J. theor. Biol. (2002) **219,** 9–17
doi:10.1006/yjtbi.3102, available online at http://www.idealibrary.com on IDEAL®

Predicting Out-of-Sequence Reassembly in DNA Shuffling

Gregory L. Moore† and Costas D. Maranas*†

†*Department of Chemical Engineering*, *The Pennsylvania State University*, 112 *Fenske Laboratory*, *University Park*, *PA* 16802, *U.S.A.*

(*Received on* 25 *February* 2002, *Accepted in revised form on* 6 *June* 2002)

We present an analysis for calculating the frequency of out-of-sequence reassembly in DNA shuffling experiments. Out-of-sequence annealing events are undesirable since they typically encode non-functional proteins with missing or repetitive regions. The approach builds on the *e*Shuffle framework for the prediction of crossover formation using equilibrium thermodynamics and complete sequence information to model the reassembly process. An *in silico* case study of a set of subtilases reveals that, as expected, the presence of significant sequence identity between distant portions of the parental sequences gives rise to out-of-sequence annealing events that upon reassembly generate sequences with missing or repetitive DNA segments. The frequency of these events increases as the fragment length decreases. Interestingly, out-of-sequence annealing events are at a minimum near the annealing temperature of 55°C used in the original DNA shuffling protocol. Neither parental sequence identity nor number of shuffled parents significantly alter the extent of out-of-sequence reassembly.

Introduction

DNA shuffling (Stemmer, 1994) is one of the most commonly used protocols for generating combinatorial DNA libraries. The DNA shuffling reaction consists of random fragmentation of a set of parental nucleotide sequences with the enzyme DNase I and reassembly of the fragments through the polymerase chain reaction (PCR) without primers (see Fig. 1). Each cycle of PCR consists of three steps: *denaturization*, in which double-stranded fragments are separated into single-stranded fragments by heating; *annealing*, in which single-stranded fragments come together to reform double-stranded DNA as the temperature is lowered; and *extension*, in which a DNA polymerase enzyme catalyses the addition of nucleotides to overhanging single-stranded regions. These cycles are repeated until full-length (parental size) sequences are reassembled. The formation of *heteroduplexes* (see Fig. 1) in the annealing step and their subsequent extension is the key mechanism for the generation of diversity. A crossover is defined as the junction point where two fragments originating from different parental sequences are linked. The extent of diversity in DNA combinatorial libraries is typically quantified as the number of crossovers per reassembled sequence.

In Moore *et al.* (2001), the *e*Shuffle framework was introduced for quantifying the distribution of crossovers in a combinatorial DNA library generated by DNA shuffling. With *e*Shuffle, the effect of fragment length, annealing temperature, sequence identity and number of shuffled

*Corresponding author. Tel.: + 1-814-863-9958; fax: + 1-814-865-7846.
E-mail address: costas@psu.edu (C.D. Maranas).

0022-5193/02/$35.00/0

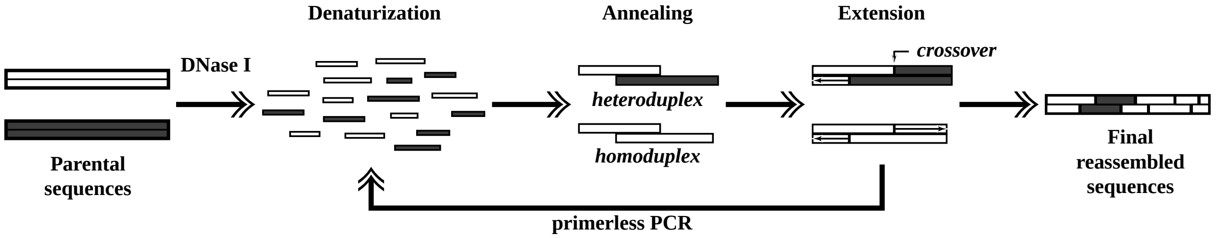


FIG. 1. The DNA shuffling protocol consists of random fragmentation of parental nucleotide sequences with DNase I and fragment reassembly through primerless PCR. Each cycle of PCR has three steps: (i) denaturization, (ii) annealing and (iii) extension. Diversity in the form of crossovers is generated when two fragments from different parents anneal (forming a *heteroduplex*) and are extended.

parental sequences on the number, type and location of crossovers along the length of reassembled sequences can be quantified. Annealing events were modeled as a network of reactions, and equilibrium thermodynamics was used to quantify the extent and selectivities of duplex formation. For the sake of simplicity, it was assumed in the *e*Shuffle algorithm that fragments anneal only in their alignment positions. Here, we relax this assumption and we allow the free energy gain or loss associated with every annealing event, either in-sequence or out-of-sequence, to determine its likelihood. This enables the quantification of out-of-sequence annealing events under different protocol setups (e.g. annealing temperature, fragmentation length, etc.). Clearly, out-of-sequence events are undesirable because they typically give rise to sequences with missing or repetitive regions, rendering the likelihood of functionality of the encoded proteins extremely low (though not zero). Although a range of sequence lengths are produced by DNA shuffling (typically seen as a "smear" in gel electrophoresis), this analysis focuses on sequences that would have ordinarily reached full length.

