Structure-based Combinatorial Protein Engineering (SCOPE)

Paul E. O’Maille¹, Marina Bakhtina² and Ming-Daw Tsai¹,²*

Presented here is the development a semi-rational protein engineering approach that uses information from protein structure coupled with established DNA manipulation techniques to design and create multiple crossover libraries from non-homologous genes. The utility of structure-based combinatorial protein engineering (SCOPE) was demonstrated by its application to two distantly related members of the X-family of DNA polymerases: rat DNA polymerase beta (Pol β) and African swine fever virus DNA polymerase X (Pol X). These proteins share similar folds but have low sequence identity, and differ greatly in both size and activity. “Equivalent” subdomain elements of structure were designed on the basis of the tertiary structure of Pol β and the corresponding regions of Pol X were inferred from homology modeling and sequence alignment analysis. Libraries of chimeric genes with up to five crossovers were synthesized in a series of PCR reactions by employing hybrid oligonucleotides that code for variable connections between structural elements. Genetic complementation in Escherichia coli enabled identification of several novel DNA polymerases with enhanced phenotypes. Both the composition of structural elements and the manner in which they were linked were shown to be essential for this property, indicating the importance of these aspects of design.

Keywords: protein engineering; sequence space; chimeragenesis; exon shuffling; X family polymerase

Introduction

The exon shuffling hypothesis asserts that nature constructs proteins by “shuffling” blocks of coding sequence (exons) mediated by recombination between non-coding regions of genes (introns).¹² Although exons generally lack correspondence to a single aspect of protein structure (they can code for autonomous domains or to simpler subdomain elements), there are many instances of the conservation of exon structure between homologous genes in different organisms.

These observations have significantly influenced approaches to structure–function studies of proteins, prompting investigators to make simple chimeras by exchanging equivalent structural elements between homologs. Substituting “modules” between hemoglobin subunits,³ or different domains in terpene cyclases⁴ or transferrin receptors⁵ has been used to evaluate their functional significance. More recently, a “module shuffling” approach has been applied to define the contribution of different structural elements to the kinetic properties of the F/10 family of xylanases.⁶,⁷ The impact of “exon shuffling” on protein engineering strategies is also evident. Using a building-block approach has guided the construction of hybrid enzymes with novel properties. Linking protein modules from N¹⁰-formyltetrahydrofolate hydrolase and glycaminide ribonucleotide transformylase,⁸ subdomain shuffling of factor Xa and trypsin,⁹ and permutation of secondary structure units of barnase to build novel globular proteins illustrate this point.

Our goal was to study the structure–function relationships between two members of the X-family of DNA polymerases: rat DNA polymerase beta (Pol β) and African swine fever virus DNA polymerase X (Pol X). Functionally, these enzymes have substantial differences in nucleotide incorporation efficiencies and fidelities (the inverse...
of the error frequency) as determined by pre-steady-state kinetic analysis.\textsuperscript{12,13} Structurally, they share similar folds but have low amino acid sequence identity. The challenge of creating functional enzymes from proteins with low identity and the absence of combinatorial methods for creating highly mosaic (multiple crossover-containing) chimeras from non-homologous genes, at the time this work was initiated, motivated the development of structure-based combinatorial protein engineering (SCOPE).

On the basis of the concept of exon shuffling, SCOPE utilizes structural information from proteins to design coding segments of genes that correspond to structural elements present in both proteins. In a series of PCR reactions, hybrid oligonucleotides (primers composed of sequence from both parents that code for variable connections between structural elements) act as surrogate introns to direct the assembly of coding segments to create hybrid genes. Iteration of the process enables the synthesis of all desired combinations of structural elements.

The effectiveness of SCOPE was demonstrated by its successful application to a pair of distantly related proteins, Pol X and Pol $\beta$. Given the low sequence identity of these proteins, flexibility was designed into the segment boundaries to accommodate the best fit between structural elements. Use of an established \textit{E. coli} genetic complementation system enabled the identification of chimeric proteins with multiple crossovers and enhanced complementation phenotypes. Our results indicate that the activities of chimeric proteins hinge upon the composition and connectivity of equivalent subdomain elements of structure. Further, the design principles and construction strategies outlined here are readily adaptable to other systems and provide a basis for creation and exploration of global protein space.

### Results

#### The system

Pol $\beta$ is a 39 kDa (335 amino acid residues) protein\textsuperscript{14} that catalyzes gap-filling DNA synthesis as a part of base excision-repair (BER) pathways.\textsuperscript{15} The protein possesses two distinct enzymatic activities: 5'-deoxyribose phosphodiesterase activity contributed by the lyase domain\textsuperscript{16} and DNA-directed nucleotidyltransferase activity supplied by the polymerase domain.\textsuperscript{17} Its structure resembles a right hand and described using the

---

Pol $\beta$ and Pol X. Sequence is colored according to the scheme used for the tertiary structures. The X-family consensus\textsuperscript{24} is underlined. Arrows or blocks above and below the alignment are used to indicate secondary structures of Pol X and Pol $\beta$, respectively.
The assignment of equivalent elements of structure between the proteins of interest can be accomplished using a variety of domain-finding algorithms, the exon structure of the genes, or visual inspection of three-dimensional structures. Here, the tertiary structure of Pol β guided the division of the core X-family polymerase domain into five structural elements of ca. 35 amino acid residues each, and the boundaries between segments were chosen to reside in turns or loops connecting elements of secondary structure. A division was made between palm and fingers, and each of these subdomains was further partitioned into three and two fragments, respectively, as illustrated in Figure 2. In the absence of tertiary structural information, the corresponding structural elements of Pol X were designed based upon sequence alignment information and homology modeling. The interchange of these structural elements produces 32 core X-family polymerase domain ensembles (i.e. five positions occupied by one of two structural elements results in $2^5 = 32$ combinations, including wild-type). The remaining residues at the N terminus (the first 150 amino acid residues of Pol β, which is the 8 kDa domain and thumb subdomain, or the first 14 amino acid residues of Pol X), which bear no structural or functional analogy to one another, constitute the sixth segment and give rise to Pol X and Pol β-sized hybrids, respectively.

