

Modeling DNA Shuffling

Fengzhu Sun

¹Department of Genetics

Emory University School of Medicine

Atlanta, GA 30322

U. S. A.

Phone (404)-727-1702, Fax (404)-727-3949,

Email: fsun@genetics.emory.edu

Running head: MODELING DNA SHUFFLING

Keywords: PCR; DNA shuffling; Recombination; *In vitro* evolution; Mathematical modeling

Abstract

In vitro evolution is a new, important laboratory method to evolve molecules with desired properties. It has been used in a variety of biological studies and drug development. In this paper, we study one important mutagenesis method used in *in vitro* evolution experiments called DNA shuffling. We construct a mathematical model for DNA shuffling and study the properties of molecules after DNA shuffling experiments based on this model. The model for DNA shuffling consists of two parts. First we apply the Lander-Waterman model for physical mapping by fingerprinting random clones to model the distribution of regions that can be reassembled through DNA shuffling. Then we present a model for recombination between different DNA species with different mutations. We compare our theoretical results with experimental data. Finally we propose novel applications of the theoretical results to the optimal design of DNA shuffling experiments and to physical mapping using DNA shuffling.

1 Introduction

In vitro evolution is a laboratory method to evolve molecules with desired properties. It has been successfully applied to a wide range of biological studies such as protein-DNA interaction, catalytic properties of RNA molecules, and catalytic properties of single-stranded DNA molecules (see Fitzwater and Polisky 1996 and Gold et al. 1995 for recent reviews). The basic principle of *in vitro* evolution experiments can be summarized as follows. First a library of random molecules—DNA, RNA or protein—is constructed. The molecular library can be composed of completely random molecules of peptides or oligonucleotides. It can also be composed of many variants of one or more parent molecules obtained through mutagenesis. The potential diversity of a random library can be huge. For a protein molecule of N residues, the potential library contains 20^N molecules (a number which can easily exceed the number of molecules in a test tube). Thus a completely random library may not be a good choice as any library can only contain a very small fraction of the potential molecules if N is relatively large, say over 50. The molecules with the desired property have a low probability of belonging to this relatively sparsely sampled library. Instead, a library generated through mutagenesis from one or a few molecules already known to have some desired property might be more useful.

Once a molecule library is constructed, some molecules in the library might have a specific function of interest. A selection or screening procedure is used to isolate those molecules. This is referred as the *selection* step. The experimental protocols for selection or screening

depend on the molecular properties of interest. The number of molecules remaining after the selection step is usually small. To increase the diversity of the molecules being explored, the selected molecules can be put through a mutation process called *mutagenesis*. The molecules generated through mutagenesis are then amplified by polymerase chain reaction (PCR) or other amplification methods. This is referred as the *amplification* step. The processes of selection, mutagenesis, and amplification form one cycle of the experiment. The experiment is repeated for multiple cycles until molecules of the desired property are selected. Irvine et al. (1991) and Sun et al. (1996) studied *in vitro* evolution not involving the mutagenesis step mathematically.

Mutagenesis is one important step in *in vitro* evolution experiments. It can be used to generate the initial random molecule library. It is also used to increase the diversity of molecule libraries after each amplification and selection step. The technique for mutagenesis should also be easily implemented in multiple cycles of *in vitro* evolution experiments. Without the mutagenesis step in *in vitro* evolution experiments, we are confined to select only molecules that were in the initial library. Because the initial molecular library usually contains only a very small fraction of the potential molecules, with low probability the library contains molecules with desired properties. Thus *in vitro* evolution of large molecules is most likely to fail without using mutagenesis. Mutagenesis methods have been and are continuing to be developed. In this paper we consider one important mutagenesis method called DNA shuffling.

DNA shuffling, also called sexual PCR, was developed by Stemmer (1994ab). It has been

successfully applied to improve the drug resistance of β -lactamase (Smith 1994, Stemmer 1994ab, Stemmer 1995, Zhang et al. 1997, Cramer et al. 1997, Patten et al. 1997), to optimize industrial enzymes (Shao and Arnold 1996, Arnold and Moore 1997, Kuchner and Arnold 1997), to help the development of pharmaceuticals and vaccines (Patten et al. 1997), and to distinguish functional from nonfunctional mutations (Zhao and Arnold 1997). Recently, DNA shuffling has been used to recombine a family of molecules from diverse species (Cramer et al. 1997).

The principle of DNA shuffling can be described as follows. A pool of closely related molecules with different point mutations is prepared either through error-prone PCR (Leung et al. 1989, Cadwell and Joyce 1992) or other mutation techniques such as oligonucleotide-directed mutagenesis (Reidhaar-Olson and Sauer 1992). 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*. There are three steps in the PCR process without primers. The first step is denaturing by increasing the temperature so that double-stranded fragments are separated completely into single-stranded ones (Figure 1a-b). The second step is annealing by lowering the temperature so that single-stranded fragments anneal to other fragments overlapping by a certain number of bases that are complementary at the overlapping region (Figure 1b-c). After the annealing step, homologous templates prime each other to form 5' and 3' overhangs. The third step is polymerase extension by increasing the temperature to the level for the DNA polymerase to

extend the 5' overhangs using the other annealed strand as template. The 3' overhangs are not changed as DNA polymerase can only extend from a 5' end (Figure 1c-d). The three steps of denaturation, annealing and polymerase extension are repeated for multiple cycles. The average fragment length is increased in each PCR cycle. After many cycles of PCR without added primers, we expect to obtain molecules of the original size. 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, we can obtain molecules with even increased function.

