Errors induced during PCR amplification

Hoda Sharifian

May 30, 2010
Abstract

The Polymerase Chain Reaction (PCR) is one of the most widely used techniques in modern molecular biology to amplify a single or few copies of a piece of DNA sequences. The PCR products are used as inputs of many other applications such as diagnosis of diseases. Therefore the accuracy of the further analysis depends on the quality of the PCR outputs. In order to study the amplification process in PCR more carefully, the mathematical modelling is essential. Since 1989 different models are presented in the literature.

In the perfect PCR all the replicated molecules are the exact copies of the original ones and the number of molecules is doubled after each replication cycle. But in practice these two conditions cannot be fulfilled. All the present mathematical models consider dissatisfaction of at least one of these conditions.

The first condition says that all the replicated sequences should be exactly the same. But in the real PCR in addition to the desired sequences, some artificial molecules are produced. These artificial sequences can be the results of mutation or they are generated because of undesired reactions between different template molecules. These artificial molecules are called “Chimera” and “Heteroduplex” in the literature. The fraction of the mutant sequences can be calculated by the distribution of number of mutations. In this thesis mutation is simulated as a stochastic process to get a distribution of number of mutations. The results show that the ratio of mutants depends on the multiplication of the sequence length and mutation rate. The mutation rate is determined by the performance of the DNA polymerase. So using an accurate polymerase, about 98% of the sequences are amplified without mutation when they are not so long (e.g. 100 bp). But the probability to have mutants is higher for longer sequences (e.g. 1000 bp). The results show that about 12% of the generated sequences have one mutation even by using an accurate enzyme.

Mutation may also change the pairwise Hamming distances between sequences when the PCR inputs are mixture of different genes. Mutation may cause the pairwise distances between different genes decrease or increase or even remain constant. When the sequences are very close to each other and the Hamming distances change a lot, it is not easily possible to detect the original genes by comparing their pairwise distances. So the changes of the pairwise Hamming distances during replication cycles are also simulated and the distribution of the pairwise Hamming distances is presented for different normalized Hamming distances between sequences. The results show that when the sequences are very distant it is more likely that a mutation happens but the pairwise distance does not change, but for closer sequences a mutation usually changes the Hamming distance. They also imply that the sequences get further when they are close to each other (normalized distances of 5% and 10%) and get closer when they are very distant (normalized distances of 90% and 95%).

The second condition is referred as a PCR efficiency, it is determined in each cycle by the proportion of the sequences which make a copy from themselves to the total sequences in that cycle. So the PCR efficiency will be 100% if all sequences make a copy from themselves. The efficiency is almost constant in the first replication cycles. But as the number of molecules grows exponentially, the environment is fully condensed by the DNA sequences and a smaller proportion of them participate in the replication process. So the efficiency or amplification rate
decreases during PCR cycles. When the PCR input sequences are mixture of different genes with different ratios the efficiency does not decrease with the same rate for all of them. This problem has been addressed in the literature as a result of the template re-annealing. Different amplification rates cause a bias towards amplifying some genes more than the others. Some other reasons are also reported which give such bias such as GC content and primer mismatch. In this thesis a mathematical model which describes template re-annealing is simulated. The results are given for some cases based on the experimental data. They show that how the genes ratios change during the PCR cycles and how the amplification rates decrease because of the template re-annealing. The simulation is run for different conditions to study the effective parameter and to find a proper PCR setting to remove this bias.
# Contents

1 Introduction ......................................................... 7
   1.1 Principles of the PCR ........................................ 7
   1.2 Modeling PCR .................................................. 9
       1.2.1 Galton-Watson process ....................................... 10
       1.2.2 Application .................................................. 12
       1.2.3 Mathematical Model ......................................... 13
   1.3 PCR Artifacts .................................................... 15
       1.3.1 Chimera ...................................................... 15
       1.3.2 Heteroduplex ................................................ 16
   1.4 PCR amplification bias .......................................... 17
       1.4.1 PCR Selection and Drift ...................................... 17
       1.4.2 GC content .................................................... 18
       1.4.3 Primer mismatch and Annealing temperature ................. 18
       1.4.4 Template Re-annealing ...................................... 19
   1.5 Summary .......................................................... 19

2 Material and Methods .................................................. 21
   2.1 Distribution of Number of Mutations ............................. 21
       2.1.1 Method ........................................................ 21
       2.1.2 Algorithm ...................................................... 24
   2.2 Pairwise Hamming Distances ...................................... 25
       2.2.1 Method ........................................................ 25
       2.2.2 Algorithm ...................................................... 27
   2.3 Simulation of Template Re-annealing ................................ 28
       2.3.1 Method ........................................................ 28
       2.3.2 Algorithm ...................................................... 30
   2.4 Summary ............................................................ 31

3 Results ........................................................................ 32
   3.1 Distribution of Number of Mutations ............................. 32
   3.2 Pairwise Hamming Distances ...................................... 36
   3.3 Template Re-annealing ............................................. 39

4 Discussion .................................................................... 48
List of Figures

1.1 Three steps of PCR, [23] .................................................. 9
1.2 Exponential amplification in PCR, [23] .................................. 9
1.3 The Galton-Watson process .................................................. 10
1.4 Formation of chimera from an incompletely extended primer, [6] ........ 16
1.5 Formation of chimera by template switching, [6] ....................... 16
2.1 The replication process by a binary tree .................................. 27
3.1 Distribution of the number of mutations for \( l = 100 \) by the analytical solution .......... 33
3.2 Distribution of the number of mutations for \( l = 1000 \) by the analytical solution .......... 33
3.3 Distribution of the number of mutations for \( l = 100 \) by the simulation algorithm .......... 34
3.4 Distribution of the number of mutations for \( l = 1000 \) by the simulation algorithm .......... 34
3.5 Distribution of the number of mutations for \( l = 100 \) by the simulation algorithm
    according to the real experiment settings ........................................ 35
3.6 Distribution of the number of mutations for \( l = 500 \) by the simulation algorithm
    according to the real experiment settings ........................................ 35
3.7 Distribution of the number of mutations for \( l = 1000 \) by the simulation algorithm
    according to the real experiment settings ........................................ 36
3.8 Hamming distance between two DNA of length 500 which are 20% different .. 36
3.9 Hamming distance between two DNA of length 500 which are 80% different .. 37
3.10 Distribution of the pairwise Hamming distances for sequence length of 50 bp .. 37
3.11 Distribution of the pairwise Hamming distances for sequence length of 100 bp .. 38
3.12 Distribution of the pairwise Hamming distances for sequence length of 1000 bp . 38
3.13 Sequences with different pairwise Hamming distances, \( l = 1000 \) bp ............... 39
3.14 Concentrations of template and primer and amplification rates for four genes with
    initial ratios of 0.7, 0.1, 0.15 and 0.05, \( k = 10 \) .......................... 40
3.15 Concentrations of template and primer and amplification rates for four genes with
    initial ratios of 0.7, 0.1, 0.15 and 0.05, \( k = 2 \) .......................... 42
3.16 Concentrations of template and primer and amplification rates for four genes with
    initial ratios of 0.5, 0.1, 0.25 and 0.15, \( k = 10 \) .......................... 43
3.17 Concentrations of template and primer and amplification rates for four genes with
    initial ratios of 0.5, 0.1, 0.25 and 0.15, \( k = 2 \) .......................... 44
3.18 Ratio between final and initial frequencies for four different genes in four PCR runs 45
3.19 Concentrations of template and primer and amplification rates for two genes with
    initial ratios of 0.8 and 0.2, \( k = 10 \) .......................... 45
3.20 Concentrations of template and primer and amplification rates for two genes with initial ratios of 0.8 and 0.2, $k = 2$  
3.21 Effect of adding more primer, $k = 10$  
3.22 Effect of adding more primer, $k = 2$  
3.23 Ratio between final and initial frequencies for two different genes in four PCR runs
Chapter 1

Introduction

In 1971 a paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method using an enzymatic assay to replicate a short DNA template with primers in vitro [9]. This early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis [16]. The polymerase chain reaction (PCR) is an in vitro method for producing large amounts of identical copies of a specific region of a DNA strand (the DNA target) from small amount of molecules [13]. This technique is widely used in human genetics, forensic science, infectious disease diagnosis, cancer research, and evolutionary and developmental biology.

1.1 Principles of the PCR

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. The reaction mixture of the PCR, in a reaction volume of 10-200 μl, contains several components:

- DNA template that contains the DNA region (target) to be amplified.
- Two primers\(^1\) that are complementary to the 3’ ends of each of the sense and anti-sense strand of the DNA target.
- DNA polymerase\(^2\) with a temperature optimum at around 70°C.
- Deoxynucleoside triphosphates (dNTPs), the building blocks from which the DNA polymerases synthesizes a new DNA strand.
- Buffer solution, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- Divalent cations, magnesium or manganese ions; generally \(Mg^{2+}\) is used.
- Monovalent cation potassium ions.

\(^1\)A primer is a strand of nucleic acid that serves as a starting point for DNA or RNA synthesis.
\(^2\)A DNA polymerase is an enzyme that catalyzes the polymerization of deoxyribonucleotides into a DNA strand.
The whole PCR process is composed of 30 to 40 repeating cycles. Each PCR cycle consists of three major steps, DNA denaturing, primer annealing, and polymerase extension. The schematic view of these three steps is reported in figure 1.1.

1. **Denaturation**

   The reaction mixtures are heated to approximately 90-95°C. During the denaturation, the double strand melts open to single stranded DNA, all enzymatic reactions stop (e.g. the extension from a previous cycle).

2. **Annealing**

   The single-stranded sequences generated by denaturing are used as templates for the primers and the DNA polymerase. The mixture is now cooled to a temperature (generally, between 40-60°C) that permits annealing of the primer to the complementary sequences in the DNA. Because DNA sequences can only grow from 5’ to 3’, the primers are oriented so that the 3’ end of each primer is directed toward the target sequence. As a rule, these sequences are located at the 3’ -ends of the two strands of the segment to be amplified.

3. **Extension**

   The temperature is raised again to 75-80°C. This is the ideal working temperature for the DNA polymerase. The primers, which are usually composed of 18-24 bases[14], already have a stronger ionic attraction to the template than the forces breaking these attractions. Primers that are on positions with no exact match, get loose again (because of the higher temperature) and do not give an extension of the fragment. The bases (complementary to the template) are coupled to the primer on the 3’ side (the polymerase adds dNTP’s from 5’ to 3’, reading the template from 3’ to 5’ side, bases are added complementary to the template). The extension step is shown more clearly in figure 1.1. The bases which are available in the reaction mixture are added to the 3’ side of the primer in a way that adenine binds to thymine and cytosine binds to guanine. This polymerization is catalyzed by the DNA polymerase.

   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, it reaches a theoretical maximum of \( 2^n \) fold amplification (Figure 1.2).