As proteins diverge during evolution, insertions and deletions accumulate while the level of sequence identity diminishes. In constructing functional chimeras from a pair of such proteins, as in the case of Pol X and Pol β, these differences give rise to uncertainty in the design of crossovers and concerns about structural element compatibility. To account for this, linkage variability was designed into segment boundaries connecting major elements of secondary structure and regions forming the active site have accumulated while catalytically essential residues and active site geometry remain intact. Therefore, producing functional chimeric enzymes by exchanging subdomain elements of structure between these proteins becomes a challenge.

**Library design**

The assignment of equivalent elements of structure between the proteins of interest can be accomplished using a variety of domain-finding algorithms, the exon structure of the genes, or visual inspection of three-dimensional structures. Here, the tertiary structure of Pol β guided the division of the core X-family polymerase domain into five structural elements of ca. 35 amino acid residues each, and the boundaries between segments were chosen to reside in turns or loops connecting elements of secondary structure. A division was made between palm and fingers, and each of these subdomains was further partitioned into three and two fragments, respectively, as illustrated in Figure 2. In the absence of tertiary structural information, the corresponding structural elements of Pol X were designed based upon sequence alignment information and homology modeling. The interchange of these structural elements produces 32 core X-family polymerase domain ensembles (i.e. five positions occupied by one of two structural elements results in $2^5 = 32$ combinations, including wild-type). The remaining residues at the N terminus (the first 150 amino acid residues of Pol β, which is the 8 kDa domain and thumb subdomain, or the first 14 amino acid residues of Pol X), which bear no structural or functional analogy to one another, constitute the sixth segment and give rise to Pol X and Pol β-sized hybrids, respectively.

As proteins diverge during evolution, insertions and deletions accumulate while the level of sequence identity diminishes. In constructing functional chimeras from a pair of such proteins, as in the case of Pol X and Pol β, these differences give rise to uncertainty in the design of crossovers and concerns about structural element compatibility. To account for this, linkage variability was designed into segment boundaries connecting

**Library design**

The assignment of equivalent elements of structure between the proteins of interest can be accomplished using a variety of domain-finding algorithms, the exon structure of the genes, or visual inspection of three-dimensional structures. Here, the tertiary structure of Pol β guided the division of the core X-family polymerase domain into five structural elements of ca. 35 amino acid residues each, and the boundaries between segments were chosen to reside in turns or loops connecting elements of secondary structure. A division was made between palm and fingers, and each of these subdomains was further partitioned into three and two fragments, respectively, as illustrated in Figure 2. In the absence of tertiary structural information, the corresponding structural elements of Pol X were designed based upon sequence alignment information and homology modeling. The interchange of these structural elements produces 32 core X-family polymerase domain ensembles (i.e. five positions occupied by one of two structural elements results in $2^5 = 32$ combinations, including wild-type). The remaining residues at the N terminus (the first 150 amino acid residues of Pol β, which is the 8 kDa domain and thumb subdomain, or the first 14 amino acid residues of Pol X), which bear no structural or functional analogy to one another, constitute the sixth segment and give rise to Pol X and Pol β-sized hybrids, respectively.

As proteins diverge during evolution, insertions and deletions accumulate while the level of sequence identity diminishes. In constructing functional chimeras from a pair of such proteins, as in the case of Pol X and Pol β, these differences give rise to uncertainty in the design of crossovers and concerns about structural element compatibility. To account for this, linkage variability was designed into segment boundaries connecting

**Library design**

The assignment of equivalent elements of structure between the proteins of interest can be accomplished using a variety of domain-finding algorithms, the exon structure of the genes, or visual inspection of three-dimensional structures. Here, the tertiary structure of Pol β guided the division of the core X-family polymerase domain into five structural elements of ca. 35 amino acid residues each, and the boundaries between segments were chosen to reside in turns or loops connecting elements of secondary structure. A division was made between palm and fingers, and each of these subdomains was further partitioned into three and two fragments, respectively, as illustrated in Figure 2. In the absence of tertiary structural information, the corresponding structural elements of Pol X were designed based upon sequence alignment information and homology modeling. The interchange of these structural elements produces 32 core X-family polymerase domain ensembles (i.e. five positions occupied by one of two structural elements results in $2^5 = 32$ combinations, including wild-type). The remaining residues at the N terminus (the first 150 amino acid residues of Pol β, which is the 8 kDa domain and thumb subdomain, or the first 14 amino acid residues of Pol X), which bear no structural or functional analogy to one another, constitute the sixth segment and give rise to Pol X and Pol β-sized hybrids, respectively.