An example of how DNA motif repeats could lead to out-of-sequence reassembly is shown in Fig. 2 where three fragments share a 5-nt motif (i.e. TACAT) repeated in different locations along the gene sequence. Specifically, the motif spans positions 46–50 in the first fragment, positions 96–100 for the second and 196–200 for the third. The out-of-sequence annealing of the first fragment to the second fragment results in the duplication of the 51–100 range in the reassembled sequence, whereas the out-of-sequence annealing of the third fragment to the second fragment yields a truncated reassembled sequence with the 101–200 range deleted. Only the in-sequence annealing of the second fragment results in a full-length sequence.

The goal of this paper is to build on the *e*Shuffle framework to examine *in silico* how out-of-sequence annealing events are affected by fragment length, parental sequence identity, annealing temperature, number of parental sequences and the presence of repeated nucleotides for sequences that would otherwise reach full length.

Algorithm

In DNA shuffling, fragments compete to anneal with growing gene templates during reassembly. Here the terms "fragment" (small) and "template" (large) refer to the respective sizes of the nucleotide oligomers participating in the shuffling reaction, not their roles in the extension step (see Fig. 3). In Moore *et al.* (2001), the competition for a particular template only included fragments in their alignment positions. However, since fragments are not "locked" into their sequence alignment, but instead anneal based on free energy gains or losses, this condition is relaxed so that the frequency of out-of-sequence reassembly can be quantified. A stepwise description of the calculations used to resolve the DNA shuffling reaction network is provided. For clarity of presentation, this analysis only considers the use of a single fragment length L for reassembly. A range of fragment lengths can be analysed in the same

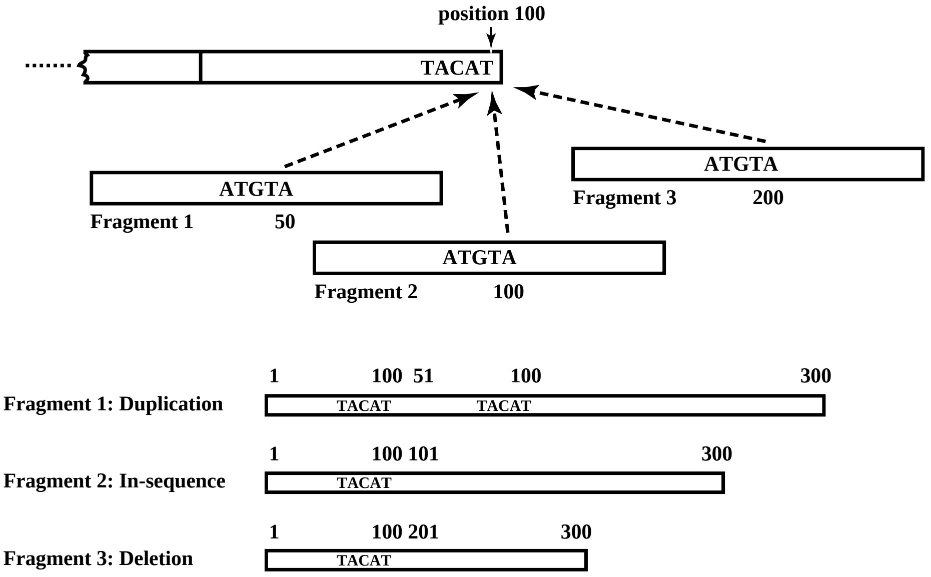


FIG. 2. The formation of missing or repetitive regions by out-of-sequence annealing. Fragment 2 is in-sequence, so a full-length sequence results. However, fragment 1 (duplication) and fragment 3 (deletion) could anneal out-of-sequence and cause the formation of duplicated or deleted sequences.

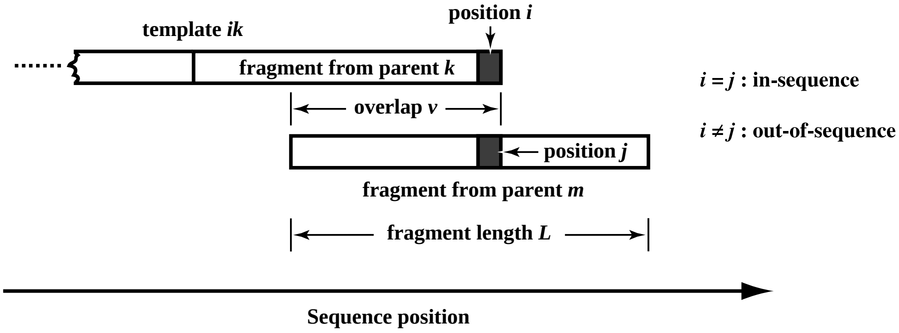


FIG. 3. Definitions of indicies i, j, k, m and v used in the analysis.

manner presented as in Moore & Maranas (2000).

