ELSEVIER

Computers and Chemical Engineering 24 (2000) 693–699

Computers & Chemical Engineering

www.elsevier.com/locate/compchemeng

Modeling and optimization of DNA recombination

Gregory L. Moore [a], Costas D. Maranas [a,*], Kevin R. Gutshall [b], Jean E. Brenchley [b]

[a] Department of Chemical Engineering, 112A Fenske Laboratory, The Pennsylvania State University, University Park, PA 16802, USA
[b] Department of Biochemistry and Molecular Biology, 209 S. Frear Building, The Pennsylvania State University, PA 16802, USA

Abstract

This paper discusses predictive models for quantifying the outcome of DNA recombination employed in directed evolution experiments for the generation of novel enzymes. Specifically, predictive models are outlined for (i) tracking the DNA fragment size distribution after random fragmentation and subsequent assembly into genes of full length and (ii) estimating the fraction of the assembled full length sequences matching a given nucleotide target. Based on these quantitative models, optimization formulations are constructed which are aimed at identifying the optimal recombinatory length and parent sequences for maximizing the assembly of a sought after sequence target. Computational results show that the recombination outcome is a 'complex' function of the recombinatory length and recombined sequences and illustrate the magnitude of improvements that can be realized. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Optimization; Bioinformatics; Directed evolution; DNA recombination

1. Introduction and background

DNA recombination techniques provide the backbone of directed evolution experiments for engineering improved proteins and enzymes. These experiments, pioneered by Stemmer (1994) and Arnold (1996), exploit natural selection in a test tube to rapidly 'evolve' enzymes with a desired property or function. Typically, a cycle of directed evolution starts with the construction of a small library of related DNA sequences that encode for enzymes exhibiting the desired property at varying levels. Next, *DNA recombination* is utilized to mix and concatenate the original library DNA sequences. This combinatorially produces a new library, thus increasing the sequence space being considered. Then the expanded library is *screened* for improved enzymes, and the coding sequences for the most greatly enhanced enzymes are *isolated* to form a new sequence library. The experimental cycle of DNA recombination, screening and sequence isolation is repeated with the newly produced DNA sequence library until the enzyme property of interest improves to the desired level. The setup of directed evolution experiments is vital to the rapid and economical production of enhanced enzymes since screening a large number of proteins for the desired property is expensive and time consuming.

Many exciting enzyme enhancements have been produced by utilizing DNA recombination in directed evolution experiments. An outline of the methodology behind these and other successes has been presented by Arnold and Moore (1997). However, despite these success stories, directed evolution experiments have largely been guided by empirical information and experience without a quantitative understanding of the recombination step and subsequent optimization of the experimental setup. The recombination step greatly influences experimental efficiency since it determines the amount of genetic diversity added to the sequence library. Therefore, optimization of the recombination process can potentially lead to a reduction in the number of experimental cycles and significant savings in screening time and costs providing the key motivation for this paper.

Currently, the recombination protocol of choice is DNA shuffling (Stemmer, 1994). A review of other DNA recombination protocols can be found in Volkov and Arnold (1999). In this paper, we focus on DNA shuffling; however, the same modeling principles apply to the other protocols.

* Corresponding author. Tel.: + 1-814-8639958; fax: + 1-814-8657846.

E-mail address: costas@psu.edu (C.D. Maranas).

0098-1354/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-1354(00)00319-7

A flowchart of DNA shuffling is shown in Fig. 1. First an initial set of parent DNA sequences is selected for recombination. The parent sequences undergo *random fragmentation*, typically by DNase I digestion. DNase I is an enzyme that catalyzes nucleotide–nucleotide bond breaking in DNA with no selectivity to nucleotide identity or nucleotide position along the chain. Double-stranded fragments within a particular size range (i.e. 50–200 base pairs) are isolated and *reassembled* by the *polymerase chain reaction* (PCR) without added primers. PCR is a cyclic three-step reaction, utilized for the amplification of small amounts of DNA that typically requires primers (small fragments of single stranded DNA of 15–30 nucleotides in length that are complementary to the ends of the amplified DNA) to proceed (A, C are complementary to T, G, respectively). Primers are unnecessary for the shuffling reaction because the fragments generated by DNase I self-prime each other. However, DNA shuffling requires a cycle of three steps just as PCR does, and this first involves denaturization of the double-stranded fragments into single-stranded ones. Next, pairs of single-stranded fragments anneal along regions overlapping by a sufficiently large number of complementary bases to form overhangs (see Fig. 2).

