Stochastic Modeling of Polymerase Chain Reaction and Related Biotechnologies

Fengzhu Sun
Department of Mathematics, University of Southern California
1042 W 36th Place, DRB 155
Los Angeles, CA90089-1113, USA
fsun@hto.usc.edu

1. Introduction

The polymerase chain reaction (PCR) is a method that uses test tubes in biological laboratories for producing large amount of identical copies of a specific gene from small amount of complex molecules. Here we present our work on the stochastic modeling and theoretical studies of PCR, sexual PCR (DNA shuffling) and their applications to *in vitro evolution*: an experimental method to evolve proteins with improved functions.

2. The Polymerase Chain Reaction

The PCR uses the mechanism of DNA replication. There are three steps in a PCR cycle. In the first step, the double-stranded DNA molecules are heated to near boiling temperature so that the double-stranded DNA molecules are separated completely into two single-stranded sequences. This process is called *denaturing*. The single-stranded sequences generated by denaturing are used as templates for the primers and the DNA polymerase. In the second step, the temperature is lowered such that the primers anneal to the templates. This process is called *annealing*. In the third step, the temperature is raised again to the temperature that is optimum for the polymerase to react. The DNA polymerases use the single-stranded sequences as templates to extend the primers that have been annealed to the templates. This process is called *polymerase extension*.

The three steps form a PCR cycle. The expeiment is repeated for n cycles. Suppose that in each PCR cycles, a fraction λ of templates make a complete copy. We assume that λ is a constant throughout the paper. A standard branching process can model the generation of the DNA molecules. PCR is not a perfect process and occasionally, DNA polymerase substitutes, adds or deletes a nucleotide to the growing DNA chain. A mutation occurs in this case. Only when a new sequence is generated from a template (with probability λ), mutations can occur along the newly generated sequence. We assume that the number of mutations along the newly generated sequence is a Poisson random variable with mean $G\mu$, where G is the length of the

target and μ is the mutation rate per base per PCR cycle. For simplicity, we assume that mutations occur at different places every time mutations occur (Sun 1995). The following results have been extended to the situation where this assumption is violated (Wang et al. 2000). Moore and Maranas (2000) considered different mutation rates at different positions.

Based on the above mathematical model, we can study the properties of the PCR products after n cycles. In particular, we have

Theorem 1 *Let M be the number of mutations of a randomly chosen sequence. Then*

i). For any $0 \le m \le G$

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

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

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

After a PCR experiment, a sample of s PCR products are sampled and sequenced. If we know the nucleotide bases of the original molecules to be amplified, we can count the number of mutations of the sampled PCR products. Let $M_1, M_2, \cdots, M_s$ be the number of mutations of the sampled sequences. The moment estimator of the mutation rate μ is given by

$$\hat{\mu_1} = \frac{(1+\lambda)\sum_{i=1}^{s} M_i}{n\lambda G s}$$

If the nucleotide bases of the original molecules to be amplified are not known, we can compare the pairwise differences among the sampled molecules. Let $H_{i,j}$ be the pairwise Hamming distance—the number of different bases between sequence i and sequence j. We propose to estimate the mutation rate using

$$\hat{\mu_2} = \frac{\sum_{i\neq j,\ i,j=1}^{s} H_{i,j}}{\binom{s}{2} ED \times G},$$

where $ED = \frac{2n\lambda}{1+\lambda} - \frac{2}{(1+\lambda)S_0+1-\lambda} + O(\frac{1}{S_0(1+\lambda)^n})$, S_0 is the initial number of molecules.

Note that the sampled sequences are correlated through a random binary tree and, thus, are not independent. The variances and the limit behavior of the above two estimators are given in Sun (1995) and Wang et al. (2000).

3. Modeling DNA Shuffling

DNA shuffling or sexual PCR is an experimental method to introduce recombination among a group of closely related molecules. The principle of DNA shuffling can be described

as follows. A pool of closely related molecules with different point mutations are prepared. The first step in DNA shuffling is to break the molecules into random fragments using DNase I which can randomly create nicks along each strand of a DNA molecule. Then fragments of lengths within a certain range are sampled. These sampled fragments go through PCR *without added primers*. In the annealing step, homologous templates prime each other to form $5'$ and $3'$ overhangs. In the polymerase extension step, the DNA polymerase extends the $5'$ overhangs using the other annealed strand as a template. The average fragment length is increased in each PCR cycle. After many PCR cycles without added primers, molecules of the original size are obtained. When a template from one molecule primes a fragment from another molecule, recombination occurs. The idea behind DNA shuffling is that, by recombining beneficial mutations from different molecules, molecules with improved function can be obtained. The idea of DNA shuffling comes from the theory of genetic algorithms.

