INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
COMBINATORIAL PROTEIN ENGINEERING
BY STRUCTURE-BASED GENE SHUFFLING

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
The Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Paul E. ÓMáille, B.S.

* * * * *

The Ohio State University
2001

Dissertation Committee:
Professor Ming-Daw Tsai, Adviser
Professor Ross E. Dalbey
Professor Richard F. Mortensen
Professor Douglas R. Pfeiffer

Approved by

Adviser
The Ohio State Biochemistry Program
ABSTRACT

Structure-based gene shuffling is proposed as a semi-rational method for directed evolution of proteins that enables construction of libraries of complex chimeras from non-homologous genes. Based on the concept of exon shuffling, coding segments of genes are designed to correspond to structural elements of proteins. Oligonucleotide primers are designed to code for segment boundaries and introduce variability between them as they direct the assembly of hybrid genes. Adapting the breeding scheme of Gregor Mendel into a strategy for DNA shuffling, hybrid genes are selectively “crossed” in an iterative process of molecular “inbreeding” until all possible combinations of coding segments are achieved.

This approach has been applied to rat DNA polymerase beta (Pol β) and African swine fever virus DNA polymerase X (Pol X), two proteins with similar folds but low sequence identity, different size, and different activities. Libraries of chimeric genes have been constructed and screened via genetic complementation in E. coli. A number of chimeric Pol X-sized DNA polymerases with enhanced in vivo activity were produced and identified in this manner. Sequence analysis of complementing hybrid polymerases indicates that variability at the boundaries of combined segments is an important element of design.
The design principles developed here outline an approach for the construction of libraries with defined complexity and without the production of wild-type genes for the directed evolution of proteins. Structure-based shuffling allows the generation of libraries with greater "effective" complexity than existing techniques by accessing larger regions of sequence space through shuffling non-homologous genes. The "inbreeding" strategy enables their controlled synthesis resulting in partitioning of the library into simpler mixtures of discrete collections of unique members with the benefit of expanding the "numerical" complexity that can be screened. These approaches are applicable to the design of any chimeric library, and whereas there is never foreknowledge that a variant with a sought after property will be produced, the probability that all variants were evaluated is known.
Dedicated to my Mother and the memory of my Father Tuathal P. ÓMáille
ACKNOWLEDGMENTS

I wish to thank my adviser, Dr. Ming-Daw Tsai, for giving me freedom to do this research, and his constant support, patients, and encouragement throughout it.

I thank my former lab mate, Tom Selby, who provided training in molecular biology techniques and many useful discussions, and my current lab mate, Alex Showalter for discussions, help with kinetic experiments, and careful review of a recently submitted manuscript relating to the work presented here.

I wish to thank Marina Bakhtina who contributed substantial work to this project and Chris Swickard for his technical assistance.

I also wish to thank Alexander Bahtin and Dmitrii Polshakov for their help with probability analysis.

The NIH Chemistry-Biology Interface pre-doctoral training grant provided financial support for me during three years of work. Also, this work was supported by a grant from the National Institutes of Health.
VITA

September 23, 1970..........................Born – Marietta, Ohio, USA

1995..............................................B.S. Chemistry, John Carroll University

1995 – present..................................Graduate Research Assistant, The Ohio State
University

PUBLICATIONS

Research Publication


FIELD OF STUDY

Major Field: Biochemistry
# TABLE OF CONTENTS

| Abstract | ii |
| Dedication | iv |
| Acknowledgments | v |
| Vita | vi |
| List of Tables | x |
| List of Figures | xi |

## Chapters:

1. Introduction........................................................................................................1
   - The Genes........................................................................................................1
   - Genetic Complementation...........................................................................5
   - Library Construction................................................................................7

2. Materials and Methods..................................................................................11
   - Materials.......................................................................................................11
   - Standard PCR Conditions..........................................................................12
   - Ligation Reactions And Bacterial Transformation....................................13
   - Incremental Truncation for the Creation of Hybrid Polymerases..............14
   - Expression and Purification of Pol Xβ1-L...................................................16
   - SDS-PAGE Analysis..................................................................................18
   - Biophysical Analysis................................................................................18
   - Polymerase Kinetics.................................................................................19
   - Rapid-Quench............................................................................................19
   - Structure-Based Gene Shuffling...............................................................20
   - Plasmid Library Amplification..................................................................21
   - Nomenclature System...............................................................................21
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Results of selection experiments from incremental truncation for the creation of hybrid polymerase libraries</td>
</tr>
<tr>
<td>2</td>
<td>The rate of correct vs. incorrect nucleotide incorporation into gapped DNA</td>
</tr>
<tr>
<td>3</td>
<td>Characteristics of structural segments</td>
</tr>
<tr>
<td>4</td>
<td>Results of selection experiments on structure-based shuffling hybrid polymerase libraries</td>
</tr>
<tr>
<td>5</td>
<td>Pol X-sized hybrids recovered from selection experiments</td>
</tr>
<tr>
<td>6</td>
<td>Pol β-sized hybrids recovered from selection experiments</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>11</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td>13</td>
<td>48</td>
</tr>
<tr>
<td>14</td>
<td>54</td>
</tr>
<tr>
<td>15</td>
<td>58</td>
</tr>
<tr>
<td>16</td>
<td>60</td>
</tr>
<tr>
<td>17</td>
<td>63</td>
</tr>
</tbody>
</table>
18 Sequence space subtended by the structure-based shuffling.........................64
19 Iterative reselection of structure-based shuffling libraries..........................67
20 Complementation of Pol X-sized hybrids from the enriched F₄ generation........................................69
21 The junctional preferences of complementing hybrids.................................74
22 Probability as a function of generation and junctional variability...............76
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CD</td>
<td>circular dichroism</td>
</tr>
<tr>
<td>cfu</td>
<td>colony forming units</td>
</tr>
<tr>
<td>CIP</td>
<td>calf-intestine alkaline phosphatase</td>
</tr>
<tr>
<td>DEAE</td>
<td>2-(Diethylamino) ethanol</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>2'-deoxyribonucleoside 5'-triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>2'-deoxycytosine 5'-triphosphate</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediamine tetraacetic acid</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl β-D-thiogalactopyranoside</td>
</tr>
<tr>
<td>ITCHY</td>
<td>incremental truncation for the creation of hybrid enzymes</td>
</tr>
<tr>
<td>MALDI-TOF</td>
<td>matrix-assisted laser desorption ionization time of flight</td>
</tr>
<tr>
<td>NA</td>
<td>nutrient agar</td>
</tr>
<tr>
<td>NB</td>
<td>nutrient broth</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate-buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PMSF</td>
<td>phenylmethylsulfonyl fluoride</td>
</tr>
<tr>
<td>Pol β</td>
<td>DNA polymerase β</td>
</tr>
<tr>
<td>Pol I</td>
<td>DNA polymerase I</td>
</tr>
<tr>
<td>Pol X</td>
<td>DNA polymerase X</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulfate</td>
</tr>
<tr>
<td>StEP</td>
<td>staggered-extension process</td>
</tr>
<tr>
<td>Tris-HCl</td>
<td>2-amino-2-(hydroxymethyl)-1,3-propanediol</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Directed evolution refers to the collection of techniques that endeavor to simulate the natural process of evolution in the laboratory in order to generate biological molecules with altered or enhanced properties. This requires a starting genetic population, a technique to create molecular diversity from which, and a system to select genes with a property of interest. The work presented here describes the directed evolution of African swine fever virus (ASFV) DNA polymerase (PolX) and rat DNA polymerase beta (Pol β) to generate new X-Family polymerases with improved functions. This required the development of new techniques to “shuffle” non-homologous genes, and chimeric enzymes produced by these methods were identified via genetic complementation in *E. coli*.

The Genes

A central reaction in a variety of biological processes is the transfer of a nucleotide to an acceptor hydroxyl. The nucleotidyltransferases that catalyze this reaction belong to more than 10 distinct superfamilies (Aravind and Koonin, 1999), five of which are involved in DNA replication and repair. One such family, the X
Family of DNA-directed DNA polymerases includes two distantly related members: rat DNA polymerase beta (Pol β) and African swine fever virus (ASFV) DNA polymerase X (Pol X).

Pol β is a 39 kDa (335 amino acid) protein (Matsukage et al., 1987) that catalyzes gap-filling DNA synthesis (Weissbach, 1979) and removes 5’-deoxyribose phosphate from apurinic or apyrimidinic (AP) sites in DNA (Sobol et al., 2000) as a part of base excision-repair processes. Its structure, in going from N to C-terminus, is composed of the 8 kDa domain with 5’-deoxyribose phosphodiesterase (lyase) activity, thumb, palm and fingers subdomains as shown in Figure 1. (A) (Arndt et al., 2001). Pol X is a 20 kDa (174 amino acid) protein (Yanez, 1995) that contains only palm and fingers subdomains and is the smallest DNA polymerase known (Oliveros et al., 1997). The structure of Pol X was unknown during the process of this work, but it has now been solved by our lab (Showalter, 2001) and others (Maciejewski, 2001) as shown in Figure 1. (B). Interestingly, detailed kinetic analysis indicated that Pol X is among the most mutagenic of polymerases, with specificity for 5 base pairs (Showalter and Tsai, 2001).