As proteins diverge during evolution, insertions and deletions accumulate while the level of sequence identity diminishes. In constructing functional chimeras from a pair of such proteins, as in the case of Pol X and Pol β, these differences give rise to uncertainty in the design of crossovers and concerns about structural element compatibility. To account for this, linkage variability was designed into segment boundaries connecting

**Library design**

The assignment of equivalent elements of structure between the proteins of interest can be accomplished using a variety of domain-finding algorithms, the exon structure of the genes, or visual inspection of three-dimensional structures. Here, the tertiary structure of Pol β guided the division of the core X-family polymerase domain into five structural elements of ca. 35 amino acid residues each, and the boundaries between segments were chosen to reside in turns or loops connecting elements of secondary structure. A division was made between palm and fingers, and each of these subdomains was further partitioned into three and two fragments, respectively, as illustrated in Figure 2. In the absence of tertiary structural information, the corresponding structural elements of Pol X were designed based upon sequence alignment information and homology modeling. The interchange of these structural elements produces 32 core X-family polymerase domain ensembles (i.e. five positions occupied by one of two structural elements results in $2^5 = 32$ combinations, including wild-type). The remaining residues at the N terminus (the first 150 amino acid residues of Pol β, which is the 8 kDa domain and thumb subdomain, or the first 14 amino acid residues of Pol X), which bear no structural or functional analogy to one another, constitute the sixth segment and give rise to Pol X and Pol β-sized hybrids, respectively.

As proteins diverge during evolution, insertions and deletions accumulate while the level of sequence identity diminishes. In constructing functional chimeras from a pair of such proteins, as in the case of Pol X and Pol β, these differences give rise to uncertainty in the design of crossovers and concerns about structural element compatibility. To account for this, linkage variability was designed into segment boundaries connecting
structural elements derived from different proteins. This concept and a few examples are described in Figure 3. For this set of experiments, uncorrelated linkage variability was used; the N terminus of a structural element is held constant while the C terminus is allowed to vary in length by seven amino acid residues resulting in ± three amino acid residues expansion or contraction about a chosen zero point at each segment boundary. Therefore, the diversity of the resulting library derives from the composition and variable linkage of structural elements.

Library construction

The idea of breeding molecules in the same manner as plants was the basis for the development of DNA shuffling. The breeding scheme used by Gregor Mendel provided the conceptual basis for this work and the corresponding notation system is used. In principle, this entails “crossing” the parent genes (P) to make an initial population of hybrid genes (with a single crossover) called the first filial generation (F1). The subsequent generation produced by “inbreeding” F1 offspring (which contains an additional crossover) is termed the F2 generation. Further iterations are performed until all possible arrangements of the structural elements are attained.

PCR-based methods have previously been developed and used to construct simple chimeras (one or two crossovers), and here a similar approach was employed and adapted to enable the synthesis of hybrid gene libraries with linkage variability at multiple crossover locations. A three-phase PCR process, indicated by roman numerals as depicted in Figure 4(a), was used to create the F1 generation of hybrid genes. On the basis of the design of structural elements and boundaries, hybrid oligonucleotides were synthesized, which span a given segment boundary and code for each linkage. Their use in independent reactions as downstream primers in conventional PCR produce a set of the seven different hybrid gene fragments (labeled as ±3 to −3) for a given segment boundary (phase I). The primers corresponding to the chosen zero point of fusion for each segment boundary are listed in Table 1, while all others are in Materials and Methods. Each product was verified on and purified from agarose gels. Since these fragments contain sequence derived from the other gene at their 3’ ends, they can serve as forward primers to direct assembly of hybrid genes during primerless PCR (phase II). A final round of conventional PCR (phase III) with the appropriate primer set (depending on the identity of the 5’ and 3’ ends of the crossover product) was used to selectively amplify the nascent hybrid genes. Once confirmed that each gene fragment was synthesized (phase I) and their ability to direct the synthesis of hybrid genes verified (phases II and III), sets of gel-purified hybrid gene fragments or hybrid genes were pooled and used as collections in various combinations for subsequent steps of library construction.

More complex chimeras can be produced from simpler ones by repeating phases II and III, using hybrid genes instead of the Pol X and Pol β, as depicted in Figure 4(b). Alternative approaches achieved the same goal. For example, pairs of F1-hybrid genes with overlapping regions of identical sequence and termini from different parental genes were “crossed” by DNA shuffling or a variation of the technique such as StEP-PCR. In another approach, restriction digestion of one hybrid gene produced fragments that served as primers while another hybrid gene served as template in a round of primerless PCR. In all cases, conventional PCR with the appropriate primer set was used to selectively amplify the nascent crossover product.
all desired structural element ensembles (segment configurations) were produced as shown in Figure 5. Each ensemble was synthesized independently totaling 62 hybrid libraries from the $F_1$ to the $F_5$ generation. Naive (unselected) libraries were analyzed at various stages of synthesis by RFLP and sequence analysis to confirm the successful construction of desired products. Hybrid libraries were independently cloned into the plasmid pTH18cr for selection experiments to give $\sim 10^5$ independent clones each.

An exponential relation defines the numerical complexity of SCOPE libraries in terms of the number and linkage of structural elements:

$$n = l^g$$  

(1)

where $n$ is the number of unique members, $l$ is the linkage variability, and $g$ is the generation ($g = \text{number of crossovers or the number of structural elements} - 1$). For a library with five generations and a seven amino acid residue linkage variability, the most shuffled generation ($F_5$) is the product of five crossovers between six segments of alternating paternal origin resulting in 16,807 (i.e. $7^5$) possible combinations as shown in Figure 5. Combining all generations, parents included, of both Pol X and Pol β-sized hybrids, gives 65,536 members.

### Selection experiments

#### Selection system

Active hybrid polymerases were selected for in an *E. coli* B/r strain SC18-12, which has the recA718 polA12 genotype, which gives rise to a temperature-sensitive phenotype (at high temperature and low cell density, the bacteria are unable to grow on rich media). The loss of Pol I function at elevated temperature renders the bacterium incapable of replicating its genome. However, introduction of a variety of exogenous DNA polymerases such as Pol β, HIV-RT, and Taq Pol I have been demonstrated to complement this deficiency and restore viability. This system has been used extensively as a tool to identify functional DNA polymerases and was employed in this work for in vivo selection of functional chimeric polymerases.

