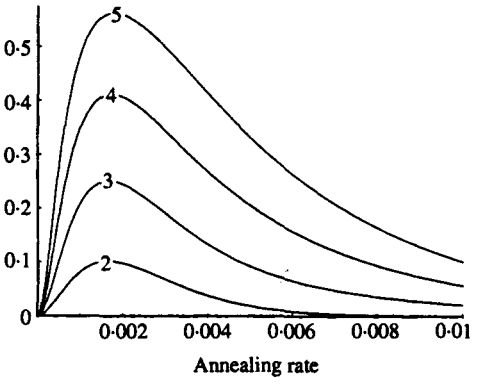

Figure 14. The probability that the target is covered by Tag sequences after 2, 3, 4, and 5 cycles with $L = 1000$, $T = 250$.

we let $L = 1000$, $T = 250$. Figure 14 shows the probability that the target is covered by Tag sequences after 2, 3, 4, and 5 PEP cycles. From Figure 14 we see that if the annealing rate is low, the coverage is also low. When the annealing rate reaches some level, the coverage reaches its maximum. Then the coverage decreases after this level. Under the above conditions, the maximum point for coverage is approximately $\lambda = 0.002$. That means that in approximately every $1/0.002 = 500$ bases there is a primer annealing to the genome. Even at this optimal annealing rate, after 5 cycles the probability that the target is covered by Tag sequences is only around 58%. The reason for this behavior is that, when the annealing rate is low, few Tag sequences are generated and thus the probability that a point is covered by Tag sequences is small. If the annealing rate is too high, the first generation sequences are more likely to replace each other and thus few second generation Tag sequences are generated but more third and higher generation sequences are generated. All the factors balance among themselves and maximum coverage is obtained. It might be noted that the optimal annealing rate here is the same as that in Section 4. That is just a coincidence. If we use different values for the parameters, the optimal annealing rate will be different.

Proof of Theorem 11. (i) We refer to Figure 15. For a fixed target of length T, let

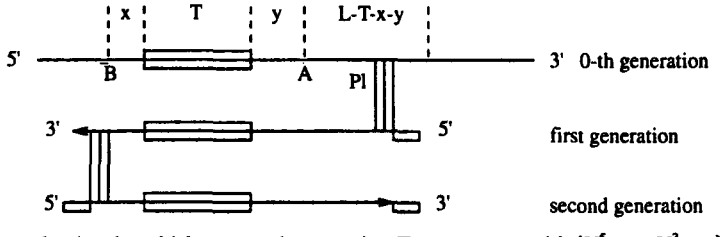

Figure 15. The mechanism by which a second generation Tag sequence with $\{Y_2^5 \cong x,\ Y_2^3 \cong y\}$ is generated

Y_2^5 and Y_2^3 be the lengths of a second generation Tag sequence covering the target at its 5′ and 3′ ends respectively. For any $0<x+y<L-T$, $x\geqq 0$, $y\geqq 0$, $\{Y_2^5\geqq x, Y_2^3\geqq y\}$ happens if and only if in the first cycle there is a primer (P1) in the interval $(y, L-T-x]$ and to the right of P1 in an interval of length L there are no primers so that the 5′ end of the extension product of P1 is tagged and the extension product of P1 covers A, B. Let the position of P1 be z. Then in the second cycle there must be a primer in $(x, L-T-z]$ which has probability $1-e^{-\lambda(L-T-z-x)}$. Therefore

$$P\{Y_2^5\geqq x, Y_2^3\geqq y\}=\int_y^{L-T-x}\lambda e^{-\lambda L}(1-e^{-\lambda(L-T-z-x)})\,dz$$
$$=e^{-\lambda L}[\lambda(L-T-x-y)-1+e^{-\lambda(L-T-x-y)}],$$
$$0\leqq x+y\leqq L-T.$$

Therefore the density of (Y_2^5, Y_2^3) is

$$f_2(x, y)=\lambda^2 e^{-\lambda(2L-T-x-y)}, \qquad 0\leqq x+y\leqq L-T.$$

(ii) The marginal density of Y_2^5 or Y_2^3 is

$$f_2(x)=\int_0^{L-T-x} f_2(x, y)\,dy=\lambda e^{-\lambda L}(1-e^{-\lambda(L-T-x)}).$$

(iii) It is easy to calculate

$$P_2(l)=\iint\limits_{l<x+y<L-T} f_2(x, y)\,dx\,dy=[\lambda(L-T)-1]e^{-\lambda L}-(\lambda l-1)e^{-\lambda(2L-T-l)}.$$

(iv) Can be seen from (ii) and (iii).