As shown in the structure-based sequence alignment in Figure 2, both proteins contain the DNA polymerase X family consensus sequence G-[SG]-[LFY]-x-R-[GE]-x-(3)-[SGCL]-x-D-[LIVM]-D-[LIVMFY]-(3)-x-(2)-[SAP] (where x is any amino acid, and brackets indicate possible identity at a given position). This motif unifies the X family of nucleotidyltransferases and is shared by all members such as terminal transferase (TdT) from vertebrates, and DNA polymerase IV (Pol IV) from yeast.
Figure 1. Tertiary structures of (A) Pol β and (B) Pol X. Ribbon diagrams depict the crystal structure of Pol β (Arndt et al., 2001) and the solution structure of Pol X (Showalter, 2001). The nucleotidyltransferase domain consists of the fingers and palm subdomains as labeled. The 8 kDa (lyase) domain and thumb subdomain represent the N terminus of Pol β as labeled.
Figure 2. Structure-based sequence alignment of Pol X and Pol β. The amino acid sequences of Pol X and Pol β (from human and rat) are aligned based on their respective structures where solid bars or arrows above and below the alignment indicate α-helix and β-sheet respectively. A rectangular box encompasses the X-Family consensus and shading indicates sequence identity.

The X family consensus forms a structural motif that plays an essential role in nucleotide binding and metal-assisted phosphoryl-transfer. Despite the presence of a similar protein fold and a common amino acid consensus, Pol β and Pol X share only 23% amino acid identity, and have no significant similarity in their gene sequences.

Genetic Complementation

A selection system that can identify functional genes out of a large population is a fundamental requirement for directed evolution studies, and in this case, DNA
polymerization is the property of interest. *E. coli* DNA Pol I catalyzes the joining of discontinuous Okazaki fragments during lagging-strand DNA synthesis in replication (Shlomai et al., 1981) and fills single-stranded gaps produced as a result of DNA repair events (Witkin and George, 1973). Strains of *E. coli* that carry the polA12 mutation, a Pol I temperature-sensitive mutation, can grow as single colonies on rich medium at 42 °C whereas strains that also carry the recA718 mutation cannot (Witkin and Roegner-Maniscalco, 1992). The temperature sensitivity of the double mutant is proposed to result from the deleterious interaction of the two mutations; the polA12 mutation results in persistent gaps in lagging-strand DNA synthesis during replication and the recA718 mutation causes a defect in recombinational repair, an important mechanism used in repairing single-stranded gaps.

An *E. coli* B/r strain, SC18-I2, has a polA12 recA718 genotype and is not viable at elevated temperature and low cell density on rich media and the 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 (Sweasy and Loeb, 1992, Kim and Loeb, 1995, Suzuki et al., 1996). This system has been used extensively as a tool to identify functional DNA polymerases (Sweasy and Loeb, 1993, Kim et al., 1996, Suzuki et al., 1996) and is employed here for *in vivo* selection of functional chimeric polymerases.
Library Construction

Creation of molecular diversity entails construction of sequence space, which is defined as the possible combinations of amino acids for a given sequence length connected by mutational moves (Smith, 1970). So, for a protein of n amino acids in length, there are $20^n$ possibilities. A variety of approaches have been and continue to be developed to create a fraction of these.

Mutation and selection is the primary mode of diversification in asexually reproducing organisms and many random mutagenesis techniques have been developed to mirror this natural process. These approaches range from PCR-based (Leung et al., 1989, Cadwell and Joyce, 1992), and chemical (Kadonaga and Knowles, 1985), to in vivo methods using mutator strains (Greener et al., 1996). Multiple rounds of mutation and selection have been effectively applied to create proteins with new or improved function (You, 1994, Melnikov and Youngman, 1999, Wan et al., 1998). These techniques can create a large number of mutant genes that often exceeds the capacity of the selection system to screen them.

DNA shuffling is an approach designed to mimic the process of homologous recombination that occurs during meiosis in sexually reproducing organisms (Stemmer, 1994). The process is depicted in Figure 3. The basis of the technique resides in the “primerless PCR” reassembly of fragments of the parental genes. Sequence identity shared between the genes allows their fragments to anneal and extension by a DNA polymerase creates “crossover” products. Conventional PCR with primers enables the chimeric genes to be amplified for subsequent selection.
experiments. Iterative rounds of shuffling and selection results in the accumulation of beneficial mutations leading to proteins with enhanced properties.

**DNA shuffling**

- **Family of related sequences**
  - Fragment
- **Pool of random DNA fragments**
  - Reassemble
- **Reassembly of random fragments**
- **Selection of improved variants**
  - Select or screen

**Figure 3.** Overview of DNA shuffling. The process begins with the random fragmentation of a population of related genes with DNaseI. This produces fragments that, after denaturation, hybridize to form an equal mixture of 5' and 3' overhangs. PCR can be used in which a DNA polymerase can “fill in” 5' overhangs. Recombination occurs when a fragment from one gene serves as a template which anneals to a fragment of another that serves as a primer which is extended by a polymerase in the reassembly of a full-length gene. Selection of improved variants can be followed by further iterations of the process.
The process can be extended to families of genes which broadens the range of sequence space that can be sampled and results in acceleration of the directed evolution process (Crameri et al., 1998). "Islands" of homology between the genes being shuffled are required to direct the assembly of their fragment by providing anchor points for crossovers. This homology dependence restricts DNA shuffling to homologous genes, which limits the range of sequence space that can be explored.

A method for combinatorial protein engineering of non-homologous genes, termed incremental truncation for the creation of hybrid enzymes (ITCHY) has been proposed and demonstrated (Ostermeier et al., 1999a, Ostermeier et al., 1999b) and explored in this work. This method allows exploration of all fusion points between two genes irrespective of homology to generate products having a single crossover. It has been proposed that DNA shuffling can be applied to ITCHY libraries to produce genes with multiple crossovers, and recently, the synthesis of chimeras with multiple crossovers from non-homologous genes has been reported (Lutz et al., 2001).

The “Exon Theory of Genes” asserts that the first genes were assembled by recombination between non-coding regions (introns) that link small coding modules (exons) (Gilbert, 1987, Gilbert, 1979, Doolittle, 1978). The process by which exons reassemble as independent elements by intron-mediated recombination is called exon shuffling (Blake, 1979, Gilbert, 1978, Blake, 1983, Blake, 1978). The combinatorial assembly of proteins in this manner solves the numerical paradox of evolution by allowing sparse sampling of large regions of sequence space and is believed to be the dominant force in protein evolution. Rather than taking mutational steps, exon-sized
strides are used to cover greater regions of sequence space, which accelerates the evolution process. This idea has been utilized to produce peptide repertoires (Fisch et al., 1996), and recently it has been suggested that this natural process be mimicked in vitro for the directed evolution of proteins (Kolkman and Stemmer, 2001).

Presented here is an approach called “structure-based gene shuffling”, which is a semi-rational method for the efficient shuffling of non-homologous genes. It’s based on the concept of exon shuffling, but it utilizes structural or sequence information to design segments of genes coding for structural elements present in both proteins. Furthermore, the construction strategy developed here allows control of the shuffling process to produce libraries of desired composition and complexity.

The work presented here demonstrates the application of this approach to the shuffling of two genes with low sequence identity: rat DNA polymerase beta (Pol β) and African swine fever virus DNA polymerase X (Pol X). Libraries of chimeric proteins with multiple crossovers have been generated and several DNA polymerase with enhanced in vivo activities have been identified using genetic complementation.
MATERIALS AND METHODS

Materials

Strain SC18-12, which is used for genetic complementation, is derived from *E. coli* B/r and has the genotype recA718 polA12 uvrA155 trpE65 lon-11 sulA1 and was a gift from Joan Sweasy. DH5-α strain [F-, φ80lacZM15, endA1, recA1, hsdR17 (rK-,mK+), supE44, thi-1, gyrA96, relA1, Δ(lacZYA-argF)U169, λ-] is used for cloning and plasmid maintenance and was purchased from Invitrogen Corporation. The protein expression strain BL21(λDE3)[pLysS] [F′ompT [lon] hsdSB
(rB−mB)] (an *E. coli* B strain) with DE3, a lambda prophage carrying the T7 RNA polymerase gene was purchased from Novagen. All restriction endonucleases and DNA modifying enzymes, with the exception of *Pfu Turbo* DNA polymerase (Stratagene), were purchased from New England Biolabs. All DNA purification kits were obtained from Qiagen. Ultra-pure dNTPs, sephacyrl-100, and DEAE sepharose were purchased from Pharmacia and [*γ-32P*] ATP were obtained from ICN Biomedicals. Talon™ metal affinity resin was purchased from Clontech. Nutrient
agar (NA) or nutrient broth (NB) were purchased from Difco and made according to
directions with the exception that 4 g/L NaCl is added. Luria-Bertani (LB) medium
(10 g/L Tryptone, 5 g/L Yeast extract, and 10 g/L NaCl) is used for culturing DH5-α.
The SC18-12 strain is cultured on or in NA and NB respectively with 4 g/L NaCl and
antibiotics. The pTH18cr plasmid (Hashimoto-Gotoh et al., 2000) which encodes for
chloramphenicol resistance is used for complementation experiments was a gift from
Tam Hashimoto-Gotoh. Pol X and Pol β genes previously cloned (Wernberg et al.,
1996) into pET17b (Novagen, Inc., Madison, WI) were used as templates for PCR.
All primers were obtained from Integrated DNA Technologies (IDT) and the
sequences are listed 5’ to 3’. Sequencing was performed at the Plant-Microbe
Genomics Facility (The Ohio State University). The pCFN-1 plasmid, which encodes
for ampicilin resistance was a gift from Allan Davidson. The plasmid pMS119HE,
which encodes for ampicilin resistance was a gift from the lab of Ross Dalbey. All
other commercially available reagents were of the highest purity.