The term overlap refers to the region where two single-stranded fragments anneal and become double-stranded. The term overhang refers to the single-stranded regions flanking the overlap region that does not align during annealing. The overhangs created are either of type 5′ or 3′ (see Fig. 2). These two different overhang types are caused by the fact that DNA nucleotides and thus strands have an inherent directionality, with the two ends labeled 5′ and 3′, respectively. Thus the overhangs also have directionality and are labeled according to the label of the overhanging end. The third step is the addition of free nucleotides via polymerase extension. Polymerases are enzymes that catalyze the addition of free nucleotides. Polymerase can only fill 5′ overhangs because it has only 5′ → 3′ activity. The three steps are repeated, and each PCR cycle increases the average fragment length. After 20–40 cycles, DNA sequences of the original length are produced.

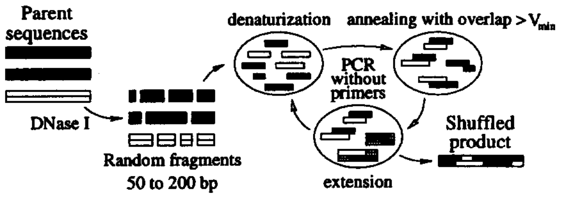

Fig. 1. The three steps of DNA shuffling.

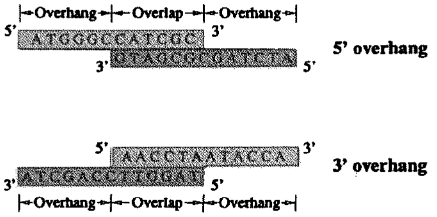

Fig. 2. Regions of overlap and overhang that form when two single-stranded fragments anneal.

The experimental efficiency of DNA shuffling is limited by key unanswered questions regarding the optimal mix and setup of initial sequence libraries 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, quantitative models are presented that describe the shuffling process. Based on this modeling base, optimization formulations are proposed to aim at maximizing the chances of meeting a recombination objective. The remainder of the paper is organized as follows. First, three models describing the DNA shuffling process are summarized. The first, (Random Fragmentation Model), describes the fragment size distribution after DNase I digestion. Given this fragment length distribution, the second model, (Fragment Assembly Model), predicts how the distribution grows for subsequent shuffling cycles. The third, (Sequence Matching Model), estimates the fraction of fully-assembled genes whose nucleotide sequence matches a target one. The first two models are described in detail in Moore and Maranas (1999). Here, two optimization formulations based on the sequence matching model are derived for optimizing the DNA recombination step. The first one is an MILP formulation that simultaneously optimizes both the parent sequence set and the recombinatory fragment length given a set of possible parent sequences. The second model is a bilinear formulation allowing for unequal parent sequence concentrations in the recombination step. An example that illustrates the improvements that can be realized with optimization is addressed throughout the sequence matching model and optimization sections.

2. DNA shuffling models

2.1. Random Fragmentation Model

This model quantifies the distribution of fragment lengths after DNase I digestion of parent sequences with lengths of B nucleotides. The probability P_{cut} of breaking a nucleotide–nucleotide bond is assumed to be constant for all $B-1$ of the bonds present. The fragment length distribution Q^0_L, which describes the fraction of fragments of length. L found in the reaction mixture after fragmentation was calculated to be as follows (see Moore and Maranas (1999) for proof):

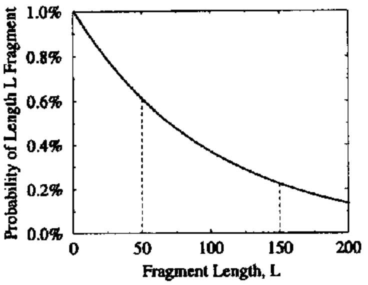

Fig. 3. Fragment length distribution after fragmentation with $P_{\text{cut}} = 1\%$.