(Note: Figure 1 is about here)

To better understand DNA shuffling and to make it applicable to more problems, a model is needed for DNA shuffling. Moore et al. (1997) studied a simple mathematical model of recombination between different molecules in DNA shuffling experiments assuming that the lengths of the fragments to be reassembled are less than the distances between the mutations. The objective of this study is to construct a mathematical model for DNA shuffling under more general situations, to analyze the properties of DNA shuffling products based on the proposed model, and to compare the theoretical results with experimental data. Finally we present the applications of the theoretical results to the optimal design of DNA shuffling experiments and propose a new method of physical mapping using DNA shuffling.

2 Modeling the Distribution of Random Fragments

In DNA shuffling, the genomic DNA is first broken into fragments by DNase I. Then fragments of lengths within a certain range are sampled and are used in DNA shuffling. This procedure is the same as in nucleic acid cloning. In nucleic acid cloning experiments, molecules are first fragmented by restriction enzymes and fragments of lengths within the cloning range of certain vectors are cloned. The distribution of the randomly cloned fragments along the genomic DNA from the nucleic acid cloning experiments was successfully modeled by Lander and Waterman when they studied the progress of physical mapping by fingerprinting random clones (Lander and Waterman 1988). In their model, it is assumed that random fragments are distributed along the genome according to a Poisson process with parameter c , where c is the coverage of the random fragments for the original genome. The parameter c can be calculated as the sum of the lengths of the sampled random fragments divided by the length of the original genome. The probability that there are k fragments with starting points within an interval of length t is $(ct)^k \exp(-ct)/k!$. As a first step, we assume that the sampled fragments are of the same length L . By rescaling, we can always assume $L = 1$.

We assume that two templates can anneal to each other and thus be extended if and only if they overlap at least a fraction, θ , of the length of the fragments and are complementary at the overlap region. In physical mapping, overlapping fragments are connected to form one contiguous region along the genome called an island. The gaps between islands are called *oceans*. The number of islands, the number of clones in an island and the expected

length of an island have been obtained under the Lander-Waterman model (Lander and Waterman 1988). These quantities are also important in the study of DNA shuffling. For example, islands represent regions that can be reassembled through DNA shuffling and oceans represent regions that can not be reassembled through DNA shuffling. The expected length of an island corresponds to the average length of regions that can be reassembled through DNA shuffling. We restate Lander-Waterman's results in the context of DNA shuffling.

Theorem 1 *Let θ be the fraction of fragment length which two fragments must share in order to anneal and be extended, N be the number of sampled fragments and c be the coverage. Then*

1. *The expected number of regions that can be reassembled through DNA shuffling is*

$$N \exp(-c(1 - \theta)).$$

2. *The expected number of regions reassembled from j fragments by DNA shuffling is*

$$N \exp(-2c(1 - \theta))(1 - \exp(-c(1 - \theta)))^{j-1}.$$

3. *The expected number of regions reassembled from at least two fragments is*

$$N \exp(-c(1 - \theta)) - N \exp(-2c(1 - \theta)).$$

4. *The expected length of a region reassembled through DNA shuffling is*

$$(\exp(c(1 - \theta)) - 1)/c + \theta.$$

One of the objectives of DNA shuffling is to recombine mutations from different DNA species. For simplicity, let us consider two DNA species with two different mutations M_1

and M_2 . Let the distance between the two mutations be t . In order to obtain molecules with both mutations M_1 and M_2 , it must be possible to reassemble fragments joining M_1 and M_2 through DNA shuffling although this is not sufficient. We call the region between M_1 and M_2 the *target*. The probability that this target can be reassembled through DNA shuffling is of interest. Unfortunately there are no explicit closed formulas for this probability. A simulation algorithm has been written to approximate this probability for given values of coverage c , overlap fraction θ , and target length t . Table 1 gives this probability for $t = 25$, c from 4 to 10 by step 1, and $\theta = 0.0, 0.25, 0.50$, respectively.

(Note Table 1 is about here)

From this table we see that the overlap fraction θ plays a very important role for the probability that the target can be reassembled. For example, at the coverage of $c = 10$, the probability that a target of length $t = 25$ can be reassembled is 0.99 when very small fraction of overlap is needed for fragments to anneal and extended compared to only 0.19 when about half the length of the fragment is needed to anneal and extended. Thus to obtain high probability that a target can be reassembled by DNA shuffling, the experimental conditions should be chosen such that short overlap is needed to anneal and extended. But there is a trade-off here. Such experimental conditions might introduce more mutations during DNA shuffling. The balance of increasing the probability that the target can be reassembled and,

at the same time, minimizing the mutation rate in DNA shuffling should be considered in designing DNA shuffling experiments.

As in Lander and Waterman (1988), we can show that, given a specific target can be reassembled through DNA shuffling, the probability that the island contains k fragments is $\exp(-c(1 - \theta))(1 - \exp(-c(1 - \theta)))^{k-1}$.

The above results can be extended to the case that, L , the length of fragments to be reassembled is a random variable as in Arratia et al. (1991) and Waterman (1995). Let the fragment length L be independent identically distributed (iid) with mean $E(L) = 1$ and cumulative distribution function $F(l)$. Define

$$\mathcal{J}(x) = P(L > x) = 1 - F(x),$$

and

$$J(x) = \exp\left\{-c \int_x^\infty \mathcal{J}(l) dl\right\}.$$

Then it can be shown that the results of Theorem 1 hold by replacing $\exp(-c(1 - \theta))$ with $J(\theta)$.