PCR Conditions

All PCR reactions were conducted with Perkin-Elmer GeneAmp PCR system
2400 using a standard set of components (2.5 U PfuTurbo DNA polymerase, 10 μL of
10x 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 5 min, followed by 25 cycles 95 °C for
30 s, 58 °C for 30 s, and 72 °C (2 min/kb) followed by 7 min at 72 °C) unless stated
otherwise. Primers that encode for Pol X or Pol b N or C-termini are labeled as forward and reverse respectively and their sequences are listed below:

Pol β forward
GGC ATA TGA TGA GCA AAC GCA AGG CGC CGC AGG
Pol X forward
GGC ATA TGA TGT TAA CGC TTA TTC AAG GAA AAA AAA TTG TAA ATC AC
Pol β reverse
GGG AGC CCA AGG ACA GGA GTG AAT GAG GTA CCG G
Pol X reverse
GGA TTT ACG TAT CGC ATA CCT AAG AAA CGT TTA TAA GGT ACC GG

Ligation Reactions And Bacterial Transformation

Unless stated otherwise, all ligation reactions were carried out using a standard set of conditions: 10 µL reactions with 1 µL of 10x reaction buffer, 1,000 U T4 DNA ligase, 50 ng vector DNA and 100 ng insert DNA at 16 °C overnight. Ligation reactions were desalted by tRNA precipitation prior to transformation by electroporation: 10 µg of tRNA and 20 ul 5M ammonium acetate is added to a 10 ul ligation reaction, mixed well, followed by addition of 100 µL absolute ethanol. Samples are frozen with N20, followed by centrifuge at > 12,000 x g for 30 min at 4 °C. The supernatant is decanted followed by washing with 60 µL of 70% ethanol. Centrifuge at > 12,000 x g for 15 min at room temperature. The supernatant is decanted and the pellet is allowed to air dry. Pellets are resuspended in 10 µL ddH2O. Typically, 1 µL is used for every transformation reaction.

All transformations were 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 msec time
constant. Cells are recovered with 0.5 to 1 mL of SOC media followed by 1 hour outgrowth at 37 °C with shaking (for DH5-α and BL21(λDE3) [pLysS]) followed by spreading onto media containing antibiotics.

Incremental Truncation for the Creation of Hybrid Polymerases

Adapted from the published method (Ostermeier et al., 1999b) as follows:

About 5 μg pM119HE-Pol X plasmid was linearized with KpnI and SnaBI (2 μL BSA, 2 μL SnaBI, 2 μL KpnI 10 μL NEB 10x buffer 1 and DNA and ddH2O to 100 μL at 37 °C for 2 hours). DNA was purified by ethanol precipitation followed by resuspension with 120 μL solution (8.8 μL 10x Exo III buffer and 111.2 μL ddH2O). The solution was equilibrated to 25 °C prior to the addition of 5 μL Exonuclease III (500 Units). The reaction was initiated at time = 0 and 20 μL aliquots were removed at 45 second intervals and added to Mung Bean solution (4 μL 10x Mung Bean Nuclease buffer and 16 μL ddH2O) and frozen on dry ice. This resulted in deletions from the C-terminal coding region of Pol X of increasing size. Once all time point are collected, tubes are warmed to 65 °C to heat inactivate for 10 minutes followed by chilling on ice for 5 minutes. To each tube, 0.6 μL of Mung Bean Nuclease (6 units) is added and incubated at 30 °C for 30 minutes. DNA is isolated with QIAGen nucleotide removal kit and incubated with SacI and calf-intestine alkaline phosphatase (CIP) (1 μL BSA, 1 μL SacI, 1 μL CIP, 5 μL NEB 10x buffer 1 and DNA and ddH2O to 50 μL at 37 °C for 2 hours). DNA is isolated with QIAGen nucleotide removal kit, quantitated by agarose gel electrophoresis and ligated with C-
terminal fragments of Pol β, which were produced by standard PCR amplification using Pol β reverse primer and 1 of the following Pol β forward primers:

- **B-G290**
  
  \[ \text{GGCTTCACAATCAATGAGTACACCATCCGCCCC} \]

- **B-I243**
  
  \[ \text{CCCCCAAGATCAGTACTACTGTGTTCTCCTACCTCAGG} \]

- **B-H222**
  
  \[ \text{CCGTTTCATTACAGATACTCGTCAAAAAGGGTGAGACAAAG} \]

- **B-A185**
  
  \[ \text{GCAGAGTCAGGGAGATATGGACCGTGCGACCC} \]

- **B-G144**
  
  \[ \text{GAGGACCTTTGAAAGAGAAATTCTCGTGAGGAGATGCTGC} \]

Amplification with the primers above produced a series of fragments that differ by ~100 bp each (from 100 to 500 bp). Deletion products from the Exonuclease III reactions were ligated to fragments of Pol β which correspond to the approximated size of deletion.

**Pol X forward primers:**

- **X-N134**
  
  \[ \text{GAATTATAAGCTAAATCAGTATGGATTATTTAAAAATCAAAACTTTAGTACC} \]

- **X-A106**
  
  \[ \text{GCCGAGGAAAAAACATACGCAATATTTCTTTACCGG} \]

- **X-I71**
  
  \[ \text{GCATAAAAGGTTCTTTTCCTGTAAAAAGTCTGCGG} \]

- **X-I34**
  
  \[ \text{CATCGTTGCTTTGGAATGAAAAGCAGCGGAAGAG} \]

- **X-MI**
  
  \[ \text{ATGATGTTAAGCTTTATCCAAGGAAAAATTGTAAATCAC} \]

The same process was performed to make the reciprocal libraries with the exception that SstI instead of SnaBI was used to linearize pMS119HE-Pol β plasmid. The hybrid genes produced in this manner were labeled Xβₙ or βXₙ, where n refers to the approximate deletion size divided by 100 (from ~100 to 500 bp) and the identity of the N and C-terminus written left to right respectively. Therefore, an Xβ₁ hybrid has
the N-terminus of Pol X, which lacks a thumb and 8 kDa domain, giving rise to Pol X-size protein, the C-terminal end of which is derived from Pol β as illustrated in Figure 4.

**Figure 4.** Schematic representation of libraries produced by ITCHY. As described, the C-terminus is incrementally deleted and the corresponding coding region from the other gene is ligated in to replace it. This process, as performed in ~100 bp increments produces βX and Xβ libraries shown from N to C-terminus. Black or white designates Pol X or Pol β-derived sequence respectively.

**Expression and Purification of Pol Xβ1-1**

Hybrid Xβ1-1 cDNA was PCR amplified, NdeI and KpnI digested, and ligated into plasmid pCFN-1 using the same restriction sites and standard conditions described above. A single BL21(λDE3) [pLysS] transformant was grown in 50 mL LB culture containing ampicilin (100 μg/ml) and chloramphenicol (35 μg/ml) for 12 hours at 37 °C with shaking. A 1:100 dilution into 1 L LB with antibiotics was performed and cells were grown at 37 °C with shaking until OD600 reached 0.6, at which time IPTG was added to 1 mM final concentration. Growth was allowed to continue for 5 hours and cells were harvested by centrifugation with Beckman JA-10
rotor at 5,000 rpm for 10 minutes.

The cell pellet was resuspended on ice in 160 ml sonication buffer (50 mM NaH$_2$PO$_4$, 10 mM Tri-HCl, and 100 mM NaCl equilibrated to pH = 8 at 4 °C) to which PMSF is added to give 100 μM final concentration. Sonication was performed with a Fisher Scientific 550 Sonic dismembrator for 35 seconds at 70 % power with 5 minute intervals was performed 4 times. The lysate was clarified by centrifugation in a JA-14 rotor at 12,000 rpm at 4 °C for 30 minutes.

About 150 mL of the resulting lysate was mixed with 7.5 mL Talon beads and incubated at 4 °C for 4 hours with gentle agitation. The mixture was placed in a 2.5 x 30 cm column and washed with 4 bed volumes of sonication buffer. Protein was eluted by adding 4 bed volumes of sonication buffer with imidazole added to 100 mM final concentration and 1 mL fractions collected and analyzed by SDS-PAGE. Fractions that contain the correct-sized protein are pooled and passed through 12.5 mL DEAE sepharose equilibrated with buffer x (50 mM Tri-HCl, 100 mM KCl, and 1 mM DTT equilibrated to pH = 8 at 4 °C) and 50 mL flow through is collected and loaded onto a column containing 30 mL of Phosphocellulose resin (P-11) equilibrated with buffer x. Protein is eluted with a KCl gradient of 100 mM to 1.4 M over 150 mL with 5 mL fractions collected. The absorbance at 280 nm of fractions was checked followed by SDS-PAGE analysis. Protein-containing fractions were pooled and concentrated using Biomax-5K NMWL concentrators (Millipore) according to manufacturers directions and separated further by gel filtration on an S-100 column equilibrated with buffer X. Protein-containing fractions were pooled, concentrated as
above, and mixed with an equal volume of glycerol a frozen with N₂0 and stored at -80 °C.