Step 1: *Template selection for annealing selectivity analysis.* The DNA fragment–template reaction network is decoupled by examining each template separately. Templates in the reaction mixture are fully characterized by indices i referring to the sequence position at the end of the partially reassembled template and k denoting the parental sequence origin of the last fragment added to the template. These definitions are clarified in Fig. 3.

It is assumed that tracking only the last annealed fragment is sufficient to determine the annealing selectivities for the next fragment choices. The calculations described in *Steps* 2–5 consider competition for only a single template A defined by indices i and k values. Thus, *Steps* 2–5 are repeated for all i and k combinations that define all possible template choices.

Step 2: *Definition of expanded DNA shuffling reaction network for a single template.* The fragment–template reaction network is expanded to consider out-of-sequence annealing events. This requires an additional index j denoting the position within the fragment where the duplex is terminated (see Fig. 3). Fragments are defined by indices v denoting the annealing overlap length and m identifying the parental sequence that yielded the fragment. It is important to note that if $i=j$, the fragment has annealed at its alignment position (i.e. in-sequence), whereas $i \neq j$ characterizes an out-of-sequence annealing event.

The competition for a particular template A by multiple fragments F is summarized by the following set of reactions:

$$A + F_{mv}^{j} \rightleftharpoons AF_{mv}^{j} \quad \forall j, m, v.$$

Here, F_{mv}^{j} refers to a fragment originating from parental sequence m that anneals with an overlap of v nucleotides so that sequence position j terminates the duplex. Also AF_{mv}^{j} refers to the fragment–template duplex formed. Note that reactions that do not yield duplexes with $5'$ overhangs or with mismatches at exactly the $3'$ end are neglected since they do not contribute to the reassembly reactions. This is because DNA polymerases have only $5' \rightarrow 3'$ activity.

Step 3: *Equilibrium thermodynamics calculations for a single template.* Nearest-neighbor parameters for DNA duplexes established by SantaLucia Jr and co-workers are utilized to estimate the enthalpy, entropy and free energy change for all annealing choices at a given temperature (Allawi & SantaLucia Jr, 1997; SantaLucia Jr, 1998; Allawi & SantaLucia Jr, 1998a–c). The free energy change is approximated as the sum of all two nucleotide (2-nt) contributions. The stabilizing effect of the 2.2 mM of magnesium ion present in DNA shuffling reactions is also considered. A logarithmic interpolation of the data in Nakano *et al.* (1999) results in an "effective" potassium ion concentration of 98 mM (including the actual $[K^+] = 50$ mM present). The duplex free energy is then adjusted using the standard correction for salt concentration (SantaLucia Jr, 1998). The incorporation of Mg^{2+} causes an approximate increase of 5–10°C in calculated DNA fragment melting temperatures, as observed experimentally by Nakano *et al.* (1999). Given this free energy predictive capability, the extent of duplex formation can be tracked at different temperatures.

The change in free energy and the equilibrium constant for the annealing reaction set proposed above are denoted as $\Delta G_{mv}^{j}(T)$ and $K_{mv}^{j}(T)$ and are linked by the following expression:

$$K_{mv}^{j}(T) = \exp\left(-\frac{\Delta G_{mv}^{j}(T)}{RT}\right) \quad \forall j, m, v.$$

Step 4: *Fragment, template and duplex equilibrium concentrations at a given temperature.* Equilibrium constants are linked to the molar concentrations (indicated by square brackets) of the species present in the reaction mixture (i.e. fragments, template and duplexes) as follows:

$$K_{mv}^{j}(T) = \frac{[AF_{mv}^{j}]}{[A][F_{mv}^{j}]} \quad \forall j, m, v.$$

First, parameter $a(T)$ is defined as the portion of templates that have annealed at temperature T (subscript 0 is used to denote initial species concentrations)

$$a(T) = \frac{[A]_0 - [A]}{[A]_0} = \frac{\sum_{j,m,v} [AF_{mv}^{j}]}{[A] + \sum_{j,m,v} [AF_{mv}^{j}]}.$$

After substituting $[AF_{mv}^{j}] = K_{mv}^{j}(T)[A][F_{mv}^{j}]$ in the above expression, we obtain

$$a(T) = \frac{\sum_{j,m,v} K_{mv}^{j}(T)[F_{mv}^{j}]}{1 + \sum_{j,m,v} K_{mv}^{j}(T)[F_{mv}^{j}]}. \qquad (1)$$

