J. theor. Biol. (2000) **205,** 483–503
doi:10.1006/jtbi.2000.2082, available online at http://www.idealibrary.com on IDEAL®

Modeling DNA Mutation and Recombination for Directed Evolution Experiments

GREGORY L. MOORE AND COSTAS D. MARANAS*

Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, *U.S.A.*

(*Received* 28 *October* 1999, *Accepted in revised form* 15 *April* 2000)

Directed evolution experiments rely on the cyclical application of mutagenesis, screening and amplification in a test tube. They have led to the creation of novel proteins for a wide range of applications. However, directed evolution currently requires an uncertain, typically large, number of labor intensive and expensive experimental cycles before proteins with improved function are identified. This paper introduces predictive models for quantifying the outcome of the experiments aiding in the setup of directed evolution for maximizing the chances of obtaining DNA sequences encoding enzymes with improved activities. Two methods of DNA manipulation are analysed: error-prone PCR and DNA recombination. Error-prone PCR is a DNA replication process that intentionally introduces copying errors by imposing mutagenic reaction conditions. The proposed model calculates the probability of producing a specific nucleotide sequence after a number of PCR cycles. DNA recombination methods rely on the mixing and concatenation of genetic material from a number of parent sequences. This paper focuses on modeling a specific DNA recombination protocol, DNA shuffling. Three aspects of the DNA shuffling procedure are modeled: the fragment size distribution after random fragmentation by DNase I, the assembly of DNA fragments, and the probability of assembling specific sequences or combinations of mutations. Results obtained with the proposed models compare favorably with experimental data.

© 2000 Academic Press

Introduction and Background

Unprecedented opportunities are now within our reach for generating novel enzymes and biocatalysts using sophisticated techniques that mutate, recombine and amplify nucleic acid sequences. Such nucleic acid manipulations are exploited within the framework of directed evolution experiments pioneered by Stemmer (1994a, b) and Arnold (1996). In directed evolution the processes of natural evolution are accelerated in a test tube for selecting proteins with the desired properties. A typical experimental cycle of directed evolution begins with the selection of a library of parent DNA sequences encoding for proteins that involve to some extent the sought after property. The diversity of sequences being explored is next increased through the *mutagenesis* step by introducing random point nucleotide mutations and/or by recombining DNA fragments. The mutagenesis and fragmentation step renders all but very few of the sequences inactive. These DNA sequences are then ligated into an expression vector and transformed into *Escherichia coli* cells. A *screening* procedure is next employed to isolate the few out of the many *E. coli* transformants containing the

*Author to whom correspondence should be addressed. E-mail: costas@psu.edu

© 2000 Academic Press

sequences encoding for active enzymes or functional proteins. These selected sequences are then *amplified* and the cycle of mutagenesis, screening and amplification is repeated multiple times until proteins with the desired property or function are found. Recently, remarkable successes of directed evolution have been reported, ranging from industrial enzymes with substantially improved activities and thermostabilities to vaccines and pharmaceuticals (Schmidt-Dannert & Arnold, 1999). These successes mark the onset of enormous possibilities for future uses of directed evolution in basic research for understanding protein function and in industry for creating new biocatalysts.

Except for the work of Moore *et al.* (1997) which examines the effect of library size and screening capacity, directed evolution experiments have largely been based on empirical information and lack quantitative description of the DNA recombination process. Although they have led to exciting successes, directed evolution methods require a large number of expensive and time-consuming mutagenesis and/or recombination experiments and often many proteins must be screened before one with the desired property is identified. The enormous potential of these methods will be better realized if the experimental design were improved to be more efficient and less expensive. This challenge provides the main motivation for this paper in which the necessary modeling framework to enable prediction of size, nucleotide sequence, and activity information in directed evolution experiments is developed.

A key step in the directed evolution experimental cycle is the introduction of new genetic diversity to the library. There are two basic ways for introducing diversity: error-prone PCR and DNA recombination. Error-prone PCR protocols were used in early directed evolution experiments (Arnold, 1996). Polymerase chain reaction (PCR) is a DNA amplification technique in which an initial small amount of DNA is replicated in consecutive cycles increasing its concentration exponentially (see Fig. 1).

The error-prone PCR replication process (Leung *et al.*, 1989; Cadwell & Joyce, 1992; Lin-Goerke *et al.*, 1997) intentionally introduces copying errors by imposing mutagenic reaction conditions (e.g. through the addition of Mn^{2+} or Mg^{2+}). The first step of PCR is the denaturization of the DNA into single strands. The second step is the annealing of a primer to the DNA single strands. Primers consist of two DNA oligonucleotides with lengths of 15–30 base pairs complementary to the ends of the amplified region. The third step is primer extension by a polymerase (typically *Taq*). Nucleotides complementary to the single-strand template are added by using the original sequence as a template, extending the complementary strands until normal DNA double strands are recovered. Unavoidable mutations occur in this step when non-complementary nucleotides are incorporated into the chain. Eckert & Kunkel (1991)

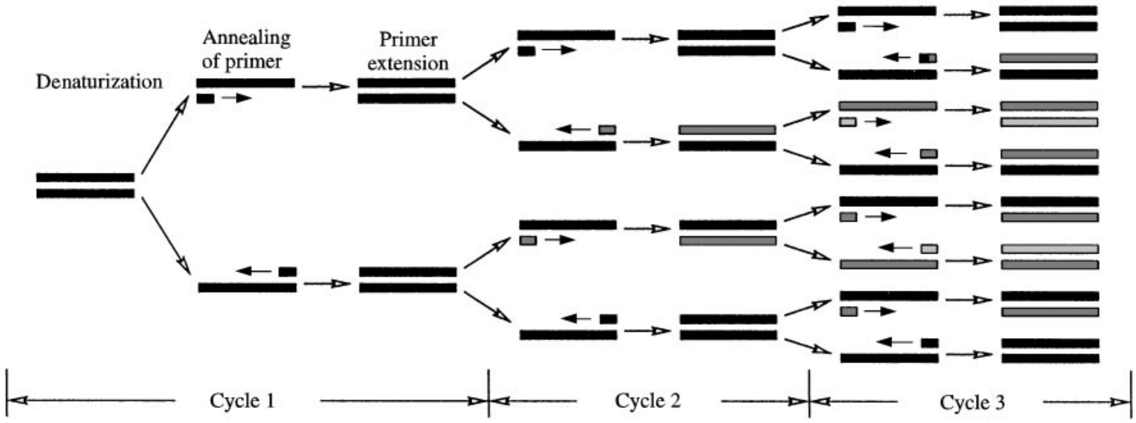

FIG. 1. Three cycles of PCR produce $2^3 = 8$ total strands after the third cycle, or 16 single strands of nucleotides. Of these 16, two are the original DNA double strand, six are the result of one extension step, six are the result of two extension steps, and two are the result of three extension steps. Strands are shown more lightly shaded as they undergo more extension steps.

report mutation rates for *Taq* ranging from 10^{-7} up to 10^{-3} mutations per nucleotide polymerized. These mutation rates are nucleotide dependent (Cadwell & Joyce, 1992; Shafikhani *et al*., 1997). The control of these highly variable (spanning four orders of magnitude) copying errors is vital for mutagenesis since the "right" number of mutations will provide just enough diversity for evolutionary advancement without producing a build-up of deleterious errors. However, the ability of error-prone PCR *alone* to successively improve a DNA sequence through continuously improving single-point mutations is somewhat limited since the build-up of deleterious mutations typically overwhelms the beneficial ones. This has been recognized by researchers and currently DNA recombination, capable of filtering out deleterious mutations while retaining the improving ones, is employed in directed evolution experiments.

Unlike error-prone PCR where no exchange of genetic material occurs between parent sequences, DNA recombination methods rely on the mixing and concatenation of genetic material from a number of parent sequences. Recombination protocols include DNA shuffling (sexual PCR) (Stemmer, 1994a, b), staggered extension process (StEP) (Zhao *et al*., 1998), random-priming recombination (RPR) (Shao *et al*., 1998), and incremental truncation (Ostermeier *et al*., 1999). A thorough review of currently employed DNA recombination protocols can be found in Volkov & Arnold (1999). Directed evolution experiments utilizing DNA recombination (shuffling) as the mutagenesis step are briefly described as follows (see also Fig. 2).

First an initial set of parent sequences sharing a number of desired traits are selected for recombination. Next, the selected sequences undergo *random fragmentation* typically using DNase I. Double-stranded fragments within a certain size range (e.g., 100–200 bp) are retained. The retained fragments are then *reassembled* by thermocycling with a DNA polymerase (PCR *without added primers*). As in regular PCR, this involves first the *denaturization* of the double-stranded fragments into single-stranded ones. Denaturing is followed by *annealing* where single-stranded fragments anneal to other fragments overlapping by a sufficiently large number of complementary bases to form 3′ or 5′ overhangs. The third step is *polymerase extension* (see Fig. 2). Note that the 3′ overhangs are not changed because DNA polymerase only possesses 5′ → 3′ activity. These three steps are repeated and the average fragment length increases after each cycle. After a number of cycles, DNA sequences of the original length are obtained. Finally, regular PCR with primers

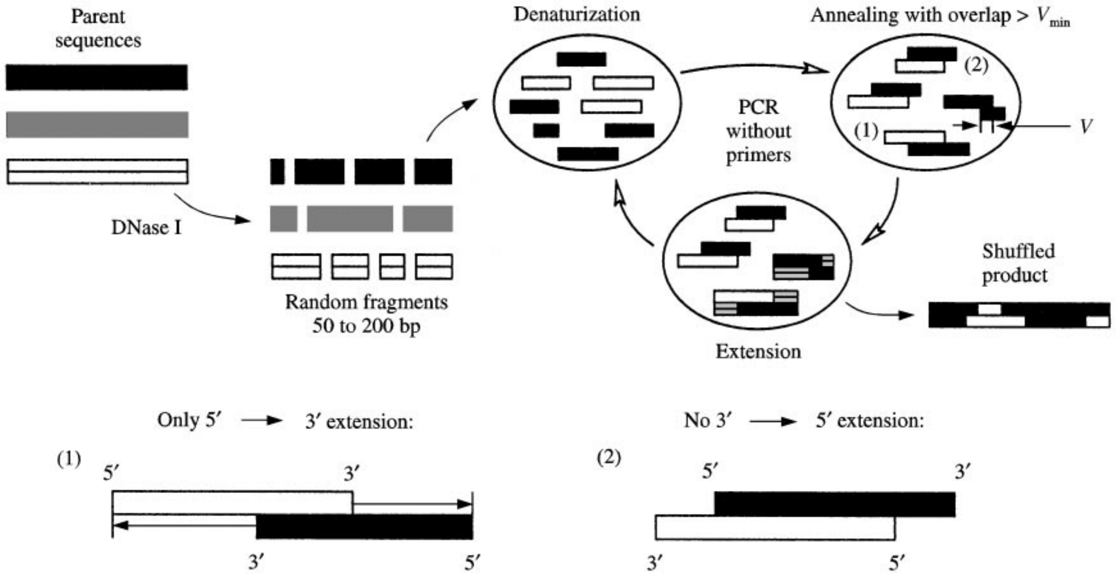

FIG. 2. DNA shuffling occurs in three steps, the most important of which is a PCR reaction without primers in which reassembly of parent sequences occurs. The product will have a combination of genetic features from all of the parent sequences.

is utilized to amplify the reassembled strands. The key advantage of DNA shuffling over error-prone PCR is that it can recombine a large number of mutations within a few selection cycles quickly yielding functional blocks with combinations of beneficial mutations.