We first evaluated the ability of Pol X to complement. Despite its low activity, Pol X has a weak ability to complement at 37°C and low cell density, but failed to complement at the non-permissive temperature of 42°C (Figure 6). Here, all selection experiments were performed at cell densities of 1000 colonies per 150 mm diameter plate at 42°C as described in Materials and Methods. Under this condition, Pol β transformants retain viability, whereas Pol X or empty vector transformants exhibit less than 0.1% survival (defined as the ratio of the no. of colonies at 42°C to that at 30°C multiplied by 100). At higher cell densities, the survival rate of Pol X and background from empty

---

Controlled synthesis of discrete hybrid populations in this manner gave rise to a library with a hierarchical structure where each consecutive generation contains an additional crossover until
vector transformants increases markedly (data not shown).

Library selection

In this example, the throughput of the selection system is limited by the cell density dependence of the complementation phenomenon. Therefore, selection experiments were performed for individual generations by plating out the number of colonies ca ten times the complexity of each generation (e.g., ca. 34,000 colonies for the F3 generation). Sampling many simple mixtures, where each represents a discrete collection, as opposed to a single complex mixture, increases the probability that each unique member is screened for the same sample size. This result can be formally expressed in equation (2), assuming that the number of copies of a gene in a collection of unique members is not limiting:

$$p(n) = 1 - \sum_{i=1}^{n-1} \frac{(-1)^{i+1}}{i!} \frac{n!}{i!(n-i)!} \left( \frac{(n-i)}{n} \right)^i$$

For $k \geq n$, where $k$ is the sample size and $n$ is the number of unique members defined here by $n = \frac{7^l}{g}$ (equation (1) and Figure 5), the probability ($p$) that a sample of size $k$ contains at least one representative of each unique member can be calculated. For $k$ as a fixed multiple of $n$ ($k = 10n$ for example) as the number of unique members of the library increases, the probability that the sample contains a copy of each gene decreases exponentially. Therefore, the limitations of the selection system and the desired probability of “covering” the library provide an upper limit on its complexity. This in turn can be related to the number of structural elements and the variability in the junctions connecting them in the design of libraries.

Table 1. Hybrid oligonucleotides used for library construction

<table>
<thead>
<tr>
<th>F1 hybrid(^a)</th>
<th>Hybrid oligonucleotides(^b)</th>
<th>aa position(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xβt-0</td>
<td>CTCAATGATGTGTAAGCCCTTTTTTTTAACGCGG</td>
<td>K138/G290</td>
</tr>
<tr>
<td>Xβu-0</td>
<td>GATAGTCTGGATTTACAAAGCCGAGGCAATGTTTGGCCAGGAG</td>
<td>L105/P261</td>
</tr>
<tr>
<td>Xβr-0</td>
<td>AGTATCTGTTAATGAAAAGCCGATGTGGCCAGGAG</td>
<td>R70/R222</td>
</tr>
<tr>
<td>Xβo-0</td>
<td>GCCCTTCGGGAAACTGGGRCAAACAGCAGATTTTGTTGA</td>
<td>G38/G179</td>
</tr>
<tr>
<td>βXt-0</td>
<td>TATCTTCATTACAGTGTGCTAAAGCCGAGGCAATGTTTGGCCAGGAG</td>
<td>L144/P154</td>
</tr>
<tr>
<td>βXu-0</td>
<td>GTATGTTTTCCTCCGGTCTTGGGATCAACCCTGAT</td>
<td>K262/A106</td>
</tr>
<tr>
<td>βXr-0</td>
<td>AGAAGACCTTTATGCGCAGTTTTTTTGACTGTTC</td>
<td>V221/R70</td>
</tr>
<tr>
<td>βXo-0</td>
<td>TTGCGTCTAAAATACCGCAGACGTGAGCGATGTA</td>
<td>C1780/C58</td>
</tr>
<tr>
<td>βXo-0</td>
<td>CAATTTTTTTCTTGTGATAAAGGAGATCTTTTCTAAAA</td>
<td>P151/L4</td>
</tr>
</tbody>
</table>

\(^a\) The name given to an F1 hybrid gene indicates its sequence composition (in going from 5' to 3'), segment boundary, and the point of fusion at the boundary (i.e., hybrid Xβt-0 is composed of Pol X and Pol β sequence at the 5' and 3' ends, fused at segment boundary I at the “zero” point of fusion). Of note, seven points of fusion were designed per segment boundary (+3 to –3) and only the zero point is shown here (the sequences of other primers are listed in Materials and Methods). A schematic representation is shown for each hybrid and Pol X-derived sequence is shaded.

\(^b\) Antisense primers (listed 5' to 3') used in phase I of F1 hybrid synthesis reactions as depicted in Figure 4. Each primer listed codes for the zero point of the specified segment boundary. The additional primers, which code for the other linkages (not shown) differ by a codon at the crossover region. Pol X-derived sequence is shaded.

\(^c\) The amino acid position indicates the location of the zero point (using Pol X (shaded) and Pol β numbering systems) of the segment boundary.