A balance on each fragment type $[F_{mv}^{j}]$ yields

$$[F_{mv}^{j}] + [AF_{mv}^{j}] = [F_{mv}^{j}]_0 \quad \forall j, m, v.$$

This expression is recast in terms of $[F_{mv}^{j}]$ and rearranged so that $[F_{mv}^{j}]$ becomes a function of only $a(T)$

$$[F_{mv}^{j}] + K_{mv}^{j}(T)[A][F_{mv}^{j}] = [F_{mv}^{j}]_0 \quad \forall j, m, v,$$

$$[F_{mv}^{j}] = \frac{[F_{mv}^{j}]_0}{1 + K_{mv}^{j}(T)[A]_0(1 - a(T))} \quad \forall j, m, v. \qquad (2)$$

The square system of equations [(1, 2)] is solved iteratively by first initializing $a(T)$ and then updating $[F_{jv}^{m}]$ using eqn (2) and $a(T)$ from eqn (1) until the system converges. This procedure finds a single solution to the network regardless of the initial guess for $a(T)$.

Step 5: *Annealing selectivities calculation at* T_{ann}. The identified equilibrium concentrations are next used to calculate all annealing selectivities for the current template (i, k). The temperature-dependent selectivity $s^j_{mv}(T)$ denotes the probability that a fragment (m, v) anneals to a particular template forming a duplex that extends up to position j at temperature T (see Fig. 4).

$$s^j_{mv}(T) = \frac{[AF^j_{mv}]}{\sum_{j',\,m',\,v'} [AF^{j'}_{m'v'}]} \quad \forall j, m, v.$$

Even though indicies i, k are not included in $s^j_{mv}(T)$ for the sake of clarity, the selectivities are implicitly defined for the template characterized by indices i, k chosen in *Step* 1. *Steps* 2–5 are repeated for all i, k combinations.

Step 6: *Fragment reassembly.* The reassembly algorithm follows closely the recursive steps introduced in Moore *et al.* (2001). Specifically, T_{ik} is defined as the probability that a template is extended to position i with a fragment that originates from parental sequence k. The boundary condition at the start is set by selecting a fragment and positioning it to completely cover the range 1 to L. This initial fragment originates from parental sequence k with probability equal to the relative parental sequence concentration C_k

$$T_{L,k} = C_k \quad \forall k,$$

$$T_{ik} = 0 \qquad \forall i < L \quad \text{and} \quad \forall k.$$

A temperature-integrated selectivity S_{mv} is employed to account for duplex formation starting at the denaturization temperature T_{den} down to the annealing temperature T_{ann}. Note that the index j is not present since for reassembly purposes all fragments are assumed to be in alignment

$$S_{mv} = \int_{T_{ann}}^{T_{den}} s_{mv}(T) \frac{\mathrm{d}a(T)}{\mathrm{d}T} \mathrm{d}T.$$

A recursive relation is then established that relates the probability T_{im} of extending a template up to position i with a fragment from parent m with all possible fragment choices in terms of overlap v and parental origin of the template end k before annealing

$$T_{im} = \sum_k \sum_v T_{i-(L-v),k} S_{mv} \quad \forall i > L \quad \text{and} \quad \forall m.$$

The resolution of this recursion provides the probabilities of formation T_{ik} for any template end position/parental origin combination.

Step 7: *Calculation of out-of-sequence reassembly probabilities.* Next the annealing selectivities s^j_{mv} (including both in-sequence and out-of-sequence annealing events) for a particular template (i, k) are combined with the frequency at which the template is reassembled T_{ik} to estimate the probability P_{ij} of annealing position i in the template with position j in the fragment. Note that ($i = j$; in-sequence annealing) whereas ($i \neq j$; out-of-sequence annealing). Two factors determine the probability P_{ij} that a template ending at position i anneals to a fragment at exactly position j: the frequency of finding templates ending at position i and the selectivities of fragments for those templates (j on i annealing). This is given by the product of T_{ik} and $s^j_{mv}(T_{ann})$ summed over all annealing choices given by k, m and v