3 Modeling Recombination

In the above section, we only model the distribution of random fragments and the corresponding regions that can be reassembled through DNA shuffling along the genome. We ignore the differences among different DNA species. In DNA shuffling, one important objective is to obtain new DNA species by recombining mutations from the available different DNA species. In this section, we model the recombination process among different DNA species.

We assume that the mutations are not extensive enough to hinder strand annealing. In modeling the recombination between different DNA species, it is necessary to make a critical assumption as to the annealing (or association) of complementary strands as with moderate and higher coverage it is possible a strand may overlap several strands. Classical reassociation studies (Britten et al. 1974) showed that the rate of reassociation of two complementary strands increases as the square root of the strand length (Wetmur and Davison 1968). If a fragment, F1, overlaps several other fragments, the fragment with the largest overlap with F1 has a relatively higher probability of annealing with F1 than those fragments with short overlaps. For simplicity, we assume that overlapping fragments of F1 (overlapping by at least θ) are equally likely to anneal to F1. Both F1 and the annealed strand are extended using the other strand as a template. We refer to this model as the equally likely model. Then any full length reassembled molecule can be regarded as concatenations of random fragments. The region between the right ends of any two consecutive fragments is assumed

to be randomly chosen from the corresponding region of the parent molecules proportional to their concentrations.

It is then important to study the distribution of the distances between the right ends of consecutive fragments forming a full length reassembled molecule. Starting from the right end of the molecule, we label the fragments consecutively. Let the distance between the right end of the $i - 1$ -st fragment and the right end of the i -th fragment be D_i , $i = 1, 2, \dots$, which are iid random variables. Denote the right end of the i -th fragment as l_i (Figure 2). We first consider only fragments with left ends in $(l_i - \theta - t, l_i - \theta]$. We will then let t tend to infinity. Assuming that there is at least one fragment with left end in $(l_i - \theta - t, l_i - \theta]$ and that each of these fragments is equally likely to anneal to the i -th fragment, then the distance from $l_i - \theta$ to the left end of the annealed fragment, X , is uniformly distributed on $(0, t]$ from the properties of Poisson processes. Let the length of the annealed fragment be L (with a cumulative distribution function $F(l)$.) Then $L = X + \theta + D_i$ (Note that $D_i < 0$ is possible) and $D_i = L - X - \theta$.

(Note: Figure 2 is about here)

$$P(D_i \leq s) = P(L \leq s + X + \theta) = \int_0^t F(s + x + \theta) \frac{dx}{t}.$$

Thus

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

We are conditional on the i -th fragment being extended, i.e. $D_i > 0$. Thus

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

Let t tend to infinity, we have

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

In order that the right fragment ends along full length reassembled molecules form a stationary renewal process, we assume that the right end of the first fragment has density function (Page 211, Lange 1997)

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

Then the distance from any fixed point to the next right fragment end along a full length reassembled molecule also has the density function $f_\infty(s)$.

Next we study the recombination probability between two mutations. suppose we have two DNA species with two different mutations M_1 and M_2 . Let t be the distance between the two mutations. We denote the locus corresponding to M_1 as 0 and the locus corresponding to M_2 as t . Let α and $1 - \alpha$ be the fractions of species 1 and species 2 molecules, respectively (Figure 3). A sampled fragment is from species 1 with probability α and from species 2 with probability $1 - \alpha$. In order that a full length reassembled molecule has both mutations M_1

and M_2 , the segment covering 0 must be from species 1 (with probability α); the next right fragment end must belong to $(0, t)$ with probability

$$\int_0^t f_\infty(s)ds = \frac{\int_0^t \int_0^\infty (1 - F(s + x + \theta)) dx ds}{\int_0^\infty x(1 - F(x + \theta)) dx} = \frac{\int_0^\infty x(F(x + \theta + t) - F(x + \theta)) dx}{\int_0^\infty x(1 - F(x + \theta)) dx};$$

and the segment covering t must come from species 2 (with probability $1 - \alpha$). Multiplying all the above three probabilities, we obtain the probability of having both mutations

(**Note: Figure 3 is about here**)

$$p_2 = \alpha(1 - \alpha) \frac{\int_0^\infty x(F(x + \theta + t) - F(x + \theta)) dx}{\int_0^\infty x(1 - F(x + \theta)) dx}. \quad (1)$$

The probability of having neither mutations is $p_0 = p_2$ and the probability of having only one of the two mutations is $p_1 = 2(1 - p_0 - p_2) = 2(1 - 2p_2)$.

Note that when the fragment length L is always less than $t + \theta$, $F(x + \theta + t) = 1$, $x > 0$.

Thus $p_2 = p_0 = \alpha(1 - \alpha)$ and the two mutations recombine independently.

Next we consider three special cases for the distribution of the lengths of the fragments used for shuffling. Firstly if $L > t + \theta$ is a constant, then D_i is uniformly distributed on $(0, L - \theta]$.

$$p_2 = \alpha(1 - \alpha) \frac{t}{L - \theta} \left(2 - \frac{t}{L - \theta} \right). \quad (2)$$

Secondly, if L is uniformly distributed on an interval $(a, b]$ and $\theta < a$, then

$$P(D_i > s \mid D_i > 0) = \begin{cases} 1 - \frac{2s}{a+b-2\theta} & \text{if } s < a - \theta, \\ \frac{(b-s-\theta)^2}{(b-a)(a+b-2\theta)} & \text{if } a - \theta \leq s < b - \theta, \\ 0 & \text{if } \geq b - \theta, \end{cases}$$

and

$$E(D_i \mid D_i > 0) = \frac{(a - \theta)^2 + (b - \theta)^2 + (a - \theta)(b - \theta)}{3(a + b - 2\theta)}.$$