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

Note that the distribution is monotonically decreasing (see Fig. 3), so that small fragments always outnumber large ones. This expression implies a mean fragment length of $1/P_{\text{cut}}$. Also, Q_L^0 is only a function of P_{cut} indicating that the gene length B only affects the spread of the final fragment length distribution.

The value of this analysis is that it provides a quantitative way to adjust the fragment size distribution by changing the time and intensity of random fragmentation. The next subsection describes how the original fragment length distribution Q_L^0 changes after each annealing/extension step.

2.2. Fragment Assembly Model

This model tracks the fragment length distribution through a given number of annealing/extension steps. The fragment length distribution after N cycles is denoted by Q_L^N. The model tracks six pathways that result in the formation of fragments of length L. The only parameters in the model are the minimum allowable nucleotide overlap $V_{\min}$ for successful annealing, the probability of failed annealing NA and the exponent α in the annealing probability expression

$$A_{L-V,L} = \frac{V^{\alpha}}{\sum_{v=V_{\min}}^{L-1} v^{\alpha}}$$

which quantifies the probability that a fragment of length L will anneal with an overlap V. Based on this analysis, it is possible to track the average fragment length through any number of shuffling cycles and thus estimate how many shuffling cycles will be needed before full length genes are assembled. These models are discussed in detail and compared to experimental data in Moore and Maranas (1999). However, this paper concentrates on the sequence matching model, which examines the likelihood of assembling a desired shuffling product.

2.3. Sequence Matching Model

This model estimates the fraction of the full length assembled genes that match a desired target sequence (e.g. GTCGGTTC) when the set of parent sequences of length B to be recombined is given. This fraction can also be interpreted as the probability of having a randomly chosen full length sequence, assembled through DNA shuffling, match the given nucleotide sequence target. The objective of this model is to quantitatively predict shuffling results so that mathematical programs can be developed that optimize experimental parameters for a specific sequence goal. To this end, let P_i be the probability of a reassembled sequence matching the target sequence from position i to position B. The goal of this model is therefore to calculate P_1 since this represents assembly of the entire target sequence. The assembly is assumed to begin at position $i = 1$, and additional fragments are assumed to anneal in the direction of increasing position with each addition independent of previous ones. The sequence matching model considers a cascade of four events that must occur for a matching sequence to be produced.

First a fragment of length L is chosen with probability Q_L^0 from the range of fragment lengths retained for shuffling (L_1, L_2). Since a number of fragment lengths can be retained for shuffling, a summation over the range (L_1, L_2) is necessary to include all possible L. Second, the fragment of length L anneals with an overlap V, which can take values from $V_{\min}$ to $L-1$ nucleotides. The multiple possible overlaps necessitate a summation over V within the summation over L. The probability that a fragment of length L will anneal with an overlap V is denoted by $A_{L-V,L}$ as defined earlier.

Wetnur and Davidson (1967) suggest that the exponent $\alpha = -1/2$ which leads to the favoring of smaller overlap values throughout assembly. The probability that governs first fragment addition is adjusted to account for fragment annealing/extension that occurs before position $i = 1$, since the B nucleotides considered are assumed to be only a portion of a much larger sequence (e.g. gene cluster). Based on a Markov Chain analysis, the following expression for $A_{L-V,L}$ at $i = 1$ is obtained. The details of the derivation are described in the Appendix of Moore and Maranas (2000).

$$A_{L-V,L}(i=1) = \frac{\sum_{v=V_{\min}}^{L-1} v^{\alpha}}{\sum_{v=V_{\min}}^{L-1} (L-v)v^{\alpha}}$$

Following the annealing step, the third event that must occur is the matching of the annealed fragment nucleotides with the target sequence. The parameter $\Delta_{i,j}$ is defined as the number of parent sequences that match the target nucleotide sequence from position i to position j. The annealing of a fragment of length L with an overlap of V nucleotides contributes $L-V$ new nucleotides starting from position i and ending with position $i+L-V-1$. The number of parent sequences that match the target between these two positions is $\Delta_{i,i+L-V-1}$, and the probability of choosing a matching parent sequence is therefore $\Delta_{i,i+L-V-1}/K$. Finally, the fourth event is the subsequent addition of the remainder of the sequence, which is assembled with probability P_{i+L-V}. This subsequent addition is assumed to be independent from previous fragment additions, so simple multiplication with the other three terms is needed. This establishes a four-term expression for P_i that must be evaluated recursively. This result is shown below.