Figure 5. Hierarchical structure of a SCOPE library. Segment configurations are displayed as filled rectangles where black and white boxes represent structural segments derived from Pol X and Pol β, respectively. Each generation is labeled, starting with the parental (P) and descending to the 5th (F5). Complexity of a given generation is defined as the product of the no. of segment configurations and $l^g$, where $l$ is the linkage variability (seven in this case), and $g$ is the generation (or no. of crossovers). For simplicity, only Pol X-sized hybrid libraries are represented by the segment configurations and complexity values shown here.
Iterative rounds of selection were performed for each generation to enrich for the most active polymerases. Isolated plasmid DNA from pooled surviving colonies was retransformed into SC18-12 and selection was repeated under the conditions described in Materials and Methods. After three rounds of selection, the percentage of surviving colonies for the most shuffled Pol X-sized chimeras (generations F3–F5) rose from close to background levels to as high as 86%, whereas no change was observed for the least complex chimeras (F1 and F2) as illustrated in Figure 7(a). For Pol β-sized hybrid libraries, the percentage of surviving colonies during the first round of selection was significantly above background (except for the F5 generation), and by the third round survival reached 90% or greater for the most complex libraries (F2–F5) as illustrated in Figure 7(b). The increase in % survival following multiple rounds of selection indicates that enrichment of the “strongest” complementors occurred. On the basis of the % survival from the first round of selection and the number of unique members in a particular generation, an estimated 1000 Pol β-sized and 100 Pol X-sized complementing hybrids were identified as summarized in Table 2.

The ability of a number of randomly picked individual hybrid polymerases from enriched populations to complement was confirmed by comparison to the parental genes and a representative is shown in Figure 6. Minimal X-family polymerases that are the product of multiple crossovers (from generations F3 and up) were able to confer survival to the SC18-12 strain at 42°C and low cell densities, whereas the same conditions were lethal for Pol X or empty vector transformants. This result indicates that Pol X-sized chimeras created by SCOPE possess an enhanced complementation phenotype relative to wild-type Pol X.

Sequence analysis

Naïve library

Library construction was monitored by RFLP analysis, which indicated that the desired chimeric genes were created and this was confirmed by sequence analysis. A total of 96 independent clones from unselected Pol X-sized hybrid libraries (from the F3, F4, and F5 generations) were sequenced. The amino acid sequences derived from this information were analyzed, and all clones were found to be unique. The intended
crossover locations, and hence the desired composition of structural elements (as designed by the oligonucleotides used for their construction), were observed. Each of the seven variable linkages for a given segment boundary were represented with the exception of +3 at segment boundary I; a hybrid Xf4-5m from selected libraries was found to have this linkage. No occurrence of any

crossover locations, and hence the desired composition of structural elements (as designed by the oligonucleotides used for their construction), were observed. Each of the seven variable linkages for a given segment boundary were represented with the exception of +3 at segment boundary I; a hybrid Xf4-5m from selected libraries was found to have this linkage. No occurrence of any

Table 2. Results of selection experiments on hybrid libraries

<table>
<thead>
<tr>
<th>Generation</th>
<th>Complexity</th>
<th>%Survival</th>
<th>Estimated no. of complementors</th>
<th>%Survival</th>
<th>Estimated no. of complementors</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>35</td>
<td>0.26</td>
<td>1</td>
<td>6.5</td>
<td>2</td>
</tr>
<tr>
<td>F2</td>
<td>480</td>
<td>0.23</td>
<td>1</td>
<td>26.8</td>
<td>132</td>
</tr>
<tr>
<td>F3</td>
<td>3430</td>
<td>0.20</td>
<td>6</td>
<td>15.1</td>
<td>515</td>
</tr>
<tr>
<td>F4</td>
<td>12,005</td>
<td>0.75</td>
<td>90</td>
<td>3.0</td>
<td>360</td>
</tr>
<tr>
<td>F5</td>
<td>16,807</td>
<td>0.11</td>
<td>17</td>
<td>0.17</td>
<td>29</td>
</tr>
</tbody>
</table>

*Generation is defined as the number of crossovers.

b Complexity is the number of unique hybrids for a particular generation as described in Figure 5.

The % survival is defined as the ratio of the no. of colonies at 42 °C to that at 30 °C multiplied by 100.

The number of complementing Pol X or Pol β-sized polymerases per generation is approximated by multiplying the fraction of survivors from the first round of selection (averaged from multiple platings) by the number of unique members of the generation.

Table 3. Segment compositions and frequency of occurrence of complementing hybrids

<table>
<thead>
<tr>
<th>Generation</th>
<th>Hybrid</th>
<th>Segment configuration</th>
<th>Frequency</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>F3</td>
<td>xf3-8g</td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf3-2a</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf3-9a</td>
<td></td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>xf3-9d</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf3-9e</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td>xf4-5m</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf4-6h</td>
<td></td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf4-4l</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf4-7b</td>
<td></td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>xf4-3n</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf4-5n</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>xf5-5e</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf5-5a</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf5-2c</td>
<td></td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>xf5-5c</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xf5-2e</td>
<td></td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Pol X-sized hybrids

| F4         | bf4-1  |                       | 1         |       |
|            | bf4-20 |                       | 1         |       |
|            | bf4-11 |                       | 1         |       |
|            | bf4-13 |                       | 1         |       |
|            | bf4-2  |                       | 2         |       |
|            | bf4-18 |                       | 1         |       |
|            | bf4-16 |                       | 1         | 19    |
|            | bf4-10 |                       | 1         |       |
|            | bf4-8  |                       | 1         |       |
|            | bf4-19 |                       | 1         |       |
|            | bf4-4  |                       | 1         |       |
|            | bf4-17 |                       | 1         |       |
|            | bf4-7  |                       | 4         |       |
|            | bf4-15 |                       | 1         |       |
|            | bf4-12 |                       | 1         |       |
| F3         | bf3-4a |                       | 10        | 10    |
| F2         | bf2-2  |                       | 1         |       |
|            | bf2-8  |                       | 5         | 10    |
|            | bf2-7  |                       | 1         |       |
|            | bf2-7a |                       | 3         |       |

Pol β-sized hybrids

a Hybrids are separated into generations on the basis of the number of crossovers which they contain as described in Figure 5.

b The unique Pol X or Pol β-sized hybrids found in selection experiments are named according to their size and the generation which they belong to (i.e. xf3-8g refers to a Pol X-sized hybrid from the F3 generation and 8g indicates the master plate coordinates of the clone).

c Shading is used to indicate blocks of sequence derived from Pol X as described in Figure 5.

d Of the 60 Pol X and 39 Pol β-sized hybrid clones sequenced, multiple copies of some were found and the number of which is listed.

e Total refers to the number of clones sequenced for a given generation.
sequences found in selected libraries or wild-type genes was discovered. Therefore, the synthesis methods employed are able to produce sufficiently random libraries of the intended composition.