   But in practice the molecules do not replicate perfectly at each cycle. One of the common errors during replication is mutation, because occasionally DNA polymerase incorporates a wrong nucleotide to the growing DNA chain. Different error rates have been reported for different DNA polymerases [7]. Another problem is that the amplification is not being done totally efficient in every cycle. Sometimes primers do not anneal to the templates or even if primers anneal to the templates, the primers can not be extended beyond the position of the primer on the opposite strand [20]. Since all the molecules cannot make a copy from themselves, the efficiency is not 100% in every cycle and the number of molecules cannot be duplicated after each cycle.

   According to the important role of the PCR in different applications, the PCR process has been studied in more detail to reveal the reasons of the bias and other errors during amplification.
Having a mathematical model of PCR helps controlling some sources of errors and extract useful information from the data. In the next chapter some different PCR models presented in the literature are described.

1.2 Modeling PCR

The replication process of PCR is described in terms of Galton-Watson process [10, 20, 26]. The Galton-Watson process is the oldest, simplest and best known branching process [8]. The branching process is a system of particles (individuals, cells, molecules, etc.) which live for a random time and, at some point during lifetime (or at the moment of death), produce a random number of progeny. This section presents Galton-Watson process and its application in modelling PCR.
1.2.1 Galton-Watson process

The Galton-Watson process can be described as follows. A single ancestor particle lives for exactly one unit of time, and at the moment of death it produces a random number of progeny according to a prescribed probability distribution. Each of the first-generation progeny behaves independently of each other, as the initial particles did. It lives for a unit of time and produces a random number of progeny. This process repeats for each of the second-generation progeny and so forth. It is different from other branching process in the life spans of the individuals. In the general branching process the waiting time until the next progeny is generated is estimated from a probability distribution, but in the Galton-Watson process the life spans of all particles are identical and equal to 1. So the process can be mathematically described using a discrete-time index, identical to the number of successive generations [8].

Mathematical Formulation Figure 1.3 shows the schematic view of the Galton-Watson process. At the beginning the ancestor generates $X$ offsprings and they initiate $X$ similar subprocesses. The number of particles in the generation $n + 1$, $(Z_{n+1})$ of the process is equal to the sum of the particle counts in the generation $n$ initiated by the first-generation offsprings of the ancestor particle. Let $Z_{n}^{(j)}$ denote the number of individuals at time $n$ in the process from the $j$th identical independent distribution (iid). Mathematically the random variable (rv) $Z_{n+1}$ is equal to the sum of $X$ random variables $Z_{n}^{(j)}$,

$$Z_{n+1} = \sum_{j=1}^{X} Z_{n}^{(j)}$$

(1.1)

Figure 1.3: The Galton-Watson process

A useful tool for handling distributions of such random sums is the probability generating function (pgf) of a distribution [8]. Probability generating function is the basic analytical tool
employed to deal with non-negative rv’s and finite and denumerable sequences (vectors) of such variables. Let us denote the set of non-negative integers by $Z_+$. Let $X$ be a $Z_+$-valued rv, such that $P[X = i] = p_i$.

**Definition** The pgf $f_X$ of a $Z_+$-valued rv $X$ is a function $f_X(s) = E(s^X) = \sum_{i=0}^{\infty} p_is^i$ of a symbolic argument $S \in U \equiv [0, 1]$.

**Theorem** Suppose $X$ is a $Z_+$-valued rv with pgf $f_X(s)$, if $Y$ is a $Z_+$-valued rv and $\{X^{(i)}, i \geq 1\}$ is a sequence of iid $Z_+$-valued rv’s independent of $Y$, then $V = \sum_{i=1}^{Y} X^{(i)}$ has the pgf $f_V(s) = f_Y[f_X^{(1)}(s)]$.

By using this theorem, the pgf of the equation 1.1 can be written as:

$$f_{n+1}(s) = f[f_n(s)] \quad (1.2)$$

Assuming $Z_0 = 1$ implies $f_0(s) = s$, which yields:

$$f_n(s) = f^{(n)}(s) = f[f[f[s]] \cdots] \quad (1.3)$$

it says that the pgf of $Z_n$ is the $n$th functional iterate of the progeny pgf $f(s)$.

**Moments** It can easily be shown that the mean of the particles can be expressed by the first derivative of $f(s)$ at $s = 1$:

$$\frac{df(s)}{ds} = \sum_{i=1}^{\infty} ip_is^{i-1} \quad (1.4)$$

$$f'(1) = \sum_{i=1}^{\infty} ip_i = E[X = i] \quad (1.5)$$

Let $m$ be the expected number of particles in the first generation, $E[Z_1] = m$. Using the above equation, $E[Z_1] = f'(1) = m$ and $E[Z_n] = f'_n(1)$. Differentiating the equation 1.3 with respect to $s$ and substituting $s = 1$, the mean of the particles in generation $n$ can be obtained using the chain rule of differentiation:

$$E[Z_n] = f'_n(1) = f'_{n-1}(1)f'(1) = \cdots = m^n \quad (1.6)$$

Similarly, the variance can be calculated using the chain rule for the second derivative:

$$Var(Z_n) = \begin{cases} \sigma^2 m^{n-1}(m^n-1), & m \neq 1 \\ \frac{m-1}{m^2}, & m = 1, \end{cases}$$

where $\sigma^2 = Var(Z_1)$ is the variance of the progeny count.
The Galton-Watson process has been extensively used to model replication process in the PCR [10, 20, 26]. The summary of some of the existing methods is presented here and then the mathematical model which is suggested in [20] is explained in more detail in section 1.2.3.

The PCR process is initiated with \( S_0 \) individuals. In the case of PCR, it refers to the \( S_0 \) initial molecules or single-stranded DNA. In every generation which is equal to one PCR cycle, any of DNA sequences can generate one copy of itself. In perfect PCR each molecule doubles in each generation. But in practice some of the sequences do not generate an offspring and if they do, they are not exactly similar to the parental sequences. The first problem refers to the efficiency of PCR and the second one is because of the mutation during replication as they were discussed in section 1.1.

The simplest model is described by Krawczak et al. in 1989 [10]. They constructed the mathematical model for the PCR mutations to count the proportion of the final products with no mutation after \( n \) PCR cycles. This is a basic and simple model which assumes the efficiency of PCR is 100% and a single-stranded sequence is falsely copied with probability \( p \) in each replication cycle. The random variable \( X \) which shows the number of correct fragments resulting from the replication is distributed as follows:

\[
P(X = 1) = p, \quad P(X = 2) = 1 - p
\]  

The distribution of \( X \) has mean \( m = 2 - p \) and using equation 1.6 the expected number of correct single-stranded fragments after \( n \) cycles is \( E[S_n] = S_0m^n \) where \( S_0 \) is the initial number of single-stranded templates. Since it is assumed that the PCR efficiency is 100%, there are \( 2^n S_0 \) replicates after \( n \) PCR cycles and the expected proportion of the correct amplified sequences among all replicates is estimated by:

\[
E\left[\frac{S_n}{2^n S_0}\right] = \left(1 - \frac{p}{2}\right)^n
\]  

However as it was mentioned before, the PCR never works 100% efficient, so one year later Hayashi [5] considered this problem by assuming that the efficiency is \( \lambda \), \( (0 \leq \lambda \leq 1) \) and that the number of mutations per single-stranded sequence per cycle of amplification is estimated by a Poisson distribution with mean \( \mu G \) where \( \mu \) is the error rate of the polymerase per nucleotide, per cycle and \( G \) is the length of the target. Then the expected fraction of the PCR products having no mutation is:

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

Sun used the same assumptions to consider efficiency and mutation in his model, but he defined a novel concept of the generation number and calculated the conditional probability to have \( M \) mutations in a randomly chosen sequence given it is a \( k^{th} \) generation sequence [20] (described in detail in section 1.2.3). In the following of his study, he proposed the distribution of the pairwise differences between two randomly chosen sequences and used it to estimate the mutation rate from a sample of the PCR products. When the nucleotide bases of the target molecules which to be amplified are not known, it is not possible to count the number of mutations of the sampled PCR products. In this case the pairwise differences among the sampled molecules can be compared. Let \( H_{i,j} \) be the pairwise Hamming distance, i.e. the number of
different bases between sequence \(i\) and sequence \(j\), then the mutation rate can be estimated as follows,

\[
\mu = \frac{\sum_{i \neq j} s_{i,j}}{\binom{s}{2} ED \times G} \quad (1.11)
\]

where \(s\) is the number of sampled sequences from the PCR products, \(ED = \frac{2n\lambda}{1+\lambda} - \frac{2}{(1+\lambda)S_0 + 1-\lambda} + O\left(\frac{1}{S_0(1+\lambda)^n}\right)\) and \(S_0\) is the initial number of molecules.

In 1995 Weiss and von Haeseler proposed another model for PCR [26]. They calculated the distribution of the number of replications between two molecules and approximated it by a binomial distribution. So the probability that two randomly chosen molecules after \(n\) PCR cycles are \(t\) replications apart is estimated by,

\[
P\{X = t\} \approx \binom{2n}{t} \frac{\lambda^t}{(1+\lambda)^{2n}} \quad (1.12)
\]

To obtain a distribution of pairwise differences, they assumed a substitution process that represents the errors made by the polymerase during amplification. They used an infinite sites model [25] and like the methods described above they modeled the process of nucleotide substitution by a Poisson process with parameter \(\mu G\). Therefore the probability to observe \(D = d\) differences between two randomly chosen molecules, given \(t\) replications between them is given by,

\[
P\{D = d|X = t\} = \frac{(\mu G t)^d}{d!} \exp(-\mu G t). \quad (1.13)
\]

The marginal probability to observe \(D = d\) differences, is obtained by using equations 1.12 and 1.13 and summing over all possible number of replications between two random molecules after \(n\) PCR cycles,

\[
P\{D = d\} = \sum_{t=0}^{2n} P\{D = d|X = t\} P\{X = t\}. \quad (1.14)
\]

Finally they approximated the number of expected pairwise differences by,

\[
2nG\mu \frac{\lambda}{1+\lambda} \quad (1.15)
\]

The summary of some basic models of PCR were reviewed in this secton, in the following section the mathematical model of the PCR using branching process is described. This model is based on the paper from Sun [20].

1.2.3 Mathematical Model

The PCR process starts with \(S_0\) identical copies of single-stranded sequences. During each cycle, every template generates a new sequence with probability \(\lambda\) and itself always remains in the products. This process is repeated for \(n\) cycles and \(S_n\) is the number of sequences in cycle \(n\). \(S_0, S_1, \cdots, S_n\) form a Galton-Watson process. The expected number of sequences generated from one template is:

\[
m = 1 \times (1-\lambda) + 2 \times \lambda = 1 + \lambda \quad (1.16)
\]

Therefore the expected number of sequences after \(n\) cycles starting from \(S_0\) sequences is:

\[
E[S_n] = S_0(1+\lambda)^n \quad (1.17)
\]

In this paper [20] it is assumed that:
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 \( \lambda \) is constant.

2. The behavior of each template does not depend on the behavior of other templates. So each template generates a complete copy independent of other templates if they generate a complete copy or not.

