Theoretical Distribution of Truncation Lengths in Incremental Truncation Libraries

Marc Ostermeier

Department of Chemical Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218; telephone: 410-516-7144; fax: 410-516-5510; e-mail: oster@jhu.edu

Received 4 June 2002; accepted 4 November 2002

DOI: 10.1002/bit.10604

Abstract: Incremental truncation is a method for constructing libraries of every one base pair truncation of a segment of DNA. Incremental truncation libraries can be created using a time-dependent nuclease method or through the incorporation of α-phosphothioate dNTPs by PCR or by primer extension (THIO(pcr) truncation and THIO(extension) truncation, respectively). Libraries created by the fusion of two truncation libraries, known as ITCHY libraries, can be created using the above methods or by the incremental truncation-like method SHIPREC. Knowing and being able to tailor the distribution of truncations in incremental truncation, ITCHY and SHIPREC libraries would be beneficial for their use in protein engineering and other applications. However, the experimental determination of the distributions would require extensive, cost-prohibitive, DNA sequencing to obtain statistically relevant data. Instead, a theoretical prediction of the distributions was developed. Time-dependent incremental truncation libraries had the most uniform distribution of truncation lengths, but were biased against longer truncations. Essentially uniform distribution over the desired truncation range (from zero to N_{max} base pairs) required that truncations be prepared up to at least 1.2–1.5 N_{max}. THIO(pcr) and THIO(extension) truncation libraries had a very nonuniform distribution of truncation lengths with a bias against longer truncations. Such nonuniformity could be significantly diminished by decreasing the incorporation rate of αS-dNTPs but at the expense of having a large fraction of the DNA truncated beyond the desired range or completely degraded. ITCHY libraries created using time-dependent truncation had the most uniform distribution of possible fusions and had the highest fraction of the library being parental-length fusions. However, the distribution of parental-length fusions was biased against fusions near the beginning/ends of genes unless the truncation libraries are prepared with a uniform distribution up to N_{max}. In contrast, SHIPREC libraries and THIO(pcr) ITCHY libraries, by the very nature of the nonuniform distributions of the truncated DNA, are ensured of having a uniform distribution of fusion points in parental-length fusions. This comes at the expense of having a smaller fraction of the library being parental-length fusions; however, this limitation can be overcome by performing size selection on the library. *Biotechnol Bioeng* **82:** 564–577, 2003.

Keywords: incremental truncation; ITCHY; SHIPREC; protein engineering; directed evolution

INTRODUCTION

The application of molecular evolution to proteins has proven to be an effective method for engineering proteins with improved properties (Arnold, 2001). Molecular evolution works through cycles of (1) creating a library of gene variations and (2) identifying by selection or screening those rare members of the library which code for proteins that have an improvement in function. Depending on the method of library construction and the methods available for selection and screening, libraries of up to 10^9 or more variants can be constructed and evaluated. However, even this seemingly large number of variants is a minutia of the possible number of the 20^{300} possible sequences that could code for an average size protein of 300 amino acids. Thus, since one can create only an infinitesimal fraction of the possible variations of a gene or genes, the more one can tailor the library to be as rich in function as possible, the higher the probability of creating and identifying a protein with improved properties.

Common methods of library construction, including error-prone PCR (Caldwell and Joyce, 1995) and DNA shuffling (Stemmer, 1994), create diversity by changing the amino acid sequences. In contrast, incremental truncation, a method for creating a library of every one base truncation of dsDNA (Fig. 1), creates diversity by changing the length of a gene (Ostermeier et al., 1999a). Incremental truncation libraries can be created by time-dependent Exo III digestions (Fig. 2a) (Ostermeier et al., 1999a) or by the incorporation of α-phosphothioate dNTPs (αS-dNTPs) (Fig. 2b, 2c) (Lutz et al., 2001a). The combination of two incremental truncation libraries, called ITCHY libraries (Ostermeier et al., 1999b) or CP-ITCHY libraries (Ostermeier and Benkovic, 2001) depending on how they are constructed, creates diversity by fusing two gene fragments (Fig. 3). Performing ITCHY on a single gene generates libraries of proteins with internal deletions and duplications while performing ITCHY between two different genes generate libraries of fusion proteins in a DNA-homology independent fashion. Both strategies, as well as an incremental truncation-like method called SHIPREC (Sieber et al., 2001) (Fig. 4), have

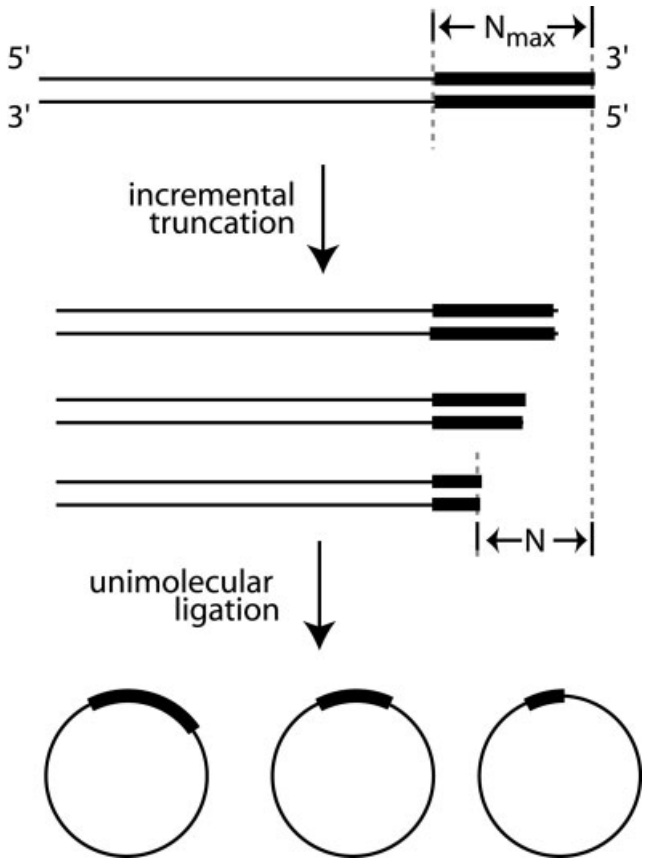

Figure 1. Schematic representation of incremental truncation. A linearized plasmid containing a segment that is to be truncated (thick lines) is truncated over a desired maximum range of truncation N_{max}. N is the number of bases truncated in an individual piece of DNA. The plasmid DNA is recircularized by unimolecular ligation.

the potential to create proteins with improved or novel properties as well as to generate artificial families for in vitro recombination in a method called SCRATCHY (Lutz et al., 2001b). In addition, incremental truncation and ITCHY have a number of demonstrated and potential applications apart from directed evolution including defining minimal functional units, identifying independent folding units, and assigning function to domains and subdomains (Ostermeier et al., 2002).

Knowing and being able to tailor the distribution of truncations in incremental truncation libraries, just as being able to control the distribution of mutations in an error-prone PCR library, would be beneficial for creating incremental truncation libraries for the above applications. The experimental determination of the distributions of incremental truncation libraries, though technically feasible, would involve extensive, cost-prohibitive, sequencing to obtain statistically relevant data. Accordingly, a theoretical prediction of the distributions is presented.

MODELING OR THEORETICAL ASPECTS

In the following theoretical treatment, N is the number of bases truncated and N_{max} is the desired maximum number of bases truncated.

Time-Dependent Truncation Libraries

A schematic representation of time-dependent ExoIII incremental truncation is shown in Figure 2A. To achieve a distribution of truncation lengths, the DNA has been truncated for various lengths of time. Typically ExoIII digestion is started at $t = 0$ and small samples are removed at regular intervals to a buffer that quenches truncation. For example a library in which $N_{max} = 300$ bp could be achieved by truncating a 90 μL DNA solution under conditions such that the ExoIII digestion rate is 10 bases/min and 1 μL samples are removed every 20 seconds for 30 minutes.

It has been experimentally determined that the distribution of truncations of digestion of dsDNA with ExoIII for a defined length of time can be described by a normal distribution (Hoheisel, 1993). Thus, for each time point collected, the distribution of truncation lengths can be related to the mean truncation length L by the standard deviation σ by Eq. (1).

$$G(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}; \qquad z = \frac{L - N}{\sigma} \tag{1}$$

Experimentally it has been determined that the standard deviation is of the form $\sigma = cL$, with $c = 0.2$ (Hoheisel, 1993); however, 9% of the truncations were found to be disproportionally far from the mean, thus distorting σ.When these 9% were removed, it was found that $\sigma = 0.075$ L. It is not known whether the 9% are experimental artifacts or a true result of ExoIII digestion. Thus, data are presented throughout this article at both $c = 0.2$ and $c = 0.075$.