Selected library

Following multiple rounds of selection, plasmid DNA from surviving colonies was isolated and analyzed by sequencing and results are summarized in Table 3. For Pol X-sized hybrid libraries, 16 unique sequences were identified from 60 independent clones and 20 unique Pol β-sized complementors were among the 39 independent clones sequenced. A total of 36 unique sequences out of an estimated 1100 complementing hybrids were obtained, and of the 99 clones analyzed, no wild-type genes were observed. For both Pol X and Pol β-sized libraries, the same combinations of structural elements connected by different linkages were observed among the isolated clones.

In the case of the F$_3$ generation of Pol β-sized hybrids, all ten sequenced clones were identical. Of 23 Pol X-sized hybrid clones from the F$_3$ generation, only two out of a possible ten segment configurations were observed. Taken together, only three out of 20 possible F$_3$ hybrids types were observed, indicating a preference for certain ensembles of structural elements.

The amino acid sequences of the five segment boundaries (segment boundaries I–V) that connect the structural elements (labeled 1–5) of the X-family polymerase domain are shown. The amino acid sequences of unique hybrids, derived from DNA sequencing data, were grouped according to size (Pol X and Pol β-sized hybrids are shown above and below the schematic in the center, respectively) and generation (the number of crossovers increases in going toward the center of the Figure). The names and order of hybrids are the same as those listed in Table 3. Pol X-derived sequence is shaded gray and the corresponding numbering system is shown at the top of the Figure, while Pol β-derived sequence is non-shaded and the corresponding numbering system is shown along the bottom. Dashes are used to indicate gaps in the aligned sequences and crossovers are indicated by transitions from shaded to non-shaded sequence (or vice versa).

![Figure 8](image_url)

**Figure 8.** Sequences of the segment boundaries of complementing hybrids. The amino acid sequences of the five crossover regions (segment boundaries I–V) that connect the structural elements (labeled 1–5) of the X-family polymerase domain are shown. The amino acid sequences of unique hybrids, derived from DNA sequencing data, were grouped according to size (Pol X and Pol β-sized hybrids are shown above and below the schematic in the center, respectively) and generation (the number of crossovers increases in going toward the center of the Figure). The names and order of hybrids are the same as those listed in Table 3. Pol X-derived sequence is shaded gray and the corresponding numbering system is shown at the top of the Figure, while Pol β-derived sequence is non-shaded and the corresponding numbering system is shown along the bottom. Dashes are used to indicate gaps in the aligned sequences and crossovers are indicated by transitions from shaded to non-shaded sequence (or vice versa).
**Discussion**

Our results demonstrate how simple PCR-based techniques, with the guidance of protein structural information, can be adapted into combinatorial approaches to create multiple crossover libraries from non-homologous genes. SCOPE was applied to a pair of distantly related proteins, Pol X and Pol β, and despite the low amino acid sequence identity and differences in 2kDa structures of equivalent structural elements (as shown in Figure 1), highly mosaic hybrid enzymes with in vivo function were produced. Selection experiments suggest that the relative phenotypic strength of hybrids increases with the number of crossovers, but they become increasingly rare in later generations. This is especially evident in the case of Pol X-sized hybrid libraries (Figure 7(a) and Table 2), where strong complementors only appear in the F3 generation or higher.

Pol β has the superior complementation phenotype of the two parents and under the selection conditions used here, Pol X is unable to complement, as evident in Figure 6. The tenfold greater number of Pol β-sized complementing hybrids suggests the thumb subdomain and the 8 kDa domain influence the complementation phenotype of the core X-family polymerase domain. However, the discovery of numerous Pol X-sized hybrids with strong complementation phenotypes indicates that the core X-family polymerase architecture is sufficient for this property. Several chimeras with the same segment configuration but different linkages (Figure 8 and Table 3) are strong complementors, which indicates that there are degenerate solutions to reconstituting a functional enzyme from the same fragments. Only a small fraction of hybrids in a given generation (about one in 1000 F3 Pol X-sized hybrids) have this property, suggesting that the interchangeability of structural elements is critically dependent upon their connectivity.

DNA shuffling and variations on the approach are powerful in vitro recombination methods, but require substantial sequence identity in the genes being shuffled, and are restricted to local regions of sequence space. The conservation of structure and divergence of sequence has motivated the development of homology-independent in vitro recombination methods. Incremental truncation for the creation of hybrid enzymes (ITCHY) and sequence homology-independent recombination (SHIPREC) allow exploration of all fusion points between two genes irrespective of homology to generate products having a single crossover. Both methods employ enzymatic digestion of genes followed by ligation of their fragments, which produces semi-correlated types of linkage variability in the resulting hybrid genes. A liability common to both of these approaches is that only a third of the initial gene-fusion library are in the same reading frame. Shuffling such libraries requires pre-selection, since the number of in-frame hybrids would decrease exponentially with the total number of crossovers. This introduces a bias through the elimination of some in-frame fusion points in the single crossover...
library that may be beneficial in the context of a multiple crossover gene. DNA shuffling has been successfully applied to ITCHY libraries, termed SCRATCHY, to produce genes with up to three crossovers.\(^{51}\) This required careful control of the ratio of the participating ITCHY libraries, generated substantial wild-type background, and no hybrids with improved phenotypes were produced.