During the synthesis of a new sequence, mutation can occur. In order to take mutation into account, not all templates have been considered as identical. For example, templates generated from the original molecules through two replications are more likely to have more mutations than template molecules generated from the original molecules through one replication. Based on this observation, he introduced a novel concept of generation number. The original sequences are called 0\(^{th}\) generation sequences and the sequences generated directly from \( k^{th}\) generation sequences are called \((k + 1)^{st}\) generation. In fact the number of mutations in each cycle is conditioned on its generation number. It has been proved by induction that the expected number of \( k^{th}\) generation sequences after \( n \) PCR cycles \( \mathbb{E}[X^n_k] \) is \( S_0 \binom{n}{k} \lambda^k \), \( 0 \leq k \leq n, n \geq 1 \). After \( n \) PCR cycles a sequence is chosen randomly, the probability to choose a \( k^{th}\) generation sequence from the total sequences after \( n \) PCR cycles is \( \mathbb{E}[X^n_k/S_n] \). Using equation 1.17 and assuming \( S_0 \) is sufficiently large, this can be approximated by:

\[
P\{K = k\} = \binom{n}{k} \frac{\lambda^k}{(1 + \lambda)^n}
\]

It is indicated that if \( \lambda > 0.85 \), this approximation is good for any \( S_0 \).

An infinite site model is used to describe the process of nucleotide substitution. That is, the number of point mutations in a target sequence of length \( G \) is a Poisson random variable with mean \( \mu G \) per PCR replication. In addition it has been assumed that mutations occur in different places whenever a mutation occurs, so that there are no back mutations. Therefore under above assumptions, the conditional probability to have \( M \) mutations in a randomly chosen sequence given it is from generation \( k \), after \( n \) PCR cycles is,

\[
P\{M = m|K = k\} = \frac{(\mu G)^m}{m!} \exp(-\mu G), \quad m \geq 0
\]

Using 1.18 and 1.19, the marginal probability to to have \( M \) mutations in a randomly chosen sequence after \( n \) PCR cycles can be calculated as follows:

\[
P\{M = m\} = \sum_{k=0}^{n} P\{M = m|K = k\} P\{K = k\}
= \sum_{k=0}^{n} \exp(-k\mu G) \frac{(\mu G)^m}{m!} \frac{\binom{n}{k} \lambda^k}{(1 + \lambda)^n}
= \frac{(\mu G)^m}{m!(1 + \lambda)^n} \sum_{k=0}^{n} \exp(-k\mu G) \binom{n}{k} k^m \lambda^k
= \frac{(\mu G)^m (1 + \lambda e^{-\mu G})^n}{m!(1 + \lambda)^n} E\left(\left. \binom{n}{m} \frac{\lambda e^{-\mu G}}{\lambda e^{-\mu G} + 1} \right| \frac{\lambda e^{-\mu G}}{\lambda e^{-\mu G} + 1} \right)^m
\]
The next step is to obtain the pgf to calculate the mean and variance of the distribution of the number of mutations. Finally the following relations have been obtained for the mean and variance,

\[ E[M] = \frac{n\mu G}{1 + \lambda} \tag{1.21} \]
\[ \text{Var}(M) = \frac{n\lambda \mu G}{1 + \lambda^2} (\mu G + 1 + \lambda) \tag{1.22} \]

It is noted that the distribution of the number of mutations in a randomly chosen sequence can be approximated by \( \text{Poisson}(\frac{n\lambda \mu G}{1 + \lambda}) \) as \( n \) tends to infinity. When \( \mu G \) is relatively small and \( n\lambda G \) is neither too small nor too large.

Using equation 1.20 the proportion of sequences without any replication error is,

\[ P\{M = 0\} = \frac{(1 + \lambda e^{\mu G})^n}{(1 + \lambda)^n} \tag{1.23} \]

which gives the result of [5], equation 1.10. If \( \lambda = 1 \) this gives the result of [10], equation 1.9.

1.3 PCR Artifacts

During the replication process in PCR, some artificial sequences can be produced. The DNA polymerase error is one of the reasons to have artifacts in the PCR products [7]. The abundance of these artifacts are equal to the number of mutants and a way to estimate it is as described in the previous section. Additionally the existence of some other artificial molecules is reported in the literature. One of them is \( \text{chimera} \) which is addressed by Brakenhoff \textit{et al.} [2], and it is also observed by others. In section 1.3.1 two kinds of mechanism for chimera formation are explained. Another artifact is \( \text{heteroduplex} \) molecules which form in the annealing step when the temperature decreases. It is reported as a major source of artifacts in mixed-template PCR [22], more explanations are given in section 1.3.2.

1.3.1 Chimera

Chimera is one of the artifacts that is produced especially at the last PCR cycles when the concentration of the products is high. Chimera is mostly formed by either “incompletely extended primer” or “template switching” [6].

\textbf{Formation of chimera from an incompletely extended primer} Figure 1.4 shows how a chimera might be produced by an incompletely extended primer.

a) The normal process when a primer anneals to its target and primer extension begins.

b) A complementary strand begins to re-anneal to the template strand before the extension is completed.

c) The incompletely extended primer is displaced by continuing re-annealing of the complementary strand. Now this incompletely extended primer may anneal to a secondary template or another incompletely extended strand and results in generating a chimeric molecule.
**Formation of chimera by template switching** Figure 1.5 shows two possible ways of generating chimera by template switching. In a-1) two partially complementary strands are partially re-hybridized and the two primers anneal to their respective targets. In a-2) primer extension begins. The extension continues until two extending strands meet in a-3), then a template-switch occurs and the continued synthesis generates chimeric molecules. The template switching may happen in another way. In b-1) a single template strand partially hybridizes two different template strands (the thick and the thin lines). Then the primer anneals to its target. In b-2) the primer is extended until it meets the branch. In b-3) the extending strand switches to the other template and a chimera is generated.

![Diagram of chimera formation](image)

**Figure 1.4: Formation of chimera from an incompletely extended primer, [6]**

**Figure 1.5: Formation of chimera by template switching, [6]**

### 1.3.2 Heteroduplex

At the denaturation temperature in a PCR cycle, all DNA will be single stranded. When the temperature decreases for annealing, three kinds of duplexes can be formed, a homoduplex between complementary strands, a heteroduplex caused by the cross-hybridization of heterologous sequences, and a duplex between primers and templates. When the primer concentration is much higher than the template concentration, the duplex between primers and templates is most abundant at the annealing step, and then, the homoduplex is produced by primer extension. Therefore, after the extension step, the homoduplex is dominant. On the other hand, in
later PCR cycles, most primers are used up, and the amplified product concentrations are very high. In these conditions, a significant amount of heteroduplex would be formed [6].

Thompson et al. [22] developed a method to eliminate heteroduplexes from mixed-template PCR products by subjecting them to “reconditioning PCR”. They diluted the PCR products 10-fold with a fresh reaction mixture and applied the mixture to PCR for three cycles. This method is based on the principle that formation of a homoduplex DNA will be favored to a heteroduplex DNA in the presence of excess primers. By restoring the initial primer concentrations during the “reconditioning PCR” a denatured DNA molecule will have a higher probability of annealing with a primer than with a heterologous strand, leading to extension of the homoduplex. This idea can also be used to reduce the formation of PCR chimeras from annealing and extension of heterologous DNA fragments by optimizing the number of amplification cycles to maintain an excess of primer through the endpoint of the reaction.

1.4 PCR amplification bias

One of the important applications of PCR is to study the characteristics of multigene families. Using degenerate primers corresponding to highly conserved regions of homologous genes, PCR can be used to detect and identify members of these gene families in samples of genomic DNA or cDNA. In this case that the PCR inputs are a mixture of different genes with different ratios, some factors give a large bias to the initial ratios during the amplification. So in this section some of the possible reasons of these biases in addition the recommended solutions to improve the PCR performance are reviewed.

1.4.1 PCR Selection and Drift

Wagner et al. suggested two major classes of processes leading to PCR bias, PCR selection and PCR drift [24]. PCR selection occurs when the reaction favors certain members of a gene family. A major contributor to PCR selection probably is differential primer affinity due to differences in primary or secondary structure of DNA at potential target sites. PCR drift is the result of random events occurring in the early cycles of the reaction. In this case, the bias will not be repeatable, i.e., separate PCR experiments in general do not produce biases towards the same member of the gene family. They showed that the effects of PCR drift increase for decreasing number of initial molecules. The easiest way to remove this stochastic bias is to repeat the experiment several times and pool the products. In this paper some factors have been mentioned that cause asymmetries in the replication rates including different melting temperatures of different primers in the reaction mixture at the elongation temperature and secondary structure formation of templates in the annealing stage of each cycle by avoiding the primer binding. Finally it has been advised to stop the reaction as early as possible or to use low annealing temperatures for early cycles, followed by a high annealing temperature for later cycles to increase the amount of products obtained. Another solution can be starting reactions with a very small number of templates so that the strong effect of the PCR drift can override the effects of PCR selection.
1.4.2 GC content

The difficulty in amplifying high G + C segments has been addressed in [3, 17]. Dutton et al. have mentioned three problems which contribute to this difficulty: (i) poor denaturation at 94°C, (ii) poor access of the PCR primer due to template secondary structure, and (iii) slow elongation due to template secondary structure [3]. They proposed a new protocol which can circumvent these problems by using four alternative solutions: (i) high temperature denaturation (98°C for one minute), (ii) five minutes of combined annealing/elongation at 70°C, (iii) very heat stable thermal polymerases and (iv) oligonucleotides of Wallace temperatures 80°C or 120°C.

Polz and Cavanaugh did some experiments with mixtures of mutagenized templates containing AT- and GC-rich priming sites [15]. They investigated the effect of varying the gene dosage and different AT-GC contents of the degenerate primers as potential major causes of PCR selection. Under the conditions of their experiment, they found that the gene dosage alone had no discernible effect on product ratios however the templates containing the GC-rich permutation were amplified with higher efficiency, indicating that different primer binding energies may be responsible for over amplification. The bias because of the PCR drift has been minimized by using high starting concentration. But the differences in the CG content at degenerate positions in the primer target sites were the main contributor of the PCR selection. The reason of this bias can be found by studying the chemical structures of the nucleotides. Because both G and C form a triple hydrogen bond, the melting temperatures of the CG-rich contents of both primers are theoretically about 2°C higher than the AT-rich permutation. Thus, at each annealing step a greater proportion of the templates containing CG complements in the priming region should hybridize to their matched primers. Finally they proposed four recommendation for limiting bias in PCR amplification: (i) avoiding degeneracies when universal primers are designed, (ii) using high template concentrations, (iii) to minimize PCR drift, several replicate PCR amplifications should be combined and (iv) using the smallest number of cycles to diminish PCR selection.

1.4.3 Primer mismatch and Annealing temperature

Primer mismatch is an inherent characteristic of the PCR with “universal” primers, that is, primers binding to phylogenetically highly conserved regions of the genes.