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

$$1 \le i < B$$

When only a single fragment length is considered for shuffling, the expression for P_i simplifies to the following:

$$P_i = \begin{cases} 1, & i > B \\ \dfrac{\Delta_{B,B}}{K}, & i = B \\ \displaystyle\sum_{j=1}^{L-V_{\min}} A_{j,L}\left(\frac{\Delta_{i,i+j-1}}{K}\right)P_{i+j}, & 1 \le i < B \end{cases}$$

Here j represents the set of possible extensions a fragment of length L can produce and runs from 1 to $L-V_{\min}$.

The relative proportions of different sequences after DNA shuffling are difficult to predict because the reassembled fragments may include none, one or multiple mutations originating from the parent sequences. This is because mutations do not recombine independently, unless the spacing between mutations is greater than maximum extension length $L-V_{\min}$. In the case of independent recombination, mutations are produced in product sequences at a rate proportional to the fraction of parent sequences containing that mutation. However, for most practical DNA shuffling, the recombinatory fragment length is greater than the mutation spacing. Two terms are used to describe what may occur, *crossover* and *linkage*. A crossover of genetic material occurs when fragments from two different parent sequences anneal and extend. This results in a larger fragment that contains genetic features from both parents. Crossovers occur more frequently when shorter recombinatory fragments are utilized. Linkage occurs when a single fragment contains two or more closely grouped mutations. This can make crossovers in areas of closely spaced mutations infrequent since the mutations tend to remain linked. The sequence matching model takes into account both crossovers and linkage, as shown in the following example.

The following example illustrates the use of the single fragment length sequence matching model. The goal is to shuffle six parent sequences each with a variety of mutations and to produce a sequence containing all 12 of the mutations. The sequences are $B=151$ nucleotides long and are shown in Fig. 4. Parameter values for this example are assumed to be $V_{\min}=5$ and $\alpha=-1/2$. For the case of independent recombination, the probability of producing the target sequence is $(2^{-8})(3^{-2})(6^{-2})=0.0012\%$, a very low success rate. In this example, the most closely spaced mutations are ten nucleotides apart, so utilizing fragments with $L<15$ nucleotides produces independent recombination. Since blocks of mutations are found from positions 0 to 20 and from 130 to 150 along with many closely spaced mutation pairs, larger fragment lengths are intuitively favored to improve the recombination frequency since this will cause the closely spaced mutations to remain linked, as desired. However, if fragments are used that are too large, the crossovers necessary for reassembling the product will occur too infrequently. The commonly used fragment length of 50 nucleotides produces a recombination probability $P_i=0.0090\%$, an improvement of close to ten-fold over independent recombina-

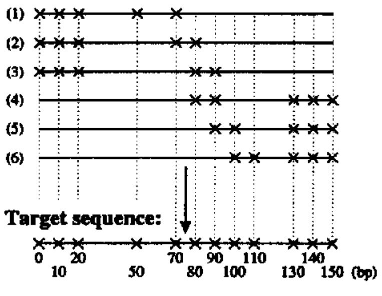

Fig. 4. The six parent sequences and the target sequence utilized in the example.

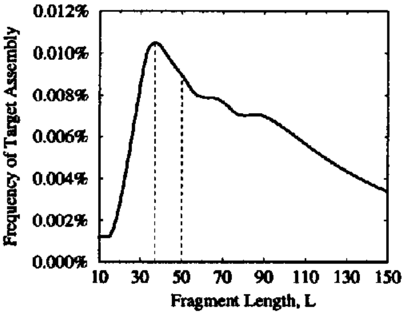

Fig. 5. A plot of recombination probability P_1 versus recombinatory fragment length L when all six parent sequences are utilized.

tion. A plot of recombination probability P_1 versus fragment length L is shown in Fig. 5. The observed optimal fragment length $L = 37$, gives $P_1 = 0.0105\%$, which is a 17% improvement over the $L = 50$ choice. As shown in Fig. 5, the recombination probability is a strong function of fragment length exhibiting a sharp maximum. These results clearly demonstrate the importance of being able to predict this 'best' fragment length.