**SDS-PAGE Analysis**

All protein samples were analyzed using a PhastSystem apparatus (Amersham Pharmacia Biotech) and pre-cast 20 % SDS-PAGE gels. Samples were prepared for electrophoresis by boiling for 5 minutes in loading dye (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 2 % SDS, 5 % β-mercaptoethanol, and 0.001 % (w:v) bromophenol blue).

**Biophysical Analysis**

All UV-Vis measurements of protein, DNA, and bacterial cultures were performed using a Beckman DU 650 spectrophotometer. Circular dichroism was conducted at the Analytical Spectroscopy Laboratory (Department of Chemistry, The Ohio State University) using an AVIV Model 202 Circular Dichroism Spectrometer. Protein samples were dialyzed into 50 mM Borate, 50 mM KCl, 1 mM DTT, pH 8 at 4 °C and were at a concentration of 20 μM as determined by UV. Mass Spectrometry of proteins was conducted at the CCIC-Mass Spectrometry Facility (The Ohio State University) using a Reflex MALDI-TOF instrument (Bruker).
Polymerase Kinetics

The rate of single nucleotide incorporation was measured by incubating “gapped G-X” $\gamma^{32P}$-labeled DNA substrate with enzyme and then rapidly mixing in Mg$^{2+}$ and dNTP. Reactions are terminated at different times and PAGE analysis enables resolution of 25 mer (substrate) from 26 mer (product) for determination of the rate of polymerization. Gapped G-X substrate:

5′-gcctgcagccgtcaaccaactca cctcgatccaatgccgtcc-3′
3′-cggagcgtcggcagttggtggtggagctagttacggcagg-5′

Rapid-Quench

Solution I (200 nM gapped G-X substrate, 200 µg/mL BSA, 2 mM DTT and 500 nM Pol Xβ1 in 180 µL Assay buffer) and Solution II (10 mM MgCl$_2$ and 2 mM dCTP in 180 µL Assay buffer) are rapidly mixed with an RQF-3 apparatus for reaction times of 5 milliseconds to 20 seconds (KinTek Instruments Corp., State College, PA). The reaction is initiated by mixing of the 2 solutions and stopped at various time points by further mixing with a quenching solution (formamide). Following the reaction, 5 µL of each time point is resolved on a 15 % PAGE (29:1 acrylamide to bisacrylamide) of 50 cm x 38 cm x .4 mm dimensions at 50 °C, 120 V, for 4 hours. The region of the gel containing radioactive product was excised with a razor blade, wrapped with suran-wrap and exposed to a phosphorscreen overnight. A Storm$^\text{®}$ phosphorimager (Molecular Dynamics) and Image QuaNT software was used to analyze results.
Structure-Based Gene Shuffling

Gene fragments were amplified from pET17b-Pol β using Pol β forward primer and a set of 7 hybrid reverse primers coding for the junctions of a given segment boundary using standard amplification conditions. Each reaction produced gene fragment sets that differed in length by a coding segment and were purified by QIAquick Gel Extraction Kit. About 1 ng of a gene fragment set and 1 ng of pET17b-Pol X were mixed in a 20 μL primerless PCR reaction using the cycling program 95 °C for 5 min, followed by 10 cycles 95 °C for 30 s, 50 °C for 30 s, and 72 °C for 1 min. Pol β forward primer and Pol X reverse primer set and 1 μL of the primerless PCR reaction enabled the synthesis of full-length hybrid genes by standard amplification conditions. StEP PCR (Zhao et al., 1998) was used for shuffling hybrid genes by mixing ~1 pg each hybrid gene and the desired primer set under standard reaction conditions using the cycling program 96 °C for 5 min, followed by 80 cycles 95 °C for 30 s, 55 °C for 5 s. A further round of standard amplification was performed and following digestion with NdeI and KpnI products are purified by QIAquick Gel Extraction Kit for cloning into pTH18cr plasmid or used for further shuffling. Alternatively, gene fragments and hybrid genes were mixed and primerless PCR followed by standard amplification with a set of primers that selects for crossover products was performed. The same procedure was carried out to make Pol X gene fragment sets and hybrids genes.
Plasmid Library Amplification

Library plasmid DNA was prepared by transformation of ligation reactions into DH5-α 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.

Nomenclature System

The boundaries between coding segments have a variability of 7 amino acids and an individual reverse primer codes for each of these 7 junctions; a “zero” point of fusion was chosen and ± 3 amino acid insertion or deletion variability is added as shown below for the βX1 boundary:

There are a total of 6 structural segments (named N, 5, 4, 3, 2, and 1 in going from N to C terminus) and the reverse primers that code for the junctions of the 5 boundaries between them are numbered 5 through 1 in going from N to C terminus. Primers that code for a junction at the boundary of segments are labeled to indicate the boundary number, the origin of gene segments N and C-terminal to the junction, and the variability. For example, the βX1-1(+) primer codes for segment boundary 1.
(indicated by the subscript 1) with Pol β and Pol X being N and C-terminal to the boundary respectively (written left to right for N and C-terminus), and the point of fusion being +1 (indicated by the number 1 after the dash followed by “+” in parentheses).

**Hybrid Reverse Primers:**

\[
\begin{align*}
Xβ_{1-3}(-) &\quad CTC\ ATT\ GAT\ TGT\ GAA\ GCC\ TAA\ CGC\ GGC\ TCG\ AAT\ TCT \\
Xβ_{1-2}(-) &\quad CTC\ ATT\ GAT\ TGT\ GAA\ GCC\ TAA\ CGC\ GGC\ TCG\ AAT\ TCT \\
Xβ_{1-1}(-) &\quad CTC\ ATT\ GAT\ TGT\ GAA\ GCC\ TTT\ TTT\ TAA\ CGC\ GGC \\
Xβ_{1-0} &\quad CTC\ ATT\ GAT\ TGT\ GAA\ GCC\ CTT\ TTT\ TTT\ TAA\ CGC\ GGC \\
Xβ_{1+1} &\quad CTC\ ATT\ GAT\ TCT\ GAA\ GCC\ ATT\ TTT\ TTT\ TAA\ CGC \\
Xβ_{1+2} &\quad CTC\ ATT\ GAT\ TCT\ GAA\ GCC\ ATA\ ATT\ TTT\ TTT\ TAA \\
Xβ_{1+3} &\quad CTC\ ATT\ GAT\ TGT\ GAA\ GCC\ ATA\ ATT\ TTT\ TTT\ TAA \\
Xβ_{2-3}(-) &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ AAA\ AAG\ ATC\ AAG\ TTG\ ATA \\
Xβ_{2-2}(-) &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ CGT\ AAA\ AAG\ ATC\ AAG\ TTG \\
Xβ_{2-1}(-) &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ AGC\ CGT\ AAA\ AAG\ ATC\ AAG \\
Xβ_{2-0} &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ TAA\ AGC\ CGT\ AAA\ AAG\ ATC \\
Xβ_{2+1} &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ GGC\ TAA\ AGC\ CGT\ AAA\ AAG \\
Xβ_{2+2} &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ CTC\ GGC\ TAA\ AGC\ CGT \\
Xβ_{2+3} &\quad GTA\ GTA\ CTG\ ATC\ TTT\ GGG\ TTC\ TTT\ CGC\ GGC\ TAA\ AGC \\
Xβ_{3-3}(-) &\quad AGT\ ATC\ TGT\ AAT\ GAA\ ACG\ GGG\ CAG\ GAC\ GTG\ TTT\ TAA\ AAG \\
Xβ_{3-2}(-) &\quad AGT\ ATC\ TGT\ AAT\ GAA\ ACG\ GTT\ GGG\ CAG\ GAC\ GTG\ TTT \\
Xβ_{3-1}(-) &\quad AGT\ ATC\ TGT\ AAT\ GAA\ ACG\ AAT\ GTT\ GGG\ CAG\ GAC\ GTG \\
Xβ_{3-0} &\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \}
AGT ATC TGT AAT GAA ACG GCG AAT GTT GGG CAG GAC
XB₃-1(+)
AGT ATC TGT AAT GAA ACG TAT GCG AAT GTT GGG CAG
XB₃-2(+)
AGT ATC TGT AAT GAA ACG CTT TAT GCG AAT GTT GGG
XB₃-3(+)
AGT ATC TGT AAT GAA ACG ACC CTT TAT GCG AAT GTT G
XB₃-1(-)
GCC TCT TCG GAA ACT GCC AAT ACG GAT GTT TTT TGA TAA
XB₂-2(-)
GCC TCT TCG GAA ACT GCC AGC AAT GAT GTT TTT TGA
XB₂-1(-)
GCC TCT TCG GAA ACT GCC AAT ACG AAC GAT GTT TTT TG
XB₂-1(+)n
GCC TCT TCG GAA ACT GCC AAT ACG AAC AAC GAT G
XB₂-2(+)
GCC TCT TCG GAA ACT GCC TAA ACT ACC AAG AAC AAG
XB₂-3(+)n
TCT CTT TTC AAA GTC TAC AAT TTT TTT TCC TTG
XB₂-2(-)
TCT CTT TTC AAA GTC CTC ATT TAC AAT TTT TTT TCC
XB₂-1(-)
TCT CTT TTC AAA GTC CTC GTG ATT TAC AAT TTT TTT
XB₂-0
TCT CTT TTC AAA GTC CTC TAA GTG ATT TAC AAT TTT
XB₂-1(+)
TCT CTT TTC AAA GTC CTC ACG TAA GTG ATT TAC AAT
XB₂-2(+)
TCT CTT TTC AAA GTC CTC GGA ACG TAA GTG ATT TAC
XB₂-3(+)
TCT CTT TTC AAA GTC CTC TCG GGA AGC TAA GTG ATT
βX₁-3(-)
AAT ATG AGA GCC CAT GCC AAT TAT AAG CTA AAT CAG
βX₁-2(-)
ATG AGA GCC CAT GCC CTG AAT TAT AAG CTA AAT CAG
βX₁-1(-)
AGA GCC CAT GCC CTG GAA AAG CTA AAT CAG
βX₁-0
GCC CAT GCC CTG GAA AAG AAT TAT AAG CTA AAT CAG
βX₁-1(+)n
CAT GCC CTG GAA AAG GGG AAT TAT AAG CTA AAT CAG
βX₁-2(+)
GCC CTG GAA AAG GGC TTC AAT TAT AAG CTA AAT CAG
βX₁-3(+)n
23
CTG GAA AAG GGC TTC ACA AAT TAT AAG CTA AAT CAG
βX2-3(-)
GTA TGG TTT TTC CTC GGC CAA CCT GAT ATC GAT TCT
βX2-2(-)
GTA TGG TTT TTC CTC GGC GAT CAA CCT GAT ATC GAT
βX2-1(-)
GTA TGG TTT TTC CTC GGC GGG GAT CAA CCT GAT ATC
βX2-0
GTA TGG TTT TTC CTC GGC TTT GGG GAT CAA CCT GAT
βX2-1(+)
GTA TGG TTT TTC CTC GGC ATC TTT GGG GAT CAA CCT
βX2-2(+)
GTA TGG TTT TTC CTC GGC CTG ATC TTT GGG GAT CAA
βX2-3(+)
AGA AAG ACC CTT TAT GCG TAA CTG TTC CAC AAC ACG
βX3-2(-)
AGA AAG ACC CTT TAT GCG TTG TAA CTG TTC CAC AAC
βX3-1(-)
AGA AAG ACC CTT TAT GCG TTT TTG TAA CTG TTC CAC
βX3-0
AGA AAG ACC CTT TAT GCG GAC TTT TTG TAA CTG TTC
βX3-1(+)
AGA AAG ACC CTT TAT GCG ACG GAC TTT TTG TAA CTG
βX3-2(+)
AGA AAG ACC CTT TAT GCG GAA ACG GAC TTT TTG TAA
βX3-3(+)
TTC GCG TTT TTC CTT GAATAAGCT TTT CAA AGT CCT CAA A
βX5-3(-)
CAA TTT TTC CTT GAA TAA GCT TTT CAA AGT CCT CAA A
βX5-2(-)
CAA TTT TTC CTT GAA TAA GTC TCT TTT CAA AGT CCT C
βX5-1(-)
Selection Experiments