(v) Just like the proof of Theorem 3, we prove this part by induction. For $n=2$, the claim is obviously true. Suppose the result is true for n. Let $X_2^n(l)$ be the number of second generation Tag sequences of length at least $T+l$ covering the target after n PEP cycles. Then we have $X_2^{n+1}(l)=X_2^n(l)+Z_2^n(l)$, where $Z_2^n(l)$ is the number of second generation Tag sequences of length at least $T+l$ generated from α_1, where α_1 is the first generation sequence generated in the first cycle. Given $Y_1^5=x$, $Y_1^3=L-T-x$, after each cycle α_1 generate a second generation Tag sequence of length at least $T+l$ covering the target if and only if there are primers in $((l-x)^+, L-T-x]$ at the 3′ end of α_1. The probability of this event is $1-\exp[-\lambda(L-T-x-(l-x)^+)]$. So after n cycles, the expected number of second generation Tag-sequences generated by α_1 of length at least $T+l$ is $n(1-\exp[-\lambda(L-T-x-(l-x)^+])$. In the proof of Lemma 1, we have proved

$$P\{Y_1^3+Y_1^5=L-T, Y_1^5>x\}=\lambda e^{-\lambda L}(L-T-x), \qquad 0<x<L-T.$$

Thus

$$-\frac{d}{dx}P\{Y_1^3+Y_1^5=L-T, Y_1^5>x\}=\lambda e^{-\lambda L}, \qquad 0<x<L-T.$$

From law of total probability we see that

$$EZ_2^n(l) = n\int_0^{L-T} (1 - \exp[-\lambda(L - T - x - (l-x)^+)])\lambda e^{-\lambda L}\, dx = nP_2(l).$$

Therefore

$$EX_2^{n+1}(l) = EX_2^n(l) + EZ_2^n(l) = EX_2^n(l) + nP_2(l).$$

By induction on n, we have $EX_2^n(l) = \binom{n}{2}P_2(l)$ and the theorem is proved.

Proof of Theorem 12. Fix a target T. Let A_n be the event that the target is covered by Tag sequences of length at least $T + l$ after n PEP cycles. Conditional on the lengths at the 5′ and 3′ ends ($Y_1^5 = x$, $Y_1^3 = y$) of the first generation sequence α_1 which is generated at first cycle, by the mechanism of PEP, A_{n+1} happens if and only if the target is covered by Tag sequences of length at least $T + l$ generated by the original sequence after n cycles, or the target is covered by Tag sequences of length at least $T + l$ generated from α_1. Let $A_n^*(x, y)$ be the last event just discussed. Then, given ($Y_1^5 = x$, $Y_1^3 = y$), A_n and $A_n^*(x, y)$ are independent and $A_{n+1} \mid (x, y) = A_n \cup A_n^*(x, y)$. Thus

$$P(A_{n+1} \mid (x, y)) = P(A_n \cup A_n^*(x, y)) = 1 - (1 - P(A_n))(1 - P(A_n^*(x, y))).$$

Therefore

$$1 - P(A_{n+1} \mid (x, y)) = (1 - P(A_n))(1 - P(A_n^*(x, y))). \tag{17}$$

Next we study $P(A_n^*(x, y))$. We consider two cases.