If $t < a - \theta$, we have

$$\int_0^t P(D_i > s \mid D_i > 0) ds = \int_0^t \left(1 - \frac{2s}{a+b-2\theta}\right) ds = t \left(1 - \frac{t}{a+b-2\theta}\right),$$

and

$$p_2 = \alpha(1 - \alpha) \frac{3t(a + b - 2\theta - t)}{(a - \theta)^2 + (b - \theta)^2 + (a - \theta)(b - \theta)}. \quad (3)$$

If $a - \theta \leq t < b - \theta$, we have

$$\int_0^t P(D_i > s \mid D_i > 0) ds = \frac{(a - \theta)^2 + (b - \theta)^2 + (a - \theta)(b - \theta) - (b - s - \theta)^3 / (b - a)}{3(a + b - 2\theta)},$$

and

$$p_2 = \alpha(1 - \alpha) \left(1 - \frac{(b - s - \theta)^3}{(b - a)((a - \theta)^2 + (b - \theta)^2 + (a - \theta)(b - \theta))}\right).$$

Thirdly, if L has an exponential distribution with mean $1/\lambda$, i.e. $F(x) = 1 - \exp(-\lambda x)$,

then

$$P(D_i > s \mid D_i > 0) = \exp(-\lambda s).$$

Thus, conditional on $D_i > 0$, D_i still has an exponential distribution with the same mean and

$$p_2 = \alpha(1 - \alpha)(1 - \exp(-\lambda t)). \quad (4)$$

In practice, we usually do not know the underlying distribution of L . It is tempting to use the mean value of L as the length of the fragments to be reassembled in DNA shuffling. Next we show that this will overestimate the probability of recombination. From Equation (??) and Equation (??), we only need to show that for $0 < t < E(L) - \theta$

$$\frac{\int_0^\infty x(F(x + \theta + t) - F(x + \theta))dx}{\int_0^\infty x(1 - F(x + \theta))dx} \leq \frac{t}{E(L) - \theta} \left(2 - \frac{t}{E(L) - \theta} \right).$$

That is

$$t(2(E(L) - \theta) - t) \int_0^\infty x(1 - F(x + \theta))dx \geq (E(L) - \theta)^2 \int_0^\infty x(F(x + \theta + t) - F(x + \theta))dx.$$

Let

$$\begin{aligned} G(t) &= (E(L) - \theta)^2 \int_0^\infty x(F(x + \theta + t) - F(x + \theta))dx - t(2(E(L) - \theta) - t) \int_0^\infty x(1 - F(x + \theta))dx \\ &= (E(L) - \theta - t)^2 \int_0^\infty x(1 - F(x + \theta))dx - (E(L) - \theta)^2 \int_0^\infty x(1 - F(x + \theta + t))dx \\ &= (E(L) - \theta - t)^2 \int_0^\infty x(1 - F(x + \theta))dx - (E(L) - \theta)^2 \int_t^\infty (x - t)(1 - F(x + \theta))dx. \end{aligned}$$

We only need to prove that $G(t) \leq 0$, $0 < t < E(L) - \theta$.

$$\begin{aligned} G'(t) &= -2(E(L) - \theta - t) \int_0^\infty x(1 - F(x + \theta))dx + (E(L) - \theta)^2 \int_t^\infty (1 - F(x + \theta))dx \\ G''(t) &= 2 \int_0^\infty x(1 - F(x + \theta))dx - (E(L) - \theta)^2 (1 - F(t + \theta)) \\ &= 2 \int_0^\infty x(1 - F(x + \theta))dx - (E(L) - \theta)^2 + (E(L) - \theta)^2 F(t + \theta) \end{aligned}$$

Next we show that $G''(t) > 0$.

$$\begin{aligned}
& 2 \int_0^\infty x(1 - F(x + \theta))dx - (E(L) - \theta)^2 \\
= & 2 \int_\theta^\infty (x - \theta)(1 - F(x))dx - (E(L))^2 + 2\theta E(L) - \theta^2 \\
= & 2 \int_0^\infty x(1 - F(x))dx - (E(L))^2 - 2 \int_0^\theta (x - \theta)(1 - F(x))dx - \theta^2 \\
= & 2 \int_0^\infty x(1 - F(x))dx - (E(L))^2 - 2 \int_0^\theta (\theta - x)F(x)dx \\
= & 2 \int_0^\infty \int_0^x 1dy(1 - F(x))dx - 2 \int_0^\infty \int_0^x (1 - F(y))dy(1 - F(x))dx - 2 \int_0^\theta (\theta - x)F(x)dx \\
= & 2 \left(\int_0^\infty \int_0^x F(y)dy(1 - F(x))dx - \int_0^\theta (\theta - x)F(x)dx \right) \\
= & 2 \left(\int_0^\infty \int_x^\infty (1 - F(y))dyF(x)dx - \int_0^\theta (\theta - x)F(x)dx \right) \\
= & 2 \int_0^\infty \left(\int_x^\infty (1 - F(y))dy - I_{\{x < \theta\}}(\theta - x) \right) F(x)dx \\
\geq & 2 \int_0^\infty (E(L) - \theta)F(x)dx \\
\geq & 0.
\end{aligned}$$

$G''(t) \geq 0$ for any t and $G'(t)$ is an increasing function. Thus, for $0 \leq t \leq E(L) - \theta$, $G(t) \leq \max(G(0), G(E(L) - \theta))$. $G(0) = 0$ and $G(E(L) - \theta) = -(E(L) - \theta)^2 \int_0^\infty (1 - F(x + E(L)))dx < 0$. Thus $G(t) \leq 0$ for any $0 \leq t \leq E(L) - \theta$.