At a particular time point, the probability p_N that DNA has been truncated a length of N is

$$p_N = \int_{z_{at\ N=N}}^{z_{at\ N=N+1}} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz = \frac{1}{\sqrt{\pi}} \int_{\xi_{at\ N=N}}^{\xi_{at\ N=N+1}} e^{-\xi^2} d\xi; \quad \xi = \frac{z}{\sqrt{2}} \tag{2}$$

that can be written as

$$p_N = \frac{1}{\sqrt{\pi}} \int_0^{\xi_{at\ N=N+1}} e^{-\xi^2} d\xi - \frac{1}{\sqrt{\pi}} \int_0^{\xi_{at\ N=N}} e^{-\xi^2} d\xi = 0.5(erf[\xi_{at\ N=N+1}] - erf[\xi_{at\ N=N}]) \tag{3}$$

It is useful to define a dimensionless truncation length N^* which normalizes the truncation length N by the maximum desired truncation length N_{max}

$$N^* = \frac{N}{N_{\max}} \tag{4}$$

so that at $N^* = 1$, the amount of bases truncated is at the maximum desired; at $N^* = 0$, no bases have been truncated; and, at $N^* = 2$, the number of bases truncated is twice the maximum desired. A dimensionless mean truncation length

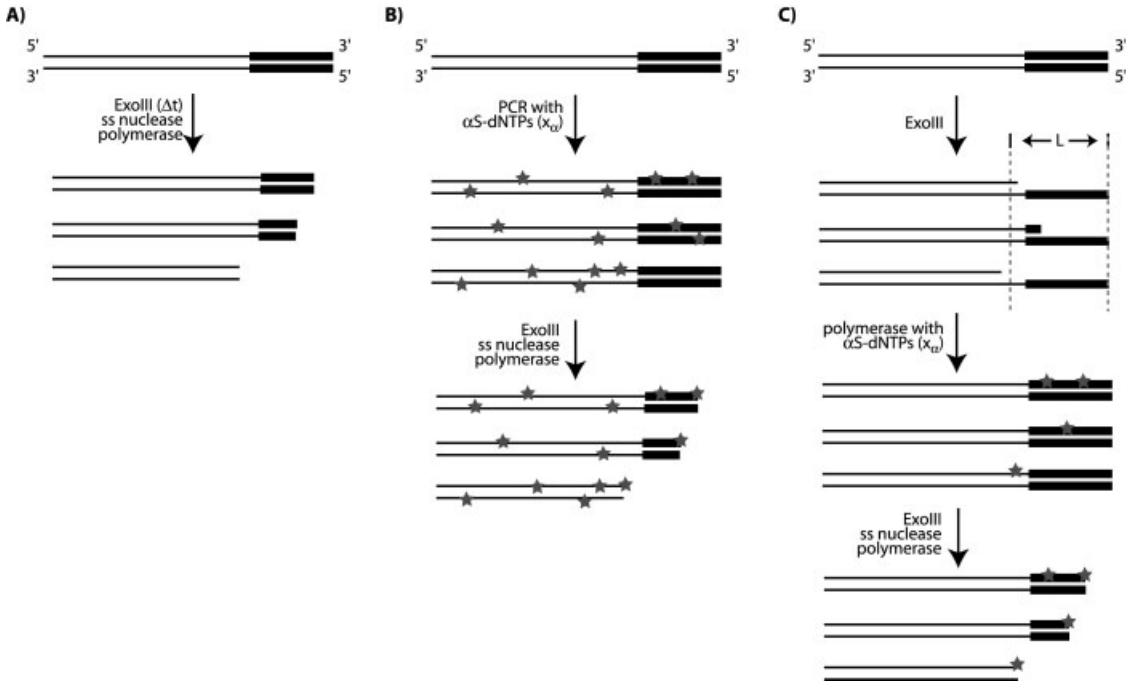

Figure 2. Schematic representation of incremental truncation methods. For all methods, the left-hand side of the DNA is protected from digestion by digesting the DNA with a restriction endonuclease that leaves a four-base 3′ overhang or by another suitable protection method. (A) In time-dependent truncation, the DNA is subjected to ExoIII digestion. During Exo III digestion, small aliquots are removed frequently and quenched by addition to a low pH, high salt buffer. Blunt ends are prepared by treatment with a single-strand nuclease and a DNA polymerase followed by unimolecular ligation to recyclize the vector. (B) In THIO(pcr) truncation, the entire plasmid is amplified by PCR using dNTPs and a small amount of αS-dNTPs. Subsequent digestion with ExoIII is prevented from continuing past the incorporated αS-dNMP. Blunt ends are prepared by treatment with a single-strand nuclease and a DNA polymerase followed by unimolecular ligation to recyclize the vector. (C) In THIO(extension) truncation, the range of incorporation of αS-dNTPs is limited by an initial truncation of average length *L*. Subsequent steps are the same as in THIO(pcr) truncation.

L^* for any time point, defined by Eq. (5), normalizes the mean truncation length L for a time point by N_{max}.

$$L^* = \frac{L}{N_{\max}} \tag{5}$$

and thus the expressions for ξ become

$$\xi_{at\ N=N} = \frac{L^* - N^*}{\sqrt{2cL^*}} \tag{6}$$

$$\xi_{at\ N=N+1} = \frac{L^* - N^* + \dfrac{1}{N_{\max}}}{\sqrt{2cL^*}} \tag{7}$$

and thus substituting Eqs. (6) and (7) into Eq. (3) we get the following expression for the probability that a piece of DNA has been truncated N^* in a particular time point.

$$p_{N^*} = 0.5\left(erf\left[\frac{L^* - N^* + \dfrac{1}{N_{\max}}}{\sqrt{2cL^*}}\right] - erf\left[\frac{L^* - N^*}{\sqrt{2cL^*}}\right]\right) \tag{8}$$

The total number of time points n_T will depend on the final mean truncation length of the final timepoint L_F, the sampling rate S (how often samples are removed and ExoIII digestion is quenched) and the rate of the exonuclease r_{exo} by Eq. (9).

$$n_T = \frac{L_F S}{r_{exo}} = \frac{L^*_F N_{\max} S}{r_{exo}} \tag{9}$$

The value of L^* at time point i will depend on r_{exo}, S, and N_{max} by Eq. (10).

$$L^*_i = \frac{i r_{exo}}{S N_{\max}} \tag{10}$$

The probability that a piece of DNA has been truncated N^* in the entire library is found by summing up the probabilities for each of the time points and dividing by the total number of time points.

$$P_{N^*} = \frac{1}{n_T}\sum_{i=1}^{n_T} p_{N^*} = \frac{1}{2n_T}\sum_{i=1}^{n_T}\left(erf\left[\frac{L^*_i - N^* + \dfrac{1}{N_{\max}}}{\sqrt{2cL^*_i}}\right] - erf\left[\frac{L^*_i - N^*}{\sqrt{2cL^*_i}}\right]\right) \tag{11}$$

THIO(pcr) Truncation Libraries

A schematic representation of incremental truncation using αS-dNTPs incorporated by PCR is shown in Figure 2B. The entire plasmid (containing the gene to be truncated) is am-

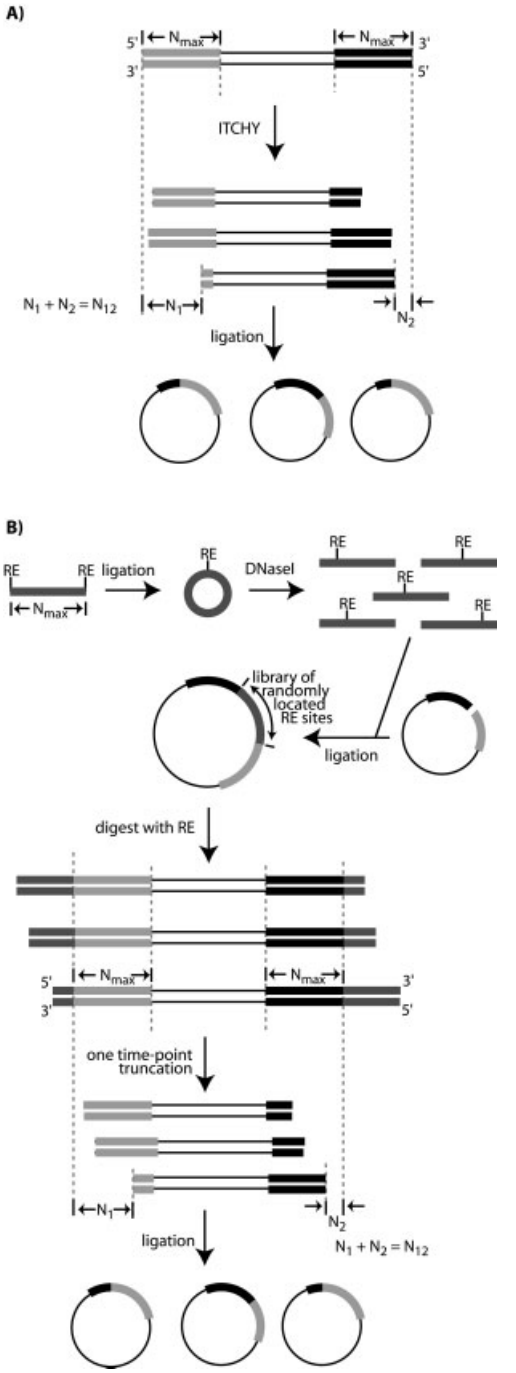

Figure 3. Schematic representations of the construction of (A) ITCHY libraries and (B) CP-ITCHY libraries. (A) Incremental truncation libraries, prepared by any of the methods in Figure 2, are prepared on two different genes and fused by blunt-end ligation. (B) For CP-ITCHY libraries, a piece of DNA of length N_{max} is created (i.e., by PCR) with identical restriction sites (RE) on the ends. The DNA is digested with RE and treated with ligase under dilute conditions such that a significant amount of closed circular DNA is formed. The closed circular DNA is linearized at random locations by digestion with very small amounts of DNase I. The randomly linearized DNA is repaired using a DNA polymerase and a DNA ligase and cloned between the genes to be truncated. This library of RE sites is the starting point for incremental truncation. The library is digested with ExoIII for the desired length of time necessary to digest N_{max} bases. The single strand overhangs are removed by mung bean nuclease, the ends are blunted with Klenow and ligation under dilute conditions results in the creation of a CP-ITCHY library.