$$P_{ij} = \sum_k T_{ik} \sum_{m,v} s^j_{mv}(T_{ann}).$$

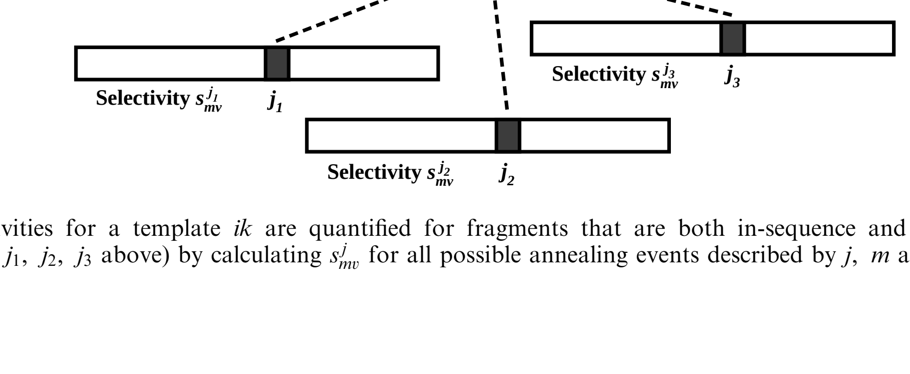

template *ik*
position *i*
fragment from parent *k*
Selectivity $s^{j_3}_{mv}$
j_3
Selectivity $s^{j_1}_{mv}$
j_1
Selectivity $s^{j_2}_{mv}$
j_2

FIG. 4. Selectivities for a template *ik* are quantified for fragments that are both in-sequence and out-of-sequence (characterized by j_1, j_2, j_3 above) by calculating s^j_{mv} for all possible annealing events described by j, m and v.

The probability p_i^{out} that an out-of-sequence annealing event occurs at position i is found by summing over all P_{ij} with $i \neq j$, neglecting the $i = j$ case since it represents fragments in their alignment positions

$$p_i^{out} = \sum_{j \neq i} P_{ij}.$$

The reassembly of the entire sequence is examined by calculating the probability P^{align} that a reassembled sequence is in alignment with the parental sequences (i.e. full length). This occurs when none of the sequence positions generate out-of-sequence annealing events.

$$P^{align} = \prod_i (1 - p_i^{out}).$$

Therefore, the probability P^{out} that a reassembled sequence includes a missing or repetitive region is $1 - P^{align}$.

$$P^{out} = 1 - \prod_i (1 - p_i^{out}).$$

Since reassembly is bidirectional, the calculations are performed independently for both forward and reverse sequences, and the results are averaged. With this computational protocol in place, we next examine the DNA shuffling of a set of subtilases as a case study.

Results and Discussion

SUBTILASE CASE STUDY

In this section, we revisit the subtilase case study first addressed in Moore *et al.* (2001). Four subtiliase genes are examined: subtilisins E and BPN′, serine protease D and proteinase K. As in an earlier DNA shuffling experiment involving subtilases (Ness *et al.*, 1999), we chose to shuffle only a 494-bp region of the four genes. Specifically, this region aligns with nucleotides 529–1022 of the mature coding region of subtilisin E and is renumbered as 1–494. We examine *in silico* how fragment length, annealing temperature, degree of parental sequence identity and number of parental sequences affect the frequency of out-of-sequence annealing events that occur during reassembly of these subtilase sequences. For this study, parameters are set to match the original DNA shuffling protocol (Stemmer, 1994): $[K^+] = 50$ mM, $[Mg^{2+}] = 2.2$ mM and total fragment concentration of 10 ng/μL.

The first two subtilases considered are subtilisin E and subtilisin BPN′, which share a nucleotide sequence identity of 80%. First, the effect of fragmentation length L on the probability of out-of-sequence annealing events occurring at specific reassembly positions is examined. Running the computational protocol with fixed $T_{ann} = 55°C$ for $L = 10$, 25, and 100-nt generates three plots of p_i^{out} vs. position i, as shown in Fig. 5. Clearly, smaller fragment

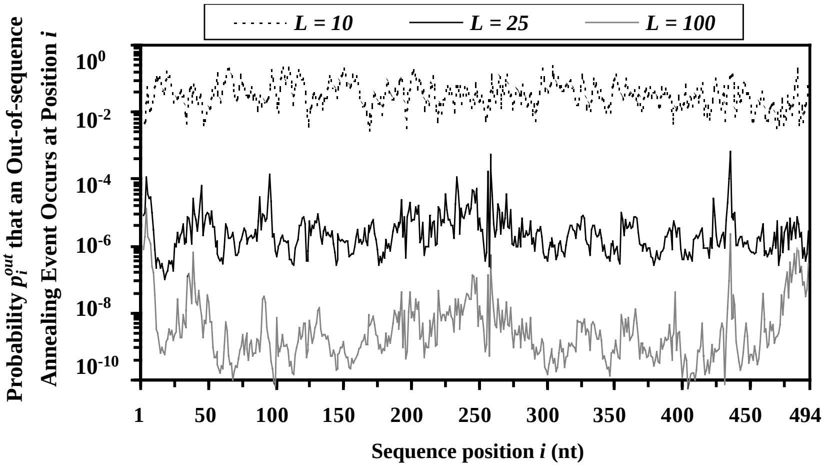


FIG. 5. The probability of an out-of-sequence annealing event as a function of position for $L = 10$, 25 and 100-nt. Out-of-sequence annealing becomes much more likely as fragment length decreases.