3.1 Alternative models

Although the above assumption that, for a given fragment F1, every overlapping fragment of F1 is equally likely to anneal to F1, is the most probable, other annealing mechanisms are also possible. We considered two alternative models: the maximum overlap model and the

minimum overlap model. Here we just present the relevant results without proof. The idea of the proof is similar to that in Lander and Waterman (1988). In the maximum overlap model, we assume that the fragment with the largest overlap with F1 (at least θ) anneals to F1. Under this model

$$P(D_i > s | D_i > 0) = \frac{\int_{\theta}^{\infty} \exp\{-c \int_x^{\infty} (1 - F(l)) dl\} c(1 - F(s + \theta + x)) dx}{1 - \exp\{-c \int_{\theta}^{\infty} (1 - F(l)) dl\}}.$$

As c tends to infinity, p_2 approximates $\alpha(1 - \alpha)$.

In the minimum overlap model, we assume that the fragment with the shortest overlap (at least θ) with F1 anneals to F1. Under this model

$$P(D_i > s | D_i > 0) = \frac{\int_{\theta}^{\infty} \exp\{-c \int_0^x (1 - F(\theta + l)) dl\} c(1 - F(s + \theta + x)) dx}{1 - \exp\{-c \int_{\theta}^{\infty} (1 - F(l)) dl\}}.$$

As c tends to infinity, p_2 approximates $\alpha(1 - \alpha) \frac{\int_0^t (1 - F(\theta + l)) dl}{\int_0^{\infty} (1 - F(\theta + l)) dl}$.

4 The Experimental Results

In the first reported experiment on DNA shuffling, Stemmer (1994a) reassembled a 1000 base pair (bp) molecule carrying the gene for the *lacZ α* fragment by reassembling 10-50 bp fragments. He recombined two different DNA species with two mutations at loci 75 bp apart within the *lacZ α* gene using equal concentrations of the two DNA species. After shuffling, $n = 386$ colonies were assayed and the fraction of reassembled fragments with both mutations was 24%, very close to the theoretical prediction of 25% according to the above recombination model. Stemmer also reassembled 14 different DNA species with 14 different mutations obtained from the above experiment. The exact locations of the 14 mutations were not specified in that study. It is reasonable to assume that these 14 mutations are distributed along the 1000 bp fragment uniformly. The average distance between adjacent markers is approximately $1000/15 \approx 67$ bp. Thus we expect that these 14 markers recombine independently. The probability that a reassembled fragment contains neither of the 14 mutations should be $(13/14)^{14} \approx 35\%$. Indeed, the observed fraction of reassembled fragments with none of the mutations was 34% ($n = 291$ colonies), very close to the theoretical prediction. Stemmer (1994a) first observed that the recombination between different positions are independent when the distances between them are shorter than the lengths of fragments used for shuffling.

Zhao and Arnold (1997) used DNA shuffling to distinguish functional from nonfunctional mutations. They recombined a wild-type subtilisin E gene with a mutant 1E2A gene with equal concentrations of the two genes. The lengths of the random shuffled fragments were

between 20 to 50 bp. Two mutations, an A to G mutation at position 995 and another A to G mutation at position 1107, are functional as to the thermoactivities of the genes. After DNA shuffling, $n = 768$ clones were sampled and their thermoactivities were measured. About 23% of the sampled clones have both mutations, very close to independent recombination.

There are scant experimental data available for the situation that the lengths of fragments are larger than the distance between the mutations of interest. Stemmer (1994a) recombined two markers 75 bp apart from random 100-200 bp fragments. 11% ($n = 328$ colonies) of the reassembled fragments contained both mutations. Independent recombination was not obtained. We assume that very short overlap is needed for the fragments to anneal, that is, $\theta = 0$. First let us consider the equally likely model. If we take the average length of the fragments, 150 bp, as the length of the fragments used in DNA shuffling, Equation (??) predicts the fraction of reassembled molecules with both mutations to be $3/16 \approx 19\%$, much higher than the observed value. If we assume that the lengths of the fragments have a uniformly distributed on $(100, 200]$, Equation (??) predicts $p_2 = 81/448 \approx 18\%$, which is still very large compared to the observed value. If we assume that the lengths of the fragments have an exponential distribution with mean 150bp, then Equation (??) predicts $p_2 = 0.25(1 - \exp(-75/150)) \approx 10\%$ which is close to the observed value. But the last assumption is not realistic as only fragments of lengths within $(100, 200]$ bp were used in the shuffling experiment.

For the maximum overlap model, we predict independent recombination between any two loci when the coverage is very high. For the minimum overlap model (assuming very

high coverage), if we take $L = 150$ bp, then $p_2 = 1/8 = 12.5\%$ which is close to the observed result. More data are needed to test the validity of the models.

5 Potential Applications

In this section we assume the equally likely model. For general cumulative distribution function $F(x)$ of the lengths of fragments to be shuffled, it is difficult to give a closed formula for the probability of a certain mutation configuration when several different DNA species are shuffled. A computer algorithm has been developed to simulate the shuffling process according to the above equally likely model. This algorithm can be used to approximate the quantities of interests, such as the fraction of DNA shuffling products having certain mutation configurations and the mean largest energy (logarithm of activity) among a sample of DNA shuffling products. In addition it is useful to explore the effects of assumptions about $F(x)$.