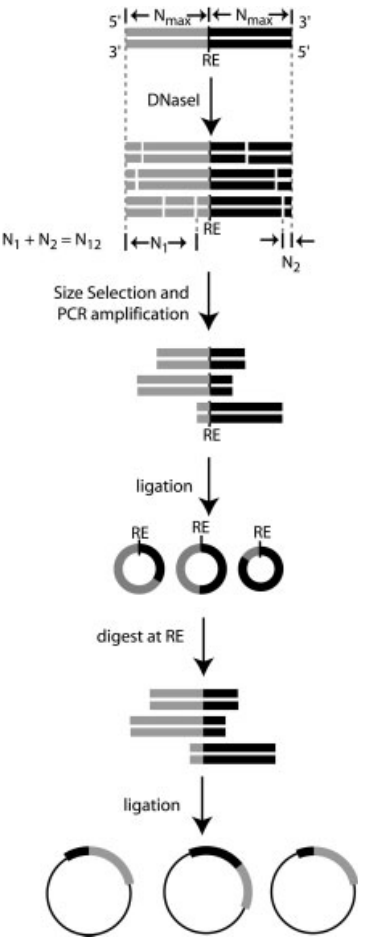

Figure 4. Schematic representation of the construction of SHIPREC libraries. The two genes are first fused end-to-end through a linker segment containing a restriction enzyme site (RE). Blunt-ended fragments are generated by DNaseI digestion and DNA of approximately the desired size can be selected and PCR amplified (not shown). The fragments are circularized by intramolecular blunt end ligation, linearized by digestion with RE, and cloned into an appropriate plasmid.

plified by PCR using a mixture of dNTPs and a small amount of αS-dNTPs, which are randomly incorporate into the DNA. ExoIII cannot remove αS-dNMPs. Thus, a distribution of truncation lengths, defined by the sites of incorporation of the αS-dNTPs, is created upon ExoIII digestion.

In a mixture of dNTPs and αS-dNTPs the ratio of incorporation rates R of dNTPs and αS-dNTPs is defined as

$$R = \frac{r_{dNTP}}{r_{\alpha S \cdot dNTP}} \tag{12}$$

where r_{dNTP} is the incorporation rate of standard dNTPs and $r_{\alpha SdNTP}$ is the incorporation rate of αS-dNTPs. To a first approximation, $R = 1$ for *E. coli* DNA polymerase and *Taq* DNA polymerase if the S-isomeric pure form of the αS-dNTPs is used (Lutz et al., 2001a). However, since the R-isomer acts as a competitive inhibitor of DNA polymerases (Burgers and Eckstein, 1979) and αS-dNTPs are only

commercially available as racemic mixtures, in practice $R > 1$ and has been empirically estimated to be 7.5 (Lutz et al., 2001a).

The mole fraction of αS-dNTPs x_α is defined as the concentration of αS-dNTPs divided by the sum of the concentration of αS-dNTPs and dNTPs. The probability P_N that a DNA molecule will have N bases truncated is:

$$P_N = \left(\frac{x_\alpha}{R}\right)\left(1 - \frac{x_\alpha}{R}\right)^N \tag{13}$$

where x_α/R is the probability that there is a αS-dNMP at the position $N + 1$ and $(1 - x_\alpha/R)^N$ is the probability that there was not a αS-dNMP at any of the previous N positions.

Next, a dimensionless truncation length N^* is introduced defined as in Eq. (4). Thus Eq. (13) becomes

$$P_{N^*} = \left(\frac{x_\alpha}{R}\right)\left(1 - \frac{x_\alpha}{R}\right)^{N^* N_{\max}} \tag{14}$$

If we normalize P_{N^*} by the probability of an ideal step function probability shown in Figure 5A where

$$P_{N^*,ideal} = N_{\max}{}^{-1}; \qquad 0 \le N^* \le 1 \tag{15}$$

we arrived at Eq. (16) for the normalized probability

$$\frac{P_{N^*}}{P_{N^*,ideal}} = N_{\max}\left(\frac{x_\alpha}{R}\right)\left(1 - \frac{x_\alpha}{R}\right)^{N^* N_{\max}} \tag{16}$$

Graphs of this normalized probability vs. N^* are very insensitive to the value of N_{max} for $100 \le N_{max} \le 2000$ bp when

$$\frac{x_\alpha}{R} = \frac{A}{N_{\max}}; \qquad 0.1 \le A \le 3 \tag{17}$$

such that

$$\frac{P_{N^*}}{P_{N^*,ideal}} = A\left(1 - \frac{A}{N_{\max}}\right)^{N^* N_{\max}} \tag{18}$$

The normalized probability for $0 \le N^* \le 1$ varies over $100 \le N_{max} \le 2000$ bp by less than 0.5% for $A \le 1$ and less than 5% for $A = 3$. Equation (17) is experimentally relevant since, to a first approximation, we desire to have the probability of incorporating an αS-dNTP (x_α/R) such that, on average, one αS-dNTP is incorporated per N_{max} bases.

To be precise, P_{N^*} in Eqs. (14) and (18) do not represent the probability of a truncation of length N^* being in the library of transformants, it represents the probability that DNA will be truncated to length N^* before transformation. This is because (1) some DNA molecules will not have a αS-dNTP incorporated, and thus will be completely degraded, and (2) some DNA molecules will get truncated to a length which makes them unable to transform bacteria (e.g., an essential portion of the origin of replication or selectable marker gets truncated). However, since the correction to the probability will be a constant for any particular library, the relative shapes of the normalized probability curves for $N^* < 1$ given by Eq. (18) will not be affected.

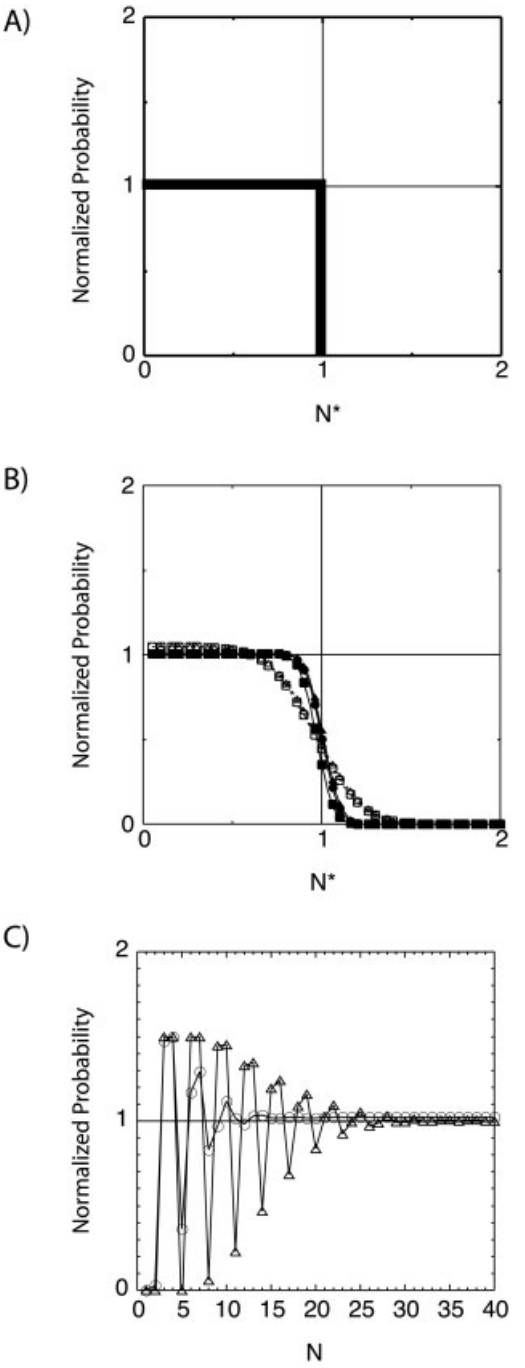

Figure 5. Distribution of truncation lengths for (A) an ideal, uniform library and (B, C) time-dependent truncation libraries. (A) The ideal, step function distribution of Eq. (15). (B) Distribution of truncation lengths in time-dependent truncation libraries as a function of the dimensionless truncation length N^* for N_{max} equal to 250 bp (triangles), 500 bp (circles), and 1000 bp (squares) and for standard deviations truncation length of 0.2 L (open symbols) and 0.075 L (solid symbols). The data is found by normalizing Eq. (11) by Eq. (15) and using the experimentally relevant values of $r_{exo} = 9$ base/min and $S = 3$ min^{-1}. For these curves the last time point was taken such that $L = N_{max}$ ($L^*_F = 1$). Only data for $N^* \ge 0.04$ is shown. (C) Distribution of early truncation lengths for standard deviations of 0.2 L (circle) and 0.075 L (triangle) as a function of the base pairs truncated with $N_{max} = 500$ bp, $r_{exo} = 9$ base/min and $S = 3$ min^{-1}.