(a) $l < x + y < L - T$. Notice that when ($Y_1^5 = x$, $Y_1^3 = y$), the first generation sequence does not have a tag at its 5′ end. Therefore the second generation sequences generated from it are not Tag sequences because the 3′ ends of the second generation sequences are not tagged. In order that $A_n^*(x, y)$ occurs, there must be a second generation sequence at some cycle j, $2 \leqq j \leqq n$, such that this second generation sequence generates tagged third generation sequences covering the target at some cycles $j + 1, j + 2, \cdots, n + 1$. Let B_j, $2 \leqq j \leqq n$, be the event just described. Then $A_n^*(x, y) = B_2 \cup B_3 \cup \cdots \cup B_n$. Therefore

$$P(A_n^*(x, y)) = P(B_2 \cup B_3 \cup \cdots \cup B_n) = 1 - \prod_{j=2}^{n} (1 - P(B_j)). \tag{18}$$

Given ($Y_1^5 = x$, $Y_1^3 = y$), B_j occurs if and only if there exists a primer at x_1, $(l - x)^+ < x_1 < y$, no primers in (x_1, y) at cycle j and in cycles $j + 1, j + 2, \cdots, n + 1$ there are primers in $((l - x_1)^+, x]$. Since in each cycle the number of primers is a Poisson process with parameter λ, the number of primers in cycles $j + 1$, $j + 2, \cdots, n + 1$ is a Poisson process with parameter $(n - j + 1)\lambda$. Therefore

$$P(B_j) = \int_{(l-x)^+}^{y} \lambda e^{-\lambda(y - x_1)}(1 - \exp[-\lambda(n - j + 1)(x - (l - x_1)^+)])\, dx_1. \tag{19}$$

From (18) and (19) we have

$$
\begin{aligned}
&P(A_n^*(x, y)) \\
&= 1 - \prod_{j=2}^{n} \left[1 - \int_{(l-x)^+}^{y} \lambda e^{-\lambda(y-x_1)} (1 - \exp[-\lambda(n-j+1)(x - (l-x_1)^+)]) \, dx_1 \right] \\
&= 1 - \prod_{i=1}^{n-1} \left[1 - \int_{(l-x)^+}^{y} \lambda e^{-\lambda(y-x_1)} (1 - \exp[-\lambda i(x - (l-x_1)^+)]) \, dx_1 \right].
\end{aligned} \tag{20}
$$

(b) $x + y = L - T$. Notice that when $(Y_1^5 = x,\ Y_1^3 = y)$, the 5′ end of the first generation sequence is already tagged. Therefore any second generation sequences generated from it are tagged at both ends. Given $(Y_1^5 = x,\ Y_1^3 = y)$, $A_n^*(x, y)$ happens if and only if there are primers in $((l-x)^+, y]$ in cycles $2, 3, \cdots, n+1$ which occur with probability

$$P(A_n^*(x, y)) = 1 - \exp[-\lambda n(y - (l-x)^+)] = 1 - \exp[-\lambda n(L - T - x - (l-x)^+)].$$

From the proof of Lemma 1 we see that the density function of $(Y_1^5,\ Y_1^3)$ is

$$f_1(x, y) = \lambda^2 e^{-\lambda(T+x+y)}, \qquad 0 < x + y < L - T,$$

and

$$P\{Y_1^5 + Y_1^3 = L - T,\ Y_1^5 \geqq x\} = \lambda e^{-\lambda L}(L - T - x), \qquad 0 < x < L - T.$$

From the law of total probability we have

$$
\begin{aligned}
P(A_n^*) = &\iint_{l < x+y < L-T} P(A_n^*(x, y)) f_1(x, y) \, dx \, dy \\
&+ \int_0^{L-T} \lambda e^{-\lambda L} (1 - e^{-\lambda n(L-T-x-(l-x)^+)}) \, dx.
\end{aligned} \tag{21}
$$

It is easy to verify that

$$
\begin{aligned}
&\int_0^{L-T} \lambda (1 - \exp[-\lambda n(L - T - x - (l-x)^+)]) \, dx \\
&= \left(\left(\lambda(L-T) - \frac{1}{n} \right) - \left(\lambda l - \frac{1}{n} \right) e^{-\lambda n(L-T-l)} \right).
\end{aligned} \tag{22}
$$

Taking the expectation of (17) and combining (20), (21) and (22), we see that Theorem 12 holds.

Acknowledgements

We would like to thank Professor N. Arnheim for explaining PEP to us. Without his help it would have been impossible for us to finish this paper. We would also like to thank Professors S. Tavaré and T. Harris for suggestions that improved the presentation of the paper. This work was partially supported by grant DMS90-05833

from the National Science Foundation and grant GM 36230 from the National Institute of Health.