Kwok et al. investigated the effects of various primer-template mismatches on DNA amplification of an HIV-1 gag region by PCR [11]. They concluded that single internal mismatches had no significant effect on PCR product yield while those at the 3’-terminal base had varied effects. Certain mismatches at the terminal 3’ position (i.e. T:C, T:G, T:T, G:G, and A:C) appear to amplify as efficiently in PCR as the fully complementary primer-template duplex, while an A:A mismatch moderately reduced PCR amplification efficiency per cycle, and A:G and C:C mismatches caused dramatic reduction of the efficiency. Their study gives an insight how to design a primer. For example, the design of primers with T rather than A or G at the 3’ terminus may increase the likelihood of extension.

Sipos et al. found a strong bias because of the primer mismatches inherent in the 63F primer, showing a preferential amplification of the template containing the perfect matching

\[ T_d = 4°C \times (G + C) + 2°C \times (A + T), \]

where A, G, C, and T are the number of occurrences of each nucleotide.

---

1 The temperature \( T_d \) at a particular salt concentration, and total strand concentration at which 50% of an oligo and its perfect filter-bound complement are in duplex. The simplest equation for \( T_d \) is the Wallace rule, \( T_d = 4°C \times (G + C) + 2°C \times (A + T) \), where A, G, C, and T are the number of occurrences of each nucleotide.
sequence [19]. They investigated the effect of primer mismatches with two widely used forward primers, namely 27F and 63F. 27F showed no mismatch with all of the target sequences, while the 63F primer had three mismatches close to the 5’ end. The reverse primer (1387R) matched all targets perfectly. As expected, the 27F primer amplified all templates without bias while two of the strains were preferentially amplified by using 63F primer. They also focused on the PCR parameters to optimize the PCR to reduce the bias caused by primer mismatch. They studied the effect of annealing temperature on primer mismatch. They mixed two DNA templates with 1:1 ratio and performed PCR using 63F and 27 primer sets and varying the annealing temperature between 47 and 61 °C. They showed that the deviation from the nearly 1:1 template ratio during amplification with the mismatch primer increases almost exponentially with increasing annealing temperature, this exponential behavior was not observed with the perfectly matching 27F primer. At the end they suggested two ways for reducing preferential amplification, (i) PCR amplification should be optimized to reach the lowest annealing temperature, where the reaction is still specific and unspecific products (mispriming) are not observed. (The cycle number can be high at this step.) (ii) PCR amplification should be repeated at the optimal temperature using parallel samples at low cycle numbers (around 25).

1.4.4 Template Re-annealing

Suzuki and Giovannoni indicated the differences in the amplification efficiencies as a factor which influences product ratios in mixed-template reactions [21]. They reported a bias towards a final ratio of 1:1 regardless of the initial ratio. They could not explain their results by simple models based on G + C bias or priming preference, therefore they developed a kinetic model which took into account the possibility that templates might re-anneal and thereby exclude primers. The central feature of this model is that as the concentrations of product molecules increase, the rate of the molecular reaction in which homologous single-stranded template molecules hybridize with each other will increase as a function of the product concentrations. Since single-stranded molecules must react with free primer to initiate extension reactions, the rate of formation of primer-template hybrids will be influenced by the proportion of template molecules in a single-stranded state. In reactions with mixed templates and high amplification efficiency, the template with the higher initial concentration in the starting mixture reaches inhibitory concentrations sooner while the second template continues to undergo amplification efficiently, and thereby the original difference in concentrations decreases until a 1:1 ratio is achieved. Since this bias happens in the later PCR cycles, it can be removed by stopping PCR earlier.

1.5 Summary

This chapter provided some background information about principles of PCR, modelling PCR and PCR errors. Simulation of the PCR models helps to study and control the effective parameters on the quality of PCR products. As it was described in section 1.2, the replication process of the PCR can be studied by the Galton-Watson process. So in the next chapter, the simulation of the PCR by the Galton-Watson process has been presented in order to find a probability distribution of the number of mutations in the final PCR products. Some other stochastic simulations have been presented to give an insight into the changes of the pairwise Hamming distances between different DNA sequences during the replication cycles. On the other
hand it is desired to find a model which can explain the PCR biases. Regarding the literature which were reviewed above, the kinetic model of Suzuki and Giovannoni [21] is an appropriate model to study the bias because of the template re-annealing. The simulation of this model is also given in the next chapter. This simulation shows how the concentrations of different genes change during the PCR cycles and how the corresponding efficiencies decrease.
Chapter 2

Material and Methods

In section 1.2.3, the mathematical model of PCR based on a paper from Sun [20] was described. Using this model, the distribution of number of mutations can be found analytically. This analytical solution is limited to the case that the amplification efficiency remains constant during the PCR cycles, but with the use of simulation, the changes of the efficiency can be taken into account more easily. So the first simulation in this chapter (section 2.1) is the simulation of the replication process in PCR and the results will be compared with the analytical solutions based on the Galton-Watson process.

The next simulation in section 2.2 shows the distribution of pairwise Hamming distances between sequences after replication cycles. This distribution has been done especially for the case that the input sequences are the mixture of some homologous genes. The question is that if the initial pairwise Hamming distances between different genes change after \( c \) replications, if they decrease or increase and how they change during the replication.

PCR bias is another factor that results in having different data as they should be in the perfect case. Modeling PCR biases helps to understand how much they affect on the PCR outputs and if it is possible to ignore them. So in section 2.3 the kinetic model of the template re-annealing which was presented in [21] has been simulated. The simulation shows how the initial ratios of different genes vary during the PCR cycles and how this bias can be removed.

2.1 Distribution of Number of Mutations

A replication process in PCR is simulated with considering mutation and variable amplification efficiency. The simulation method is explained in section 2.1.1 and the computational algorithm is described step by step in section 2.1.2. The results are presented in section 3.1.

2.1.1 Method

The simulated replication process starts with \( N_0 \) initial molecules (molecules, sequences or templates, all refer to the single-stranded DNA sequences). Because the efficiency is not 100%, not all of the \( N_0 \) molecules make a copy of themselves. For example 80% of them are selected for replication. Let’s explain the algorithm by using hypothetical numbers. Assume \( N_0 = 100 \), so 80 molecules will be selected to make a copy from themselves. Because of the polymerase error, they cannot make an exact copy from themselves and some of them generate a mutated copy.
Now the question is how many sequences replicate without any error and how many mutations occur in rest of the sequences. Sun [20] and others [5, 26] used a Poisson distribution to estimate the number mutations in each replication. So in this case, 80 samples should be taken from the Poisson distribution with parameter $\mu l$, where $\mu$ is the mutation rate and $l$ is the sequence length. Another way is to take one sample from a multinomial distribution. It is assumed that each molecule replicates independently, so the number of new molecules which will be generated in each cycle can be replaced by the $n$ independent trials which results in exactly one of the $l + 1$ possible outcomes with probabilities $p_i$, $i = 0, 1, \ldots, l$. $p_i$ is equal to the probability to have $i$ mutations in a replicated sequence with length $l$, obviously $p_0$ refers to the probability to produce a wild type sequence. $p_i$ can be obtained from equation 2.1.

$$p_i = \binom{l}{i} \mu^i (1 - \mu)^{l-i}, \quad i = 0, 1, \ldots, l \quad (2.1)$$

A sample from a multinomial distribution is a vector of size $l + 1$ whose $0^{th}$ element is the number of sequences without any mutation, and the $i^{th}$ element is the number of sequences with $i$ mutations.

Let’s return to the example with $N_0 = 100$ and amplification rate of 80%. At the first cycle 80 new molecules will be generated. So a sample is taken from the multinomial distribution with parameters $n = 80$ and probabilities $p_0, p_1, \ldots, p_l$. This sample can be a vector like $s = [70, 7, 3, 0, \ldots, 0]$ which means 70 sequences will be generated without any mutations, 7 with one mutation, 3 with two mutations and none of the sequences has more than two mutations. So after the first cycle the sequences pool is composed of 180 sequences in which there are 170 wild type sequences (100 initial molecules plus 70 new generated ones), 7 sequences with one mutation and 3 with two mutations. The next cycle starts with 180 sequences and $180 \times 0.8 = 144$ new sequences will be generated. Again a sample is taken from the multinomial distribution with parameters $n = 144$ and the same probabilities $(p_0, p_1, \ldots, p_l)$. Assume that this sample is $s = [135, 8, 1, 0, \ldots, 0]$. It says that 135 sequences should be replicated without any error.

Now the question is how to choose 135 sequences out of 180 sequences. Since there are different sequences (wild type, 1-mutation and 2-mutations) with different ratios ($\frac{170}{180}$, $\frac{7}{180}$ and $\frac{3}{180}$), they do not have the same chances to be selected for the replication. In this case the sequences can be sampled from the multinomial distribution with probabilities $[\frac{170}{180}, \frac{7}{180}, \frac{3}{180}]$. So one sample is taken from the multinomial distribution with parameter $n = 135$, it can be $s1 = [130, 3, 2]$. So 130 sequences are chosen out of wild type sequences, 3 out of 1-mutation sequences and 2 out of 2-mutations sequences. Since all of them should be replicated without error, there will be 130 more wild type sequences, 3 more sequences with one mutation and 2 more with two mutations. Now the sequence pool should be updated as follows:

$$\text{sequence pool} = \{170 + 130 = 300, \ \text{wild type sequences}, \$$
$$7 + 3 = 10, \ \text{1-mutation},$$
$$3 + 2 = 5, \ \text{2-mutations}\}$$

Regarding the sample $s = [135, 8, 1, 0, \cdots, 0]$, 8 sequences will be replicated with one error. First one sample should be taken from the multinomial distribution with $n = 8$ and the same
probabilities \( \left( \frac{1}{170}, \frac{7}{150}, \frac{3}{150} \right) \). It is assumed that all the sequences are sampled simultaneously so the probability vector does not change after each sampling. This new sample is \( s2 = [7, 1, 0] \) which means 7 sequences are chosen out of wild type sequences to be replicated with one error and there will be 7 more sequences with one mutation. Only 1 sequence is chosen out of 1-mutation sequences and replicates with one error, so there will be 1 more sequences with two mutations. As it was assumed by Sun [20], there is no back mutation and one position does not mutate two times. Then the sequence pool changes to:

\[
\text{sequence pool} = \{170 + 130 = 300, \ \text{wild type sequences}, \]
\[
7 + 3 + 7 = 17, \ \ 1 - \text{mutation},
\]
\[
3 + 2 + 1 = 6, \ \ 2 - \text{mutations}\}
\]

One more sequence should be selected to be replicated with two errors. It is very likely to select a wild type sequence, so \( s3 = [1, 0, 0] \) and the new sequence pool is:

\[
\text{sequence pool} = \{170 + 130 = 300, \ \text{wild type sequences}, \]
\[
7 + 3 + 7 = 17, \ \ 1 - \text{mutation},
\]
\[
3 + 2 + 1 + 1 = 7, \ \ 2 - \text{mutations}\}
\]

Therefore at the end of the second replication cycle the number of mutations is distributed as follows:

<table>
<thead>
<tr>
<th>Number of mutations</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.93</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
</tr>
<tr>
<td>3 and more</td>
<td>0</td>
</tr>
</tbody>
</table>

The same procedure repeats until the given number of cycles. 