THIO(Extension) Truncation Libraries

In the incorporation of αS-dNTPs by PCR, the αS-dNTPs are incorporated evenly throughout the plasmid. However, in the primer extension method (Fig. 2C), αS-dNTPs can only be incorporated in the region exposed by the first ExoIII digestion. For this reason, the probability of truncating N bases P_N is

$$P_N = P_{thio} P_{dig>N} \tag{19}$$

where P_{thio} is the same as P_N in Eq. (13) and

$$P_{dig>N} = 1 - \int_0^N P_{dig=N} dN = 1 - \int_{Z_{(at\ N)}}^{Z_{(at\ N=0)}} P_{dig=N} dZ \tag{20}$$

where $P_{dig\ =\ N}$ is given by the normal distribution of Eq. (1). Converting this equation in terms of ξ as in Eq. (2), Eq. (20) becomes

$$P_{dig>N} = 1 - \frac{1}{\sqrt{\pi}} \int_{\xi_{at\ N}}^{\xi_{at\ N=0}} e^{-\xi^2} d\xi \approx 1 - \frac{1}{\sqrt{\pi}} \int_{\xi_{at\ N}}^{\infty} e^{-\xi^2} d\xi \tag{21}$$

which can be rewritten in terms of the error function so that the probability of Eq. (19) becomes

$$P_N = 0.5 \left(\frac{x_\alpha}{R} \right) \left(1 - \frac{x_\alpha}{R} \right)^N \left[1 + erf\left(\frac{L - N}{\sqrt{2} cL} \right) \right] \tag{22}$$

Introducing the dimensionless parameters N^* and L^* of Eqs. (4) and (5) (in this case L^* represents the mean length of truncation in the initial ExoIII digestion step normalized by N_{max}) and normalizing by $P_{N^*,ideal} = N_{max}^{-1}$ results in Eq. (23).

$$\frac{P_{N^*}}{P_{N^*,ideal}} = 0.5 N_{\max} \left(\frac{x_\alpha}{R} \right) \left(1 - \frac{x_\alpha}{R} \right)^{N_{\max} N^*} \left[1 + erf\left(\frac{L^* - N^*}{\sqrt{2} cL^*} \right) \right] \tag{23}$$

Graphs of this normalized probability vs. N^* are very insensitive to the value of N_{max} for $100 \le N_{max} \le 2000$ bp when

$$\frac{x_\alpha}{R} = \frac{A}{N_{\max}}; \qquad 0.1 \le A \le 3 \tag{24}$$

and thus Eq. (23) becomes

$$\frac{P_{N^*}}{P_{N^*,ideal}} = 0.5A \left(1 - \frac{A}{N_{\max}} \right)^{N_{\max} N^*} \left[1 + erf\left(\frac{L^* - N^*}{\sqrt{2} cL^*} \right) \right] \tag{25}$$

Again, Eq. (24) is experimentally relevant since, to a first approximation, we desire to have the probability of incorporating an αS-dNTP (x_α/R) such that, on average, one αS-dNTP is incorporated per N_{max} bases. P_{N^*} in Eq. (25) does not represent the probability of a truncation of length N being in the library of transformants, it represents the probability that DNA will be truncated to length N before transformation. However, as was the case for truncation by incorporation of αS-dNTP by PCR, the correction to the probability will be a constant for any particular library. Thus, the relative values of the normalized probability given by Eq. (25) will not be affected.

Fraction of Truncation Libraries in Desired Range

For any of the above three methods, the fraction of the library in the desired range of $0 \le N^* \le 1$ is found by the following equation

$$f = \frac{\int_0^1 P_{N^*} dN^*}{\int_0^\infty P_{N^*} dN^*} \tag{26}$$

In practice, however, truncations where N^* is too much greater than one will delete an essential part of the plasmid and not produce a transformant. This will of course depend on the plasmid, the length of the gene and N_{max}. As an example, if we say that truncations where $N^* > w$ will not produce a transformant, the fraction of the library in the desired range is found by

$$f_w = \frac{\int_0^1 P_{N^*} dN^*}{\int_0^w P_{N^*} dN^*} \tag{27}$$

The fraction of DNA that is capable of transforming bacteria is

$$f_{trans} = \frac{\int_0^w P_{N^*} dN^*}{\int_0^\infty P_{N^*} dN^*} \tag{28}$$

The probability $P_{N^*,\ trans}$ that a truncation of length N^* is in the transformed library is

$$P_{N^*,trans} = \frac{P_{N^*}}{\int_0^w P_{N^*} dN^*} \tag{29}$$

ITCHY Libraries

ITCHY libraries are the random fusion of two incremental truncation libraries as shown in Figure 3A. For the fusion of two truncation libraries over the same size range (N_{max} is the same for both), the probability P_{N12} of having a total truncation of $N_{12} = N_1 + N_2$ bases in the fusion gene is given by Eqs. (30) and (31).

$$P_{N_{12}} = \sum_{i=0}^{N_{12}} P_N|_{N=i} P_N|_{N=N_{12}-i}; \quad for\ 0 \le N_{12} \le N_{\max} \tag{30}$$

$$P_{N_{12}} = \sum_{i=0}^{2N_{\max}-N_{12}} P_N|_{N=N_{12}-N_{\max}+i} P_N|_{N=N_{\max}-i}; \quad for\ N_{\max} \le N_{12} \le 2N_{\max} \tag{31}$$

P_{N12} excludes fusions where one or both of the genes has been truncated beyond N_{max}. This does not introduce any error in the distribution of P_{N12} for $0 \le N_{12} \le N_{max}$ but Eq. (31) will underestimate the probability of DNA with a truncation size of $N_{12} > N_{max}$ since it doesn't include fusions where one gene has been truncated more than N_{max}. However, since we are not interested in having fusions where one or both of the truncations has gone past N_{max}, the probabilities of Eq. (31) is most appropriate since it represents the probability of desired fusions. Note that at $N_{12} = N_{max}$, the fusion genes created by ITCHY will be the same size as the original genes (provided the original genes were of the same size and truncations were done over the same regions of both genes) which will be referred to as parental-length fusions.

For ideal ITCHY libraries in which each incremental truncation library has $P_{N^*} = 1/N_{max}$ for $0 \le N^* \le 1$ and $P_{N^*} = 0$ for $N^* > 1$, the summations of Eqs. (30) and (31) reduce to the following.

$$P_{N_{12}} = \frac{N_{12}+1}{(N_{\max})^2}; \qquad for\ 0 \le N_{12} \le N_{\max} \tag{32}$$

$$P_{N_{12}} = \frac{2N_{\max} - N_{12} + 1}{(N_{\max})^2}; \quad for\ N_{\max} \le N_{12} \le 2N_{\max} \tag{33}$$

Time-Dependent ITCHY Libraries

The probabilities are found by substituting Eq. (34) [analogous to Eq. (11)] into Eqs. (30) and (31) to solve for P_{N12}.

$$P_N = \frac{1}{n_T}\sum_{j=1}^{n_T} p_N = \frac{1}{2n_T}\sum_{j=1}^{n_T}\left(erf\left[\frac{jr_{exo} - SN + S}{\sqrt{2}cjr_{exo}}\right] - erf\left[\frac{jr_{exo} - SN}{\sqrt{2}cjr_{exo}}\right]\right) \tag{34}$$

THIO(pcr)-ITCHY Libraries

For ITCHY libraries created using αS-dNTPs incorporated by PCR, Eqs. (13) and (17) are used for P_N and the summations reduce to

$$P_{N_{12}} = (2N_{\max} + 1 - N_{12})\left(\frac{A}{N_{\max}}\right)^2\left(1 - \left(\frac{A}{N_{\max}}\right)\right)^{N_{12}}; \quad for\ 0 \le N_{12} \le N_{\max} \tag{35}$$

$$P_{N_{12}} = (1 + N_{12})\left(\frac{A}{N_{\max}}\right)^2\left(1 - \left(\frac{A}{N_{\max}}\right)\right)^{N_{12}}; \quad for\ N_{\max} \le N_{12} \le 2N_{\max} \tag{36}$$

The fraction of DNA that is in this desired range ($N_1 \le N_{max}$ and $N_2 \le N_{max}$) can be calculated from Eq. (37) using Eqs. (26) or (27).

$$f_{12} = (f)^2 \tag{37}$$

THIO(Extension)-ITCHY Libraries

The primary benefit of using primer extension is that the range that the αS-dNTPs can be incorporated can be limited, depending on L^* and c. For the ideal, limiting case where c is very small and $L^* = 1$, the incorporations will be limited to $0 \le N \le N_{max}$ and thus

$$P_{N_{12}} = \frac{P_{N_{12}}(from\ Equations\ 35\ and\ 36)}{f_{12}} \tag{38}$$

CP-ITCHY Libraries

CP-ITCHY eliminates the need for extensive time-point sampling by the insertion of a library of circularly permuted DNA that contains a unique restriction site and is of length N_{max} (Fig 3b). Whereas time-dependent truncation creates diversity in length by digesting from one starting point for a variety of time lengths, CP-ITCHY achieves its distribution of truncation lengths by digesting DNA from a variety of locations for one length of time. In doing so, CP-ITCHY creates a bias towards fusions of approximately the same size as the original genes (Ostermeier and Benkovic, 2001).