In recent years, directed evolution principles have been successfully applied to enhance a number of protein properties. These include enhancements in enzyme thermostability (Arnold & Moore, 1997; Kuchner & Arnold, 1997; Zhao & Arnold, 1997b; Giver *et al*., 1998; Matsumura & Ellington, 1999; Lin *et al*., 1999) and psychrophilicity (Taguchi *et al*., 1998); alterations in substrate specificity (Zhang *et al*., 1997; Kumamaru *et al*., 1998; Hansson *et al*., 1999) and foreign media activity (Chen & Arnold, 1993; Moore & Arnold, 1996); improved stereoselectivity (Reetz *et al*., 1997; Bornscheuer *et al*., 1998); development of pharmaceuticals and vaccines (Patten *et al*., 1997); bioremediation of polychlorinated biphenyls (Wackett, 1998); detoxification of an arsenate pathway (Crameri *et al*., 1997); augmentation of the stability of folded antibody fragments (Proba *et al*., 1998; Martineau *et al*., 1998); and increased sensitivity to AZT for HIV research (Christians *et al*., 1999).

At the same time that new directed evolution success stories are published and the potential for discovering truly novel biocatalysts is gaining acceptance, it is becoming apparent that the process is limited by key unanswered questions regarding the optimal mix, scheduling and setup of error-prone PCR and DNA recombination steps; the optimal selection of parent sequences for recombination; and the effect of parameters such as recombinatory fragment length, annealing temperature and number of shuffling cycles on the assembly of full length product sequences. To answer these questions, a set of quantitative models are introduced. The remainder of the paper is organized as follows. In the next section, a model of error-prone PCR is presented, and the predictions are compared to experimental data. Then, three models describing the DNA shuffling process are discussed. The first (random fragmentation model) describes the fragment size distribution after treatment with DNase I. The second (fragment assembly model) predicts the fragment size distribution after each annealing/extension step. The third (sequence matching model) estimates the fraction of fully assembled genes whose nucleotide sequence matches a target one. For all models, examples are provided along with comparisons with experimental data.

Modeling Error-prone PCR

While lately error-prone PCR has been largely replaced by DNA recombination as the mutagenesis step, modeling single-point mutations is still important since they will occur within any recombination protocol. Quantitative studies of PCR have so far addressed PCR efficiency (Weiss & Haeseler, 1995), reaction kinetics (Hsu *et al*., 1997), effect of annealing temperatures (Rychlik *et al*., 1990), and primer lengths (Wu *et al*., 1991; Sakuma & Nishigaki, 1994). Eckert & Kunkel (1991) proposed the following simple equation:

$$f = \frac{Np}{2}$$

for predicting the overall error rate f after N PCR cycles given that the per cycle error rate is p. This relation does not account for the fact that copying errors depend on the nucleotide being replicated. For example, A miscopies to C, G or T with different probabilities (Shafikhani *et al*., 1997; Cadwell & Joyce, 1992; Lin-Goerke *et al*., 1997). This omission thus may yield inaccurate estimates.

In the proposed model, mutations that occur during the extension step when nucleotides are added via polymerase are treated as being nucleotide dependent. A per cycle mutation matrix $\mathscr{M}$ is defined that models these different mutation rates with elements M_{ij} representing the probability of nucleotide i mutating to nucleotide j:

$$\mathscr{M} = \begin{pmatrix} M_{AA} & M_{AT} & M_{AC} & M_{AG} \\ M_{TA} & M_{TT} & M_{TC} & M_{TG} \\ M_{CA} & M_{CT} & M_{CC} & M_{CG} \\ M_{GA} & M_{GT} & M_{GC} & M_{GG} \end{pmatrix}.$$

These values depend on the experimental conditions. The per cycle mutation rate matrix $\mathscr{M}$ can then be used to identify the mutation rate matrix $\mathscr{C}^n$ after n extension steps. This matrix measures the mutation rates of a sequence obtained after n extension events starting from the original sequence. Because the occurrence of mutations in one extension step is independent of mutations that occurred in previous extension steps a recursive relation for $\mathscr{C}^n$ is derived as follows:

$$C^n_{ij} = \begin{cases} \delta_{ij}, & n = 0, \\ M_{ij}, & n = 1, \\ \sum\limits_{k=A,C,T,G} M_{kj} C^{n-1}_{ik}, & n \geqslant 2, \end{cases}$$

where δ_{ij} equals one if $i = j$ and zero otherwise.

However, after N PCR cycles not all sequences in the reaction mixture result after exactly N extensions of the original sequence. This is due to the fact that after a sequence is formed, it remains in the mixture to serve as a template in subsequent extension steps. For example, after three PCR cycles (see Fig. 1), 16 single strands of DNA are produced, of which two are the original DNA double strand ($n = 0$), six are the result of one extension step ($n = 1$), six are the result of two extension steps ($n = 2$), and two are the result of three extension steps ($n = 3$).

This result is generalized for N PCR cycles (see Fig. 3). In Appendix A, it is proven by induction that after N PCR cycles the number of sequences which are the product of exactly n extensions of the original DNA strand is equal to

$$Z_{N,n} = 2\binom{N}{n}.$$

The total number of single-stranded sequences present in the reacting mixture after N PCR cycles is equal to

$$2 \cdot 2^N$$

since every PCR cycle doubles their number. Therefore, the fraction of the sequences present in the reaction mixture after N PCR steps that are the result of n extension events is equal to

$$\frac{1}{2^N}\binom{N}{n}.$$

This relation is used in conjunction with matrix $\mathscr{C}^n$ to construct matrix $\mathscr{P}^N$ with elements P^N_{ij} representing the probability of nucleotide i mutating to nucleotide j after N PCR cycles:

$$P^N_{ij} = \frac{1}{2^N}\sum_{n=0}^{N}\binom{N}{n} C^n_{ij}.$$

By exploiting the assumption that mutations at different locations along the sequence are independent of each other, the probability $\prod^N_{S^0,S}$ of assembling a sequence S through successive single point mutations on an original sequence S^0 after N PCR cycles is given by

$$\Pi^N_{S^0,S} = \frac{1}{2^N}\sum_{n=0}^{N}\binom{N}{n}\prod_{j=1}^{B}\,[P^n]_{s^0_j,s_j},$$

where B is the length of the two sequences and s^0_j and s_j are the nucleotides at position j for sequences S^0 and S, respectively. This relation provides the quantitative means to *a priori* estimate the fraction of the sequences obtained after

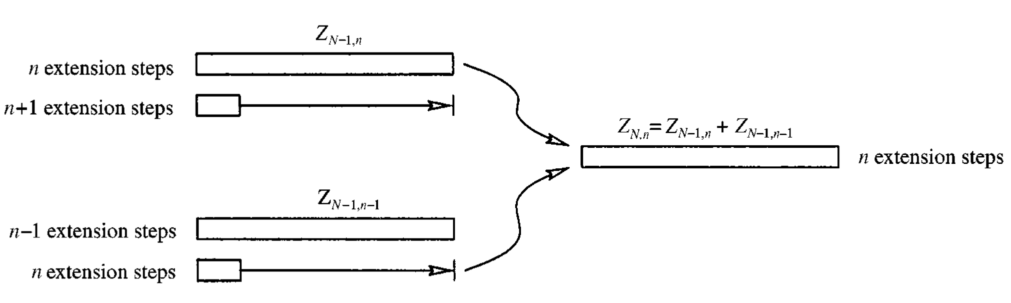

FIG. 3. After N PCR cycles, the reaction mixture contains a number of strands that have been through n extension steps. These strands ($Z_{N,n}$) originate from either (i) old templates that have been through n extension steps prior to the N-th PCR cycle ($Z_{N-1,n}$); or (ii) new strands extended from templates that had already been through $n - 1$ extension steps ($Z_{n-1,n-1}$).

TABLE 1

An example of mutation matrix calculation given reported mutation bias for zero Mn^{2+} concentration

PCR mutation matrix after 13 cycles (Shafikhani *et al.*, 1997)

$$\mathscr{P}^{13} = \left[\begin{array}{cccc} 99.522\% & 0.227\% & 0.046\% & 0.205\% \\ 0.227\% & 99.522\% & 0.205\% & 0.046\% \\ 0.046\% & 0.137\% & 99.817\% & 0.000\% \\ 0.137\% & 0.046\% & 0.000\% & 99.817\% \end{array}\right] \begin{array}{c} (A) \\ (T) \\ (C) \\ (G) \end{array}$$

Calculated mutation matrix

$$\mathscr{M} = \left[\begin{array}{cccc} 99.926\% & 0.035\% & 0.007\% & 0.032\% \\ 0.035\% & 99.926\% & 0.032\% & 0.007\% \\ 0.007\% & 0.021\% & 99.972\% & 0.000\% \\ 0.021\% & 0.007\% & 0.000\% & 99.972\% \end{array}\right] \begin{array}{c} (A) \\ (T) \\ (C) \\ (G) \end{array}$$

Average per-cycle mutation rate calculated = 0.016%
Reported per-cycle mutation rate (Ling *et al.*, 1991) = 0.02%

N PCR steps that conform to some target sequence S given the mutation matrix $\mathscr{M}$. Therefore, by adjusting the reaction conditions to control the mutation rate, an experimenter can control the probability of achieving a desired target sequence.