While for short sequences exhaustive calculation for all possible fragment sizes is feasible this becomes impractical for larger sequences. Mathematical programs for calculating this optimum without resorting to exhaustive calculation are presented in the next section.

3. Optimization framework

The goal of this section is to formulate mathematical programs for optimizing the selection of recombinatory fragment length and parent sequence set. The parent DNA sequences as well as the target sequence are assumed to be known, so the framework of the sequence matching model is utilized. The objective of the desired mathematical program is the maximization of P_1, which defines the probability of matching the target sequence.

First, the binary variable y_k is introduced to represent the inclusion of parent k in the shuffling mixture. For a parent sequence that is selected for shuffling, $y_k = 1$, otherwise $y_k = 0$. The number of parent sequences available for shuffling is denoted by K_{tot}.

Two parameters in the sequence matching model depend on the selection of the parent set. The first, $\Delta_{i,j}$, is now expressed as

$$\Delta_{i,j} = \sum_{k=1}^{K_{\mathrm{tot}}} y_k \Delta_{i,j}^k$$

The parameter $\Delta_{i,j}^k$ equals one if the specific parent sequence k matches the target sequence from positions i to j, and equals zero otherwise. The inclusion of the binary variable y_k in the summation ensures that if a parent sequence is not selected for recombination ($y_k = 0$), it does not contribute to the matching probability value. Second, the parameter K that represents the number of parent sequences being recombined is expressed in a similar manner:

$$K = \sum_{k=1}^{K_{\mathrm{tot}}} y_k$$

The next step is to introduce these two expressions into the single fragment length sequence matching model for $1 \leq i < B$.

$$P_i = \sum_{j=1}^{L-V_{\min}} A_{j,L} \left(\frac{\sum_{k=1}^{K_{\mathrm{tot}}} y_k \Delta_{i,i,+j-1}^k}{\sum_{k=1}^{K_{\mathrm{tot}}} y_k} \right) P_{i+j}$$

Rearranging to eliminate the denominator yields the following:

$$\sum_{k=1}^{K_{\mathrm{tot}}} y_k P_i = \sum_{j=1}^{L-V_{\min}} A_{j,L} \left(\sum_{k=1}^{K_{\mathrm{tot}}} \Delta_{i,i+j-1}^k y_k P_{i+j} \right)$$

This expression contains the nonlinear products $y_k P_i$ and $y_k P_{i+j}$. Since this product consists of a binary variable and a continuous variable, it can be expressed equivalently with two set of linear inequalities (Glover, 1975). Four additional constraints are introduced, and the continuous variable $w_{i,k}$ replaces the product $y_k P_i$.

$$\begin{aligned} 0 \leq \quad & w_{i,k} \leq y_k \\ P_i - K_{\mathrm{tot}}(1 - y_k) \leq \quad & w_{i,k} \leq P_i \end{aligned}$$

$$\sum_{k=1}^{K_{\mathrm{tot}}} w_{i,k} = \sum_{j=1}^{L-V_{\min}} A_{j,L} \left(\sum_{k=1}^{K_{\mathrm{tot}}} \Delta_{i,i+j-1}^k w_{i+j,k} \right)$$

For $y_k = 0$, the above constraints yield $w_{i,k} = 0$ while for $y_k = 1$, $w_{i,k} = P_i$ is recovered.