The probability of having a total truncation length of N_{12} is

$$P_{N_{12}} = \sum p_{N_1} p_{N_2} \tag{39}$$

where the summation is for all p_{N1} p_{N2} pairs where

$$N_{12} = N_1 + N_2 = N_{\max} N^* \tag{40}$$

and the individual probabilities p_{Ni} are found from Eq. (41) which is Eq. (8) with $L^* = 1$ (i.e., the mean truncation length $L = N_{max}$).

$$P_{N_i} = 0.5\left(erf\left[\frac{1 - \frac{N_i}{N_{\max}} + \frac{1}{N_{\max}}}{\sqrt{2}c}\right] - erf\left[\frac{1 - \frac{N_i}{N_{\max}}}{\sqrt{2}c}\right]\right) \tag{41}$$

SHIPREC Libraries

A schematic representation of the construction of SHIPREC libraries is shown in Figure 4. The probability P_N that one side of the DNA fusion-molecule will have been cleaved by DNaseI N bases from the free end is

$$P_N = (F)(1 - F)^{N_{\max} - N} \tag{42}$$

where F is the frequency of double-stranded breaks (the inverse of the average distance between double stranded breaks) and $(1 - F)^{N_{max} - N}$ is the probability that there was not a double-stranded break in the base pairs between N and N_{max}.

To a first approximation, we desire to have the probability of a double-stranded break to be $1/N_{max}$ such that, on average, one double-stranded break occurs every N_{max}

bases. As in the development of the αS-dNTP models, we thus replace F with

$$F = \frac{A}{N_{\max}} \tag{43}$$

$$0.1 \le A \le 3 \tag{44}$$

and our probability P_N becomes

$$P_N = \left(\frac{A}{N_{\max}}\right)\left(1 - \frac{A}{N_{\max}}\right)^{N_{\max}-N} \tag{45}$$

and thus the summations Eqs. (30) and (31) reduce to

$$P_{N_{12}} = (N_{12}+1)\left(\frac{A}{N_{\max}}\right)^2\left(1 - \left(\frac{A}{N_{\max}}\right)\right)^{2N_{\max}-N_{12}}; \quad for\ 0 \le N_{12} \le N_{\max} \tag{46}$$

$$P_{N_{12}} = (2N_{\max} - N_{12} + 1)\left(\frac{A}{N_{\max}}\right)^2\left(1 - \left(\frac{A}{N_{\max}}\right)\right)^{2N_{\max}-N_{12}}; \quad for\ N_{\max} \le N_{12} \le 2N_{\max} \tag{47}$$

The probabilities in Eqs. (46)–(47) are for the DNaseI digested DNA prior to size selection, not the final SHIPREC libraries in which size selection has occurred. As in the case where size selection is incorporated into ITCHY libraries, size selection reduces the probability of sizes not selected to zero with a concomitant increase in the probabilities for those sizes selected (see Discussion).

RESULTS

Time-Dependent Truncation

Equation (11) was solved using the approximate expression for erf(x) developed by Hastings (1955) which is accurate to within $\pm\ 1.5 \times 10^{-7}$. Time-dependent truncation produces the most uniform distribution of truncation lengths. Figure 5B shows that the dimensionless truncation profile for $N^* > 0.04$ (P_{N^*} of Eq. (11) normalized by the ideal step-function distribution of Eq. (15)). The curves use the experimentally relevant values of $r_{exo} = 9$ bases/min and $S = 3\ min^{-1}$. The normalized probability is a very weak function of N_{max} over typical ranges of N_{max} (250–1000 bp). As expected, a narrower distribution of truncation lengths for the individual time points ($c = 0.075$) results in a distribution that more closely resembles a step function. The prediction that longer truncations are less prevalent is in agreement with experimental observations (Ostermeier et al., 1999a; 1999b). The curves shown are when $L^*_F = 1$ (i.e., that last time point is taken when $L = N_{max}$). An increase in L^*_F would shift the curves to the right and extend the region with a uniform distribution to a larger fraction of the desired range. To ensure a uniform distribution over $0 \le N^* \le 1$, a L^*_F on the order of 1.2 (for $c = 0.075$) or $L^*_F = 1.5$ (for $c = 0.2$) is required. In other words, the final time point has an average truncation length 20–50% larger than N_{max}.

For $N^* < 0.04$, the distribution of truncation lengths is a dampening oscillating function (Fig. 5C) since the standard deviation of truncation varies linearly with the mean truncation length. This oscillation is a strong function of the standard deviation. For the experimentally relevant values of $R = 9$ bases/min and $S = 3\ min^{-1}$, this oscillating profile disappears at 15–25 bases into the truncation. Designing vectors for incremental truncation such that truncation starts about 20 bases upstream from the gene to be truncated [as in pDIMN2 and pDIMC6 (Ostermeier et al., 1999a)] prevents this bias of truncation lengths in the gene. However, this comes at the expense of having library members in which truncation has stopped before reaching the gene. Serendipitously, such prematurely truncated variants can be functional (Ostermeier et al., 1999a).

THIO(pcr) Truncation

The normalized probability of truncating N^* bases [Eq. (18)] depends almost exclusively on the actual incorporation frequency of the αS-dNTP ($x_\alpha/R = A/N_{max}$). The profile of αS-dNTP(pcr) truncation libraries is heavily biased towards short truncations (Fig. 6a). This bias can be alleviated by a low incorporation rate (lower A), but at the expense of having a large portion of the library outside the desired truncation range.

THIO(Extension) Truncation

The distribution of truncations for THIO(extension) truncation libraries depends almost exclusively on the actual incorporation frequency of the αS-dNTP ($x_\alpha/R = A/N_{max}$) and the standard deviation of the initial ExoIII truncation. Figure 6B and C uses a typical experimental value of N_{max} of 500 bp and a value of L^* chosen so that that the distributions would begin to drop off faster at approximately $N^* = 1$. At high incorporation rates of αS-dNTPs the distribution is heavily biased towards short truncations. However, the incorporation of αS-dNTPs in a limited range allows one to arrive at more uniform truncation profiles at lower αS-dNTP incorporation rates without the trade off of having a majority of the library truncated beyond N_{max}. This is because at lower incorporation rates the large fraction of DNA that does not have an αS-dNTP incorporated in the desired range $0 \le N \le N_{max}$ will be completely digested by ExoIII.

Time-Dependent ITCHY

Unlike the probabilities for the incremental truncation libraries, the distributions for these ITCHY libraries are an appreciable function of N_{max}. However, because the individual truncation distributions (Fig. 5B) closely resemble an ideal step function (Fig. 5A) the distributions of P_{N12} will approximately be that of the ideal case in Figure 7A with slightly lower values. An examination of ITCHY libraries made using αS-dNTPs, in which the deviations from a step function for the incremental truncation libraries are much more severe helps to illustrate this (see below).

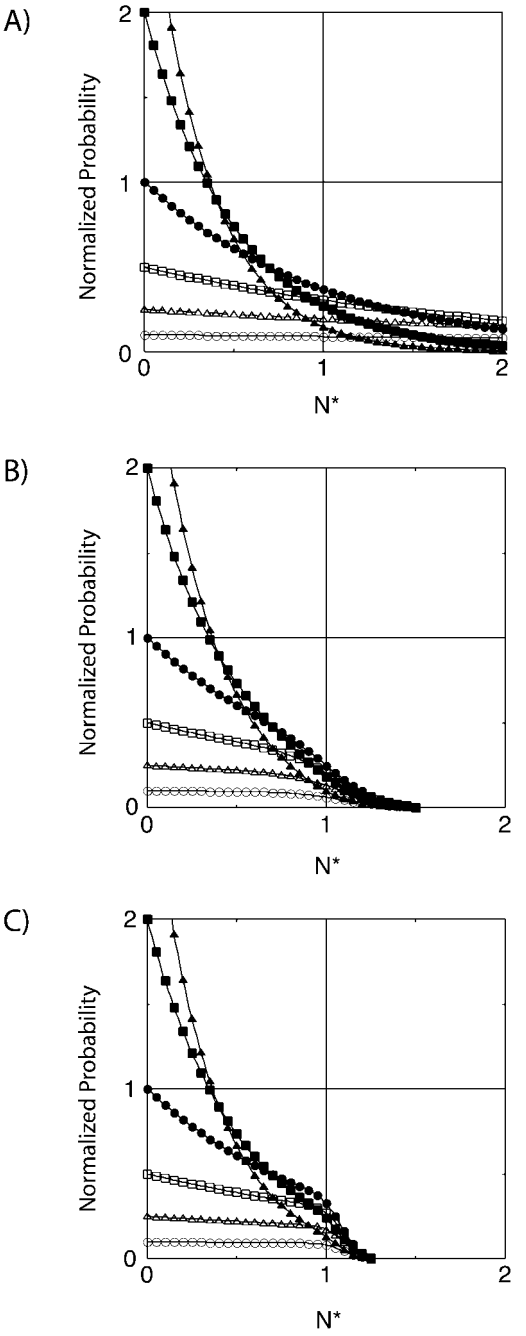

Figure 6. Distribution of truncation lengths for (A) THIO(pcr) truncation (B,C) THIO(extension) truncation as a function of the dimensionless truncation length N^* for A equal to 3 bases (solid triangles), 2 bases (solid squares), 1 base (solid circles), 0.5 bases (open squares), 0.25 bases (open triangles), and 0.1 bases (open circles). In all curves N_{max} is 500 bp, though the deviations for $100 \leq N_{max} \leq 2000$ bp is minimal (see text). Equation (18) was used for THIO(pcr) truncation. Equation (25) was used for THIO(extension) truncation with $L^* = 1.1$ and the standard deviation of the initial truncation being (B) 0.2 L and (C) 0.075 L. In (B) and (C) the probability of truncations of a certain length being in the DNA before transformation is presented. As the incorporation rate decreases, a larger fraction of the DNA is completely degraded. This degraded DNA figures into the probability but is not depicted in the graphs.