Next, the proposed model is verified by calculating the per cycle mutation rate matrix $\mathscr{M}$ given the mutation rate matrix $\mathscr{P}^{13}$ reported by Shafikhani *et al.* (1997) after 13 PCR cycles (see Table 1). The average per-cycle mutation rate, assuming an equal concentration of each type of nucleotide throughout the sequence, is calculated to be 0.016%. Note that the data presented by Shafikhani *et al.* (1997) correspond to experimental conditions similar to the ones reported by Ling *et al.* (1991). In the latter PCR study, an average per-cycle mutation rate of 0.02% is reported, which is very close to the value 0.016% that the proposed model predicts. Fig. 4 illustrates the effect of the sequence GC content on the total number of mutations expected after 12 PCR cycles. Data from error-prone PCR with no Mn^{2+} added (Shafikhani *et al.*, 1997) is used to derive the per-cycle mutation matrix. As shown in Fig. 4, a GC rich strand can reduce the number of mutations produced by almost one-half.

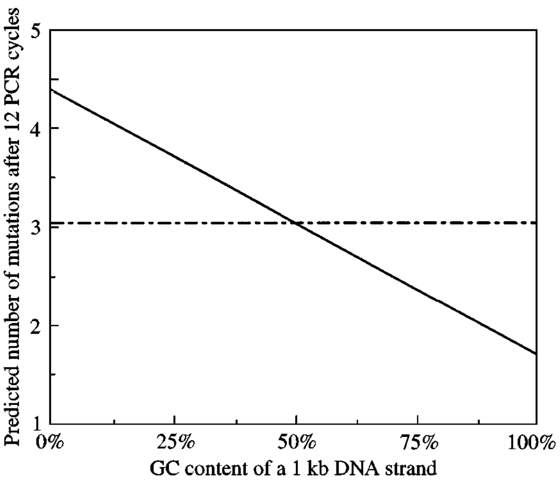

FIG. 4. The GC content of a DNA strand can significantly alter the number of mutations produced by error-prone PCR. Data shown here is for a 12-cycle PCR with no Mn^{2+} added. Moore & Maranas (——), $f = Np/2$ (Eckert & Kunkel, 1991) (————-).

In the proposed model, the PCR efficiency is assumed to be 100% meaning that the amount of DNA present doubles from one cycle to the next. In practice, this is not always true since a lack of excess primer or nucleotides may result in incomplete amplification. This assumption affects both the calculation of the amount of DNA present after N cycles and $Z_{N,n}$. For a PCR efficiency ε an amplification of $(1 + \varepsilon)^N$ instead of 2^N is achieved. The calculation of $Z_{N,n}$ also needs to be changed. Furthermore, it is assumed that no mutational "hot spots", or positions in the sequence with an increased mutation rate, are produced. The lack of "hot spots" is reported by Cadwell & Joyce (1992) and also by Shafikhani *et al.* (1997).

Finally, nucleotide insertions and/or deletions are not modeled because such events are reported to comprise less than 5% of all mutations (Shafikhani *et al.*, 1997). Nevertheless, by augmenting the mutation matrix $\mathscr{M}$ to include deletions and insertions in addition to nucleotide mutations such events can be accommodated at the expense of increased dimensionality.

Modeling DNA Recombination

The modeling of three different aspects of the DNA recombination process is addressed:

1. *Random fragmentation model.* In this model the size distribution of the DNA fragments after treatment of the parent sequences with DNase I is examined. This provides the necessary quantitative information regarding fragment size distribution necessary for modeling the subsequent DNA shuffling step.
2. *Fragment assembly model.* Given the initial fragment size distribution, the objective here is to model the fragment size distribution after each annealing/extension step. This allows tracking of how effectively the recombination protocol assembles full length genes without regard to sequence or function of the assembled sequences.
3. *Sequence matching model.* After all shuffling cycles have been completed, the fraction of fully-assembled genes whose nucleotide sequence matches a given target (e.g., AGGTCC) is quantified.

RANDOM FRAGMENTATION MODEL

After a gene of length B is treated with DNase I (random fragmentation), a random distribution of nucleotide fragments is obtained. Random fragmentation implies that each one of the $B-1$ nucleotide-nucleotide bonds has an equal probability P_{cut} of being broken. The resulting fragment size probability distribution denoted by Q_L^0 is desired to describe the fraction of fragments of different lengths L present in the reaction mixture.

First the special case $L = B$ is addressed. The only possible way for a fragment of length B to result is if none of the $B-1$ bonds are cut. The probability of a single bond remaining intact is $(1 - P_{cut})$. The random nature of fragmentation implies that bond-breaking events are independent therefore

$$Q_B^0 = (1 - P_{cut})^{B-1}.$$

While the generation of a fragment of length B requires that all $B-1$ bonds must remain intact, a fragment of length L can be formed after having different numbers of bonds being broken. The total number of broken bonds cannot exceed $B-L$ because in that case at least one of the $L-1$ bonds in a fragment of length L must break. Therefore, the calculation of Q_L^0 requires enumerating all possible ways of generating a fragment of length L after breaking $s = 1, \ldots, B-L$ bonds. Mathematically, this implies that Q_L^0 is equal to the sum of the products of the conditional probabilities $P_{L|s}$ of generating a fragment of length L given that s bonds are broken times the probability P_s of breaking s bonds:

$$Q_L^0 = \sum_{s=1}^{B-L} P_s P_{L|s}, \quad L = 1, \ldots, B-1.$$

There exist $\binom{B-1}{s}$ alternatives for breaking s out of $B-1$ bonds. Because bond cutting and bond preservation are independent events, each one of these alternatives has a probability

$$(P_{cut})^s (1 - P_{cut})^{B-1-s}$$

of occurring. By combining these two results we obtain

$$P_s = \binom{B-1}{s} P_{cut}^s (1 - P_{cut})^{B-1-s}.$$

Random fragmentation implies that the order in which fragments are produced does not affect their respective probabilities of occurrence. For example, two cuts that produce fragments of lengths a, b, and c occur with the same probability as two cuts that produce fragments of lengths c, a, and b. This greatly simplifies the analysis by allowing the placement of the fragment of length

L at the beginning without any loss of generality. Specifically, given that after breaking s bonds a fragment of length L is formed, the formation of the fragment of length L can be assumed to occur first without any loss of generality. This means that there exists

$$\binom{B-1-L}{s-1}$$

alternatives to form the remaining $s-1$ cuts. Each one of these alternatives signifies a way of generating a fragment of length L. Because there exist

$$\binom{B-1}{s}$$

ways of creating s cuts, the conditional probability $P_{L|s}$ is equal to

$$P_{L|s} = \binom{B-1-L}{s-1} \Big/ \binom{B-1}{s}.$$

By combining the expressions for P_s and $P_{L|s}$ the following result for Q_L^0 is obtained:

$$Q_L^0 = \sum_{s=1}^{B-L} \binom{B-1-L}{s-1} P_{cut}^s (1-P_{cut})^{B-1-s}.$$

After rearranging terms,

$$Q_L^0 = P_{cut}(1-P_{cut})^{L-1} \times \left[\sum_{s=1}^{B-L} \binom{B-1-L}{s-1} P_{cut}^{s-1} (1-P_{cut})^{B-L-s} \right]$$

and invoking the binomial distribution properties this expression simplifies further to $Q_L^0 = P_{cut}(1-P_{cut})^{L-1}$. Therefore, the fragment size probability distribution after random fragmentation is

$$Q_L^0 = \begin{cases} P_{cut}(1-P_{cut})^{L-1} & \text{for } 1 \leqslant L \leqslant B-1, \\ (1-P_{cut})^{B-1} & \text{for } L = B. \end{cases}$$

It is interesting to note that the resulting expressions for $L \leqslant B-1$ are independent of the length B of the original gene. Furthermore, it can be shown (see Appendix B) that for small values of P_{cut}, Q_L^0 approaches the exponential distribution $P_{cut}\exp(-P_{cut}L)$ (see also Table 2) with a mean of $1/P_{cut}$. A graph of the expected fragment size distribution after treatment with DNase I is shown in Fig. 5. Typically, only a range of fragments between L_1 and L_2 are retained (e.g., $L_1 = 50$, $L_2 = 150$) in subsequent DNA shuffling experiments. In this case, Q_L^0 must be renormalized by dividing by $\sum_{L=L_1}^{L_2} Q_L^0$. Note also that Q_L^0 is a monotonically decreasing function of L implying that irrespective of the size of B and the fragmentation intensity, quantified by P_{cut}, "small" fragments are always more ubiquitous than "large" ones.

Comparisons of the proposed model predictions with the bands obtained after agarose gel electrophoresis requires converting the fragment size distribution to corresponding signal intensities. The intensity of an agarose gel band, composed of fragments of length L, is proportional to the amount of intercalated ethidium bromide. This is approximately proportional to fragment length since ethidium bromide stains DNA sequences evenly. Therefore, the relative intensity of a band I_L^0 is proportional to the particular size

TABLE 2

Comparison of discrete model vs. exponential approximation for fragment size probability calculation

P_{cut}	Q_{100}^0, discrete model	Q_{100}^0, exponential approximation
10^{-4}	0.00990%	0.00990%
10^{-3}	0.0906%	0.0905%
10^{-2}	0.370%	0.368%
10^{-1}	0.000295%	0.000454%

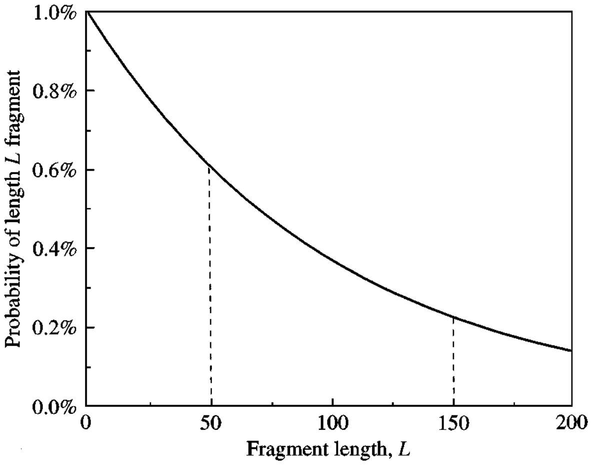

FIG. 5. Fragment size distribution after a 1000 bp gene is fragmented with DNase I with $P_{cut} = 0.01$ resulting in a mean fragment length of 100 bp. The dotted lines indicate that only a portion of these fragments are retained for shuffling.

fragment distribution Q^0_L times the number of nucleotides L in the fragment. Thus, the following expression describes the relative intensity distribution:

$$I^0_L = \begin{cases} LP_{cut}(1 - P_{cut})^{L-1} & \text{for } 1 \leqslant L \leqslant B - 1, \\ B(1 - P_{cut})^{B-1} & \text{for } L = B. \end{cases}$$

Unlike Q^0_L which is monotonically decreasing, I^0_L exhibits a sharp maximum in intensity for $L = 1/P_{cut}$. It is interesting that the location of the peak depends only on the bond-breaking probability P_{cut}.