In the special case that $F(x) = 1 - \exp(-\lambda x)$, $x > 0$, where $1/\lambda$ is the average length of the fragments used in the shuffling experiments, theoretical studies become possible. From Equation (??), we see that the distances between the right ends of the consecutive fragments along a full length reassembled molecule form a Poisson process. In the rest of this section, we make this assumption.

5.1 Calculating the fraction of molecules with certain mutation configurations

Suppose there are n different types of DNA species which will be referred to as parent molecules. There are m polymorphic loci with mutation M_i , $i = 1, 2, \dots, m$, at the i -th locus along a region of interest. Let the distance between the i -th locus and the $i + 1$ -st locus

be t_i . Each molecule can be represented as a $(0, 1)$ -array of size m with index 1 at position i if the molecule has mutation M_i and 0 otherwise. The j -th type parent molecules will be represented as $(c_{j1}, c_{j2}, \dots, c_{jm})$. In the shuffling experiment, let α_j , $j = 1, 2, \dots, n$ be the fraction of j -th type parent molecules. Denote $P(C_i C_{i+1} \dots C_m)$ as the probability that a randomly chosen reassembled molecule has index $C_k \in \{0, 1\}$, $k \geq i$ at the k -th locus and $P_j(C_i C_{i+1} \dots C_m)$ the corresponding probability given the the $i - 1$ -st and the i -th loci are covered by one fragment coming from an j -th type molecule. In order that a randomly sampled molecule has mutation configuration $C_1 C_2 \dots C_m$, the fragment covering the first locus must come from a parent molecule whose index at the first locus is C_1 . The parent molecule is an j -th type molecule with probability α_j . If the fragment covering the first locus ends after the second locus (with probability $\exp(-\lambda t_1)$), the first and the second loci are on the same fragment. If the fragment covering the first locus ends before the second locus (with probability $1 - \exp(-\lambda t_1)$), the rest of the molecule is randomly reassembled from the parent molecules. Because of the properties of the Poisson processes, we have the following recursive formula

$$P(C_1 C_2 \dots C_m) = \sum_{j=1}^n \alpha_j I_{\{C_1=c_{j1}\}} \left[\exp(-\lambda t_1) P_j(C_2 C_3 \dots C_m) + (1 - \exp(-\lambda t_1)) P(C_2 C_3 \dots C_m) \right].$$

$$P_j(C_2 C_3 \dots C_m) = I_{\{C_2=c_{j2}\}} \left[\exp(-\lambda t_2) P_j(C_3 \dots C_m) + (1 - \exp(-\lambda t_1)) P(C_3 \dots C_m) \right].$$

From the above formula, it is possible to calculate the fraction of reassembled molecules having a certain mutation configuration. In particular, when we shuffle equal concentrations of m different types of molecules, of which the i -th type molecules have a mutation at the

i -th locus, we have

$$P(11 \cdots 1) = \prod_{i=1}^{m-1} (1 - \exp(-\lambda t_i)) / m^m.$$

5.2 Maximizing the mean largest energy among a sample of DNA shuffling products

In some applications, the need for screening rather than selection for the biological function of interest severely limits our ability to screen very large number of colonies. If only a limited number of colonies, say S , are screened after DNA shuffling, it is not always optimal to recombine all the available molecules of interest (Moore et al. 1997). They studied the number of molecules to be recombined to maximize the largest energy (the natural logarithm of the activity) among a sample of S molecules using simulation approaches assuming the mutations of interest can recombine independently and equal concentration of different molecules in DNA shuffling.

If the energies of different species of molecules are known *a priori*, it is reasonable to put different concentrations of the available molecules in DNA shuffling experiments to maximize the largest energy among a sample of S shuffling products. It is also reasonable to choose the length of fragments to be reassembled to achieve this goal. Next we formulate this problem into an optimization problem. Suppose the energy of the molecules with mutation M_n is e_n . We also assume that the energies are additive.

Using the above formula, we can calculate the probability of a certain mutation configu-

ration of a randomly chosen DNA shuffling product. Define the random variable

$$X_s = \sum_i e_i I_s\{M_i\},$$

where $I_s\{M\} = 1$ if the molecule has mutation M and equals 0 if it does not have the mutation. The target function is

$$T = E(\max_{1 \leq s \leq S} X_s).$$

The determination of the length of fragments to be shuffled and the concentrations of different molecules to maximize the target function is a challenging computational problem.

5.3 Physical mapping using DNA shuffling

There are many physical mapping methods such as radiation hybrid mapping (Cox et al. 1990), optical mapping (Schwartz et al. 1993), STS mapping, and physical mapping by fingerprinting random clones (see Lander and Waterman 1988 for some references to classical experiments). It is also possible to order marker loci using genetic mapping methods such as pedigree analysis and sperm typing (Boehnke et al. 1989). DNA shuffling provides an alternative method of physical mapping. Compared to radiation hybrid mapping, DNA shuffling is experimentally simple. It does not need the enzyme hypoxanthine phosphoribosyl transferase (HPRT) to rescue the hybrid cells. This property might make DNA shuffling a preferred technology to do physical mapping for some organisms that lack enzymes to distinguish hybrid and non-hybrid cells. As shown below, physical mapping using DNA

shuffling needs DNA molecules heterozygous for the marker loci of interest. In radiation hybrid mapping, any DNA molecules, not necessarily heterozygous for the marker loci, can be used. Sperm typing can be applied to make high resolution genetic maps of human chromosomes due to the availability of large number of sperm that can be typed. When the marker loci are close, many sperm need to be typed to obtain accurate estimation of the order of marker loci and the distances between them. For example, to estimate a recombination fraction of $\theta = .01$ with maximum probable error of $.25\theta$, over 6,000 sperm need to be typed (Boehnke et al. 1989). In DNA shuffling, by suitably choosing the lengths of fragments to be reassembled, we can control the probability of crossovers between the marker loci of interest and thus greatly reduce the number of clones that need to be typed to estimate the distances between loci.