THIO(pcr)-ITCHY

Unlike the probabilities for the THIO truncation libraries, the distributions for THIO(pcr)-ITCHY libraries are an appreciable function of N_{max}. Figure 7A shows the values of P_{N12} for $N_{max} = 500$ and $0.1 \leq A \leq 3$. The asymmetry about $N^* = 1$, most apparent at $A = 1$, is a result of Eq. (31) including only those fusions that have both individual truncations in the desired range ($N_1 \leq N_{max}$ and $N_2 \leq N_{max}$). Thus, P_{N12} of Eq. (36) represents the probability of desired fusions, which is most germane. Higher incorporation rates result in libraries that are biased towards larger fusions. This bias can be mitigated by lower incorporation rates but at the expense of having little of the library fused in the desired range. The fraction of DNA that is in this desired range ($N_1 \leq N_{max}$ and $N_2 \leq N_{max}$) can be calculated from Eq. (37) using Eqs. (26) or (27) and is shown as a function of A in Figure 7B [using Eq. (26)].

THIO(Extension)-ITCHY

The primary benefit of using primer extension is that the range that the αS-dNTPs can be incorporated can be limited, depending on L^* and c. As can be shown in Figure 7C, for $A < 1$, the distributions very closely match that of the ideal time-dependent ITCHY libraries. In reality, THIO(extension)-ITCHY libraries (or THIO(pcr)-ITCHY libraries that have the correct positioning of essential plasmid elements proximal to the truncation range) will have distributions somewhere between Figure 7A and C.

Optimum Incorporation Rate of αS-dNTPs

ITCHY library members that have parental length ($N^* \approx 1$) are most likely to be functional since the fusions points will be at structurally homologous locations. This is true if the genes are the same size and are devoid of significant gaps in their alignment. Thus, it is advantageous to optimize the value of P_{N12} at $N_{12} = N_{max}$ ($N^* = 1$).

Figure 7D shows P_{N12} (at $N^* = 1$) as a function of A for THIO(pcr)-ITCHY libraries and the limiting, ideal case of THIO(extension)-ITCHY libraries. For THIO(pcr)-ITCHY libraries, the optimum value of A is 2; thus, αS-dNTPs should be incorporated at a frequency of $2/N_{max}$ to maximize P_{N12} at $N^* = 1$. This optimum value of A is independent of N_{max}.

For the limiting case of ITCHY libraries using αS-dNTPs incorporated by primer extension, the frequency of parental-length fusions approaches that of the time-dependent ExoII method as A decreases. The frequency is 92% of the optimum at $A = 1$ and 98% of the optimum at $A = 0.5$. However, in practice, the curves for ITCHY libraries using αS-dNTPs incorporated by primer extension lie between the two curves in Figure 7D and depend on the values of L* and c.

CP-ITCHY

CP-ITCHY libraries are biased towards parental-length fusions. However, due to the large standard deviation in

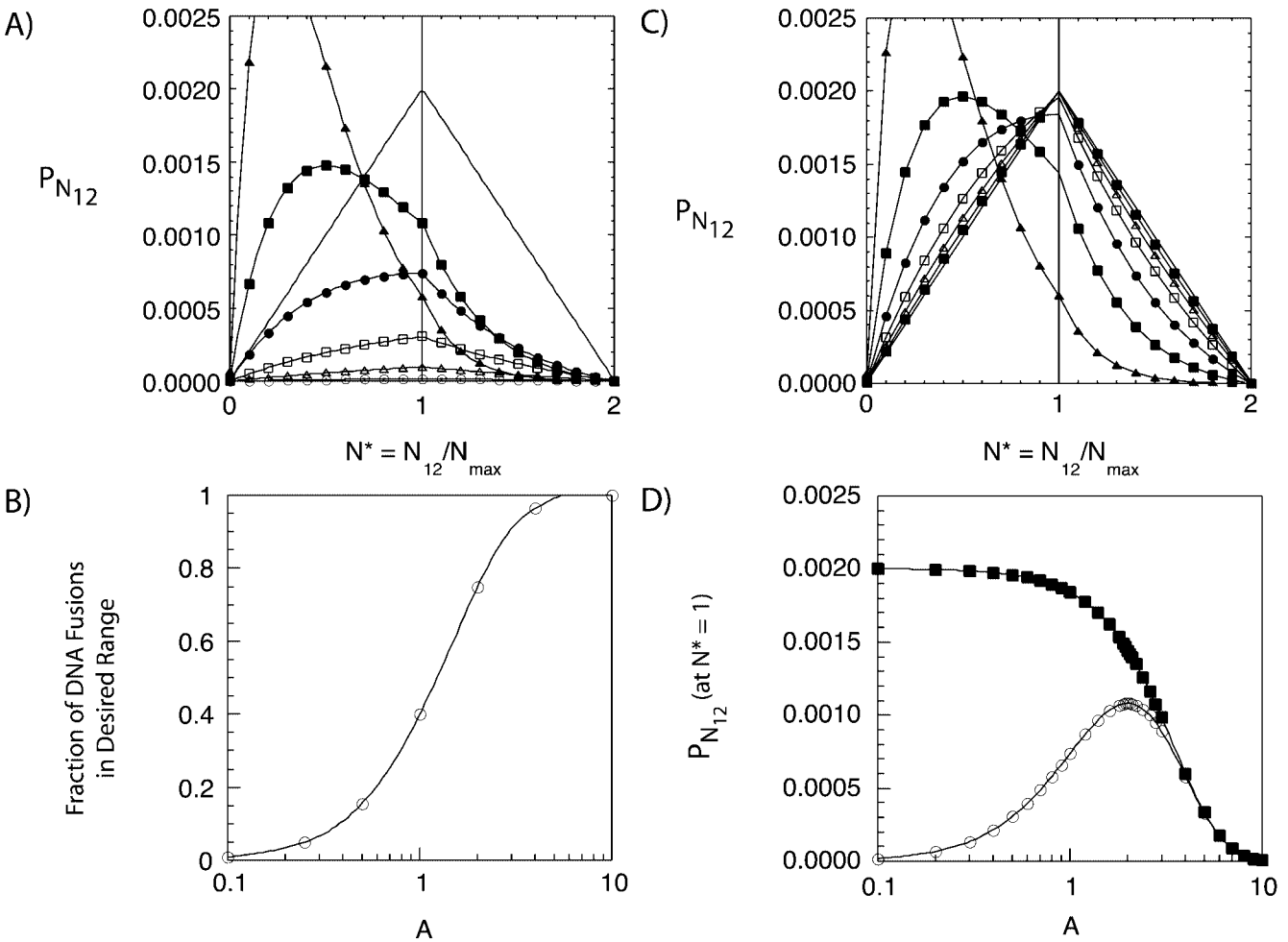

Figure 7. Distribution of total truncation lengths for ITCHY libraries. (A) Distribution of truncation lengths for an ideal ITCHY library (solid line) and for THIO(pcr)-ITCHY libraries from Eqs. (32)–(33) and (35)–(36), respectively. N_{max} is 500 base pairs and A is equal to 3 bases (solid triangles), 2 bases (solid squares), 1 bases (solid circles), 0.5 bases (open squares), 0.25 bases (open triangles), and 0.1 bases (open circles). (B) Fraction of the THIO(pcr)-ITCHY library in the desired range ($N_1 < N_{max}$ and $N_2 < N_{max}$) as a function of the actual incorporation rate of αS-dNTPs from Eq. (37). (C) Distribution of truncation lengths for ideal THIO(extension)-ITCHY libraries from Eq. (38). N_{max} is 500 base pairs and A is equal to 3 bases (solid triangles), 2 bases (solid squares), 1 base (solid circles), 0.5 bases (open squares), 0.25 bases (open triangles), and 0.1 bases (open circles). (D) Probability of parental-length fusions for THIO(pcr)-ITCHY libraries (open circles) and ideal THIO(extension)-ITCHY libraries (solid squares) when $N_{max} = 500$ bp.

ExoIII truncation, the frequency of parental-length fusions only increases 1.5–4-fold (depending on σ) for truncations where $N_{max} = 500$ bp (Fig. 8). This magnitude of this predicted biasing towards parental-length fusions qualitatively matches that observed experimentally. In CP-ITCHY libraries (in which $N_{max} = 548$ bp) sequences that were $0.8 \leq N^* \leq 1.2$ were 2–3-fold more frequent than in time-dependent ITCHY libraries (Ostermeier and Benkovic, 2001). However, this increase in parental-length fusions is inversely proportional to N_{max} since the standard deviation is a function of the average truncation length. Thus, the parental-length size biasing will be more pronounced for libraries where truncation is designed to occur over a smaller range.