Next, a binary variable x_L is introduced to represent whether or not a given fragment length is selected for recombination. Since only the single optimal fragment length is desired, only one of the binary variables will be equal to one, resulting in the following constraint.

$$\sum_{L=L_{\min}}^{L_{\max}} x_L = 1$$

where $L_{\min}$ and $L_{\max}$ define the range of fragment lengths being considered. Note that in the single fragment length sequence matching model, the variable L only appears in the summation index. The binary variable x_L is then introduced in the LHS and RHS of the above inequality such that linearity is preserved:

$$\begin{aligned} & K_{\mathrm{tot}}(x_L - 1) \\ & \leq \sum_{k=1}^{K_{\mathrm{tot}}} w_{i,k} - \sum_{j=1}^{L-V_{\min}} A_{j,L} \left(\sum_{k=1}^{K_{\mathrm{tot}}} \Delta_{i,i+j-1}^k w_{i+j,k} \right) \\ & \leq K_{\mathrm{tot}}(1 - x_L) \end{aligned}$$

Note that for $x_L = 1$, the expression reduces to the equality presented in the sequence matching model. For $x_L = 0$, the expression is constrained by the range $[-K_{\mathrm{tot}}, +K_{\mathrm{tot}}]$, effectively inactivating the constraint for that value of L, since the expression contained between the inequalities is bounded within $[-K_{\mathrm{tot}}, +K_{\mathrm{tot}}]$.

Given the above expression, the problem of maximizing P_1 with respect to both x_L and y_k can be posed as the following MILP:

max P_1, subject to;

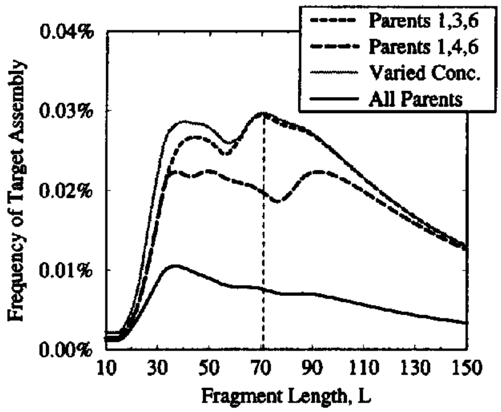

Fig. 6. A plot of recombination probability P_1 versus recombinatory fragment length L for different parent sequence sets.

$$K_{\mathrm{tot}}(x_{L-1}) \le \sum_{k=1}^{K_{\mathrm{tot}}} w_{i,k} - \sum_{j=1}^{L-V_{\min}} A_{j,L}\left(\sum_{k=1}^{K_{\mathrm{tot}}} \Delta^k_{i,i+j-1} w_{i+j,k}\right)$$
$$\le K_{\mathrm{tot}}(1-x_L), \quad 1 \le i < B$$

$$\sum_{k=1}^{K_{\mathrm{tot}}} w_{B,k} = \sum_{k=1}^{K_{\mathrm{tot}}} y_k \Delta^k_{B,B}, \quad i = B; \quad P_i = 1, \quad i > B$$

$$0 \le w_{i,k} \le y_k$$
$$P_i - K_{\mathrm{tot}}(1-y_k) \le w_{i,k} \le P_i$$

$$\sum_{L=L_{\min}}^{L_{\max}} x_L = 1$$

This MILP allows selection of parent sequences that are present in equal concentrations in the recombination mixture. Next an optimization formulation is constructed for selecting the optimal relative concentrations for each of the parent sequences given the recombinatory fragment length.

The continuous variable C_k is defined as the relative concentration (mole fraction) of parent sequence k. The use of relative concentrations implies that the sum of all C_k will be equal to one. Parameters $\Delta_{i,j}$ and K are now expressed as

$$\Delta_{i,j} = \sum_{k=1}^{K_{\mathrm{tot}}} C_k \Delta^k_{i,j} \quad K = \sum_{k=1}^{K_{\mathrm{tot}}} C_k = 1$$

The newly defined parameters can be directly substituted into the single fragment length sequence matching model. This produces the following bilinear NLP formulation which is solved once for each L in the range being considered to find the optimal recombinatory fragment length. Despite the presence of nonconvexities no local optima are observed in the studied examples.