Weiss and von Haeseler [27] estimated cycle specific efficiencies by a stepwise function using the following formula,

\[
\frac{E[S_j]}{E[S_{i-j}]} = (1 + \lambda_i)^j, \quad i \geq j
\]

(2.2)

where \( \lambda \) is the efficiency and \( E[S_j] \) is the expected number of sequences after cycle \( j \). They applied this formula to the published dataset by Saiki et al. [18] and found these values:

\[
\lambda_i = \begin{cases} 
0.872, & 1 \leq i \leq 20, \\
0.743, & 21 \leq i \leq 25, \\
0.146, & 26 \leq i \leq 30.
\end{cases}
\]

(2.3)

In this simulation the variable efficiency is considered with the above relation.
2.1.2 Algorithm

A Python code has been developed to derive a distribution of number of mutations in PCR products. The input parameters of this code are as follows:

- $n$: total number of initial molecules,
- $l$: length of the DNA sequence,
- $c$: number of cycles,
- $u$: mutation rate.

In order to make the simulation closer to the real experiment, a small change has been done with respect to the method described in the previous section. In this PCR experiment (which is going to be studied by simulation), the input sequences are the RNA sequences which should be converted to the cDNA sequences before amplification by PCR. These conversion from RNA to cDNA may also be done with error. In this step the copy number does not change, but the sample sequences may mutate before entering the replication cycles [Francesca Di Giallonardo, personal communication]. So this code has been developed in a way to simulate this process.

With considering the cDNA synthesis as cycle 0, the simulation is being done following these steps:

1. Read the input parameters from the command line,
2. cDNA synthesis
   (a) Set the cycle index ($CI$) equal to 0, $CI = 0$,
   (b) Calculate the probabilities ($pr_{mutation}$) using equation 2.1,
   (c) Take a sample from the multinomial distribution with parameters $n$ and $pr_{mutation}$,
   (d) Save the number of sequences with their respective number of mutations in the sequence pool,
   (e) Calculate the probabilities ($pr_{selection}$) by calculating the sequences ratios for selecting any kind of existing sequences (wildtype, 1-mutation, ...) in the pool.
3. DNA replication
   (a) $CI = 1$,
   (b) Calculate the probabilities ($pr_{mutation}$) using equation 2.1, these probabilities may be different from those for cDNA synthesis because of different mutation rate.
   (c) Calculate the number of new sequences which are generated in this cycle, $\text{new} = \text{int}^1(\text{eff} \times n)$, $\text{eff}$ is the efficiency of the current cycle which is calculated using equation 2.3
   (d) Take a sample from the multinomial distribution with parameters $\text{new}$ and $pr_{mutation}$, this sample (seq.num) has $m$ members, $i^{th}$ member shows the number of sequences with $i$ mutation.

$^1$Returns the integer part of the input argument
(e) For all $m$ members, take a sample from the multinomial distribution with parameters $\text{seq.num}[i]$ and $\text{pr.selection}$, $i = 0, \ldots, m$,

(f) Update the sequence pool by saving the number of sequences with their respective number of mutations,

(g) Update $n$ by $n = n + \text{new}$,

(h) Update $\text{pr.selection}$ by calculating the sequences ratios,

(i) Increase the number of cycles by one, $\text{CI} = \text{CI} + 1$,

(j) Return to step (3c) for the given number of cycles.

4. Return the sequence pool and final number of sequences,

5. Calculate the frequencies for the members of the sequence pool,

6. Plot the frequencies which give the distribution of the number of mutations.

This algorithm simulates one example from the Galton-Watson process. It can be matched to the Galton-Watson process in this way; a particle lives for exactly one unite of time which refers to one PCR cycle, and at the moment of death it produces a random number of progeny according to a prescribed probability distribution. A random number of progeny is 0 or 1, but it is possible to assume that one individual dies and then a random number of progeny is 1 or 2. In other words there is no extinction in the PCR. The number of molecules which will be produced at each cycle is dependent on the the efficiency and the current individuals in that cycle. In fact the Galton-Watson process is used to count the total number of individuals. Any progeny can be a mutant individual, so the mutation should be superimposed on the Galton-Watson process. Of course mutation does not change the total number of individuals, but the number of mutant progenies can be calculated by using the Galton-Watson process.

2.2 Pairwise Hamming Distances

Mutation in the case of multi-template PCR can be more complicated. In the multi-template PCR, the input sequences might be very close to each other. There would be two or more different DNA sequences which can be less than 10% divergent. (The percentage of diversity refers to the normalized Hamming distances between sequences.) The wrong base pairing can be a major problem if it changes the Hamming distances between sequences such that it is not possible to recognize between different genes in the final PCR product. This can be studied using the distribution of number of mutations. If the distribution shows that the probability to have more than 5 mutations is zero, then the mutation does not cause this problem for the sequences which are different in e.g. 20 positions. But when the sequences are closer to each other, the distribution of the pairwise Hamming distances should be studied more carefully.

2.2.1 Method

Assume that there are two sequences with $n$ nucleotides which $D$ of them are different, so the normalized Hamming distance between them is $D/n = d$. If one mutation happens (with mutation rate $\mu$) in any of $D$ different nucleotides and it mutates exactly to the nucleotide
which is on the same position of the other sequence, then the Hamming distance between them reduces. The probability to have such mutation is denoted by $P_{\text{dec}}$ and calculated by,

$$P_{\text{dec}} = \frac{1}{3} D \mu (1 - \mu)^{n-1} \quad (2.4)$$

On the other hand the probability that a mutation increases the Hamming distance between two sequences is $P_{\text{inc}}$, there can be one mutation in any of $D - n$ similar nucleotides or one of the $D$ different nucleotides can mutate to either of two other nucleotides which do not exist in either of sequences, so

$$P_{\text{inc}} = \mu (1 - \mu)^{n-1} \left( \frac{n-D}{n} + \frac{2D}{n} \right) \quad (2.5)$$

From these equations it is observed that the probability of decrease is smaller than increase of distance,

$$\frac{P_{\text{dec}}}{P_{\text{inc}}} = \frac{D}{n - \frac{D}{3}} = \frac{d}{\frac{d}{3} + \frac{D}{3}} = \frac{d}{3 - d} \quad (2.6)$$

So the distribution is not symmetric around the initial Hamming distances.

But one sequence may mutate several times until the last PCR cycle. It is desired to see how the distances between sequences change during the replications. For this, the following simulation has been performed.

To see the changes of the Hamming distance during the replications, it is not necessary to simulate a PCR process exactly. This simulation will be done faster and easier if the number of molecules does not increase after each replication. What is going on in the PCR is that one sequence makes a copy from itself and in the next cycle both of them can participate in the following replication. So after $c$ PCR cycles, one molecule is at most $c$ replications apart from the original molecule. Figure 2.1 makes it more clearly. It explains a three cycles replication process by a binary tree. The original molecule is colored yellow. After the first replication it replicates to a blue molecule but it stays to make another copy in the next cycle (the red one). So the blue and red molecules are one replication apart from the original one. All the molecules do the same as the original one, they replicate to a new molecule and participate in the next replication cycle. The last line of the figure shows the number of replications to produce each molecule after three cycles. It can be seen that the purple molecule is generated by three replications, which is the maximum number of replications. Mutation happens only during replication. So it is more likely to have more mutations when the molecules are generated through more replications. Therefore to study the critical case, it is enough to simulate the process which yields to the purple molecule. The original molecule (yellow one) replicates to the blue one, then the blue one replicates to the green one and finally the green one replicates to the purple molecule. Each replication can be done with error. The Hamming distance between the purple and yellow molecule shows the maximum distance between the PCR product and the original molecule.

When there are two random sequences with the given Hamming distance, each of them is replicated independently. So one binary tree like figure 2.1 can be assumed for each sequence. For each sequence the replication process is simulated independently by following the process which yields to the purple molecule as described above. The distance between two purple molecules belonging to two sequences shows how many bases the sequences are different from each other after replication cycles.

This process can be simulated by taking a random DNA sequence with a given length and let it to replicate during the PCR cycles. In each replication cycle a random number is chosen.
and compared with the probability to have a mutant, this probability is calculated by equation 2.1 which gives the probability to have $i$ mutations in a sequence of length $l$. If the random number was smaller, the replication is done with error, otherwise the same sequence is delivered to the next cycle.

### 2.2.2 Algorithm

In this section, the algorithm of a computer code is explained which generates two random DNA sequences with the given distance between them. The changes of the Hamming distance between these two sequences are calculated during the replication cycles. This code takes four parameters as inputs:

- $l$: length of the DNA sequence,
- $d$: normalized Hamming distance between two DNA sequences,
- $c$: number of cycles,
- $u$: mutation rate,

It works simply by following these steps:

1. Read the input parameters from the command line,
2. Calculate the probability to have a mutant by equation 2.1,
3. Generate two random DNA sequences with the given length $l$ and Hamming distance $d$, $S_1$ and $S_2$,
4. Start the first replication cycle
5. Take a random number and compare it with the probability to have a mutant, if it is less, then update $S_1$ with the mutated sequence,
6. Repeat the previous step for $S_2$,
7. Calculate the Hamming distance between $S_1$ and $S_2$, 
8. Repeat steps (5) to (7) for the given number of cycles $c$.

9. Plot the Hamming distances versus the number of cycles.

This algorithm can be run for more than one sequence (of each type). Then it can be seen that how the final Hamming distances are distributed between sequences. This distribution shows that how many sequences get closer or farther to each other or how many stay with their initial distances.

### 2.3 Simulation of Template Re-annealing

The kinetic model of the template re-annealing based on a paper from Suzuki and Giovannoni [21] is simulated in this section. This simulation helps to understand the bias in template-to-product ratios as a result of template re-annealing.

#### 2.3.1 Method

Let’s start by writing the chemical reactions in the annealing stage. As it was pointed out by Suzuki and Giovannoni, two reactions happen in the annealing step at each PCR cycle. One reaction is the desired one, which is the reaction between a primer and a template. In this reaction a primer anneals to one template, then the polymerase can extend the complementary strand from the primer binding site. Let’s call it “priming”. Another reaction is between to complementary DNA strands. When the concentration of the templates is high enough, a template can compete with a primer to anneal to its complementary strand. This reaction is undesired because it excludes primers and prevents the extension step in which a polymerase makes a new copy. It is called “re-annealing”. These two reactions are described by relations 2.7 and 2.8, where $T_1$ and $T_2$ are complementary templates and $P_1$ refers to a primer. $[T_1P_1]$ is a template-primer duplex and $[T_1T_2]$ is a homoduplex which is made be re-annealing two complementary strands. $K_P$ and $K_A$ are respectively the reaction constants of the priming and re-annealing and $[.]$ denotes the concentration.

\[
T_1 + P_1 \xrightleftharpoons{K_P} T_1P_1, \quad \rightarrow \quad K_P = \frac{[T_1P_1]}{[T_1][P_1]} \quad (2.7)
\]

\[
T_1 + T_2 \xrightleftharpoons{K_A} T_1T_2, \quad \rightarrow \quad K_A = \frac{[T_1T_2]}{[T_1][T_2]} \quad (2.8)
\]

The efficiency of each cycle is defined by the fraction of the new generated templates to the initial templates in that cycle. If all the initial templates make a copy from themselves, the amounts of the new generated templates is equal to the initial templates and the efficiency is equal to 1 in the ideal case.

\[
\text{Efficiency} = \frac{\text{concentration of new templates}}{\text{concentration of initial templates}} \quad (2.9)
\]

Now the corresponding values for the concentration of new and initial templates should be found from the chemical reactions. The available sequences at the beginning of each cycle participate in two reactions, priming and re-annealing. So the number of initial sequences is
equal to the sum of those which bind to primers (primer-template duplexes) and those which 
re-anneal to their complementary strands (homoduplexes). It is assumed that only homologous 
strands can anneal to each other, so this model does not include heterodoxy formation [21].