SHIPREC

The distribution of probabilities for the DNaseI digested DNA used in SHIPREC libraries is shown in Figure 9 and is the mirror image of THIO(pcr)-ITCHY libraries (Fig. 7A). This distribution is not that of the final SHIPREC libraries in which size selection has occurred (see Discussion).

Distribution of Parental-Length Fusions

An important consideration is the distribution of fusions along the sequence at $N^* = 1$. In other words, what is the probability of having truncated N bases on one gene and $N_{max} - N$ bases on the other gene as a function of position along the gene? If the answer is a function of N, then the distribution of gene fusions where $N_{12} = N_{max}$ will be uneven.

Time-dependent ITCHY libraries will have a bias against fusions nearest the ends of the truncation range if the average truncation length of the last time point is N_{max} (Fig. 10A). However, this bias can be eliminated if the truncation

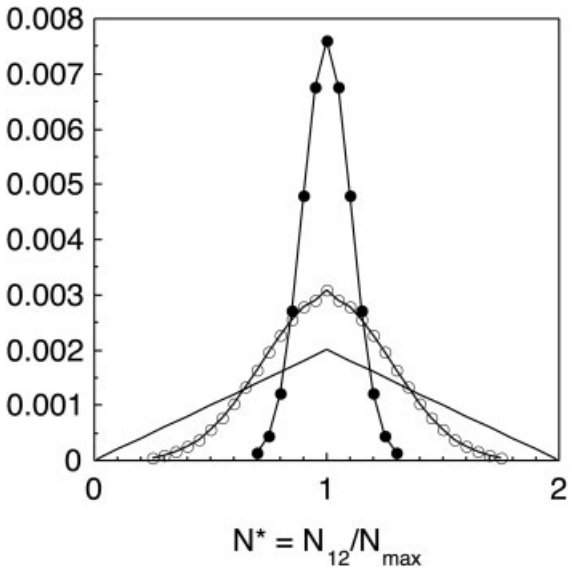

Figure 8. Distribution of total truncation lengths for CP-ITCHY libraries for N_{max} of 500 bp and the standard deviation of ExoIII truncation of 0.2L (open circles) and 0.075L (solid circles) from Eq. (39). The distribution of an ideal ITCHY library (solid line) taken from Eqs. (32) and (33) is shown for comparison.

is performed such that the average truncation length of the last time point greater than 1.2–1.5 N_{max}. THIO(pcr)-ITCHY libraries and SHIPREC will have an even distribution, since the probability of having truncated N-bases on one gene and $N_{max} - N$ bases on the other gene is not a function of N.

$$P_N P_{N_{\max}-N} = \left(\frac{A}{N_{\max}}\right)^2 \left(1 - \frac{A}{N_{\max}}\right)^{N_{\max}} \tag{48}$$

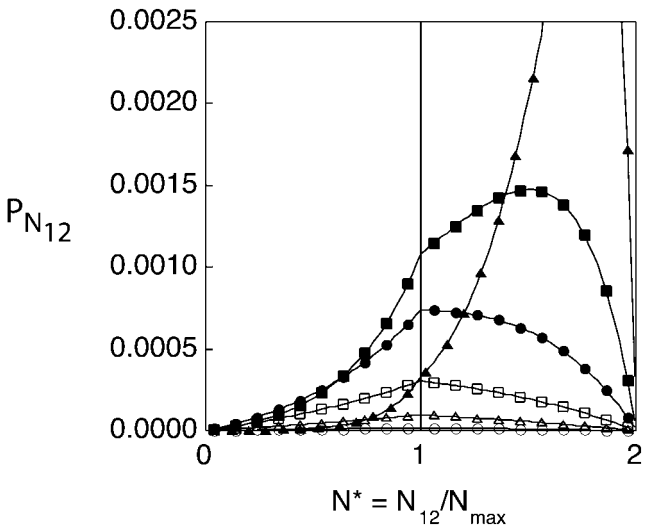

Figure 9. Distribution of truncation length of DNase I digested DNA in the preparation of SHIPREC libraries from Eqs. (46)–(47). N_{max} is 500 base pairs and A is equal to 4 bases (solid triangles), 2 bases (solid squares), 1 base (solid circles), 0.5 bases (open squares), 0.25 bases (open triangles), and 0.1 bases (open circles).

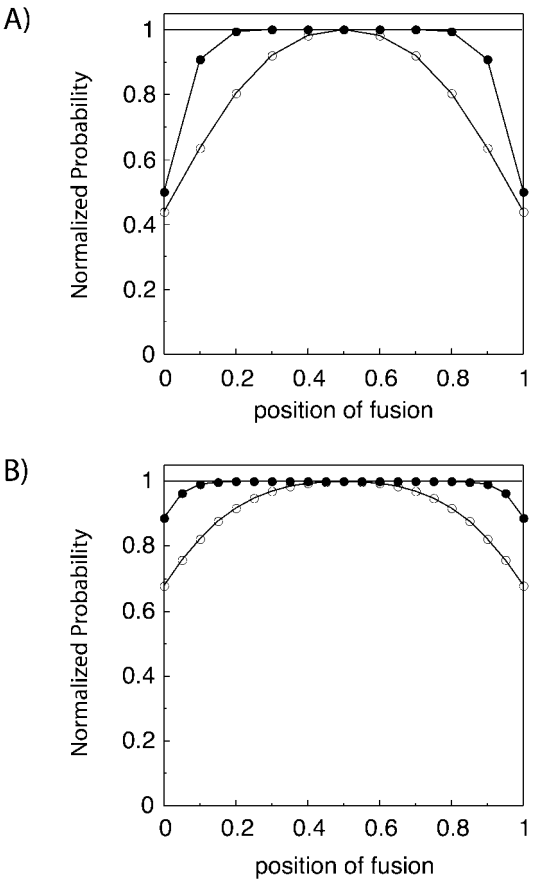

Figure 10. Distribution of fusion points of parental-length fusions in ITCHY libraries. (A) Time-dependent ITCHY libraries with $L^*_F = 1$ and the standard deviation of ExoIII truncation of 0.2 L (open circles) and 0.075 L (solid circles). (B) Distributions for THIO(extension)-ITCHY libraries when $L^*_F = 1$ and the standard deviation of the initial ExoIII truncation is 0.2 L (open circles) or 0.075 L (solid circles).

The expressions for P_N and P_{Nmax-N} in Eq. (48) are found by Eqs. (13) and (17) for THIO(pcr)-ITCHY and Eq. (45) for SHIPREC. In contrast, THIO(extension)-ITCHY libraries will have the probabilities at the ends diminished to the degree to which $P_{dig > N} < 1$ when $N < N_{max}$. For the values used in Figure 6B and C, this effect is minimal; however, if the initial truncation is not far enough, for example when $L^*_F = 1$, the probability will become much more uneven (Fig. 10B).

DISCUSSION

The theoretical prediction of the distribution of truncation lengths suggests a number of advantages and disadvantages for the different methods that are summarized in Table I. Time-dependent truncation produces the most even distribution of truncation lengths; however, there is a bias against longer truncations (Fig. 5B). This can be overcome by performing truncations over a range of DNA that is 1.2–1.5 N_{max} but at the expense of having more of the library outside the desired truncation range. The values of the normalized

Table I. Comparison of incremental truncation methods.

Method	Advantage	Disadvantage
Time-dependent truncation[a]	• Most uniform distribution • Highest control over range of truncations	• Multiple time-point sampling
THIO(pcr) truncation[b]	• Experimental convenience • Option of including random mutagenesis in library construction	• Distribution biased towards short truncations • Limited control of truncation range
THIO(extension) truncation[b]	• Experimental convenience • Better control over range of truncations than THIO(pcr) truncation • Option of including random mutagenesis in library construction	• Distribution biased towards short truncations • Less convenient than THIO(pcr) truncation
Time-dependent ITCHY[c]	• Most uniform distribution of possible fusions • Highest control over range of truncations	• Multiple time-point sampling • Risk of uneven distribution of parental-length fusions
THIO(pcr)-ITCHY[b]	• Experimental convenience • Inherent uniform distribution of parental-length fusions • Option of including random mutagenesis in library construction	• Lower probability of desired fusions
THIO(extension)-ITCHY[b]	• Experimental convenience • Frequency of desired fusions higher than THIO(pcr)-ITCHY • Option of including random mutagenesis in library construction	• Risk of uneven distribution of desired fusions • Less convenient than THIO(pcr)-ITCHY
CP-ITCHY[d] SHIPREC[e]	• Highest probability of parental-length fusions • Inherent uniform distribution of parental-length fusions • Size selection can be conveniently performed prior to ligation • Option of including random mutagenesis in library construction	• Difficult library construction • Difficult library construction • PCR amplification of the size selected libraries can create bias • Implementation on subsections of gene is cumbersome

[a]Ostermeier et al., 1999a.
[b]Lutz et al., 2001a.
[c]Ostermeier et al., 1999b.
[d]Ostermeier and Benkovic, 2001.
[e]Sieber et al., 2001.

probabilities at shorter truncation lengths are marginally greater than one due to the fact that later time points (with a longer average truncation length) have a larger standard deviation that serves to contribute back to shorter truncation lengths. This is supported by the fact that this effect is greater for a standard deviation of 0.2 L than for 0.075 L.