Fifty μL aliquots of SC18-12 cells are transformed by electroporation with 1 μL of 5 ng/μL of pTH18cr-library plasmid followed by a 500 μL NB recovery and 2 hr incubation at 30 °C. Serial dilutions are plated onto NA\textsubscript{TCl} (TCl = 12.5 μg/mL tetracycline, 35 μg/mL chloramphenicol, and 1 mM IPTG) and incubated at 30 °C for 18 hrs while the remainder of the transformation outgrowth is kept at 4 °C overnight. The titre of transformation is calculated from counting colonies from the serial dilutions and based on which, ~1000 cfu are plated onto 150 mm NA\textsubscript{TCl} plates and incubated at 42 °C for 20 hrs. Survivors were isolated by transferring single colonies to NB\textsubscript{TCl} media and culturing ~16 hours at 37 °C followed by QIAquick plasmid miniprep for purification of pTH18-hybrid plasmid DNA. Reselection was performed by harvesting all survivors by suspension with 1x PBS, recovering plasmid with Qiagen QIAspin miniprep kit, and retransforming and replating.

Complementation Assay

Single SC18-12 transformants were transferred to NB\textsubscript{TCl} liquid culture and grown to mid-log phase at 30 °C. Cultures were serially diluted with 1x PBS and
plated in duplicate. Following a 24-hour incubation at 30 °C or 42 °C colonies are counted and complementation calculated. Complementation is defined as the percentage of surviving colonies at the non-permissive temperature (colonies at 42 °C/colonies 30 °C) x 100. All trials are performed at least in triplicate.

**Rotary Streak**

Single SC18-12 transformants were transferred to NB_{TCI} liquid culture and grown to mid-log phase at 30 °C. An OD_{600} of 0.226 corresponds to approximately 2 x 10^8 cfu/mL and ~10^4 cfu in 10 μL was deposited in the center of an NA_{TCI} 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 or 37 °C for 24 hours.
CHAPTER 3

GENETIC COMPLEMENTATION

Introduction

Genetic complementation is used as a tool to select for functional polymerases created by the protein engineering techniques developed in this work. The survival of the *E. coli* strain SC18-12 under conditions of high temperature and low cell density on rich media provides the basis for selection as described in Chapter 1. *E. coli* Pol I is required in the initiation of replication at the ColE1 origin (Witkin and Roegner-Maniscalco, 1992) which is present in most commonly used vectors. The plasmid pTH18cr, which is a derivative of pHSG576, has a pSC101 origin of replication that is Pol I-independent, a feature that is required for these experiments (Hashimoto-Gotoh et al., 2000). The cDNA insert in this vector is under regulation of the *lac* promoter rendering protein expression inducible by addition of isopropyl β-D-thiogalactopyranoside (IPTG) to the growth medium.
Complementation by Pol X and Pol β

Initially, the ability of the parental polymerases to complement as a function of cell density and temperature was investigated and compared. This was done qualitatively by the rotary streak method, and quantitatively by the complementation assay as described in Chapter 2.

Rotary Streak

This technique allows examination of the cell density dependence of complementation for a given temperature on the same plate. In brief, the SC18-12 strain was transformed with pTH18cr plasmid that contained Pol β, Pol X, or no insert. Single colonies were transferred to liquid culture containing the appropriate antibiotics and IPTG and were allowed to grow to mid-log phase at 30 °C. Based on OD<sub>600</sub> measurements, serial dilutions were made and ca. 10<sup>4</sup> cells were placed in the middle of a rich media plate containing antibiotics and IPTG, followed by streaking to the perimeter as the plate is spun on a turn-table which produces serial dilutions in going from the center to the edge. This was performed in duplicate and plates were placed at 30 °C or 37 °C. Results are shown in Figure 5.
Figure 5. Rotary streak assay for genetic complementation. Approximately $10^4$ SC18-12 transformants are deposited in the center of a NA plate in duplicate and diluted as sterile loop streaks from the center to the perimeter followed by 24-hour incubation at the indicated temperature. Survival of bacteria at the edge of the 37 °C plate constitutes complementation.
Complementation Assay

In brief, mid-log phase cultures were grown as described and serial dilution were plated in duplicate onto rich media containing antibiotics and IPTG and incubated at 30 °C and 37 °C or 42 °C. Following a 24-hour incubation, colonies were counted and the % survival, which is the # of colonies (colony forming units or cfu) at the non-permissive temperature (37 °C or 42 °C) divided by the # of colonies at the permissive temperature (30 °C) multiplied by 100, was determined. Plotting the % survival versus cell density gives a quantitative measure of complementation at the non-permissive temperature of 37 °C or 42 °C as shown in Figure 6. (A) and (B) respectively.
Figure 6. Complementation assay of Pol X and Pol β. The parental polymerases were tested for their ability to complement SC18-12 as a function of cell density and the non-permissive temperature of (A) 37 °C and (B) 42 °C. Transformants were cultured in the presence of IPTG to mid-log phase followed by plating serial dilutions onto NA_{TCT} plates in duplicate and incubated at 30 °C and the indicated non-permissive temperature.
Conclusion

The results of the complementation studies can be summarized as follows: According to the rotary streak assay, both Pol β and Pol X are capable of complementation at 37 °C and low cell density on rich media as apparent from growth at the plate perimeter. According to the quantitative complementation assay, Pol X has a weak ability to complement, which is consistently around 30 to 40 % at the lowest cell densities tested, whereas Pol β is able to fully complement at all cell densities at 37 °C. When the temperature is shifted to 42 °C, Pol β maintains its ability to complement at all densities examined, whereas Pol X transformants become inviable at densities below 500 cfu/100 mm plate. Therefore, Pol X possesses weak complementation ability, and under certain conditions (on rich media at ≤ 500 cfu/100 mm plate and 42 °C), is incapable of conferring survival to the E. coli strain SC18-12.
CHAPTER 4

INCREMENTAL TRUNCATION FOR THE CREATION OF HYBRID POLYMERASES

Introduction

DNA shuffling is limited to genes that share significant homology for efficient recombination to occur. As a result, the location of crossovers in the resulting chimeric genes is restricted to regions of high sequence similarity between the genes being shuffled. The goal of this work is to evolve new X Family polymerases with altered and enhanced properties from Pol X and Pol β proteins for structure-function analyses. Since the genes encoding for these proteins have no significant similarity, development of an alternative strategy for their chimeragenesis was required. As described in the introduction, the incremental truncation strategy is not restricted by a sequence homology requirement and a variation of the technique was explored here.