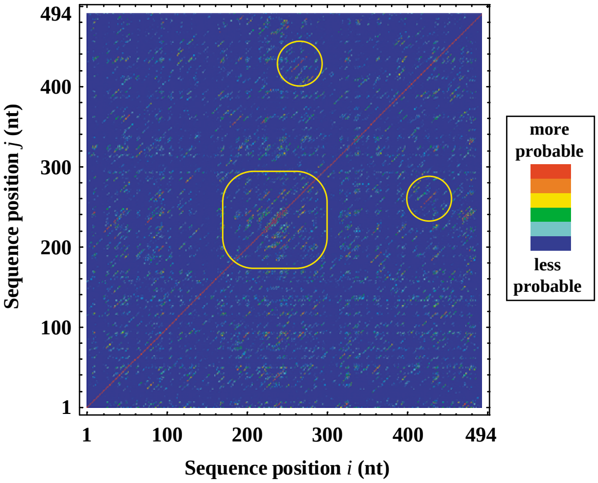
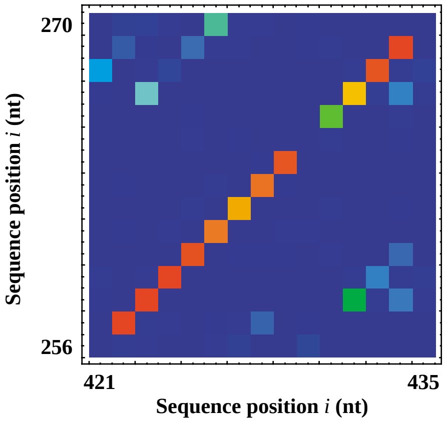
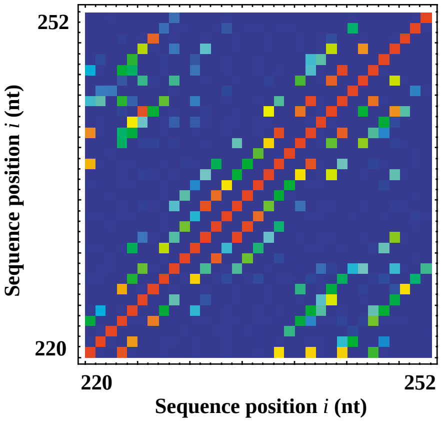


FIG. 6. Density plot of P_{ij} for the subtilisin E/BPN′ system for $L = 25$-nt and $T_{ann} = 55°C$.



```
Subtilisin E     421   GGAGGCACTTACGGC   435
Subtilisin BPN'  256   GAAGGCACTTCCGGC   270
                       * ******** ****
```

FIG. 7. A close-up of Fig. 6 focusing on $i = 421$–435 (subtilisin E) and $j = 256$–270 (subtilisin BPN′). These two DNA motifs have a sequence identity of 87%, causing a tendency for them to anneal out-of-sequence.



```
SerSerGlyIleValValAlaAlaAlaAlaGly

TCCAGCGGTATCGTCGTTGCTGCCGCAGCCGGA

GCATCCGGCGTCGTAGTCGTTGCGGCAGCCGGT

AlaSerGlyValValValValAlaAlaAlaGly
```

FIG. 8. A close-up of Fig. 6 focusing on sequence positions 220–252. A number of "near" codon repeats beginning with guanine are present (boxed with gray), causing out-of-sequence shifts by multiples of 3-nt.

lengths increase the frequency of out-of-sequence annealing events by many orders of magnitude. This occurs because small fragments are more likely to share significant identity with some other portion of the sequence than large ones. In addition, the low melting temperatures of small fragments implies that those that match perfectly with the template are less able to dominate the annealing competition since the melting temperatures of out-of-sequence annealing events have been reached. However, despite the great differences in magnitude for the three curves, the pattern of peaks and valleys remains fairly consistent. This indicates that a portion of gene sequence that shares identity with other parts of the gene sequence retains these annealing tendencies regardless of fragment length. Even though the sequence positions where out-of-sequence reassembly is most likely to occur are identified in Fig. 5, information as to what pairs of sequence positions are likely to anneal out-of-sequence is not provided. For example, one may question whether the two largest peaks found at positions 258 and 435 correspond to complementary out-of-sequence events. It is the probability matrix P_{ij} that contains the answer to this question.