Number of initial molecules = Number of template–primer duplexes 
+ Number of homoduplexes \[ (2.10) \]

According to the definition of molarity\(^1\), both sides of the above equation are divided to the 
volume to come up with the relation between concentrations.

Concentration of initial templates = \([T_1P_1] + [T_1T_2]\) \[ (2.11) \]

Obviously only template-primer duplexes result in new templates, so,

Concentration of new templates = \([T_1P_1]\) \[ (2.12) \]

By plugging equations 2.11 and 2.12 in equation 2.9, the efficiency of each cycle is defined 
by,

Efficiency = \[\frac{[T_1P_1]}{[T_1P_1] + [T_1T_2]} = \frac{1}{1 + \frac{[T_1T_2]}{[T_1P_1]}}\] \[ (2.13) \]

Now the values of \([T_1P_1]\) and \([T_1T_2]\) in each cycle should be obtained from the concentrations 
of primer and template at the beginning of that PCR cycle. Let’s focus on the chemical reactions 
2.7 and 2.8. Assume \([T_1] = [T_2] = T\) and \([P_1] = P\), then the equations of \(K_P\) and \(K_A\) change to,

\[K_P = \frac{[T_1P_1]}{T \cdot P}\] \[ (2.14) \]

\[K_A = \frac{[T_1T_2]}{T^2}\] \[ (2.15) \]

The constant \(k\) is defined by Suzuki and Giovannoni which is the ratio between the rate 
constants of the re-annealing and priming reactions. So from the above equations \(k\) is equal to,

\[k = \frac{K_A}{K_P} = \frac{P \cdot [T_1T_2]}{T \cdot [T_1P_1]}\] \[ (2.16) \]

By plugging equation 2.16 in equation 2.13, the efficiency of cycle \(n\), \(f(n)\) is obtained as a 
function of \(k\) and concentrations of primer and template at the beginning of cycle \(n\), \(P(n)\) and 
\(T(n)\).

\[f(n) = \frac{P(n)}{P(n) + kT(n)}\] \[ (2.17) \]

\(k\) in the above equation represents the template re-annealing, if \(k = 0\), then the efficiency 
does not decrease because of this problem. In addition Suzuki and Giovannoni assumed an 
initial efficiency (\(f_0\)) which is constant during PCR and it is independent of the re-annealing 
problem. So finally they proposed the following equation for the efficiency,

---

\(^1\) The molar concentration or molarity is expressed as the number of moles of solute per liter of solution. The 
number of moles is calculated by dividing the number of molecules to the Avogadro constant, approximately 
6.023 \(\times 10^{23}\) \text{mol}^{-1}
\[ f(n) = f_0 \frac{P(n)}{P(n) + kT(n)} \] (2.18)

This equation needs primer and template concentrations at the beginning of each cycle, then the efficiency of that cycle is calculated by equation 2.18 and it is used to calculate the concentration of new generated templates in the next cycle. Assume that PCR process starts with \( T_0 \) molar of templates, they will be replicated with efficiency \( f_0 \), so the concentration of the new generated templates is \( f_0 T_0 \), and total concentration increases to \( T_0 (1 + f_0) \) which should be replaced by \( T(n) \) in equation 2.18 to calculate the efficiency \( f(n) \) for the first cycle, \( n = 1 \). The number of primer molecules decreases as many as the total number of templates increases because primer sequences are used to generate new templates. This relation between the number of molecules can be changed to the relation between the molarities, so the primer concentration decreases as much as the template concentration increases. If \( P_0 \) denotes the initial primer concentration, \( P(1) \) will be \( P_0 - f_0 T_0 \) after the first replication cycle. At the end the efficiency of the first cycle, \( f(1) \) is obtained by substituting the values of \( f_0, P(1), T(1) \) and \( k \) in equation 2.18. This efficiency will be used in the next cycle to calculate the concentration of new generated templates.

This study can be easily extended to the case that the template is a mixture of different genes. In in this case a universal primer is used whereas there are different genes with different concentrations. So the primer concentration reduces as much as the total concentration of different genes increases. But the efficiencies are calculated independently for each gene because it is assumed that each template can re-anneal only to its homologous complement and does not inhibit the priming reaction of the other templates [21]. In summary the following recurrence equations should be solved successively,

\[ T(n) = T(n-1)(1 + f(n-1)) \] (2.19)

\[ T_{total} = \sum T_i, \quad i = 1, \ldots, \text{number of genes} \] (2.20)

\[ P(n) = P(n-1) - (T_{total}(n) - T_{total}(n-1)) \] (2.21)

then the efficiency will be updated using equation 2.18.

It should be noted that this model assumed that enough nucleotides exist for each replication, in other words the concentration of new templates does not change because there are not enough nucleotides to make a new copy.

With this background, an algorithm for simulating template re-annealing is described in the next section.

### 2.3.2 Algorithm

These parameters should be defined as the input values:

- \( c \): number of cycles,
- \( t \): the molarity of the templates, (total molarity of different genes),
• $p$: the molarity of the primer,
• $e$: the initial efficiency which was denoted by $f_0$,
• $k$: the ratios between the rate constants of the re-annealing and priming reactions for different genes,
• $f$: the initial frequencies of different genes,

The simulation is being done by following these steps:

1. Read the input parameters,

2. Calculate the initial concentrations of different genes using the initial frequencies, $t[i]=t\times f[i]$, $i=1, \ldots, \text{number of genes}$,

3. Construct a vector $\text{eff}$ for efficiencies of different genes and initiate it by $e$, $\text{eff}[i]=e$, $i=1, \ldots, \text{number of genes}$,

4. Set cycle index ($CI$) equal to 1, $CI=1$,

5. For each gene, add the concentration of new generated templates to the primary concentration in the current cycle using equation 2.19,

6. Calculate the total concentration of the new generated templates ($t_{\text{new}}$) by sum over all genes,

7. Update the primer concentration by subtracting $t_{\text{new}}$ from the primary concentration of the primer in the current cycle,

8. Check if the primer concentration is larger than zero, otherwise stop replication,

9. For each gene, update the efficiency, $\text{eff}[i]$, by equation 2.18, they are used for the next replication cycle,

10. Increase the number of cycles by one, $CI = CI + 1$,

11. Return to step (5) for the given number of cycles,

12. Plot the values of efficiencies, primer and templates concentrations versus the number of cycles.

### 2.4 Summary

In this chapter three different algorithms were presented which simulate different phenomena in PCR. The first two simulations consider the stochastic event of mutation. The effect of mutation were studied in two ways, by calculating the distribution of the number of mutations and by studying the pairwise Hamming distances between DNA sequences. On the other hand, the third algorithm simulates a PCR bias in a deterministic way. Template re-annealing was modeled by solving the chemical equations. In the next chapter these methods are applied to simulate a PCR with some real experimental data. These results from the simulations not only will help to interpret the results from the real experiments but also provide some knowledge to establish better experimental conditions in order to reduce PCR errors and biases.
Chapter 3

Results

In this chapter the results of three simulation algorithms for the PCR process are presented. The first simulation shows the distribution of number of mutations, the results of this simulation will be compared with the analytical solution by Sun [20]. The simulation results in section 3.2 show the variations of pairwise Hamming distances during the replication cycles, in addition some distributions of pairwise Hamming distances after a PCR in some different conditions are presented. The results which are presented in sections 3.1 and 3.2 are from simulating stochastic processes and each run may give different results. The algorithms were run several times and since the results were similar only the results of one run are given. The last section is dedicated to the results of the template re-annealing simulation. The simulation algorithm has been run for different cases to compare the effects of this bias in different situations.

3.1 Distribution of Number of Mutations

First step is to compare the simulation algorithm which was described in section 2.1.2 with the analytical solution from the paper by Sun [20]. In order to have a same setting for both methods, the simulation algorithm should be run for constant efficiency and the cDNA synthesis step should be ignored.

The distribution of number of mutations has been obtained for some typical PCR inputs by both methods. In general 30 cycles should be sufficient for a usual PCR reaction [1]. As it was reported in the literature, the error rate of the Taq DNA polymerase \(^{1}\) is in the order of \(10^{-4}\) per base replication [7]. Assume that PCR works with 85% efficiency. The results of the analytical solution are shown in figures 3.1 and 3.2 for two sequence lengths, \(l = 100\) and \(l = 1000\) nucleotides.

The same inputs have been entered in the simulation algorithm. Here the number of initial molecules can also be set as an input parameter. Sun has approximated the analytical solution for a large number of initial molecules [20]. Now it is possible to check that in which cases this approximation works fine. So the algorithm has been run one time for 10 and one time for 1000 initial molecules.

Figures 3.3 and 3.4 show that the simulation results approve the analytical solution. As it was expected the simulation result is much closer to the analytical solution when the number of

\(^{1}\) Thermus aquaticus, a thermostable DNA polymerase
initial molecules is larger, especially for longer sequences (Figure 3.4).

Now that both analytical and simulation methods came up with similar results, only the simulation method is used to study other effective parameters in the distribution of number of mutations. In the rest of the simulations the number of initial molecules ($n$) is set to 1000 and the efficiency is calculated by equation 2.3. As it was described in section 2.1.2, cDNA synthesis is simulated in cycle 0 before starting replication. In the real experiment this process is being done with the *Roche RT* enzyme, which has an error rate of $10^{-4}$ per base. But the DNA polymerase which is being used in the replication process is the *FastStart Hifi* enzyme which has an error rate of $3.25 \times 10^{-6}$ per base replication [Francesca Di Giallonardo, personal communication]. The number of cycles is one of the PCR parameters which can be set to get higher yield with high quality. So the distribution of number of mutations is plotted for different number of cycles. The simulation is run for three sequences with three different lengths, 100, 500 and 1000 nucleotides.