THIO(pcr) truncation and THIO(extension) truncation libraries are prone to having nonuniform distributions of truncation lengths that can only be overcome by decreasing the incorporation rate of αS-dNTP. For THIO(pcr) truncation, this results in a large fraction of the library outside the desired truncation range. THIO(extension) libraries can alleviate this if the length of the initial ExoIII digestion is carefully controlled. It is worth noting that THIO(pcr) truncation libraries can also be "forced" to behave in this manner by positioning essential plasmid elements (e.g., origin of replication or an antibiotic resistance gene) just outside the desired truncation range. This can be best achieved in truncations from the 5′ to the 3′ end of the gene, since the essential element can be placed just outside the end of the gene. Truncations in the opposite direction will usually have to have the essential element position further away, outside the promoter.

Paradoxically, the nature of the size bias in THIO(pcr) truncation libraries necessarily results in an even distribution of fusion points along the gene of parental-length fusions when implemented in ITCHY libraries. In contrast, unless time-dependent truncation libraries are prepared over a range of greater than 1.2–1.5 N_{max}, time-dependent ITCHY libraries will be biased against fusion points near the ends of genes for parental-length fusions (Fig. 10A). Such biasing could be useful in some circumstances (i.e., fusions near the ends are more likely to have the properties similar to the gene with the larger fragment) or in the case where computational methods [such as schema disruption (Voigt et al., 2001)] predict central regions of the gene as more likely to tolerate fusion. However, in the absence of any rationale for biasing, the inherent even distribution of THIO(pcr)-ITCHY is preferable. In particular this is true for when the ITCHY libraries are to be homologously recombined in the creation of SCRATCHY libraries (Lutz et al., 2001b).

The experimental convenience and even distribution of THIO(pcr)-ITCHY libraries comes at the expense of having a lower frequency of parental-length fusions in the library compared to time-dependent ITCHY. The frequency of parental-length fusions is predicted to be at a maximum for THIO(pcr)-ITCHY libraries when the αS-dNTP are incorporated at a frequency of $2/N_{max}$. Since the relative incorporation rates of Eq. (12) are not known, one cannot at this

point determine the mole fraction of αS-dNTPs to use to achieve a frequency of $2/N_{max}$. However, the optimum can still be determined experimentally by testing different mole fractions and determining which achieves the truncation profile of $A = 2$ in Figure 6. At $A = 2$, the ratio of the amount of DNA not truncated to that that has been truncated N_{max} should be 7.4.

The degree to which THIO(extension) truncation can limit the distribution to the desired range will shift the optimum incorporation rate of αS-dNTPs to lower values. Although Figure 7D shows the maximum frequency of parental-length fusions is highest at lower incorporation rates, one must balance optimizing the frequency of parental-length fusions with the amount of transformable DNA.

SHIPREC libraries are analogous to THIO(pcr)-ITCHY libraries in that (a) the nature of the bias of truncations produced by DNaseI digestion in the construction of libraries ensures an even distribution of parental-length fusions across the gene, and (b) the optimum frequency of double-stranded breaks induced by DNaseI will be $2/N_{max}$. The view that SHIPREC distinguishes itself from ITCHY by primarily creating parental-length fusions (Sieber et al., 2001) is somewhat of a misconception. ITCHY libraries can be and have been (Ostermeier and Benkovic, unpublished) selected for size simply by subcloning the library, by size selection on the plasmid prior to ligation, or by size selection after ligation (with PCR amplification of the desired size products). Size selection was not performed on published ITCHY libraries (Lutz et al., 2001a; Ostermeier and Benkovic, 2001; Ostermeier et al., 1999b) simply because the strong selection of auxotrophic complementation made it unnecessary. While size selection can be performed as needed in ITCHY libraries, SHIPREC libraries require it to avoid complications with the DNA fragments cleaved off the ends interfering with subsequent steps in the library construction. The size-selection step in SHIPREC libraries is followed by a seemingly necessary PCR amplification to obtain enough material for subsequent steps; however, PCR amplification of multiple templates has been observed to introduce bias (Lutz et al., 2001b; Sugimoto et al., 1993).

CONCLUSIONS

The different methods for creating incremental truncation and ITCHY libraries are theoretically predicted to have different distributions of truncation lengths. For producing truncations of a single gene, time-dependent truncation is preferable due to its relatively uniform distribution of truncations. However, for ITCHY libraries, a uniform distribution of parental-length fusions is most readily prepared using THIO(pcr)-ITCHY.

I thank Stephan Lutz and Gurkan Guntas for critically reading the manuscript prior to submission.

NOMENCLATURE

A	average number of incorporations of αS-dNTPs per N_{max} (bases)
c	standard deviation factor $= \sigma/L$
f	fraction of DNA truncated in desired range ($0 \le N \le N_{max}$ or $0 \le N^* \le 1$)
f_{12}	fraction of DNA fusions where $N_1 \le N_{max}$ and $N_2 \le N_{max}$
f_{trans}	fraction of DNA capable of transforming bacteria ($0 \le N^* \le$ w)
F	frequency of double strand breaks by DNaseI (base pairs^{-1})
$G(z)$	Gaussian distribution function
L	mean truncation length of Exo III digestion (bases)
L_F	mean truncation length of final time-point (bases)
L^*	dimensionless mean truncation length for a time point
L^*_F	dimensionless mean truncation length for the final time point
n_T	total number of time points
N	length of truncation (base pairs)
N_1, N_2	length of truncation of gene 1 and gene 2, respectively (base pairs)
N^*	dimensionless truncation length
N_{12}	total length of truncation of both genes in ITCHY $= N_1 + N_2$ (base pairs)
N_{max}	desired maximum length of truncation (base pairs)
r	rate of incorporation of αS-dNTPs or dNTPs (base pairs/min)
R	ratio of incorporation rate of dNTPs to incorporation rate of αS-dNTPs
r_{exo}	rate of ExoIII digestion (base pairs/min)
x_α	mol fraction of αS-dNTPs = [αS-dNTPs]/([αS-dNTPs] + [dNTPs])
p_N	probability that a DNA molecule has been truncated N bases in an individual time point
$P_{dig > N}$	probability that a DNA molecule has been truncated more than N base pairs
P_N	probability that a DNA molecule will have N base pairs truncated
P_{N12}	probability that a DNA fusion molecule has a total truncation length of N_{12} and that neither gene was truncated more than N_{max}
P_{N^*}	probability that a DNA molecule will have been truncated N^*
$P_{N^*, ideal}$	probability that a DNA molecule of truncation length N^* in an ideal library with a flat distribution
$P_{N^*, trans}$	probability that a DNA molecule with truncation length N^* is in the transformed library
P_{thio}	probability that there is αS-dMTP incorporated at position N and there is not a αS-dMTP at any position less than N
S	sampling rate (min^{-1})
w	dimensionless truncation length beyond which the DNA is not transformable
z	standard score
σ	standard deviation of ExoIII digestion equal to cL (bases)

References

Arnold FH. 2001. Combinatorial and computational challenges for biocatalyst design. Nature 409:253–257.

Burgers PM, Eckstein F. 1979. A study of the mechanism of DNA polymerase I from *Escherichia coli* with diastereomeric phosphorothioate analogs of deoxyadenosine triphosphate. Carl Dieffenbach, ed. J Biol Chem 254:6889–6893.

Caldwell RC, Joyce GF. 1995. Mutagenic PCR. In: PCR Primer: A laboratory manual. Plainview, NY: Cold Spring Harbor Laboratory Press.

Hastings C. 1955. Approximations for digital computers. Princeton, NJ: Princeton University Press.

Hoheisel JD. 1993. On the activities of *Escherichia coli* exonuclease III. Anal Biochem 209:238–246.

Lutz S, Ostermeier M, Benkovic SJ. 2001a. Rapid generation of incremental truncation libraries for protein engineering using α-phosphothioate nucleotides. Nucleic Acids Res 29:e16.

Lutz S, Ostermeier M, Moore G, Maranas C, Benkovic SJ. 2001b. Creating

multiple-crossover DNA libraries independent of sequence identity. Proc Natl Acad Sci USA 98:11248–11253.

Ostermeier M, Benkovic SJ. 2001. Construction of hybrid gene libraries involving the circular permutation of DNA. Biotechnol Lett 23: 303–310.

Ostermeier M, Lutz S, Benkovic SJ. 2002. Incremental truncation for hybrid enzymes/combinatorial shuffling. In: Golemis E, editor. Protein–protein interactions: A molecular cloning manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Ostermeier M, Nixon AE, Shim JH, Benkovic SJ. 1999a. Combinatorial protein engineering by incremental truncation. Proc Natl Acad Sci USA 96:3562–3567.

Ostermeier M, Shim JH, Benkovic SJ. 1999b. A combinatorial approach to hybrid enzymes independent of DNA homology. Nat Biotechnol 17: 1205–1209.

Sieber V, Martinez CA, Arnold FH. 2001. Libraries of hybrid proteins from distantly related sequences. Nat Biotechnol 19:456–460.

Stemmer WPC. 1994. DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. Proc Natl Acad Sci USA 91:10747–10751.

Sugimoto T, Fujita M, Taguchi T, Morita T. 1993. Quantitative detection of DNA by coamplification polymerase chain reaction: A wide detectable range controlled by the thermodynamic stability of primer template duplexes. Anal Biochem 211:170–172.

Voigt CA, Mayo SL, Arnold FH, Wang ZG. 2001. Computationally focusing the directed evolution of proteins. J Cell Biochem 37Suppl: 58–63.