A plot of relative gel intensities I^0_L after the random fragmentation of a 1 kb gene for $P_{cut} = 0.01, 0.02$ and 0.04 is shown in Fig. 6. As P_{cut} increases the peak migrates to smaller fragment lengths and the relative intensity distribution broadens. Density plots of the relative intensity shown in Fig. 7 simulate the appearance of an agarose gel after DNase I fragmentation of a 2 kb gene. Distributions for $P_{cut} = 0.002, 0.004,$ 0.01, 0.04 and 0.1 are shown (top to bottom), which produce intensity peaks at $L = 500$, 250, 100, 25 and 10 bp, respectively. The horizontal length scale shown is logarithmic due to the typical rate of DNA migration through a gel. These plots conform to the qualitative features exhibited by agarose gels.

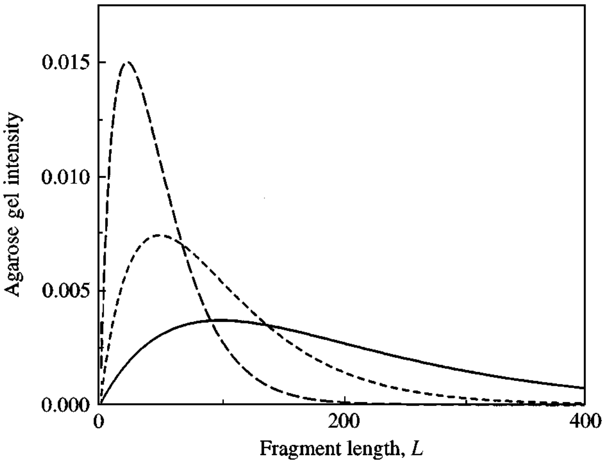

FIG. 6. Calculated agarose gel intensities for $P_{cut} = 0.01$, 0.02 and 0.04 for a 1 kb gene. $P_{cut} = 0.01$ (——); $P_{cut} = 0.02$ (- - - -); $P_{cut} = 0.04$ (– – –).

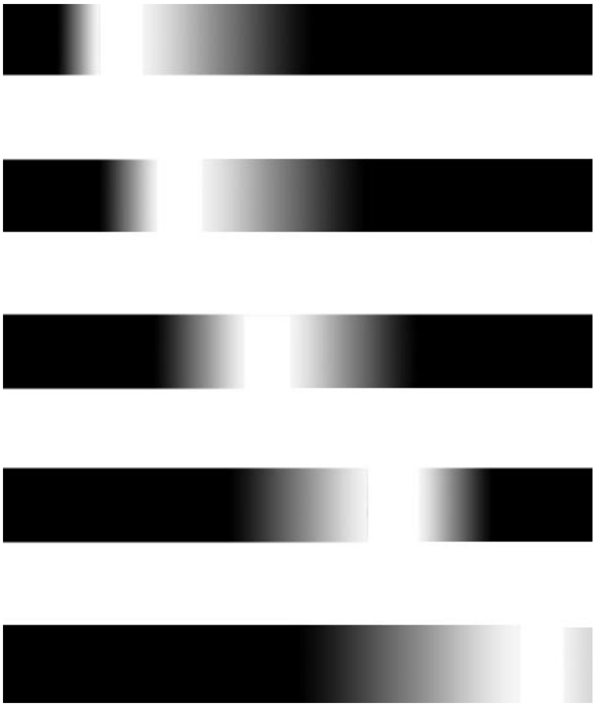

FIG. 7. Calculated agarose gel intensities for $P_{cut} = 0.002, 0.004, 0.01, 0.004$ and 0.2 (top to bottom). The gel runs from a maximum of $L = 2000$ at the left down to $L = 1$ at the right.

These predictions are next compared with agarose gel data quantifying the fragment size distribution at different points in time. Table 3 summarizes the location of the intensity peak at different digestion times observed on an agarose gel for a system examined by Volkov & Arnold

TABLE 3
Random fragmentation reaction progress (Volkov & Arnold, 1999)

Digestion time (min)	Fluorescence maximum, $1/P_{cut}$	P_{cut}
0.5	600 bp	0.17%
1	300 bp	0.33%
2	120 bp	0.83%
3	70 bp	1.4%
5	40 bp	2.5%

(1999). The proposed model predicts that the peak intensity must occur at $1/P_{cut}$ (bp). This implies that based on the experimentally observed peak intensities a model-based estimate of P_{cut} can be derived (see Table 3).

P_{cut} can alternatively be expressed as the extent of digestion

$$P_{cut} = \frac{C_b^0 - C_b}{C_b^0},$$

where C_b equals the concentration of unbroken nucleotide-nucleotide bonds and C_b^0 equals the initial concentration of bonds. C_b^0 can be represented as $C_{gene}B$, where C_{gene} is the concentration of the gene in solution. Because DNase I is in excess, a first-order rate expression can be used to fit the rate of digestion:

$$C_b = C_b^0 \exp(-kt).$$

This leads to the following expression for P_{cut}:

$$P_{cut} = 1 - \exp(-kt).$$

After substituting the model predictions for P_{cut} a straight line is obtained after plotting $-ln(1 - P_{cut})$ vs. t as shown in Fig. 8. The slope of this straight line is equal to the rate constant of $0.320\ \mathrm{hr}^{-1}$ verifying the model predictions.

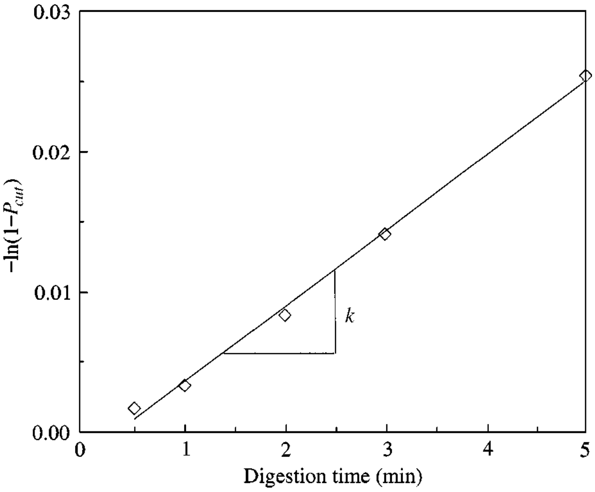

FIG. 8. First-order kinetics of DNase I digestion.

FRAGMENT ASSEMBLY MODEL

The goal of this model is to quantitatively describe how the fragment size distribution changes after a shuffling step. The value of this analysis is two-fold: first, it identifies how may shuffling cycles are necessary for reassembling the full-length gene. Second, by modeling fragment size distribution, which is experimentally accessible, it provides a unique way of matching experimental with modeling results quantifying important parameters in the model. Such experimental studies are currently under investigation. In DNA shuffling, fragments are assembled by a PCR-like reaction without added primers. Denatured fragments prime each other during the annealing step creating regions of *overlap*, where annealing has taken place, and *overhangs*, where the fragments do not align. The overhangs then serve as templates for *Taq*-catalysed extension.

In the proposed model it is assumed that tertiary collisions are not important and that annealing only occurs between pairs of fragments. In compliance with *Taq* polymerase function, fragment assembly only occurs in the direction from 5′ to 3′. Sequences of length no greater than that of the original gene are assembled since the fragments are assumed to anneal only along areas of high sequence identity. This requires that the gene does not have a high amount of repetition. The fraction of fragments that fail to anneal during each annealing step is represented by parameter NA which is assumed to depend on reaction conditions such as concentration and temperature. Fragment annealing is assumed to be governed by second-order kinetics so that the probability of a fragment of length X and a fragment of length Y annealing is proportional to the product of their relative concentrations. The proportionality constant, denoted by $A(X, Y, V)$, is assumed to be a function of only overlap (V) and annealed fragment lengths (X, Y). A minimum overlap of V_{min} nucleotides is assumed to be necessary for annealing. V_{min} depends on the degree of identity shared by the parent sequences and reaction conditions and it is usually between 5 and 15 nucleotides (Stemmer, 1994a). Fragments with an overlap smaller than V_{min} are assumed to denature before extension takes place.

Given the original fragment size distribution Q_L^0 obtained after random fragmentation, the next step is to quantify how this distribution will be reshaped after a shuffling step. The fragment probability size distribution after N shuffling cycles is denoted by Q_L^N. During the shuffling step pairs of DNA fragments randomly anneal and subsequently extend giving rise to successively larger DNA fragments from one shuffling cycle to the next. The fragment growth depends on the allowable overlap choices between fragments and their respective chances of annealing and extending. The allowable range of overlap for successful annealing between two fragments of lengths X and Y, respectively, is illustrated in Fig. 9. The maximum possible overlap is equal to the length of the smaller of the two fragments, or $\min(X, Y)$. Every overlap value from V_{min} up to $\min(X, Y) - 1$ occurs twice, once for each of the two fragment overhang orientations (5′ and 3′). The maximum overlap $\min(X, Y)$, however, occurs for $|X - Y| + 1$ internal annealing choices.