Figure 3.5 shows the distribution of number of mutations for a sequence with 100 nucleotides. Thanks to the accurate *FastStart Hifi* enzyme only a few mutant sequences exist in final PCR products. But for longer sequences (500 and 1000 bp) it is more likely to have mutants in the products. The output data (they are not shown here, only the plots are presented in the report) state that about 5 to 8 percent of the final sequences have only one mutation when the sequences
are 500 bp, figure 3.6. Whereas for longer sequences (1000 bp) about 11 to 13 percent have one mutation and 1 or 2 percent have 2 mutations, figure 3.7. In all cases it is nearly impossible to have more than three mutations. This implies that if the input sequences are heterogeneous and they are different in more than three positions, mutation does not cause the ambiguity to distinguish between different sequences. The positive effect of this accurate enzyme is more obvious by comparing figures 3.4 and 3.7. The distribution which is shown in figure 3.4 has a peak for one mutation. It shows that most of the products have one mutation using a less accurate enzyme with the error rate of $10^{-4}$ and only about 26% replicate without error. On the other hand the distributions are dependent on the sequence lengths. Amplification of the shorter sequences can be done with much less errors. The analytical solution (equation 1.20) shows that the distribution of number of mutations depends on the multiplication of a mutation rate and sequence length. So using the *FastStart Hifi* enzyme which is 100 times more accurate than *Taq* polymerase (with error rate of $10^{-4}$) for sequence of 10 kbp will result in the same distribution as using a *Taq* polymerase for sequence of 100 bp (100 times shorter).

Figures 3.5 to 3.7 also show the effect of number of cycles on the distribution of number of mutations. Comparing the effects of the mutation rate and the sequence lengths, number of cycles does not change the distribution so much. Figure 3.7 shows that if PCR works for 10
more cycles, the amount of wild type sequences will reduce about $2\%$.\footnote{Normally PCR does not run for more than 40 cycles.}

Finally the results of this section imply that:

1. Using an accurate enzyme the amount of the PCR artifacts because of the mutation reduces nearly to zero for shorter sequences, for longer sequences (e.g. 1kbp) this reduction reaches to about $12\%$.

2. In the case of heterogeneous samples, mutation does not cause a problem in recognizing different genes if they are not very close to each other (In this example the sequences of 1000 bp should be different in more than 3 or 4 positions which means more than $0.4\%$).

3. The number of cycles does not have a large effect on the distribution of number of mutations.

Figure 3.5: Distribution of the number of mutations for $l = 100$ by the simulation algorithm according to the real experiment settings

Figure 3.6: Distribution of the number of mutations for $l = 500$ by the simulation algorithm according to the real experiment settings
3.2 Pairwise Hamming Distances

The simulation results from the previous section imply that only a few mutant sequences will be produced using a DNA polymerase enzyme with low error rate. But in a case that such an accurate enzyme is not available there will be more mutants and it is worth to study how the Hamming distances between sequences change during replication.

The simulation is done with respect to the prescribed algorithm in section 2.2.2. Figure 3.8 shows the variations of the pairwise distances for two sequence of 500 bp which are 20% different. At the beginning they are different in 100 positions. For the lowest mutation rate ($10^{-6}$) the initial distance does not change after 30 replications, but for higher mutation rates it increases with the number of cycles. For the highest error rate the pairwise distance increases from the primary cycles. It is interesting to compare this result to a simulation with two sequences of 500 nucleotides which are different in 400 positions (figure 3.9). In this case the sequences get closer to each other because there are more possibilities to change one of different nucleotides and mutate it to a common one which happens in one third of mutations.
Figure 3.9: Hamming distance between two DNA of length 500 which are 80% different

But mutation is simulated as a random process and the above graphs show the variations of the Hamming distances only between two sequences, so every run gives a slightly different result. Running this simulation for more sequences shows that how the pairwise distances are distributed between sequences. The distribution of pairwise Hamming distances states how often the sequences get closer or farther to each other or stay with their initial distances. This distribution is shown in figures 3.10 to 3.12 for three different sequence lengths, 50, 100 and 1000 bp. In each case the distributions are plotted for four different mutation rates. The mutation rates of $10^{-4}$, $10^{-5}$ and $10^{-6}$ are relevant to the error rates of the DNA polymerase, but normally it does not have a high mutation rate of 0.001, nevertheless this case can be compared with lower error rates. It is assumed that the input sequences are from five different types, and there are 1000 sequences of each type. The normalized Hamming distances between type 1 and other sequences is respectively 5%, 10%, 90% and 95%. The results show the distribution of pairwise distances between type 1 and other sequences after 30 replications.

Figure 3.10: Distribution of the pairwise Hamming distances for sequence length of 50 bp
It can be concluded from these simulations that,

1. using an accurate enzyme most of the sequences stay with their initial Hamming distances,

2. the final pairwise distances change for more sequences when they are longer, this is similar to the conclusion from the distribution of number of mutations. As it can also be seen in figures 3.7 and 3.5, about 12% of sequences of 1000 bp have one mutation whereas this proportion is at most 2% for sequences of 100 bp,

3. when the sequences are very distant it is more likely that a mutation happens but the pairwise distance does not change, but for closer sequences a mutation usually changes the Hamming distance. A mutation can occur in any of the different positions and that
Figure 3.13: Sequences with different Hamming distances, \( l = 1000 \text{ bp} \)

nucleotide can mutate to either of two nucleotides which does not exist in either of two sequences (which is more likely for more distant sequences). So when the input sequences are a mixture of different genes the distribution of the pairwise Hamming distances may give different information from the distribution of the number of mutations.

4. these distributions show that the sequences get further when they are close to each other (normalized distances of 5% and 10%) and get closer when they are very distant (normalized distances of 90% and 95%). It can be seen better in the closer view of the distribution for sequences of 1000 bp in figure 3.13. This had been concluded from the figures 3.8 and 3.9.

3.3 Template Re-annealing

The simulation of template re-annealing helps to study the PCR bias towards amplifying some genes more than the others. This simulation runs with six input parameters which described in section 2.3.2. The number of cycles is set to 30 (it is a typical value in real experiments). The template and primer concentrations are adjusted according to the data from the real experiment [Karin J. Metzner, personal communication]. The template concentration is fixed to \( 10^{-14} \text{ M} \) in every simulation run but the primer concentration has different values between \( 0.5 \times 10^{-6} \text{ M} \) and \( 5 \times 10^{-5} \text{ M} \). The initial efficiency is always set to 85%. Two other parameters are \( k \) (the ratios between the rate constants of the re-annealing and priming reactions for different genes) and \( f \) (the initial frequencies of different genes). The value of \( k \) depends on the sequence length and the conditions of the chemical reaction. Wetmur and Davidson studied the kinetics of renaturation of DNA [28]. They showed that the reaction ratio of renaturation depends on the temperature and sequence length. It increases as the temperature decreases below \( T_m \), then it reaches a broad flat maximum from 15 to 30°C below \( T_m \) which is about the temperature at the annealing step of PCR (40-60°C). But the reaction rate decreases with a further decrease in temperature. They also mentioned that the reaction rate increases slightly with the CG content of the DNA. They suggested different effective parameters such as solvent-pH, ionic strength and viscosity. Since in a PCR experiment all the conditions are equal for different genes, it is expected that the reaction rate of the template re-annealing is equal for all templates. However it might be slightly different if the templates have very different CG contents. Finally they proposed that the reaction rate is proportional to \( \sqrt{L/N} \) where \( L \) is the average number of nucleotides per

\( ^1 \text{Melting temperature of DNA, in PCR it is set to 90-95 °C at the denaturation step} \)
single strand of the denatured DNA preparation and $N$ is the complexity of the DNA or the total number of DNA base pairs in non-repeating sequences in a virus, bacterium or mammalian cell.

From this study it can be concluded that the ratio between the rate constants of the re-annealing and priming reactions is proportional to the ratio between template and primer lengths. As it was also mentioned in a paper from Mehra and Hu [12], for renaturation of single strands of primer and template to form a primer-template complex, the annealing rate (which is denoted by $K_P$ in equation 2.14) is proportional to the square root of length of the shorter strand (which is a primer strand). Similarly the annealing rate of the renaturation of two DNA strands (two complementary templates) is proportional to the square root of length of the template. Therefore the value of $k$ can be estimated by the square root of the ratio between template and primer lengths ($k \approx \sqrt{L_T/L_P}$).

The simulation has been done for eight different settings. In four cases the input template is a mixture of four different genes with different ratios. Four other simulations have been run for two different genes with different ratios and the effect of the primer concentration is studied for different values of $k$. The results of the first simulation are shown in figure 3.14. The replication starts with four different genes with initial frequencies of 0.7, 0.1, 0.15 and 0.05 of the total template concentration. The primer concentration is set to $0.5 \times 10^{-6} \ M$ and $k$ is equal to 10 (regarding the approximated relation, it corresponds to the amplification of sequences of 2 kbp because assuming the primer is 20 bp the template sequences are 2 kbp).

![Figure 3.14: Concentrations of template and primer and amplification rates for four genes with initial ratios of 0.7, 0.1, 0.15 and 0.05, $k = 10$](image)

The third plot shows the variations of the amplification rate (or efficiency) versus the number of cycles. It can be seen that after 20 replication cycles, these four different genes do not amplify with the same efficiency. As it was explained in section 1.4.4 template re-annealing
happens in later PCR cycles when the template concentration is high enough to compete with the primer concentration, then instead of a primer, a complementary template anneals to a template strand and inhibits the extension step of the PCR. This event results in the reduction of the amplification rate. But when there are mixed templates from different genes with different ratios, the one with higher initial frequency reaches to this inhibitory stage sooner than the others. Therefore in this simulation the more abundant template (lets call it type A) with initial frequency of 70% do not replicate as much as the others. Its amplification rate drops after 20 cycles and it reduces to 15% at the last cycle. Because of this considerable reduction of the amplification rate of gene A, its frequency decreases from 70% to 46% in the final products.

On the other hand the primers which did not anneal to the type A templates, can anneal to other templates (from other genes) and they can continue replicating. The efficiency decrease of the other templates starts nearly three cycles after type A. So the final frequencies of other templates (from other genes) increase from their initial values. The final frequencies of each template is written in the first row of table 3.1.

<table>
<thead>
<tr>
<th>[T] (M)</th>
<th>[P] (M)</th>
<th>k</th>
<th>Initial frequencies</th>
<th>Final frequencies</th>
<th>Additional comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>10</td>
<td>0.7, 0.1, 0.15, 0.05</td>
<td>0.46, 0.19, 0.23, 0.12</td>
</tr>
<tr>
<td>2</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>2</td>
<td>0.7, 0.1, 0.15, 0.05</td>
<td>0.54, 0.16, 0.21, 0.09</td>
</tr>
<tr>
<td>3</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>10</td>
<td>0.5, 0.1, 0.25, 0.15</td>
<td>0.36, 0.17, 0.26, 0.21</td>
</tr>
<tr>
<td>4</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>2</td>
<td>0.5, 0.1, 0.25, 0.15</td>
<td>0.40, 0.14, 0.27, 0.19</td>
</tr>
<tr>
<td>5</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>10</td>
<td>0.8, 0.2</td>
<td>0.64, 0.36</td>
</tr>
<tr>
<td>6</td>
<td>$10^{-14}$</td>
<td>$0.5 \times 10^{-6}$</td>
<td>2</td>
<td>0.8, 0.2</td>
<td>0.7, 0.3</td>
</tr>
<tr>
<td>7</td>
<td>$10^{-14}$</td>
<td>$10^{-5}$</td>
<td>10</td>
<td>0.8, 0.2</td>
<td>0.76, 0.24</td>
</tr>
<tr>
<td>8</td>
<td>$10^{-14}$</td>
<td>$10^{-5}$</td>
<td>2</td>
<td>0.8, 0.2</td>
<td>0.79, 0.21</td>
</tr>
</tbody>
</table>

As it can be seen from equation 2.18, the efficiency of each cycle decreases with increasing $k$. The higher the value of $k$ is, the more dominant the template re-annealing is (equation 2.16). So to study the effect of the template re-annealing, this simulation was run with a lower value of $k$ but the same initial frequencies. In this case $k$ was equal to 2. If $k$ is determined only by the sequence length, $k$ equal to 2 refers to shorter sequences of 80 bp (given the primer is 20 bp). Figure 3.15 shows the results of this simulation. The amplification rate again decreases but this time descends later so the final and initial frequencies are closer to each other compared to the previous simulation (second row of table 3.1).