Next we indicate how to make physical maps using DNA shuffling. As in genetic mapping using sperm typing, we first need individuals heterozygous for at least three loci. The number of individuals needed to achieve this goal was given in Boehnke et al. (1989). The DNA from such individuals are collected and used in DNA shuffling. Suppose an individual is heterozygous for m marker loci. Let the alleles at one strand be A_i , $i = 1, 2, \dots, m$ and the alleles at the other strand be B_i , $i = 1, 2, \dots, m$. We denote the A allele as 1 and the B alleles as 0. As above, let t_i be the distance between locus i and locus $i + 1$. Certainly, here, equal concentrations of the two types of molecules are used for DNA shuffling. After DNA shuffling, we sample many molecules and type each molecule at the m loci. Each sampled molecule may have alleles A_i , B_i or ? at the i -th locus, where ? means locus i is not typed

or ambiguous.

To order these m loci, we can use the criterion of minimizing the number of obligate breaks for these m molecules (Cox et al. 1990). The number of obligate breaks of a molecule is defined as the number of times runs of A's and B's change. The order of the marker loci is taken as the order that have the minimum number of obligate breaks across all the molecules. Because in DNA shuffling the crossover probability increases as the distance between two loci increases, this criterion is statistically consistent, that is, as the number of typed molecules tends to infinity, the estimated order converges to the true order (Lange 1997). The criterion of minimum obligate breaks was used in radiation hybrid mapping and algorithms have been developed to find the marker orders having the minimum number of obligate breaks (Boehnke and Lange 1991, Lange 1997).

As in radiation hybrid mapping, the advantage of the minimum number of breaks criterion is that it does not depend on the model of crossovers between different species. The disadvantage is that it does not provide estimates of physical distances between loci nor the comparisons of likelihoods of competing orders. To overcome the disadvantages, a maximum likelihood approach should be used. We denote the A alleles as 0 and the B alleles as 1. The DNA molecules with all A alleles as type 0 DNA molecules and DNA molecules with B alleles as type 1 DNA molecules. Using the above general recursive formula, here we have

$$P(C_1 C_2 \cdots C_m) = \frac{1}{2} \left[\exp(-\lambda t_1) P_{C_1}(C_2 C_3 \cdots C_m) + (1 - \exp(-\lambda t_1)) P(C_2 C_3 \cdots C_m) \right].$$

$$P_{C_1}(C_2 C_3 \cdots C_m) = I_{\{C_2=C_1\}} \left[\exp(-\lambda t_2) P_{C_2}(C_3 \cdots C_m) + (1 - \exp(-\lambda t_2)) P(C_3 \cdots C_m) \right].$$

From this formula we can calculate the likelihood of a particular reassembled molecule. The log-likelihood of many shuffled molecules is the sum of the log-likelihoods across all the molecules. It is then possible to obtain the maximum likelihood estimation of the order and the distances between the loci. The likelihood function in this situation is different from that in radiation hybrid mapping, and new algorithms need to be developed.

6 Discussion

In this paper, we construct a probabilistic model for DNA shuffling. First we model the distribution of random fragments along the genome using the Lander-Waterman model for physical mapping by fingerprinting random clones. Based on this model, we give corresponding results for the distribution of regions that can be reassembled through DNA shuffling. These results correspond to the Lander-Waterman results for the distribution of islands and contigs along the genome. Then we model the recombination between two markers. We show that if the lengths of fragments to be reassembled are smaller than the distance between the two markers, recombination is complete, *i.e.*, the two markers recombine independently. On the other hand when the lengths of random fragments are larger than the distance between the two markers, the two markers do not recombine independently. We derive a formula for the probability of recombination between two loci. Based on this formula, we give the probability that a reassembled fragment has two, one and none of the two mutations. We compare this model with experimental results. This recombination model can easily be extended to multiple markers. Based on this model, we can study many problems related to DNA shuffling. We give three applications of the theoretical results: calculating the fraction of DNA shuffling products having a certain number of mutations, maximizing the mean largest energy among a sample of DNA shuffling products, and constructing physical maps using DNA shuffling.

Recently two new alternative shuffling methods: random-priming *in vitro* recombination

(RPR) (Shao et al. 1998) and staggered extension process (StEP) *in vitro* recombination (Zhao et al. 1998), were proposed. In RPR, instead of using DNase I to randomly break each DNA strand, PCR with random primers was used to prepare fragments used in DNA shuffling. The recombination models presented in this paper can be directly applied to RPR. In StEP, two fixed primers flanking the target region were used and very short extension time was carried out in each PCR cycles. This process will generate many growing strands and these growing strands anneal randomly to the parent template strands in the following PCR cycles. Finally full length reassembled molecules can be obtained. Here, as in the above model, each reassembled molecule is a concatenation of fragments randomly selected from the template sequences. The distribution of the lengths of the consecutive fragments equals the distribution of the lengths of extension at each PCR cycle.

DNA shuffling has been used in many *in vitro* evolution experiments. Mutagenesis, in particular DNA shuffling, is only one component in *in vitro* evolution experiments. The problem of incorporating the model of DNA shuffling into models of *in vitro* evolution is a project for future research.