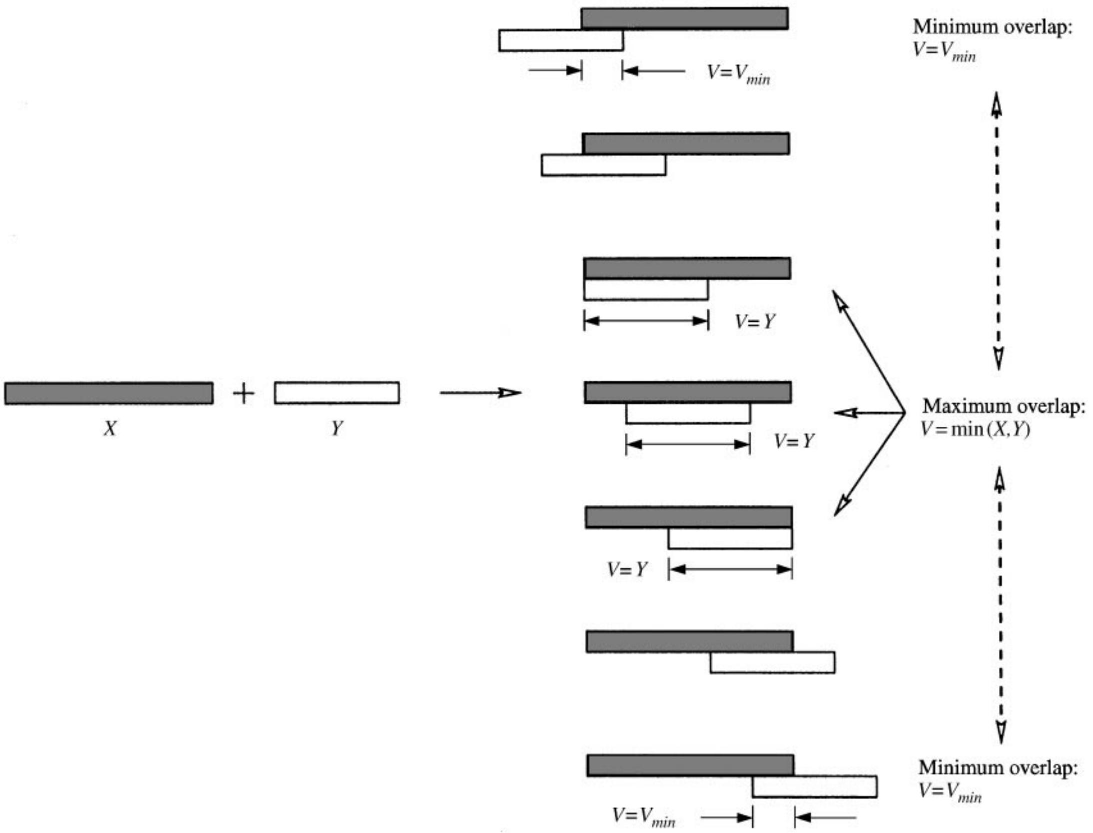

FIG. 9. Possible overlap alternatives between two annealed sequences.

This means that the multiplicity (degeneracy) d_V for different overlap values V is as follows:

$$d_V = \begin{cases} 2 & \text{for } V_{min} \leqslant V \leqslant \min(X, Y) - 1, \\ |X - Y| + 1 & \text{for } V = \min(X, Y). \end{cases}$$

The probability of observing a particular annealing choice shown in Fig. 9 depends on the extent of overlap. The following annealing probability model is postulated where high or low overlap values are favored depending on the sign of the exponent α:

$$A(X, Y, V) = d_V V^{\alpha} \Bigg/ \sum_{V = V_{min}}^{\min(X,Y)} d_V V^{\alpha}.$$

For $\alpha = -0.5$ this annealing probability becomes inversely proportional to the square root of the overlap length as Wetmur & Davidson (1967) suggest, thus favoring shorter overlap values. They assumed DNA annealing to be a two-step process, an initial rate determining nucleation step and a fast "zippering" step. In their analysis, nucleation is taken to be an elementary second-order reaction, thus supporting the second-order assumption above. The inverse square-root dependence is caused by an excluded volume effect which can be verified by approximating the DNA by an ideal random coil.

After establishing an annealing probability model the next step is to identify all mechanisms that generate a fragment of a particular length after a single annealing/extension cycle is completed. Six different pathways for producing a fragment of length L are considered which exhaustively enumerate all possibilities (Fig. 10). An fragment of length L can be produced by (i) the extension of smaller fragments to length L (first two pathways); (ii) a fragment of length L that fails to extend after annealing (next three pathways); or (iii) a fragment of length L that fails to anneal (last pathway). The first five pathways listed above require two fragments to collide and anneal. These collision pathways depend on three probability terms. First, the fragments must anneal, and this occurs with probability $(1 - NA)$ where NA denotes the probability of having a failed annealing. Second, the collision probability between two fragments of lengths X and Y is proportional to the product of their relative concentrations (or size probability distributions):

$$Q_X^{N-1} Q_Y^{N-1}.$$

Because many fragment combinations can combine to form a fragment of a particular length L, a summation over all X and Y values that give fragments of length L after extension is necessary. Third, the annealing probability $A(X, Y, V)$ multiplying the product of the fragment size probability distributions is assumed to be a function of the fragment lengths X, Y and the nucleotide overlap V. These three factors govern the collision and annealing of two fragments. Each one of the five possible collision pathways are next examined in detail.

The first pathway (outer extension) describes the 5′ → 3′ successful annealing and extension of two fragments whose lengths X, Y are smaller then L and their overlap $V = X + Y - L$ is such that two single-stranded fragments of length L are recovered after denaturing. The length of the first fragment X may vary between L_1 and L while the second fragment Y is bounded between $L - X + V_{min}$ and L. The three probability terms listed above result in the following expression for the size distribution of fragments of length L obtained through the outer extension pathway after the N-th shuffling cycle:

$$Q_L^N(\text{outer extension})$$

$$= (1 - NA) \sum_{X = L_1}^{L} Q_X^{N-1} \sum_{Y = L - X + V_{min}}^{L} Q_Y^{N-1} A(X, Y, X + Y - L)$$

The second pathway (inner extension) considers the case when a smaller fragment anneals completely within a fragment larger than L. Given an appropriate placement the smaller fragment can then be extended to produce a fragment of length L. Similarly, the corresponding size probability distribution term accounting for the

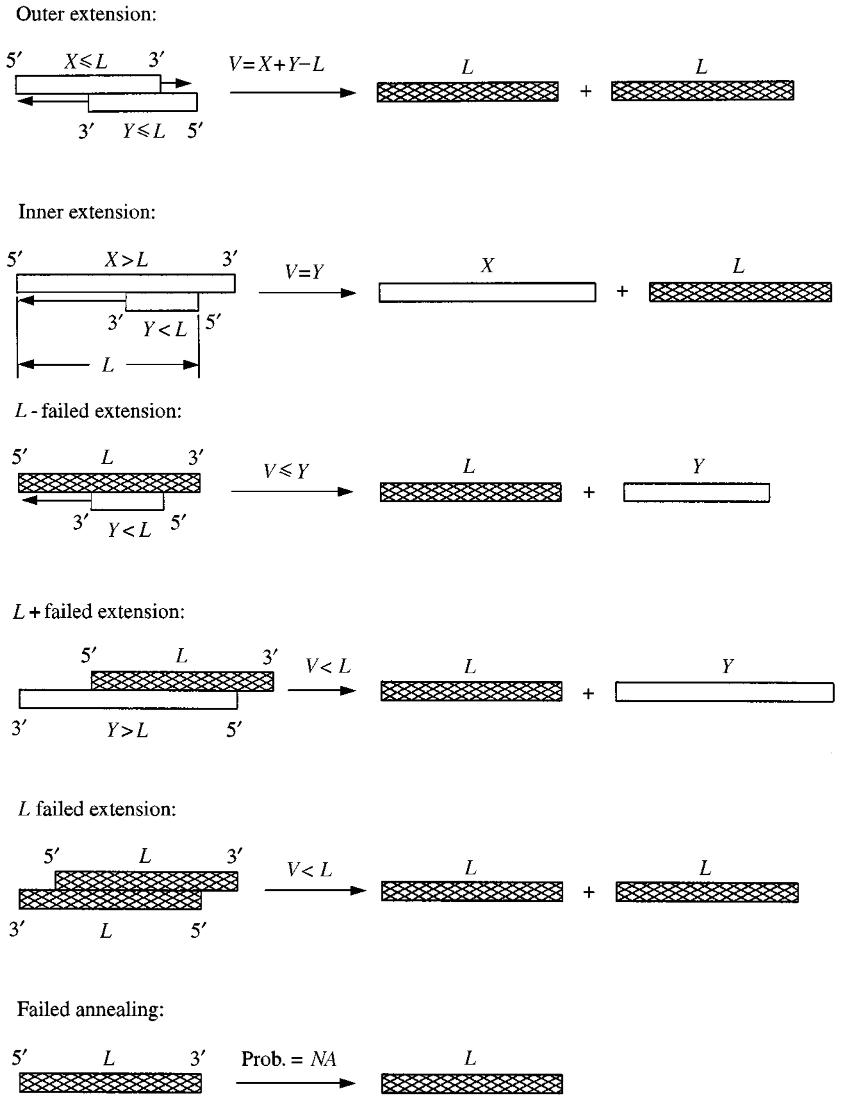

FIG. 10. The six pathways for producing a fragment of length L by extension, failed extension and failed annealing.

inner extension pathway is

Q_L^N(inner extension)

$$= (1 - NA) \sum_{X=L+1}^{B} Q_X^{N-1} \sum_{Y=L_1}^{L-1} Q_Y^{N-1} A(X, Y, Y).$$

The third, fourth and fifth pathways describe cases when fragments of length L are retained after annealing but unsuccessful extension. This occurs when a 3′ overhang is created, causing the *Taq*-catalysed extension to fail. The three failed extension pathways refer to the case where the second fragment is smaller than L (L^- failed extension); larger than L (L^+ failed extension); or equal to L (L failed extension). The following probability terms quantify the contribution of the third, fourth and fifth pathways to Q_L^N:

$Q_L^N(L^-$ failed extension)

$$= (1 - NA) Q_L^{N-1} \sum_{Y=L_1}^{L} Q_Y^{N-1}$$

$$\times \left(\sum_{V=V_{min}}^{Y-1} A(L, Y, V) + (L - Y) A(L, Y, Y) \right),$$

$$Q_L^N(L^+ \text{ failed extension}) = (1-NA)Q_L^{N-1} \sum_{Y=L+1}^{B} Q_Y^{N-1} \sum_{V=V_{min}}^{L} A(L,Y,V),$$

$$Q_L^N(L \text{ failed extension}) = (1-NA)Q_L^{N-1} Q_L^{N-1} \sum_{V=V_{min}}^{L-1} A(L,L,V).$$

Finally, fragments of length L may remain in the reaction mixture after failing to anneal. Failed annealing occurs with a probability of NA, so the following expression represents the portion of fragments of length L that remain unchanged after failed annealing:

$$Q_L^N(\text{failed annealing}) = (NA)Q_L^{N-1}.$$

The sum of the contributions of the six pathways generates a recursive model for Q_L^N that tracks the fragment size distribution from one shuffling cycle to the next. An internal consistency check verifies that $\sum_L Q_L^N = 1$ is preserved. The only adjustable parameters in this model are the minimum-allowable overlap V_{min}, the probability of failed annealing NA, and the exponent α in the annealing probability expression. Resolving the recursion requires going back shuffling steps, eventually encountering as an input the original fragment size distribution Q_L^0 obtained after random fragmentation.