References

ARNHEIM, N. AND ERLICH, H. A. (1992) PCR strategy. *Ann. Rev. Biochem.* **61,** 131–56.

ARNHEIM, N., LI, H. AND CUI, X. (1990a) PCR analysis of DNA sequences in single cells: single sperm gene mapping and genetic disease diagnosis. *Genomics* **8,** 415–419.

ARNHEIM, N., WHITE, T. AND RAINEY, W. E. (1990b) Application of PCR: organismal and population biology. *BioScience* **40,** 174–82.

DEAR, P. H. AND COOK, P. R. (1993) Happy mapping—linkage mapping using a physical analog of meiosis. *Nucl. Acids Res.* **21,** 13–20.

DURRETT, R. (1991) *Probability: Theory and Examples.* Wadsworth and Brooks, New York.

ERLICH, H. A. AND ARNHEIM, N. (1992) Genetic analysis using the polymerase chain reaction. *Ann. Rev. Genet.* **26,** 479–506.

GROTHUES, D., CANTOR, C. R. AND SMITH, C. L. (1993) PCR amplification of megabase DNA with tagged random primers (T-PCR). *Nucl. Acids Res.* **21,** 1321–1322.

HARRIS, T. E. (1963) *The Theory of Branching Processes.* Springer, Berlin.

JEFFREYS, A. J., TAMAKI, K., NEIL, D. L. AND MONCKTON, D. G. (1991) Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354,** 204–209.

KINZLER, K. W. AND VOGELSTEIN, B. (1989) Whole genome PCR: Applications to the identification of sequences bound by gene regulatory proteins. *Nucl. Acids Res.* **17,** 3645–3653.

KRAWCZAK, M., REISS, J., SCHMIDTKE, J. AND ROSLER, U. (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucl. Acids Res.* **17,** 2197–2201.

KRISTJANSSON, K., CHONG, S. S., VAN DEN VEYVER, I. B., SUBRAMANIAN, S., SNABES, M. C. AND HUGHES, M. R. (1994) Preimplantation single-cell analyses of dystrophin gene deletions using whole genome amplification. *Nature Genet.* **6,** 19–23.

LUDECKE, H., SENGER, G., CLAUSSEN, U. AND HORSTHEMKE, B. (1989) Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification. *Nature* **338,** 348–350.

MULLIS, K. B. AND FALOONA, F. A. (1987) Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. *Meth. Enzymol.* **155,** 335–51.

ROSS, S. M. (1971) *Applied Probability Models with Optimization Applications.* Holden-Day, New York.

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* **230,** 1350–54.

SAIKI, R., GELFAND, D. H., STOFFEL, S., SCHARF, S. J., HIGUCHI, R., HORN, G. T., MULLIS, K. B. AND ERLICH, H. A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239,** 487–491.

SUN, F. (1995) The polymerase chain reaction and branching processes. *J. Comp. Bio.* **2,** 63–86.

SUN, F., ARNHEIM, N. AND WATERMAN, M. S. (1995) Whole genome amplification of single cells: mathematical analysis of PEP and tagged PCR. *Nucl. Acids Res.* **23,** 3034–3040.

TELENIUS, H., CARTER, N. P., BEBB, C. E., NORDENSKJOLD, M., PONDER, B. A. AND TUNNACLIFFE, A. (1992) Degenerate oligonucleotide-primed PCR—general amplification of target DNA by a single degenerate primer. *Genomics* **13,** 718–725.

WEISS, G. AND von HAESELER, A. (1995) Modeling the polymerase chain reaction. *J. Comp. Bio.* **2,** 49–62.

WHITE, T. J., ARNHEIM, N. AND ERLICH, H. A. (1989) The polymerase chain reaction. *Trends Genet.* **5,** 185–89.

ZHANG, L., CUI, X., SCHMITT, K., HUBERT, R., NAVIDI, W. AND ARNHEIM, N. (1991) Whole genome amplification from a single cell: Implications for genetic analysis. *Proc. Nat. Acad. Sci.* **89,** 5847–5851.