We assume that overlapping fragments (overlapping by at least θ) are equally likely to anneal. A full-length reassembled molecule can be regarded as concatenations of many fragments. The $3'$ ends of the fragments form a renewal process. Starting from the $5'$ end of the molecule, we label the fragments consecutively. Let the distance between the $3'$ end of the $i-1$-st fragment and the $3'$ end of the i-th fragment be $D_i,\ i = 1, 2, \cdots$, which are independent identically distributed random variables. Let $F(x)$ be the cumulative distribution function of the fragments to be shuffled. Then

$$P(D_i > s \mid D_i > 0) = \frac{\int_0^\infty (1 - F(s + x + \theta))dx}{\int_0^\infty (1 - F(x + \theta))dx}.$$

In order that the $3'$ ends along full length reassembled molecules form a stationary renewal process, the $3'$ end of the first fragment should have density

$$f_\infty(s) = \frac{P(D_i > s \mid D_i > 0)}{E(D_i \mid D_i > 0)}.$$

With the above model, we are able to calculate the recombination probabilities at different locations.

4. Applications to *in vitro* evolution

In vitro evolution is a laboratory method to evolve molecules with desired properties based on the evolution of natural organisms by mutation and selection. The two most widely used mutagenesis techniques are error-prone PCR and DNA shuffling. The objective of this study is to build theoretical models for *in vitro* evolution together with mutagenesis technique and to optimize experimental conditions. We want to know the optimal mutation rate during error-prone PCR, when to use error-prone PCR and DNA shuffling, the optimal concentrations of the different molecules and the length of the fragments for DNA shuffling.

REFERENCES

Moore, G. L. and Maranas, C. D. (2000) Modeling DNA mutation and recombination for directed evolution experiments, *J. Theor. Biol.*, **205**, 483-503.

Sun, F. Z. (1995) The polymerase chain reaction and branching processes. *J. Comp. Biol.*, **2**, 63-85.

Sun, F. Z. (1999) Modeling DNA shuffling. *J. Comp. Biol.*, **6**, 77-90.

Wang, D. et al. (2000) Estimation of the mutation rate during error-prone polymerase chain reaction. *J. Comp. Biol.*, **7**, 143-158.

RESUME

In vitro evolution is an important experimental method to evolve molecules with improved or new functions. It uses mutagenesis techniques such as error-prone polymerase chain reaction (PCR) and DNA shuffling during the experiment. We develop mathematical models for error-prone PCR, DNA shuffling and *in vitro* evolution. Error-prone PCR can be modeled as a branching process which generates a random binary tree. Mutations are then superimposed onto the random tree. We develop statistical methods to estimate the mutation rate using the number of mutations or pairwise differences for a sample of PCR products. In DNA shuffling, every reassembled molecule can be modeled as a concatenation of random fragments. The connecting points form a renewal process. We characterize the length of renewal with respect to the distribution of the fragment length used in DNA shuffling. The models have been validated using experimental data and can be used to optimize *in vitro* evolution experiments.

L'évolution *in vitro* est une méthode expérimentale importante pour faire évoluer des molécules possédant des fonctions nouvelles ou améliorées. Elle utilise lors de l'expérience des techniques de mutagénèse telles que la réaction en chaîne de polymérase (RCP) sujette à erreur et le réarrangement d'ADN. Nous développons des modèles mathématiques pour la RCP sujette à erreur, le réarrangement d'ADN et l'évolution *in vitro*. La RCP sujette à erreur peut être modelée par un processus de ramification qui engendre un arbre binaire aléatoire. Les mutations sont alors superposées sur l'arbre aléatoire. Nous développons des techniques statistiques pour estimer le taux de mutations et de différences en paires pour un échantillon de produits de RCP. Dans le réarrangement d'ADN, chaque molécule réassemblée peut être modelée comme une concaténation de fragments aléatoires. Les points de connexions forment un processus de renouvellement. Nous caractérisons la longueur du renouvellement par rapport à la distribution de la longueur du fragment utilisé lors du réarrangement d'ADN. Les modèles ont été validés à l'aide de données expérimentales et peuvent être utilisés pour optimiser les expériences d'évolution *in vitro*.