Figure 11 illustrates the fragment size distribution predicted by the model after 5, 10 and 15 shuffling cycles. The original 1 kb gene is first randomly fragmented and only fragments with sizes between 50 and 150 bp are retained for shuffling. After only 5 shuffling steps the signature of the original fragment pool is still evident in the form of a sharp peak. After 10 cycles this sharp peak is nearly eliminated and a single broad maximum can be found in the fragment size distribution. Finally, after 15 cycles this maximum has migrated to reach the end of the length range and a large portion of the fragments have assembled into full length genes.

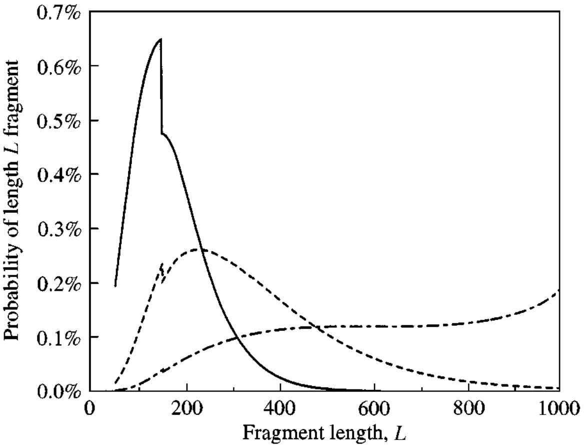

FIG. 11. Fragment size distributions after $N = 5, 10$ and 15 shuffling cycles of a ($L_1 = 50$, $L_2 = 100$) random fragment pool of a 1000 bp gene ($NA = 50\%$, $\alpha = -0.5$, $V_{min} = 15$). $N = 5$ (——); $N = 10$ (- - - -); $N = 15$ (— — — -).

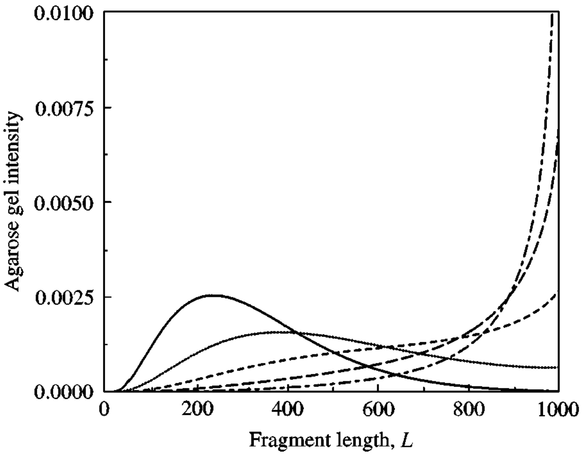

FIG. 12. Fragment size distributions after $N = 25, 30, 35$, 40 and 45 shuffling cycles of a ($L_1 = 10$, $L_2 = 50$) random fragment pool of a 1000 bp gene ($NA = 70\%$, $\alpha = -0.5$, $V_{min} = 5$). $N = 25$ (——); $N = 30$ (···); $N = 35$ (- - - -); $N = 40$ (— — — —); $N = 45$ (— — —).

Comparisons with experimental data are encouraging. Stemmer (1994b) initially studied the assembly of a 1 kb gene. The experiment began with random fragmentation to an approximate mean fragment length of 100 bp verified on an agarose gel implying a value for P_{cut} of 1%. Then fragments sized from 10 to 50 bp were assembled, and aliquots taken after $N = 25, 30, 35, 40$ and 45 shuffling steps were analysed on a gel to monitor the progress of the reaction. After 25 cycles, an intensity peak could be seen at approximately $L = 250$. After 30 cycles, a peak could be seen near $L = 450$. As the assembly progressed further, the fluorescence broadened, and full-length genes were reassembled. The proposed model matches these experimental observations as illustrated in Fig. 12. Parameter values of $P_{cut} = 1\%$, $L_1 = 10$ and $L_2 = 50$ are selected to match the ones employed in Stemmer's work. An α value of -0.5 was chosen (Wetmur & Davidson, 1967). Furthermore, the last two parameters were set at $NA = 70\%$ and $V_{min} = 5$.

SEQUENCE MATCHING MODEL

In the fragment assembly model the process of recovering full-length sequences was analysed without regard to the nucleotide sequence of the assembled genes. In the target sequence matching model the goal is to relate the nucleotide sequence of the fully assembled genes, obtained after recombination, to the nucleotide sequence and concentration of the parent sequences and experimental conditions. Specifically, given the precise nucleotide sequence of the parent sequences available for recombination, the objective is to find the fraction of the fully assembled sequences whose nucleotide sequence matches a prespecified target (e.g. ATTGG). This target can be (i) sequence identity, (ii) percent sequence homology or (iii) a desired number of crossovers. The work presented here focuses on matching the nucleotide sequence identity of a prespecified target. Moore *et al.* (1997) study a simplified model assuming that the lengths of the fragments to be reassembled are less than the distances between mutations. Later, Sun (1998) considers larger fragment lengths and addresses the case of single (Sun, 1998) and multiple (Sun, 1999) mutations. Also, Bogarad & Deem (1999) model molecular evolution with Monte Carlo simulations. By building on these contributions, this modeling effort addresses the general case of multiple mutations per DNA strand and arbitrary selections for the fragment lengths.

In our analysis, the nucleotide sequence of only complete DNA products of full length is analysed. The fraction of the sequences achieving full length can be estimated based on the results presented in the previous section. Also, the parent sequences are assumed to have a high degree of homology so that fragment annealing is possible along the entire gene length. As in the fragment assembly model, a minimum overlap of V_{min} nucleotides is assumed to be necessary for annealing and subsequent assembly, and assembly is assumed to proceed only from 5′ to 3′. Furthermore, it is assumed that the assembly process from a position i until the end B of the sequence is independent of assembly that has occurred before position i. In other words, the annealing of a fragment is independent of all prior fragment annealing that occurred in previous shuffling cycles. Therefore, if P_i is the probability of reproducing the portion of a target sequence between positions i and its end B then P_i is independent of all P_j where $j < i$.

The correct assembly of a target sequence is achieved if and only if a cascade of four independent events occurs, as shown in Fig. 13. Each one of these events contributes a probability term to P_i. The first step is to choose a fragment of length L to add to the sequence. Assuming random fragmentation, a fragment of length L is chosen with probability Q_L^0 discussed earlier. The second step in the assembly process is the annealing of the fragment of length L to the rest of the previously assembled sequence. The overlap must be at least V_{min} nucleotides. Thus, the non-overlapping portion of the fragment adds at most $L - V_{min}$ new nucleotides to the sequence. Therefore, there are $L - V_{min}$ possible ways for a fragment to align itself during annealing with overlaps V ranging from V_{min} to $L - 1$. The probability of adding $L - V$ new nucleotides with a fragment of length L is denoted as $A_{L-V,L}$ and is defined identically with the annealing probability $A(X, Y, V)$ described in the previous section:

$$A_{L-V,L} = V^{\alpha} \Bigg/ \sum_{V=V_{min}}^{L-1} V^{\alpha}.$$

After summing up over all possible overlap values this contributes a factor of $\sum_{V=V_{min}}^{L-1} A_{L-V,L}$ to P_i.

The third step is to calculate the probability that the extended sequences will contribute nucleotides that match the ones in the target sequence. Starting from a nucleotide at position i and assuming that a fragment of length L has annealed with an overlap of V nucleotides, the probability of matching the target nucleotide sequence from i to position $i + (L - V) - 1$ is equal to the fraction of the parent sequences that exactly match the target sequence from position

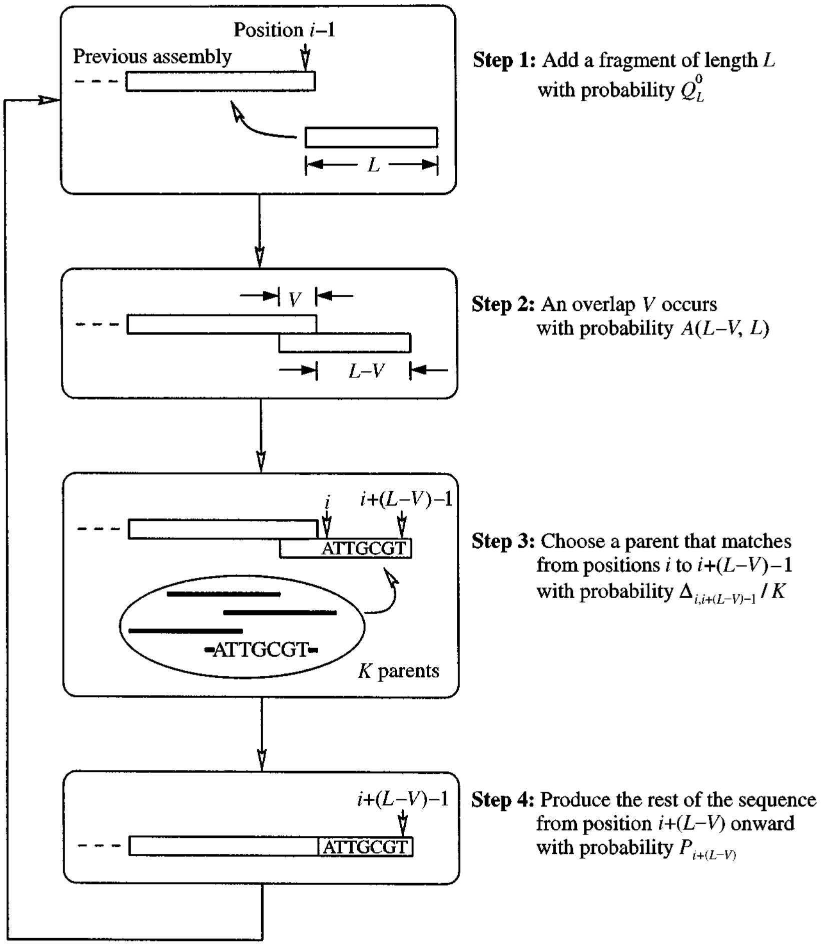
Position i–1
Previous assembly
L
Step 1: Add a fragment of length L with probability Q_L^0
V
L–V
Step 2: An overlap V occurs with probability $A(L–V, L)$
i
i+(L–V)–1
ATTGCGT
-ATTGCGT-
K parents
Step 3: Choose a parent that matches from positions i to i+(L–V)–1 with probability $\Delta_{i,i+(L-V)-1}/K$
i+(L–V)–1
ATTGCGT
Step 4: Produce the rest of the sequence from position i+(L–V) onward with probability $P_{i+(L-V)}$