Acknowledgments

I thank Dr. Stemmer and Dr. Arnold for their comments on a preliminary version of the paper and for making their preprints available to me. I also thank Dr. Arnheim and Dr. Speed for suggestions that improved the presentation of the manuscript. I am particularly indebted to Dr. Waterman for introducing me to the problem of DNA shuffling and for his advice and input during the preparation of this manuscript. Without his help, this work could never have been finished. This paper is supported in part by a grant from the Research Council of Emory University and NIH FIRST award from NIDDK R29DK53392.

References

1. Arnold FH, Moore JC (1997) Optimizing industrial enzymes by directed evolution. *Advances in Biochemical Engineering-Biotechnology* 58:1-14
2. Arratia R, Lander ES, Tavare S, Waterman MS (1991) Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* 11:806-827
3. Boehnke M, Arnheim N, Li H, Collins FS (1989) Fine structure genetic mapping of human chromosomes using the polymerase chain reaction on single sperm: experimental design considerations. *Am J Hum Genet* 45:21-32
4. Boehnke M, Lange K, Cox D (1991) Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet* 49:1174-1188
5. Britten RJ, Graham DE, Neufeld BR (1974) Analysis of repeating DNA sequences by reassociation. *Methods in Enzymology*. 29:363-418
6. Cadwell RC, Joyce GF (1992) Randomization of genes by PCR mutagenesis. *PCR Method Applic* 2:28-33
7. Cramer A, Raillard S, Bermudez E, Stemmer WPC (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391:288-291
8. Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) A somatic cell genetic method for constructing high resolution maps mammalian chromosomes. *Science* 250:245-250

9. Fitzwater T, Polisky B (1996) A SELEX primer. *Methods in Enzymology* 267:275-301
10. Gold L, Polisky B, Uhlenbeck O, Yarus M (1995) Diversify of oligonucleotide functions. *Annual Review of Biochemistry* 64:763-797
11. Irvine D, Tuerk C, Gold L (1991) SELEXION: Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *J Mol Biol* 222:739-761
12. Kuchner O, Arnold FH (1997) Directed evolution of enzyme catalyts. *Trends in Biotechnology* 15:483-531
13. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231-239
14. Lange K (1997) *Mathematical and statistical methods for genetic analysis*. Springer-Verlag, New York
15. Leung DW, Chen E, Goeddel DV (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* 1:11-15
16. Moore CM, Jin HM, Kuchner O, Arnold FH (1997) Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J Mol Biol* 272: 336-347
17. Patten PA, Howard RJ, Stemmer WPC (1997) Applications of DNA shuffling to pharmaceuticals and vaccines. *Current Opinion in Biotechnology* 8:724-733

18. Reidhaar-Olson JF, Sauer RT (1988) Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science* 241:53-57
19. Schwartz D, Li X, Hernandez L, Ramnarain S, Huff E, Wang Y (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* 262:110-114
20. Shao Z, Arnold FH (1996) Engineering new functions and altering existing functions. *Current Opinion in Structural Biology* 6:513-518
21. Shao Z, Zhao H, Giver L, Arnold FH (1998) Random-priming *in vitro* recombination: an effective tool for directed evolution *Nucleic Acids Res* 26:681-683
22. Smith GP (1994) The progeny of sexual PCR. *Nature* 370:424-325
23. Stemmer WPC (1994a) DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc Natl Acad Sci USA* 91:10747-10751
24. Stemmer WPC (1994b) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370:389-391
25. Sun F, Galas D, Waterman MS (1996) A mathematical analysis of *in vitro* molecular selection-amplification. *J Mol Biol* 258:650-660
26. Stemmer WPC (1995) Searching sequence space. *Bio/technology* 13:549-553
27. Waterman MS (1995) *Introduction to Computational Biology*. Chapman & Hall, London, UK

28. Wetmur JG, Davidson N (1968) Kinetics of renaturation of DNA. *J Mol Biol* 31:349-70
29. Zhang JH, Dawes G, Stemmer WPC (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc Natl Acad Sci USA* 94:4504-4509
30. Zhao H, Arnold FH (1997) Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc Natl Acad Sci* 94:7997-8000
31. Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nature Biotechnology* 16:258-261

Table 1. The probability that a target of length $t = 25$ can be reassembled through DNA shuffling with fragment length 1 for different values of coverage c and overlap fraction θ .

	Coverage c						
θ	4	5	6	7	8	9	10
0.00	0.15	0.44	0.69	0.86	0.94	0.98	0.99
0.25	0.00	0.01	0.18	0.41	0.62	0.77	0.87
0.50	0.00	0.00	0.00	0.00	0.02	0.08	0.19

Figure Legends

Figure 1 PCR without primers. a) two double-stranded DNA fragments; b) The two double-stranded fragments are separated into single-stranded ones through denaturing; c) Overlapping fragments anneal to each other; d) 5' overhangs are extended by polymerase extension and 3' overhangs are not changed.

Figure 2 The extension of the i -th fragment along a reassembled molecule. l_i : the position of the right end of the i -th fragment; D_i : the distance from the right end of the $i - 1$ -st fragment to the right end of the the i -th fragment; X : the distance from $l_i - \theta$ to the left end of the annealed fragment; and L : the length of annealed fragment.

Figure 3 Recombination of two DNA species in DNA shuffling. a) The two DNA species with different mutations (*); b) Break by DNase I; c) Fragments anneal to each other; d) Recombine to obtain a molecule with both mutations.