Values of P_{ij} for the subtilisin E/BPN′ system for $L = 25$-nt and $T_{ann} = 55°\mathrm{C}$ are plotted in Fig. 6 in the form of a density plot (lighter shades represent annealing events that are more likely). The large values found along the main $i = j$ diagonal indicate that fragments annealing in their alignment positions are strongly favored during reassembly. A number of other interesting features can be observed in the space off of the diagonal. Most prominent are the short 45° diagonal lines running parallel to the $i = j$ line. These represent sequence regions that have sufficiently high sequence identity to cause out-of-sequence annealing events. For instance, the connection between positions 258 and 435 can be verified by examining the graph near both $(258, 435)$ and $(435, 258)$. A short-line segment is found at both these points, and a closer examination of the sequence reveals that this occurs because subtilisin BPN′, between positions 256–270, and subtilisin E, between positions 421 and 435, have an 87% sequence identity (see Fig. 7).

Another interesting feature is the multicolor area found very near the $i = j$ line at positions 220–252 indicating a series of "near" codon repeats. As shown in Fig. 8, a number of valine, alanine, and glycine residues are found in this range. These repeated residues enhance out-of-sequence annealing events because their codons have a guanine in the first position. The multiple guanines align to produce out-of-sequence annealing events, demonstrated by the separation of the streaks by multiples of 3-nt. In addition, the GC content of the region is over 66%, encouraging non-specific annealing since free energy penalties for regions that are GC rich are less severe.

Next, the effect of changing the annealing temperature on the total probability of reassembling out-of-sequence products with missing or repetitive regions is studied. In Fig. 9, six curves ($L = 10, 15, 25, 50, 100$ and a range from 10 to 50-nt) of P^{out} as a function of T_{ann} are plotted. Note that utilizing fragment sizes less than 15-nt for reassembly results in a library in which the majority of the members are out-of-sequence, regardless of the choice of annealing temperature. This result is consistent with experimental evidence.

As in Fig. 5, these results are extremely sensitive to fragment length, and larger fragments ameliorate the formation of reassembled sequences with repetitive or deleted regions. In addition, all curves pass through a minimum that is slightly less than the melting temperature of homoduplexes (perfect identity) for that fragment length. This temperature is a function of fragment length, GC content of the parental sequences and DNA/cation concentrations. Lowering the annealing temperature below the one yielding a minimum increases out-of-sequence reassembly since more annealing alternatives become available. On the other hand, raising the annealing temperature above the one associated with the minimum greatly decreases the total amount of annealing events as well as their specificity leading to more out-of-sequence events. At these higher temperatures, the enthalpic contribution to the free energy change, responsible for annealing specificity, is overwhelmed by the entropic portion. Interestingly, for the fragment length range of 10–50-nt used in the original DNA shuffling protocol (Stemmer, 1994), out-of-sequence annealing events are minimized near 55°C matching the experimentally chosen annealing temperature. Remarkably, the best annealing temperature found *in silico* in terms of mitigating out-of-sequence annealing events is the same as the one employed in the protocol.

Finally, we examine the effect of low sequence identity and multiple parental sequences on the extent of out-of-sequence reassembly. The genes for serine protease D and proteinase K with a much lower nucleotide sequence identity of 47% are considered next. The results for this low-sequence identity pair are contrasted against the high-sequence identity pair by superimposing plots of P^{out} vs. fragment length for an annealing temperature of 55°C in Fig. 10. No significant differences are observed between the two cases despite the dramatic differences in sequence identities. A similar analysis for the family DNA shuffling (Crameri *et al.*, 1998) of all four subtilases also does not reveal any new trends (Fig. 10). These results are not surprising in light of the fact that the great majority of annealing events yield homoduplexes irrespective of the sequence identity and size of the parental sequence set. As annealing temperature is

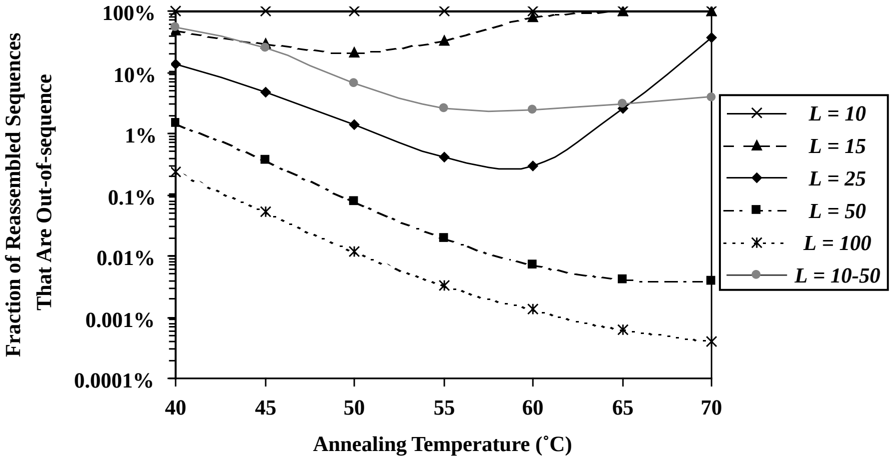


FIG. 9. The fraction of reassembled sequences that are out-of-sequence plotted as a function of annealing temperature for ($L = 10, 15, 25, 50, 100$ and 10 to 50-nt).