max P_1, subject to:

$$\sum_{k=1}^{K_{\mathrm{tot}}} C_k = 1$$

$$P_i = \begin{cases} 1, & i > B \\ \sum_{k=1}^{K_{\mathrm{tot}}} C_k \Delta^k_{B,B}, & i = B \\ \sum_{j=1}^{L-V_{\min}} A_{j,L}\left(\sum_{k=1}^{K_{\mathrm{tot}}} C_k \Delta^k_{i,i+j-1}\right) P_{i+j}, & 1 \le i < B \end{cases}$$

The six-parent example discussed earlier can now be solved for both the MILP and the bilinear formulations. First, when all parents are selected for recombination (achieved by fixing all $y_k = 1$), the optimal recombination probability P_1 of 0.0105% for a fragment length L of 37 nucleotides is confirmed. However, when the complete MILP is solved for both x_L and y_k, the subset 5 of parent sequences 1, 3 and 6 is revealed to be the optimum recombinatory choice with a recombination probability of 0.0294%, an almost three-fold improvement. Note that the new optimal length is $L = 70$ nucleotides, almost twice the length of the previous optimum implying that the selection of the optimal fragment length strongly depends on the selection of the parent set. A plot of P_1 versus L for different parent sequence recombination sets is shown in Fig. 6. These results suggest a surprising complexity in the shape and form of the P_1 versus L plots for different parent choices. Specifically, the multimodal characteristics of these curves reveal narrow fragment length regions for which favorable recombination results are obtained.

Next, the bilinear formulation is solved, producing the result shown in Fig. 6. The optimal recombination probability is equal to 0.0297% at $L = 71$. The optimal parent sequence concentrations for this fragment length are $C_1 = 0.362$, $C_3 = 0.339$, $C_6 = 0.299$, with all other $C_k = 0$, which are fairly close to the equal relative concentration solution. These results indicate that utilizing these formulations can produce a substantial increase in recombination probability.

4. Summary

In this paper, three predictive models for quantifying DNA recombination were presented. The Random Fragmentation Model and the Fragment Assembly Model provided a method for tracking the fragment length distribution in the reaction mixture during fragmentation and DNA shuffling, respectively. The sequence matching model was established to calculate the probability of producing a prespecified target sequence. Optimization formulations based on the framework of the sequence matching model were then developed to optimize experimental setup parameters for a specific sequence objective. The experimental parameters examined were recombinatory fragment length, parent DNA sequence set and parent DNA sequence concentration. Computational results indicated that by systematically optimizing parent sequence selec-tion and recombinatory fragment length significant improvement on typical experimental can be realized. Specifically, in the example presented, the use of an optimal fragment length for parent sequences at optimal concentrations produced an improvement in recombination probability of approximately three times over the commonly used fragment length of 50 nucleotides.

Nevertheless, DNA recombination processes such as DNA shuffling that require single-strand annealing can be limited in some cases since they require a high degree of sequence similarity in the parent sequence set for successful annealing. The modeling and optimization framework presented is currently being extended to combine DNA shuffling with the new technique of incremental truncation libraries (Ostermeier, Nixon & Benkovic, 1999) to create a new recombination technique that does not depend solely on parent sequence similarity.

Acknowledgements

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

References

Arnold, F. H. (1996). Directed evolution: creating biocatalysts for the future. *Chemical Engineering Science*, *51*, 5091–5102.

Arnold, F. H., & Moore, J. C. (1997). Optimizing industrial enzymes by directed evolution. *Advanced Biochemical Engineering*, *58*, 1–14.

Glover, F. (1975). Improved linear integer programming formulations of nonlinear integer problems. *Management Science*, *22*, 455.

Moore, G. L., & Maranas, C. D. (1999). Modeling DNA mutation and recombination for directed evolution experiments. *Journal of Theoretical Biology*, in press.

Moore, G. L., & Maranas, C. D. (2000). Optimization of DNA recombination protocols. Journal of Comparative Biology, in preparation.

Ostermeier, M., Nixon, A. E., & Benkovic, S. J. (1999). Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorganic & Medicinal Chemistry*, *7*, 2139–2144.

Stemmer, W. P. C. (1994). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 10747–10751.

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

Wetnur, J. G., & Davidson, N. (1967). Kinetics of renaturization of DNA. *Journal of Molecular Biology*, *31*, 349–370.