FIG. 13. Four necessary steps of the annealing process as described in Sequence Matching Model.

i to position $i + (L - V) - 1$. Let parameter $\Delta_{a,b}$ denote the number of parent sequences that match the target sequence from positions a to b. Matching between positions i and $i + (L - V) - 1$ then occurs with probability $\Delta_{i,i+L-V-1}/K$ where K is the number of parents available for recombination. The fourth and final step is to calculate the probability of reproducing the remainder of the target sequence after adding $L - V$ new nucleotides. Because the annealing of additional fragments is independent of prior additions, simple multiplication by P_{i+L-V} suffices. This establishes a function for P_i that must be evaluated recursively. These four steps result in the expression for P_i shown below, where B is the sequence length in nucleotides, L_1 and L_2 are the smallest and largest recombinatory fragments, V_{min} is the minimum annealing overlap, and K is the number of parent sequences:

$$P_i = \begin{cases} 1, & i > B, \\ \frac{\Delta_{i,i}}{K}, & i = B, \\ \sum_{L=L_1}^{L_2} Q_L^0 \left[\sum_{V=V_{min}}^{L-1} A_{L-V,L} \left(\frac{\Delta_{i,i+L-V-1}}{K}\right) P_{i+L-V} \right], & i < B. \end{cases}$$

The above recursive formula calculates the probability P_i of obtaining as assembled sequence that is identical with some target sequence S after nucleotide position i. Therefore, P_1 is equal to the probability of assembling a sequence identical to the target. This target may be either a specific pattern or an entire gene.

In this analysis, the target sequence is assumed to be the entire assembled sequence. If only a portion of the assembled sequence is to be analyzed, the probability of annealing for the first fragment at $i = 1$ must be adjusted to include previous fragment additions ($i < 1$). In Moore & Maranas (2000), a renewal probability analysis is performed to account for this.

The predictions of the sequence matching model are consistent with experimental data. Stemmer (1994a) recombined two markers 75 bp apart from random fragments of size between 100 and 200 bp and reported that only 11% of the reassembled fragments contained both mutations. Note that independent assembly of the two mutations would have predicted a 25% value. Assuming a required minimum overlap for annealing of $V_{min} = 15$ and $\alpha = -1/2$, this model estimates this probability for the average fragment size of $L = 150$ to be 12.4%, which is very close to the experimentally observed one.

Next the possibility of increasing the probability of containing both mutations in the recombined sequences by appropriately choosing the fragment length is examined. The estimated probability of assembling a two-mutation sequence is plotted as a function of fragment length in Fig. 14. As shown in Fig. 14, this probability is a strong function of fragment length exhibiting a sharp maximum at around $L = 110$ bp of 21.4%. These results clearly demonstrate the importance of being able to predict this "right" fragment length.

Further comparisons with experimental results (Zhao & Arnold, 1997a) are shown in Tables 4 and 5. Zhao & Arnold (1997a) shuffle two 1.3 kb sequences, one with no mutations (wild-type) and the other with multiple-point mutations. In Table 4, the results of modeling an 83 bp portion of this sequence are compared with the experimental results. The experimental method is used to parameterize the model, so that $P_{cut} = 0.83\%$ (2 min DNase I digestion from Table 3), $(L_1, L_2) = (30, 50)$ (fragments less than 50 bp), $V_{min} = 15$, $\alpha = -0.5$ (standard annealing conditions). The modeling results match the trends found experimentally. The variations are most likely due to the small number of sequenced products reported. In addition, the modeling results confirm the experimentally observed tendency of the mutations at positions 35 and 47 to be "linked". The results shown in Table 5 more clearly demonstrate this tendency by examining only the recombination of the closely spaced mutations.

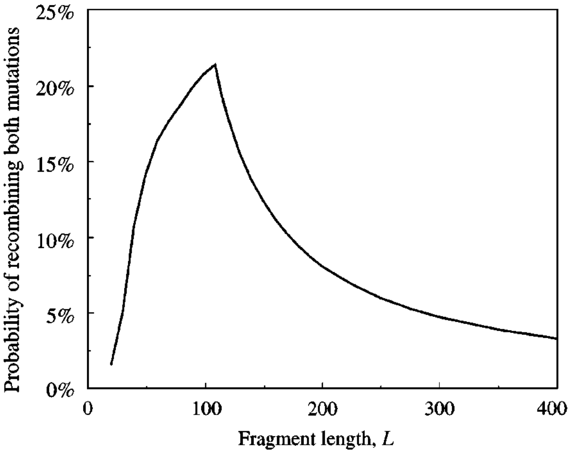

FIG. 14. Probability of recombining two markers 75 bp apart as a function of the fragment length L.

Summary and Conclusions

In this paper, quantitative models for predicting the outcome of DNase I fragmentation, error-prone PCR and DNA shuffling experiments were introduced. Specifically, the random fragmentation model and the fragment assembly model provided the quantitative means of tracking the size probability distribution of fragments in the reacting mixture during DNase I fragmentation and DNA shuffling respectively. On the other hand, the PCR model and the sequence matching model establish a formalism for estimating the probability of matching a prespecified nucleotide target. These models can be used in combination with optimization algorithms based on mixed-integer linear technologies to "home in" on the optimum fragment length and parent set without resorting to exhaustive enumeration of all alternatives (Moore & Maranas, 2000).

The predictions of these models were tested against experimental data available in the open literature. Unfortunately, such published data on directed evolution experiments do not contain sufficient detail on the size and nucleotide order of the recombined sequences to allow for a complete model validation and optimization. Currently, research is being conducted to overcome this limitation by designing directed evolution

TABLE 4
DNA shuffling calculationsfor $L_1 = 30, L_2 = 50, P_{cut} = 0.83\%,$ *and* $V_{min} = 15$

Parent sequences (2)

```
1        35   47      83
×————————×————×———————×
———————————————————————
```

Shuffled sequence	Calculated probability	Reported frequencey (Zhao & Arnold, 1997a)
×———————×—×————————×	8.2%	20%
×———————×—×————————	8.2%	10%
×———————×——————————×	4.3%	0%
×———————×———————————	4.3%	0%
×—————————×————————×	4.3%	0%
×—————————×—————————	4.3%	0%
×——————————————————×	8.2%	0%
×———————————————————	8.2%	0%
————————×—×————————×	8.2%	20%
————————×—×—————————	8.2%	0%
————————×——————————×	4.3%	0%
————————×———————————	4.3%	10%
——————————×————————×	4.3%	0%
——————————×—————————	4.3%	0%
———————————————————×	8.2%	20%
————————————————————	8.2%	20%

TABLE 5
DNA shuffling calculations for $L_1 = 30,$ $L_2 = 50,$ $P_{cut} = 0.83\%,$ *and* $V_{min} = 15$

Parent sequences (2)

```
1                     13
×———————————————————————×
—————————————————————————
```

Shuffled sequence	Calculated probability	Reported frequencey (Zhao & Arnold, 1997a)
×——————————————————×	32.8%	50%
×———————————————————	17.2%	10%
———————————————————×	17.2%	0%
————————————————————	32.8%	40%

experiments on a test gene to specifically provide data for our modeling effort. These experiments will provide information on fragment sizes to validate and parameterize the proposed models. In addition, work is underway to apply the modeling framework presented to other recombination protocols, particularly the new technique of incremental truncation (Ostermeier *et al.*, 1999). The combination of theoretical, experimental and analytical approaches will lead to the improved application of directed evolution methods yielding higher success rates and lower costs.

Financial support by the PSU Innovative Biotechnology Research Fund, NSF Career Award CTS-9701771 and computing hardware support by the IBM Shared University Research Program 1996, 1997 and 1998 is gratefully acknowledged.

REFERENCES

Arnold, F. (1996). Directed evolution: creating biocatalysts for the future. *Chem. Eng. Sci.* **51,** 5091–5102.

Arnold, F. & Moore, J. (1997). Optimizing industrial enzymes by directed evolution. *Advan. Biochem. Eng.* **58,** 1–14.

Bogarad, L. & Deem, M. (1999). A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2591–2595.

Bornscheuer, U., Altenbuchner, J. & Meyer, H. (1998). Directed evolution of an esterase for the stereoselective resolution of a key intermediate in the synthesis of epothilones. *Biotechnol. Bioeng.* **58,** 554–559.

Cadwell, R. & Joyce, G. (1992). Randomization of genes by PCR mutagenesis. *PCR Meth. Appl.* **2,** 28–33.

Chen, K. & Arnold, F. (1993). Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 5618–5622.

Christians, F., Scapozza, L., Crameri, A., Folkers, G. & Stemmer, W. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat. Biotechnol.* **17,** 259–264.

Crameri, A., Dawes, G., Rodriguez Jr., E., Silver, S. & Stemmer, W. (1997). Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nat. Biotechnol.* **15,** 436–438.

Eckert, K. & Kunkel, T. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR Meth. Appl.* **1,** 17–24.

Giver, L., Gershenson, A., Freskgard, P. & Arnold, F. (1998). Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. U.S.A.* **95,** 12 809–12 813.