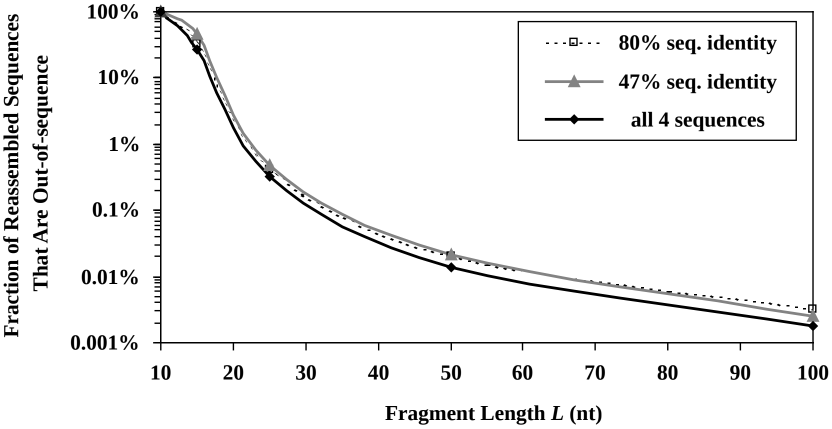


FIG. 10. Extent of out-of-sequence reassembly for subtilisin E/BPN′ (80% sequence identity), serine protease D/proteinase K (47% sequence identity) and the family DNA shuffling of all four sequences.

lowered in a quest for more crossovers, the frequency of heteroduplex formation and out-of-sequence events both increase. Thus, selecting an annealing temperature involves a careful trade-off between the number of crossovers generated and the amount of starting material that is lost to reassembled sequences with missing or repetitive regions.

Conclusion

In this paper, by building on the *e*Shuffle system, a systematic framework is introduced for quantifying the effect of out-of-sequence reassembly in PCR-based directed evolution protocols. As discussed in Moore *et al.* (2001), a potential strategy for boosting the number of crossovers in DNA shuffling combinatorial libraries is to decrease fragment size and annealing temperature. Here we identified *in silico* quantitative limits for the minimum fragmentation size and annealing temperature beyond which out-of-sequence annealing events completely overwhelm the reassembly process, yielding almost entirely out-of-sequence DNA libraries. In addition, we detected specific problematic regions that have a high probability of annealing out-of-sequence. Note that these regions are not always detectable by simple motif finding tools that do not consider the effect of GC stabilization and neighboring nucleotides. Even though no quantitative comparisons with experimental data are available the general prediction trends agree with experimental observations.

We thank Professor Stephen Benkovic and Dr Stefan Lutz for helpful discussions. Financial support from National Science Foundation Award BES0120277 is gratefully acknowledged.

REFERENCES

Allawi, H. & SantaLucia Jr, J. (1997). Thermodynamics and NMR of internal G · T mismatches in DNA. *Biochemistry* **36,** 10581–10594.

Allawi, H. & SantaLucia Jr, J. (1998a). Nearest neighbor thermodynamic parameters for internal G · A mismatches in DNA. *Biochemistry* **37,** 2170–2179.

Allawi, H. & SantaLucia Jr, J. (1998b). Nearest-neighbor thermodynamics of internal A · C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* **37,** 9435–9444.

Allawi, H. & SantaLucia Jr, J. (1998c). Thermodynamics of internal C · T mismatches in DNA. *Nucleic Acids Res.* **26,** 2694–2701.

Crameri, A., Raillard, S., Bermudez, E. & Stemmer, W. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391,** 288–291.

Moore, G. & Maranas, C. (2000). Modeling DNA mutation and recombination for directed evolution experiments. *J. theor. Biol.* **205,** 483–503, doi:10.1006/jtbi.2000.2082.

Moore, G., Maranas, C., Lutz, S. & Benkovic, S. (2001). Predicting crossover generation in DNA shuffling. *Proc. Natl Acad. Sci. U.S.A.* **98,** 3226–3231.

Nakano, S., Fujimoto, M. & Sugimoto, H. H. N. (1999). Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Res.* **27,** 2957–2965.

Ness, J., Welch, M., Giver, L., Bueno, M., Cherry, J., Borchert, T., Stemmer, W. & Minshull, L. (1999). DNA shuffling of subgenomic sequences of subtilisin. *Nature Biotech.* **17,** 893–896.

SantaLucia Jr, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest neighbor thermodynamics. *Proc. Natl Acad. Sci. U.S.A.* **95,** 1460–1465.

Stemmer, W. (1994). DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. U.S.A.* **91,** 10747–10751.