In another experiment the input templates are mixed with different ratios, 0.5, 0.1, 0.25 and 0.15. These input templates are tested with the same values of $k$, 10 and 2. The results are shown in figures 3.16 and 3.17. Since the initial ratios are closer to each other compared with the previous simulations (the ratio between the most abundant and less abundant is 0.5/0.1 = 5 whereas in the previous examples the ratio was 0.7/0.05 = 14), the amplification rates of all genes drop nearly at the same time. In this experiment the more abundant genes are not so dominant to exclude primers. So more primers are used by the templates especially for $k = 2$ and the concentration of primer decreases to $0.3 \times 10^{-8}M$ (figure 3.17). In this case lack of primer molecules results in replicating less DNA strands and the amplification rate reaches to 41.
Figure 3.15: Concentrations of template and primer and amplification rates for four genes with initial ratios of 0.7, 0.1, 0.15 and 0.05, $k = 2$

zero at the end of cycle 30. So in this case running PCR for more than 30 cycles needs more primer. The final frequencies of different genes are written in the third and fourth rows of table 3.1 for $k = 10$ and $2$.

In order to compare the bias in the final to initial template ratios in these experiments, the ratio between final and initial frequencies are calculated for each case. The values are shown in a bar plot in figure 3.18. A, B, C and D refer to the four genes with different frequencies. In the ideal case this ratio should be one, so the closest ratio to one implies that the bias has not been so large. This chart illustrates that the strong bias is not only for high value of $k$, but also depends on the diversity of the initial ratios. When the genes frequencies are in a more diverse range, the template re-annealing starts sooner for more abundant genes and the ratios between all genes change so much in the following replications. This can be found by comparing blue and yellow bars for $k = 10$ or orange and green bars for $k = 2$.

The following simulations focus on the amplification of two genes with initial ratios of 80% (type A) and 20% (type B) of the total template concentration. These two genes are selected to study this bias more deeply and find a way to remove it. The ratios are chosen such that one is more abundant than the other to simulate the worst case as it was founded from the previous simulations. The results of the simulation with $k = 10$ is depicted in figure 3.19. The second plot shows the templates concentrations in a logarithmic scale. It shows that the concentration of gene A (the blue line) increases almost linearly until cycle 25 when its amplification rate decreases to about 50%. In the last five cycles gene A enters into the saturation phase which means the concentration becomes nearly constant. While the concentration of gene A does not increase so much, the primers are used by the other gene (type B) and the concentration grows linearly until cycle 28. This causes the frequency of type B increases from its initial value and
the frequency of type A decreases. This is what was mentioned in [21] as a bias towards a 1:1 ratio in the final products. Finally type B also reaches to the saturation phase in the last two cycles.

The same experiment was done by using \( k = 2 \), both genes (A and B) enter to the saturation phase later than the previous experiment (figure 3.20). Fifth and sixth rows of table 3.1 show the final frequencies of genes A and B in these two experiments.

Now the question is how to remove this bias. One way is to avoid template re-annealing by setting the reaction conditions. As it was indicated by Wetmur and Davidson DNA renaturation decreases by decreasing temperature below \( T_m - 30^\circ C \). Generally the annealing temperature is between 40-60\(^\circ C\), so setting the temperature as low as possible can prevent template re-annealing, however this must be verified by an experiment. Here using the simulation it is possible to test other parameters if the constant \( k \) cannot be reduced by setting the reaction conditions. The only effective parameter which can be set in this simulation is the concentration of primer. As it was mentioned before and shown in the previous examples, template re-annealing starts when the template concentration is comparable with the primer concentration. So if the amplification begins with more primer molecules, then they are always in excess of the templates. To verify this claim the following simulations have been done with 20 times more primer molecules. Figure 3.21 shows the results for \( k = 10 \). The linear grows of the template concentration continues until the last cycle. Although the amplification rates decrease in the last five cycles but still more than 50% of the templates continue with replicating. So in this case the final ratios are more closer to the initial ratios (seventh row of table 3.1).

Another simulation with \( k = 2 \) shows that this bias can be removed by using more primer molecules (figure 3.22). In this case the final and initial frequencies are nearly equal (eighth row of table 3.1).
Figure 3.17: Concentrations of template and primer and amplification rates for four genes with initial ratios of 0.5, 0.1, 0.25 and 0.15, \( k = 2 \)

row of table 3.1). So if it is not possible to avoid template re-annealing by controlling PCR conditions, it can be reduced by running PCR with more primer molecules. It should be noted that this solution does not mean that 20 times more primers are used for the amplification. The primer concentration decreases as much as the total template concentration increases. In the simulation with \( 5 \times 10^{-7} \) M primer and \( k = 10 \), \( 2 \times 10^{-7} \) M primer is used and the same amount of templates (of both genes) is produced. In the other simulation with \( 1 \times 10^{-5} \) M primer and \( k = 10 \), \( 7.75 \times 10^{-7} \) M primer is used to generate the same amount of templates. It means that still \( 9.225 \times 10^{-6} \) M primer is available which is enough for another PCR run. The higher initial concentration of primer is to control the conditions such that the primer molecules are always in excess of the template molecules.

Similar analysis is done by comparing the ratios between final and initial frequencies for genes A and B in four previous examples. The effect of adding more primer molecules is clear by comparing blue and pink bars with green and purple bars in figure 3.23. In both cases (\( k = 2 \) and \( k = 10 \)) running PCR with more primers causes the final ratios become closer to the initial ratios.

In summary the following conclusions can be drawn from the simulations of this section:

1. The bias because of the template re-annealing is stronger when the template is a mixture of different genes from a diverse range of ratios, more precisely when the template mixture is mostly dominated by one of genes,

2. The higher value of \( k \) results in decreasing the amplification rates in the earlier PCR cycles. The higher \( k \) implies that the templates have more tendency to anneal each other instead of the primer sequences and then they do not make a copy from themselves,
Figure 3.18: Ratio between final and initial frequencies for four different genes in four PCR runs

Figure 3.19: Concentrations of template and primer and amplification rates for two genes with initial ratios of 0.8 and 0.2, \( k = 10 \)

3. All the simulations showed that PCR works fine in the first 20 cycles, the amplification rates do not drop and the templates concentrations grow exponentially. The saturation phase appears in the last cycles and for lower \( k \) it happens later,

4. There are some ways to remove the bias because of the template re-annealing. One way is to control the reaction conditions such as temperature to avoid DNA renaturation,

5. Running PCR with more primer molecules causes that primers always remain in excess of templates, therefore it is more likely that primers anneal to templates,

6. Stopping PCR earlier is another solution but then less products will be produced.
Figure 3.20: Concentrations of template and primer and amplification rates for two genes with initial ratios of 0.8 and 0.2, $k = 2$

Figure 3.21: Effect of adding more primer, $k = 10$
Figure 3.22: Effect of adding more primer, $k = 2$

Figure 3.23: Ratio between final and initial frequencies for two different genes in four PCR runs
Chapter 4

Discussion

PCR is a method which is being used since more than twenty years to produce large amounts of target sequences. During these years many sophisticated techniques have improved the performance of PCR. Some of the PCR troubles were mentioned in this report. They were divided into two major parts, PCR artifacts and PCR bias.

Mutation is known as one reason of having artifacts. These artifacts which are mutant sequences differ from the original products in some bases. The distribution of number of mutations which was explained in section 2.1 helps to estimate the percentage of the wild type sequences in total products. The simulation results in section 3.1 showed that using an accurate enzyme more than 90% of the sequences (with 100 or 500 nucleotides) replicate exactly without any error. For longer sequences with 1000 nucleotides the probability to have mutants increases to about 12%. On the other hand the distribution of pairwise Hamming distances gives some other information. It may show that only 5% of sequences have different Hamming distances compared to their initial distances but it does not necessarily mean that only 5% of sequences are mutant, because mutation may happen in some positions that are different between two genes and the pairwise distance remains constant.

Chimeric and heteroduplex molecules are known as the other PCR artifacts. Both of them are generated in the last PCR cycles when the concentration of products is high. Thompson et al. [22] proposed an effective way to eliminate heteroduplexes by reconditioning PCR.

The PCR bias because of the template re-annealing has been covered in this project. This bias was studied using a kinetic model of chemical reactions between template and primer molecules. Two more detailed kinetic models of PCR have been proposed in [4] and [12]. In these two papers all the reactions during the PCR steps (denaturation, annealing and extension) are modeled using the differential equations. But in this project a very simple model was simulated which describes the static behavior of the molecular reactions in the annealing step. In fact the more exact way to simulate DNA replication is to solve all the differential equations which explain the dynamic behavior of the molecular reactions. But to study the effective parameters in PCR bias and how to control them, this detailed modelling was not necessary. In addition both methods (simple model based on static modelling and the others based on differential equations) predict similar behaviors for the concentrations and efficiency during the PCR cycles.

Simulation template re-annealing exposed some facts which were addressed before by real experiments. The bias towards the 1:1 ratio disregarding the initial ratios was shown and this was more stronger for higher $k$ (the ratio between the rate constants of the re-annealing and...
priming reactions) and more diverse gene ratios. Since the re-annealing reaction was dominant when the template concentration is comparable with primer concentration, the effect of adding more primer molecules was tested by simulation. The results showed that keeping primers always in excess of templates prevents the problem of template re-annealing. On the other hand using much more primers may cause another problem of primer re-annealing which has been mentioned in [12]. But this problem does not cause a bias towards amplifying one gene more than the others, because all genes are amplified using a universal primer. This problem results in decreasing efficiency as it has been shown in [12], but the efficiency decreases with the same rate for all genes.

Finally it should be stated that different PCR troubles have been reported in literatures. Some of them such as CG content, primer mismatch, template re-annealing have been mentioned here. These problems are really dependent on the template sequences, primer sequences and reaction conditions like temperature. So it is possible to have contradictory observations from PCR experiments. For example Polz and Cavanaugh reported that the gene dosage alone had no discernible effect on product ratios [15] which is inconsistent with the simulation results in section 3.3. This is because they have adjusted the reaction parameters to avoid template saturation phase. However it is approved by most of experiments that multi-template PCR should be carried out using several replicates and stopped at an early stage. The PCR products in the replicates should be mixed before further analysis [6].
Bibliography