Hansson, L., Bolton-Grob, R., Massoud, T. & Mannervik, B. (1999). Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling. *J. Mol. Biol.* **287,** 265–276.

Hsu, J., Das, S. & Mohapatra, S. (1997). Polymerase chain reaction engineering. *Biotechnol. Bioeng.* **55,** 359–366.

Kreyszig, E. (1993). *Advanced Engineering Mathematics*, 7th Edn., p. 1164. New York: Wiley.

Kuchner, O. & Arnold, F. (1997). Directed evolution of enzyme catalysts. *Trends Biotechnol.* **15,** 523–530.

Kumamaru, T., Suenaga, H., Mitsuoka, M., Watanabe, T. & Furukawa, K. (1998). Enhanced degradation of polychlorinated biphenyls by directed evolution of biphenyl dioxygenase. *Nat. Biotechnol.* **16,** 663–666.

Leung, D., Chen, E. & Goeddel, D. (1989). A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* **1,** 11–15.

Lin, Z., Thorsen, T. & Arnold, F. (1999). Functional expression of horseradish peroxidase in *E. coli* by directed evolution. *Biotechnol. Prog.* **15,** 467–471.

Lin-Goerke, J., Robbins, D. & Burczak, J. (1997). PCR-based random mutagenesis using manganese and reduced dNTP concentration. *Biotechniques* **23,** 409–412.

Ling, L., Keohavong, P., Dais, C. & Thilly, W. (1991). Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and Vent DNA polymerases. *PCR Meth. Appl.* **1,** 63–69.

Martineau, P., Jones, P. & Winter, G. (1998). Expression of an antibody fragment at high levels in the bacterial cytoplasm. *J. Mol. Biol.* **280,** 117–127.

Matsumura, I. & Ellington, A. (1999). *In vitro* evolution of thermostable p53 variants. *Protein Sci.* **8,** 731–740.

Moore, J. & Arnold, F. (1996). Directed evolution of a *para*-nitrobenzyl esterase for aqueous–organic solvents. *Nat. Biotect.* **14,** 458–467.

Moore, G. & Maranas, C. (2000). Modeling and optimization of DNA recombination. *Comput. Chem. Eng.*, (in press).

Moore, J., Jin, H., Kuchner, O. & Arnold, F. (1997). Strategies for the *in vitro* evolution of protein function: enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272,** 336–347.

Ostermeier, M., Nixon, A. & Benkovic, S. (1999). Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg. Med. Chem.* **7,** 2139–2144.

Patten, P., Howard, R. & Stemmer, W. (1997). Applications of DNA shuffling to pharmaceuticals and vaccines. *Curr. Opin. Biotechnol.* **8,** 724–733.

Proba, K., Worn, A., Honegger, A. & Pluckthun, A. (1998). Antibody scFv fragments without disulfide bonds made by molecular evolution. *J. Mol. Biol.* **275,** 245–253.

Reetz, M., Zonta, A., Schimossek, K., Liebeton, K. & Jeger, K. (1997). Creation of enantioselective biocatalyses for organic chemistry by *in vitro* evolution. *Angew. Chem. Int. Ed. Engl.* **36,** 2830–2835.

Rychlik, W., Spencer, W. & Rhoads, R. (1990). Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucl. Acids. Res.* **18,** 6409–6412.

Sakuma, Y. & Nishigaki, K. (1994). Computer prediction of general PCR products based on dynamical solution structures of DNA. *J. Biochem.* **116,** 736–741.

Schmidt–Dannert, C. & Arnold, F. (1999). Directed evolution of industrial enzymes. *Trends Biotechnol.* **17,** 135–136.

Shafikhani, S., Siegel, R., Ferrari, E. & Schellenberger, V. (1997). Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques* **23,** 304–310.

Shao, Z., Zhao, H., Giver, L. & Arnold, F. (1998). Random-priming *in vitro* recombination: an effective tool for directed evolution. *Nucl. Acids Res.* **26,** 681–683.

Stemmer, W. (1994a). Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370,** 389–391.

Stemmer, W. (1994b). DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 10 747–10 751.

Sun, F. (1998). Modeling DNA shuffling. *Proc. 2nd Ann. Inte. Conf. Comput. Mol. Biol.*, p. 251.

Sun, F. (1999). Modeling DNA shuffling. *J. Comput. Biol.* **6,** 77–90.

Taguchi, S., Ozaki, A. & Momose, H. (1998). Engineering of a cold-adapted protease by sequential random mutagenesis and a screening system. *Appl. Environ. Microbiol.* **64,** 492–495.

Volkov, A. & Arnold, F. (1999). Methods for *in vitro* DNA recombination and chimeragenesis. *Meth. Enzymol.*, (in press).

Wackett, L. (1998). Directed evolution of new enzymes and pathways for environmental biocatalysis. *Ann. NY Acad. Sci.* **864,** 142–152.

Weiss, G. & Haeseler, A. (1995). Modeling the polymerase chain reaction. *J. Comput. Biol.* **2,** 49–61.

Wetmur, J. & Davidson, N. (1967). Kinetics of renaturization of DNA. *J. Mol. Biol.* **31,** 349–370.

Wu, D., Ugozzoli, L., Pal, B., Qian, J. & Wallace, R. (1991). The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol.* **10,** 233–238.

Zhang, J., Dawes, G. & Stemmer, W. (1997). Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 4504–4509.

Zhao, H. & Arnold, F. (1997a). Optimization of DNA shuffling for high fidelity recombination. *Nucl. Acids Res.* **25,** 1307–1308.

Zhao, H. & Arnold, F. (1997b). Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc. Natl. Acad. Sci. U.S.A.* **94,** 7997–8000.

Zhao, H., Giver, L., Shao, Z., Affholter, J. & Arnold, F. (1998). Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nat. Biotechnol.* **16,** 258–261.

APPENDIX A

Calculation of $Z_{N,n}$

Let $Z_{N,n}$ represent the number of strands that have been through n extension steps after N PCR cycles. Information about the value of $Z_{N,n}$ can be discerned for some values of N and n. Initially, as stated above, two single strands of DNA are present, so $Z_{0,0}=2$. These two strands are the only two that are not the product of an extension step; therefore, $Z_{N,0}=2$ for all N. Also, after N cycles no DNA strands will be the result of more extension steps than N ($Z_{N,n}=0$ for $n>N$). After the N-th PCR cycle, a strand that is produced after n extension steps is either one that was just produced in the N-th PCR cycle or one that was already in the reaction mixture before the N-th PCR cycle began. In the first case, this implies that a sequence that has undergone $(n-1)$ extension steps after $(N-1)$ PCR cycles served as the template to produce the sequence in question. In the second case, the sequence in question has already undergone n extensions by the $N-1$ PCR cycle. See Fig. 3 for an illustration of these two cases. This implies that $Z_{N,n}=Z_{N-1,n}+Z_{N-1,n-1}$. Based on this relation a proof by induction of $Z_{N,n}=2\binom{N}{n}$ is constructed. First, this result is shown to be valid for $n=0$, 1 and 2.

For $n=0$,

$$Z_{N,0}=2=2\binom{N}{0}.$$

For $n=1$,

$$\begin{aligned}Z_{N,1}&=Z_{N-1,1}+Z_{N-1,0}=Z_{N-1,1}+2\\&=(Z_{N-2,1}+Z_{N-2,0})+2\\&=(Z_{N-2,1}+2)+2=Z_{N-2,1}+2(2)\\&=Z_{N-3,1}+2(3)\\&=Z_{N-k,1}+2k,\quad \forall 0\leqslant k\leqslant N.\end{aligned}$$

To resolve the recursion, set $k=N$:

$$Z_{N,1}=Z_{0,1}+2N=0+2N=2N=2\binom{N}{1}.$$

For $n=2$,

$$\begin{aligned}Z_{N,2}&=Z_{N-1,2}+Z_{N-1,1}\\&=(Z_{N-2,2}+Z_{N-2,1})+Z_{N-1,1}\\&=Z_{1,1}+Z_{2,1}+\cdots+Z_{N-1,1}\\&=\sum_{k=1}^{N-1}Z_{k,1}=\sum_{k=1}^{N-1}2k\\&=2\left[\frac{N(N-1)}{2}\right]=2\binom{N}{2}.\end{aligned}$$

After the postulated result is shown to be valid for $n=0$, 1, and 2, it is assumed that $Z_{N,n}=2\binom{N}{n}$. To complete the proof by induction, this assumption is utilized to prove that $Z_{N,n+1}=2\binom{N}{n+1}$:

$$\begin{aligned}Z_{N,n+1}&=Z_{N-1,n+1}+Z_{N-1,n}\\&=(Z_{N-2,n+1}+Z_{N-2,n})+Z_{N-1,n}\\&=(Z_{N-3,n+1}+Z_{N-3,n})+Z_{N-2,n}+Z_{N-1,n}\end{aligned}$$

$$= Z_{n,n} + Z_{n+1,n} + \cdots + Z_{N-1,n}$$

$$= \sum_{k=n}^{N-1} Z_{k,n} = \sum_{k=n}^{N-1} 2\binom{k}{n}.$$

Let $s = k - n$, then

$$Z_{N,n+1} = 2\sum_{s=0}^{(N-n)-1}\binom{n+s}{n}.$$

This expression can equivalently be rewritten as (Kreyszig, 1993)

$$Z_{N,n+1} = 2\sum_{s=0}^{(N-n)-1}\binom{n+s}{n} = 2\binom{(N-n)+n}{n+1}$$

$$= 2\binom{N}{n+1}.$$

APPENDIX B

Approximation of Q_L^0 with the Exponential Distribution

$$Q_L^0 = P_{cut}(1-P_{cut})^{L-1}$$

$$= \frac{P_{cut}}{1-P_{cut}}\left[\left(1-\frac{1}{1/P_{cut}}\right)^{-1/P_{cut}}\right]^{-P_{cut}L}.$$

For small values of P_{cut} we can write

$$\frac{P_{cut}}{1-P_{cut}} \approx 1, \quad \text{and} \quad \left(1-\frac{1}{1/P_{cut}}\right)^{-1/P_{cut}} \approx \exp(1).$$

Therefore, $Q_L^0 \approx P_{cut}\exp(-P_{cut}L)$ for small values of P_{cut}.