Modeling the DNA shuffling process is currently an active area of research.\(^{32-34}\) A recently published analysis of shuffled gene libraries indicates that, as expected, crossovers occur mainly in regions of high sequence identity, which in turn limits the accessible genetic diversity (sequence space) of these approaches.\(^{52}\) As an alternative, SCOPE can be used to create multiple crossover libraries irrespective of the sequence identity between the genes, and the experimenter can determine the number, location, and linkage type of crossovers (Figure 3). At the cost of additional labor in library construction relative to other approaches, an experimentally defined region of sequence space can be synthesized for subsequent exploration.

The possibility of shuffling exons as nature does for directed evolution of enzymes has been previously suggested\(^{56}\) and one group have reported the development of PCR-based approaches that could be applied in such a manner. Their approach, called random multi-recombinant PCR (RM-PCR), takes a block shuffling approach to constructing protein libraries.\(^{57}\) The method consists of a PCR step to amplify a given building block followed by blunt-end ligation of various pairs of them. The desired building block dimers are PCR amplified from isolated plasmids, gel purified, and mixed and assembled into full-length products in a single PCR reaction. This process has successfully generated random shuffling and alternate splicing libraries. The diversity of proposed libraries produced by these methods derive from the number and combination of modules with fixed boundaries. Our results suggest that the activities of chimeras, especially those created from distantly related proteins, are influenced by the connectivity of their structural elements; an element of design neglected by these proposed approaches.

In conclusion, in light of accumulating data from structural biology efforts and the desire to engineer the activities of proteins for commercial, therapeutically, and fundamental scientific pursuits, the SCOPE approach will likely become of increasing utility. Functional proteins are estimated to be sufficiently abundant in sequence space, (occurring \(\sim 1 \text{ in } 10^{11}\)).\(^{56}\) Searching large regions of sequence space using a block-wise or exon shuffling approach with distantly related proteins and an intelligent bias from structural information narrows the search considerably. Current models of evolution predict that the rate-limiting event in the acquisition of new protein folds and functions is the exchange of low energy secondary structures providing a theoretical underpinning.\(^{59}\) One caveat, as it has been suggested, is that proteins made by building-block approaches may require further optimization by exploration of local sequence space by random mutagenesis and DNA shuffling.\(^{11}\) In sum, the controlled construction of chimeras by SCOPE provides a means of exploring global protein space with a knowable probability of searching it.

### Materials and Methods

#### Strains, media, and plasmids

Strain SC18-12, which was used for genetic complementation, was derived from *E. coli* B/r and has the genotype *recA718 polA12 uvrA155 tryEpE65 lon-11 sulA1* and was a gift from Joan Sweasy. DH5-α strain \(\left(\Phi 80d lac Z Y A - m c a l A 1, c o n d A 1, r e c A 1, l h s d 8 1 7 (r s r - m c a l +,\right.\)

#### Sequencing

All sequencing was performed at the Plant-Microbe Genomics Facility (Ohio State University).

#### PCR conditions

All PCR reactions were conducted with Perkin-Elmer GeneAmp PCR system 2400 using a standard set of components (2.5 units *Pfu* Turbo DNA polymerase (Stratagene), 10 μl of 10× cloned *Pfu* reaction buffer, 1 μl 10 mg/ml BSA, and 0.8 μl dNTP mix (25 mM each) with primers at 0.1 μM and 1–10 ng template per 100 μl reaction volume) and cycling conditions (cycling program 95°C for five minutes, followed by 25 cycles of 95°C for 30 seconds, 58°C for 30 seconds, and 72°C (two minutes/kb) followed by seven minutes at 72°C) unless stated otherwise. Pol X and Pol β genes previously cloned into pET17b (Novagen, Inc.) were used as templates for PCR.

All primers were obtained from Integrated DNA Technologies (IDT). Primers coding for Pol X or Pol β N or Ctermini were labeled as forward and reverse, respectively. Their sequences (5′ to 3′) were as follows, Pol β forward: GGGCATTGATGGACACGGAAGGGCCGAGG, Pol X forward: GGGCATTGATGGTTACGGCTTATTCAGGAAAAACATTTGAAATC, and Pol β reverse: GGAGGCCAAAGGAGAGATGAGGTTACC, and Pol X reverse: GGGATTTACGATTGACTTACCTTGAGAGACGCTTTTTTTATATTGACCCGG. The primers that code for the zero point of fusion for the segment boundaries are listed in Table 1 and the remainder of...
Library construction

Genes fragments were released from the plasmid pET17b-Pol β using Pol β forward primer and one of seven hybrid oligonucleotides (βX) as reverse primers for a given segment boundary under standard amplification conditions (phase I). Products were verified on and purified from agarose gels by QIAquick Gel Extraction Kit (Qiagen). About one of a gene fragment and 1 ng of pET17b-Pol X were mixed in a 20 μl primerless PCR reaction using the cycling program 95 °C for five minutes, followed by ten cycles of 95 °C for 30 seconds, 50 °C for 30 seconds, and 72 °C for one minute (phase II). Subsequently, 1 μl of the primerless PCR reaction and Pol β forward and Pol X reverse primer set were mixed and standard amplification conditions enabled the synthesis of full-length F1, hybrid genes (phase III). The same process was used to make Pol X-sized F1 hybrid genes, but using the appropriate templates and primers (phase I; pET17b-Pol X, Pol X forward and X reverse primers; phase II: gene fragments and pET17b-Pol β; phase III: Pol X forward and Pol β reverse primers). Repeating phases II and III using hybrid genes and hybrid genes (collections of seven at ~1 pg each) enabled synthesis of more complex combinations. Alternatively, restriction digestion of hybrid genes produced by gel purification for use in primerless reactions (described above). STEP PCR29 was used for shuffling hybrid genes by mixing ~1 pg of each hybrid gene and the desired primer set under standard reaction conditions using the cycling program 95 °C for five minutes, followed by 80 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds. A further round of standard amplification was performed to achieve a sufficient amount of insert for subsequent cloning.