Library Design

The C-terminal coding region of one gene was incrementally truncated by nuclease digestion and replaced by the corresponding coding region of the other gene.
that was amplified by conventional PCR. Gene fragments produced by PCR ranged in size from 100 to 500 bp with 100 bp increments. The nuclease digestion reactions were terminated at time points corresponding to deletion products in 100 bp increments. Ligation of gene fragments and deletion products produces a series of hybrid libraries. Use of PCR rather than nuclease digestion to generate the C-terminal fragments is a deviation from the published method, which enabled the partitioning of libraries into groups based on the extent of substitution as depicted in Figure 4 and described in Chapter 2.

Library Construction

The protocol for the synthesis of incremental truncation libraries is described in detail in Chapter 2 and depicted in Figure 7. In brief, the pMS119HE plasmid containing Pol X or Pol β was linearized with KpnI and SnaBI (Styl for Pol β), which cut at the 3’ end of the gene. Exonuclease III treatment selectively digests the SnaBI (or Styl) end (5’ overhang) whereas the KpnI end (3’ overhang) is not susceptible. The digestion reaction is quenched at various time points and Mung Bean Nuclease was used to remove the remaining single-stranded DNA for blunt end cloning. Further, the truncated pMS119HE-Pol X plasmid and Pol β gene fragments digested with SacI for directional cloning. The hybrid genes were amplified by PCR using Pol X forward and Pol β reverse primer set and product of the anticipated size was produced and cloned into pTHI8cr for complementation experiments. Libraries were constructed and amplified individually to give ~10,000 individual clones each.
Figure 7. Incremental truncation for the creation of hybrid polymerases. Construction occurred in 2 phases. In phase 1, a plasmid containing Pol X or Pol β (pMS119HE labeled here as pMS-pol) is linearized with 2 restriction enzymes, one cuts internally at the 3' (C-terminal-coding) end of the gene and has a 5' overhang (Styl or SnaBI) and the other cuts adjacent to, but down stream of the gene to give a 3' overhang (KpnI). Exonuclease III digestion degrades the 5' overhang end and hence the gene producing a set of deletion products when terminated at different time points during the reaction. In phase 2, C-terminal gene fragments of the other gene are amplified by conventional PCR. The fragments and digestion products are subsequently paired and ligated.
**Diagram Description:**

- **pMS-pol** is shown with a circular structure.
- **Styl or SmaI** site is marked as (1).
- **Kpnl** site is marked as (1).
- **SacI** site is marked as (2).

**Exonuclease Digestion:**

- Digestion with ExoIII/Mung Bean Nuclease leads to the following pieces:
  - **Δ100 bp**
  - **Δ200 bp**
  - **Δ300 bp**
  - **Δ400 bp**
  - **Δ500 bp**

**PCR Reaction:**

- PCR products after digestion are shown as follows:
  - 100 bp
  - 200 bp
  - 300 bp
  - 400 bp
  - 500 bp
Library Screening

Initially, selection experiments were attempted with Pol X-sized hybrids on rich media at ≤ 500 cfu/100 mm plate and 42 °C but no hybrids were isolated that could reproducibly confer survival under these conditions. Selection was performed on individual libraries of pTH18cr-XP1, 2, and 3 or βX1, 2, and 3 hybrid plasmids and as described in Chapter 2 with the exception that conditions of 37 °C and ≤ 200 cfu/100 mm plate were used. Approximately 1,000 transformants were plated out for each library and the average results of selection experiments are summarized in Table 1.

Several surviving colonies from selection experiments were cultured and their plasmid isolated. The reproducibility of their complementation behavior was assessed by retransformation of the SC18-12 strain and retesting by the rotary streak method. A few examples are shown in Figure 8. Reproducibly complementing hybrids identified in this manner were further analyzed for expression and solubility in *E. coli* for purification and *in vitro* characterization.
<table>
<thead>
<tr>
<th>Plasmid</th>
<th>cfu @ 30 °C</th>
<th>cfu @ 37 °C</th>
<th>% Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>pTH18cr</td>
<td>120</td>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>pTH18-X</td>
<td>37</td>
<td>15</td>
<td>40.5</td>
</tr>
<tr>
<td>pTH18-Xβ₁</td>
<td>65</td>
<td>8</td>
<td>12.3</td>
</tr>
<tr>
<td>pTH18-Xβ₂</td>
<td>43</td>
<td>5</td>
<td>11.6</td>
</tr>
<tr>
<td>pTH18-Xβ₃</td>
<td>180</td>
<td>8</td>
<td>4.4</td>
</tr>
<tr>
<td>pTH18-β</td>
<td>74</td>
<td>75</td>
<td>98.6</td>
</tr>
<tr>
<td>pTH18-βX₁</td>
<td>34</td>
<td>20</td>
<td>58.8</td>
</tr>
<tr>
<td>pTH18-βX₂</td>
<td>276</td>
<td>116</td>
<td>42</td>
</tr>
<tr>
<td>pTH18-βX₃</td>
<td>111</td>
<td>15</td>
<td>13.5</td>
</tr>
</tbody>
</table>

*Each value presented is the average of 5 trials.

**Table 1.** Results of selection experiments from incremental truncation for the creation of hybrid polymerase libraries
Figure 8. Rotary streak of Xβ hybrid polymerases. Approximately $10^4$ SC18-12 transformants with Pol X-sized hybrid polymerases are deposited in the center of a NA_{TCT} plate in duplicate and diluted as sterile loop streaks from the center to the perimeter followed by 24-hour incubation at the indicated temperature. Survival of bacteria at the edge of the 37 °C plate constitutes complementation.
Expression and Purification

A total of 12 hybrid proteins, 6 hybrids of Pol X size and 6 of Pol β sized (2 from each class; $X_{β_{1,2,3}}$ and $β_{X_{1,2,3}}$ respectively) were cloned into pET17b for overexpression in the E. coli host BL21(λDE3)[pLysS]. Detectable levels of expression were observed for 1 Pol b-sized and 2 Pol X-sized hybrids, but no soluble protein was apparent. Expression of a Pol X-sized hybrid, $X_{β_{1,1}}$, as a fusion protein with an N-terminal 6-Histidine affinity and flag epitope tags from the pCFN-1 plasmid produced soluble protein as shown in Figure 9. The amino acid sequence of Pol $X_{β_{1,1}}$ is presented below and consists of the N-terminal 135 amino acids of Pol X and the C-terminal 46 amino acids of Pol β (underlined) giving rise to a 21 kDa (181 amino acid) protein:

```
MLTLIQGKKTVNHLRSRLAFEYNGQLKILSKNIVAVGSLRREEKMLNDV
DLLIVPEKLLLKVLPNIRIKLGSFSVKVCGERKCVLFIWEKKTYQLD
LFTALAEKPYAIFHTGPVSYLIRIRAALKKNYGFTINEYATRPLGVT
GVAEGPLPVDEQDIFDYIQWRYREPDRSE
```

Purification was attempted and achieved by using affinity (Talon resin), anion-exchange (DEAE sepharose), cation exchange (phosphocellulose P-11), and size exclusion (gel filtration) chromatography as described in Chapter 2.
Figure 9. Expression and solubility of a Pol X-sized hybrid Xβ₁₋₁. A hybrid polymerase expressed in BL21(λDE3)[pLysS] from the plasmid pCFN-1 has an N-terminal 6-histidine and FLAG epitope tag. Samples were analyzed on a 20 % SDS-PAGE gel: Lane 1, total cell lysate; 2, clarified lysate following sonication; 3, Pol X marker; 4, no induction; 5, 2 hours post-induction; 6, 5 hours post-induction. The arrow points to the band which corresponds to the Xβ₁₋₁ fusion protein.

Affinity chromatography with soluble native protein using Talon™ resin resulted in enrichment of target protein, but a substantial amount of contaminant proteins co-eluted as apparent in Figure 10. Consecutive anion and cation-exchange chromatography with DEAE and P-11 respectively, removed most all other proteins as seen in Figure 11. Xβ₁₋₁ eluted from the P-11 column at about 760 mM KCl whereas Pol X elutes at about 1 M. The last step was gel filtration, which yielded approximately 150 μg of greater than 95 % pure protein as estimated by UV and SDS-PAGE analysis.
Figure 10. XB1-1 affinity chromatography. Six-histidine affinity resin (Talon\textsuperscript{TM}) was used to capture 6-his-tagged protein. Binding and washing was followed by elution with 100 mM imidazole and 1 mL fractions were collected and analyzed on a 20 % SDS-PAGE gel. Lanes 6 and 21, Pol X; lanes 1 through 5, 7 through 20, and 22 through 24 are the 1 mL fractions collected during the elution. The arrow points to the band which corresponds to the XB1-1 fusion protein.
Figure 11. Ion exchange chromatography of Xβ₁₋₁. The eluent collected from affinity chromatography was passed through a DEAE (anion exchange) column and bound to a P-11 (cation exchange) followed by elution with a KCl gradient. Shown in (A) is the elution profile from the P-11 column. A 20 % SDS-PAGE gel (B) was used to analyze fractions and lane 1, fraction 16; 2, fraction 17; 3, fraction 18; 4, the sample loaded on the P-11 column; 5, fraction 20; 6, fraction 21; and 7, fraction 22. The arrow points to the band which corresponds to the Xβ₁₋₁ fusion protein.
Biophysical Analysis

The mass of purified protein was determined by mass spectrometry using a MALDI-TOF instrument and was within 1% of the mass predicted from the amino acid sequence. The secondary structure content was verified by circular dichroism (CD) and found to possess both α-helical and β-sheet content and representative spectra is shown in Figure 12.