Ligation reactions and bacterial transformation

All hybrid gene inserts were digested for directional cloning with NdeI/KpnI (New England Biolabs Inc.) using manufacturer suggested conditions followed by QIAquick Gel Extraction prior to cloning into pTH18cr plasmid cut at these sites. A standard set of conditions was used: 10 μl reactions with 1 μl of 10× reaction buffer, 1000 units T4 DNA ligase (New England Biolabs Inc.), 50 ng linearized pTH18cr vector DNA and 100 ng insert DNA at 16 °C overnight. Ligation reactions were conducted by RNAse precipitation technique prior to electrophoresis. Desalting was as follows: 10 μg of tRNA and 20 μl 5 M ammonium acetate was added to a 10 μl ligation reaction, mixed well, followed by addition of 100 μl absolute ethanol. Samples were frozen and centrifuged at >12,000g for 15 minutes at room temperature. The supernatant was decanted and the precipitate washed with 60 μl of 70% (v/v) ethanol. After further centrifugation, the pellet was washed with 10 μl ddH₂O, and typically, 1 μg of DNA was used for every transformation reaction. All transformations were performed with competent DH5α (Stratagene) cells.
conducted with a BioRad Gene Pulser II apparatus using the following instrument settings and parameters: 0.1 cm cuvette gap, 1.8 kV voltage, 25 mF capacitance, 200 Ω (Pulse Controller) resistance, and ~5 ms time constant. Cells are recovered with 0.5–1 ml of SOC media followed by either one hour outgrowth at 37 °C (for DH5-α and BL21 (DE3) [pLysS]) or two hours at 30 °C (for SC18-12) with shaking followed by spreading onto media containing the appropriate antibiotics.

Plasmid library amplification

Library plasmid DNA was prepared by transformation of ligation reactions into DH5-α cells by electroporation (as described above) followed by plating the transformation outgrowths onto LB media containing antibiotics and incubating the plates at 37 °C overnight. The number of independent clones in a library was estimated by counting the number of colonies from serial dilutions. Biomass was harvested in 1 × PBS and the resulting pellet subjected to plasmid miniprep to isolate plasmid DNA.

Selection experiments

Aliquots (50 μl) of SC18-12 cells were transformed by electroporation with 1 μl of ~5 ng/μl of pTH18cr-library plasmid followed by a 500 μl NB recovery and two hours incubation at 30 °C. Serial dilutions are plated onto NAC media and incubated at 30 °C for 18 hours while the remainder of the transformation outgrowth was kept at 4 °C overnight. The titer of transformation is calculated from counting colonies from the serial dilutions and on the basis of which, ~1000 colony-forming units (cfu) were plated onto 150 mm diameter NAC plates and incubated at 42 °C for 20 hours. Survivors were isolated by transferring single colonies to NB-TCI liquid culture and grown to mid-log phase at 30 °C, 37 °C, or 42 °C. Serial dilutions are plated on NCI plates and incubated at 30 °C overnight. The number of independent clones in a library was estimated by counting the number of colonies from serial dilutions. Biomass was harvested in 1 × PBS and the resulting pellet subjected to plasmid miniprep to isolate plasmid DNA.

Rotary streak

Single SC18-12 transformants were transferred to NCI liquid culture and grown to mid-log phase at 30 °C. An A600 of 0.226 corresponds to approximately 2 × 10^8 cfu/ml and ~10^10 cfu in 10 μl was deposited in the center of an NAC plate which was rotated as an inoculation loop was dragged from the center to the perimeter producing a serial dilution on the same plate. Duplicate plates were incubated at 30 °C, 37 °C, or 42 °C for 24 hours.

Atomic coordinates

All hybrid polymerase sequences obtained from selection experiments (as listed in Table 3 and Figure 8) have been deposited to the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database and accession numbers are as follows: AJ439649, b4-1; AJ439650, b4-2; AJ439651, b4-4; AJ439652, b4-7; AJ439653, b4-8; AJ439654, b4-10; AJ439655, b4-41; AJ439656, b4-12; AJ439657, b4-13; AJ439658, b4-15; AJ439659, b4-16; AJ439660, b4-17; AJ439661, b4-18; AJ439662, b4-19; AJ439663, b4-20; AJ439664, b4-3a; AJ439665, b4-2; AJ439666, b4-8; AJ439667, b4-7a; AJ439668, b4-7; AJ439669, x5-2b; AJ439670, x5-5e; AJ439671, x5-3a; AJ439672, x5-2e; AJ439673, x5-5c; AJ439674, x3-9d; AJ439675, x3-9e; AJ439676, x3-8g; AJ439677, x3-2a; AJ439678, x3-9a; AJ439679, x4-6h; AJ439680, x4-7b; AJ439681, x4-4l; AJ439682, x4-5m; AJ439683, x4-3n; AJ439684, x4-5n.

Acknowledgments

We thank Chris Swickard for technical assistance, Alexander Bahtin and Dmitrii Polshakov for consultations, Tam Gotoh for the pTH18cr plasmid, Joan Sweasy for the SC18-12 strain, and Linda Dixon for the plasmid pL ID20 containing ASFV Pol X. This work, was supported by a grant from the National Institutes of Health (GM43268).

References


Edited by A. R. Fersht

(Received 21 March 2002; received in revised form 3 July 2002; accepted 3 July 2002)