Figure 12. Circular dichroism of 6-His-FLAG-tagged-Pol Xβ1-1. The optical rotation produced by a solution of purified fusion protein was obtained with an AVIV model 202 circular dichroism spectrometer. Ellipticity (rotation of plane polarized light) as a function of wavelength produces a plot that is consistent with α-helical and β-sheet character.
Kinetic Characterization

The rate of single nucleotide incorporation of a correct and incorrect Watson-Crick base pair was measured using rapid-quench and manual quench techniques respectively. As seen in Figure 13. (A), incorporation of cytosine opposite guanine in a gapped DNA substrate occurs on the timescale of milliseconds and when the formation of product (26 mer) is plotted versus time, a “burst” phase or single exponential is observed Figure 13. (B).
Figure 13. Single nucleotide incorporation by Pol Xβ1-1. Rapid quench (on the millisecond scale) or manual quench (on the second to minute scale) for the correct or incorrect nucleotide incorporation event was monitored by autoradiography of $\gamma^{32}$P-ATP-labeled DNA on the left and right of panel (A) respectively. (B) A single exponential curve was derived from Image QuaNT densitometry measurements of the relative intensity of bands.
A.

G:C  G:G

26 mer  
25 mer

(msec)  (sec)

B.

Fitting parameters:
\[ R^2 = 0.998 \]
\[ A = 94 \pm 2 \text{ nM} \]
\[ k_{\text{on}} = 16 \pm 1 \text{ s}^{-1} \]
The rate of incorporation is compared to Pol X and Pol β, and remarkably, despite being composed largely of the former has catalytic activity more consistent with the latter. Results are compared to the parental polymerases using the gapped G-X substrate a described in Chapter 2. In the base-pair notation, X:Y, X refers to the templating base, and Y refers to the incoming nucleotide. Results are summarized in Table 2.

<table>
<thead>
<tr>
<th>Polymerase</th>
<th>$k_{obs}$ G:C (s$^{-1}$)</th>
<th>$k_{obs}$ G:G (s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pol X</td>
<td>0.19</td>
<td>0.029</td>
</tr>
<tr>
<td>Pol β</td>
<td>12.5</td>
<td>0.0073</td>
</tr>
<tr>
<td>Pol Xβ₁₋₁</td>
<td>16 ± 1</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

Table 2. The rate of correct vs. incorrect nucleotide incorporation into gapped DNA.

The same result for the G:C pair was consistently obtained in 3 separate experiments using 2 independent enzyme preparations. When mis-incorporation was analyzed using the G-G mispair, no incorporation is observed during the time scale of seconds to minutes for any enzyme preparation. More extensive kinetic analysis was attempted but the enzyme proved to be unstable: whereas 2 independent purifications produced active protein, complete loss of activity occurred during storage in 50 % glycerol at -20 °C and -80 °C over the course of days. Further purifications and
alternative storage conditions such as temperature, additives (BSA and glycerol), and lyophilization failed to produce active protein.

Conclusion

The hybrid protein Pol Xβ₁₋₁, found in preliminary complementation experiments, posses Pol X-like complementation behavior but has Pol β-like catalytic properties based on the limited analysis performed here. The instability of the protein in vitro is a property of Pol Xβ₁₋₁, which precluded its detailed characterization.

The ITCHY method produced hybrid polymerases with in vivo activities less than or equal to the parental polymerases. No hybrid polymerases were discovered that complement at 42 °C and low cell density. It is possible that hybrid polymerases that are the product of multiple crossovers could posses improved activities. Since certain pairs of hybrid genes produced in the process share regions of overlapping identity, they can be shuffled.

Consideration of shuffling of ITCHY libraries gives rise to some practical issues. At every crossover point, there is a 1 in 3 chance that the coding regions appended are in the same reading frame. This means that 1/3^x, where x is the number of crossovers, is the fraction of in-frame shuffled hybrids that would be produced. For a population of genes that are the product of 5 crossovers, only 1 in 243 (i.e. 1/3^5) hybrids are in frame. The selection system used here is restricted to low cell densities which makes the goal of screening these libraries unfeasible. Alternatively, in-frame hybrids from selection experiments could be further shuffled. Inactive hybrids would
be excluded from this population, and whereas the point of fusion is non-functional in that context, it may be beneficial in a more shuffled protein. These considerations motivated the development of an alternative strategy to shuffle non-homologous genes.
CHAPTER 5

COMBINATORIAL PROTEIN ENGINEERING BY STRUCTURE-BASED GENE SHUFFLING

Introduction

Since screening capacities are often exceeded by the ability of various techniques to create diversity, control of library construction and complexity become important determinants for successful directed evolution of proteins. The structure-based gene shuffling approach was developed to allow the controlled synthesis of complex chimeras from non-homologous genes. This approach enables the construction of defined regions of sequence space for subsequent systematic exploration of its functional content.

Design of Structural Elements

The tertiary structure of Pol β, modeling studies of Pol X, and a sequence alignment of the two protein sequences guided division of the fingers and palm subdomains into five structural segments of ~35 amino acids as shown in Figure 14 (A through C).
**Figure 14.** Structural elements of the nucleotidyl-transferase domain. (A). A ribbon diagram of Pol β derived from the crystal structure is colored to illustrate the location of structural segments used for shuffling as indicated by alternating dark and light gray shading (from N to C terminus, the 8 kDa domain and thumb (light gray), palm (dark, light, dark gray) and fingers (light and dark gray) subdomains). (B). A ribbon diagram of Pol X derived from the solution structure using the same shading scheme used for Pol β. (C). Sequence alignment of Pol β and Pol X (using the program Clustal X 1.8 where ‘*’, ‘:’ and ‘.’ below the alignment indicate positions that have ‘fully’, ‘stronger’, or ‘weaker’ conserved residues according to the Gonnet Pam250 Matrix). Sequence is shaded according to the scheme above (sequence corresponding to dark and light gray shading in the structure is depicted by black shading with white letters and light gray shading with black letters respectively) to represent structural segments as defined in Table 3. The X-family consensus is underlined and in bold. The boundaries between segments are indicated above the alignment (5 through I). (D). The Primary structure of segment boundary I is shown to illustrate the nature of the 7 amino acid variability. A "zero" fusion point was chosen, and 3 amino acid expansion and contraction variability relative to which was introduced.
The boundaries of coding segments are chosen to reside in turns or loops connecting elements of secondary structure. The amino acid position, size, and homology of these structural elements are listed in Table 3. The remaining residues at the N-terminus (first 150 amino acids of Pol β which is the thumb and 8 kDa domain or first 14 amino acids of Pol X) constitute the sixth segment (N) and give rise to Pol X or Pol β-sized hybrids, respectively. The N-terminal segment of Pol X bears no structural or functional analogy to the thumb and 8 kDa domain of Pol β and its assignment was a compromise between the modeled structure and the sequence alignment data. Furthermore, additional variability of 1—7 amino acids is introduced at the C-terminus of each structural segment as shown in Figure 14. (D). As a result, the size of a structural element can vary by ±3 amino acids from what is listed in Table 3, which allows for structural elasticity at each junction.
<table>
<thead>
<tr>
<th>Structural Segment</th>
<th>Pol β</th>
<th>Pol X</th>
<th>Homology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Size</td>
<td>Position</td>
</tr>
<tr>
<td>N</td>
<td>1-143</td>
<td>143</td>
<td>1-14</td>
</tr>
<tr>
<td>5</td>
<td>144-178</td>
<td>34</td>
<td>3-37</td>
</tr>
<tr>
<td>4</td>
<td>179-220</td>
<td>42</td>
<td>38-69</td>
</tr>
<tr>
<td>3</td>
<td>221-262</td>
<td>42</td>
<td>70-105</td>
</tr>
<tr>
<td>2</td>
<td>263-289</td>
<td>27</td>
<td>106-133</td>
</tr>
<tr>
<td>1</td>
<td>290-335</td>
<td>46</td>
<td>134-174</td>
</tr>
</tbody>
</table>

*Sequence analysis preformed using The Sequence Manipulation Suite @ http://www.ualberta.ca/~stothard/javascript/index.html

Table 3. Characteristics of structural segments

Library Construction

Gregor Mendel’s breeding scheme was adapted into a strategy for shuffling non-homologous genes called iterative molecular inbreeding (IMI). The parent genes are “crossed” to make an initial population of hybrid genes. The “progeny” of the parental generation (P) is called the first filial generation (F₁), and the subsequent generation produced by “breeding” together the F₁ offspring is termed the F₂ generation.

The construction methods are described in Chapter 2, illustrated in Figure 15, and outlined here. In brief, oligonucleotides designed to correspond to the boundaries
between structural segments and encode for junctional variability direct the assembly of the first generation of hybrid genes through a series of template switching reactions Figure 15. (A) and (B). This involves the synthesis of N-terminal gene fragments of one parental gene using a set of 7 reverse primers that have sequence of the other gene at their 5’ termini. These fragments are then mixed with the other parental gene where they serve as primers to synthesize full-length hybrid genes followed by amplification with a primer set that selects for products that code for the N-terminus of one protein and the C-terminus of the other. Hybrid genes having different termini, that contain overlapping regions of identity, can be selectively “crossed” by DNA shuffling or variations of the technique like StEP (Zhao et al., 1998) to produce the next generation (F2) of hybrids which contain an additional crossover Figure 15. (C) and (D). Alternatively, the gene fragments produced in (A) can be used as in (B) but with a hybrid gene in lieu of a parental gene. Further iterations produce a library with a hierarchal structure in which successive generations are increasingly interbred until all possible arrangements of the structural elements are attained Figure 16. This systematic approach allows for controlled synthesis of discrete hybrid populations without the production of wild-type genes. Each class of each generation was synthesized independently and cloned into pTH18cr to give ≥ 10 times the number of independent clones as unique members.
Figure 15. Library Construction by IMI. (A). Fragments of one parental gene (Pβ) are amplified using a complementary forward primer and a collection of reverse hybrid primers containing the engineered junction by conventional PCR. (B). The gene fragments generated in A. are mixed with the other parental gene (Pχ) and used to synthesize a set of hybrid genes by primerless PCR. A second conventional PCR, with primers directed toward opposite termini, is used to amplify full-length hybrid genes. The product is called a first generation (F₁) hybrid. (C). Two F₁ hybrid genes with different termini that have overlapping regions of identity can be shuffled using primerless PCR followed by amplification with a primer set that selects for crossover products to produce a collection of second-generation (F₂) hybrids. (D) Repetition of the process to generate the F₃ generation of hybrids.
Figure 16. Hierarchal structure of an IMI library. Segment configurations are displayed as closed rectangles with black or white boxes within that represent the structural segments derived from Pol X or Pol β respectively. Each generation is labeled starting with the parental (P) and descending to the 5th (F5). The # of classes in each generation refers to the number of segment configurations in that generation. Complexity is defined for each generation and is the product of the # of classes and \( l^g \) where \( l \) is the junctional variability (7 for this case), and \( g \) is the generation. Segment configurations, # classes and complexity are presented for half the library; segment configuration is the inverse of this for the other half whereas classes and complexity are identical.
<table>
<thead>
<tr>
<th>Segment Configurations</th>
<th>Generation</th>
<th>Classes</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>P</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F_1</td>
<td>F_1</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>F_2</td>
<td>F_2</td>
<td>10</td>
<td>490</td>
</tr>
<tr>
<td>F_3</td>
<td>F_3</td>
<td>10</td>
<td>3430</td>
</tr>
<tr>
<td>F_4</td>
<td>F_4</td>
<td>5</td>
<td>12005</td>
</tr>
<tr>
<td>F_5</td>
<td>F_5</td>
<td>1</td>
<td>16807</td>
</tr>
</tbody>
</table>
Library Complexity

The numerical complexity for the most shuffled generation can be expressed with the exponential relation:

\[ n = l^g \]  \hspace{1cm} (1)

where \( n \) is the number of unique members, \( l \) is the junctional variability, and \( g \) is the generation (\( g = \# \text{ of structural segments} - 1 \)). For a library with 5 generations and a 7 amino acid junctional variability, the most shuffled generation (\( F_5 \)) is the product of 5 crossovers between 6 segments resulting in 16,807 (i.e., \( 7^5 \)) possible combinations. Combining all generations, including the parental, of both Pol X and Pol \( \beta \)-sized hybrids gives a total numerical complexity of 65,536 members.

Consider the differences in the sizes of "equivalent" structural elements as listed in Table 3 and the size variability introduced at their C-termini. For Pol X-sized hybrids, the largest class of hybrid is in the \( F_3 \) generation and is 224 ± 9 amino acids, the smallest class is in the \( F_2 \) generation and is 167 ± 6 amino acids, and the most shuffled generation (\( F_5 \)) is 196 ± 15 amino acids. Therefore, Pol X-sized hybrids range from 161 to 233 amino acids in length.

Sequence space is defined as the possible combinations of amino acids for a given sequence length connected by mutational moves. Multiple sequence alignment data can be used to construct phylogenies, which depicts the evolutionary proximity of genes within a family of homologues and hence the sequence space subtended by them (Page, 1996). The protein sequences for all of the possible segment
configurations of Pol X-sized polymerases were aligned using the program ClustalX (Jeanmougin et al., 1998) and a phylogenetic analysis was conducted. A cladogram was produced and illustrates the relationship between different members of the Pol X-sized hybrid “family” as produced in shuffling reactions shown in Figure 17. The unit of measure for sequence space is defined as # of substitutions per position. A comparison of the sequence space covered by the library here to one created by random mutagenesis is depicted in Figure 18. The sequence of Pol X was randomly mutated in silico and the resulting sequences were treated in an identical manner as the shuffled Pol X-sized hybrid sequences. There is approximately 100-fold difference between the radius of sequence spanned by the random mutant and structure-based shuffling library. The end of each branch corresponds to a segment configuration shown in Figure 16. Therefore, the size heterogeneity and the range of sequence space covered constitute the “effective” complexity of the library.
Figure 17. Cladogram of the Pol X-sized hybrid family produced by structure-based gene shuffling. The program ClustalX 1.8 was used to align all possible segment configurations and a Bootstrap N-J Tree was generated from the alignment data from which a cladogram is derived (where Pol X is labeled as Pol and hybrids labeled by generation). Sequence names are followed by the name of the generation. The distance scale at the bottom left is in units of substitutions per position.
Figure 18. Sequence space subtended by the structure-based shuffling. The Bootstrap N-J Tree generated from the alignment data of possible Pol X-sized hybrids was used to generate an unrooted phylogenetic tree (where Pol X is labeled as Pol and hybrids labeled by generation). The distance scale at the bottom left is in units of substitutions per position and below it is an unrooted tree produced from a virtual population of random mutants of Pol X.
Screening of Libraries

On the basis of previous analyses described in Chapter 3, it was decided to perform all selection experiments at cell densities of 1,000 colonies per 150 mm plate (ca. 450 cfu/100 mm plate) at 42 °C. Under this condition Pol β transformants retain viability whereas Pol X or empty vector transformants exhibit less than 1 % survival.

Selection experiments were performed for individual generations by plating out the number of colonies ca. 10 times the complexity of each generation (e.g., ca. 34,000 colonies for the F3 generation). About 700 150 mm plates were required, and at ~0.75 inches/plate produces a stack of plates about 44 feet high (about the height of Johnston lab). Since the percentage of complementors during the first round of selection is near background levels, iterative reselection was performed to enrich for active polymerases. A total of 3 rounds of reselection were performed for Pol b or Pol X-sized hybrid generations as illustrated in Figure 19. (A and B respectively). An estimated 1,000 Pol β-sized and 100 Pol X-sized complementing hybrids were produced and identified, as a result of selection experiments as summarized in Table 4.
Figure 19. Iterative reselection of structure-based shuffling libraries. Each hybrid generation was plated out individually, and the resulting survivors were pooled, plasmid DNA isolated from them, followed by retransformation and replating. A total of 3 iterations, including the first round of selection, were performed. The % survival was plotted for each generation of Pol β (A) or Pol X-sized (B) hybrids as a function of the round of selection (iteration).
A.

Iterative Reselection of Pol β-sized hybrids

B.

Iterative Reselection of Pol X-sized hybrids
The number of complementing Pol X or Pol β-sized polymerases per generation is approximated by multiplying the fraction of survivors with the number of unique members of the generation.

Table 4. Results of selection experiments on structure-based shuffling hybrid polymerase libraries.

<table>
<thead>
<tr>
<th>Generation</th>
<th>Complexity</th>
<th>Pol X-sized</th>
<th></th>
<th>Pol β-sized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Survival</td>
<td># of Complementors</td>
<td>% Survival</td>
<td># of Complementors</td>
</tr>
<tr>
<td>F₁</td>
<td>35</td>
<td>0.26</td>
<td>1</td>
<td>6.5</td>
</tr>
<tr>
<td>F₂</td>
<td>490</td>
<td>0.23</td>
<td>1</td>
<td>26.8</td>
</tr>
<tr>
<td>F₃</td>
<td>3,430</td>
<td>0.20</td>
<td>6</td>
<td>15.1</td>
</tr>
<tr>
<td>F₄</td>
<td>12,005</td>
<td>0.75</td>
<td>90</td>
<td>3.0</td>
</tr>
<tr>
<td>F₅</td>
<td>16,807</td>
<td>0.11</td>
<td>17</td>
<td>0.17</td>
</tr>
</tbody>
</table>

The ability of a number of randomly picked individual hybrid polymerases from enriched populations to complement was confirmed by comparison to the parental genes. Pol X-sized polymerases that are the product of multiple crossovers (from generations F₃ and up) are able to confer greater than 50% survival at 42 °C and cell densities ≤ 400 cfu/100 mm plate, whereas under the same conditions, Pol X or empty vector transformants are completely inviable. The complementation behavior of 2 representative Pol X-sized hybrid polymerases from the enriched F₄ generation is
shown in Figure 20. This result indicates that shuffled products, with the size of Pol X, posses enhanced catalytic properties in vivo.

Figure 20. Complementation of Pol X-sized hybrids from the enriched F₄ generation. Two Pol X-sized hybrid polymerases were assayed for their complementation ability by the complementation assay at 42 °C. Each trial was performed in triplicate and the % survival as a function of cell density was plotted along side Pol X, Pol β, or empty plasmid SC18-12 transformants.

Sequence Analysis

Sequence analysis of 60 complementors from F₃ to F₅ generations of Pol X-sized libraries was performed and 16 unique sequences representing 6 different segment configurations were identified as displayed in Table 5. A total of 39 Pol β-sized complementors were sequenced from F₂ to F₄ generations and 20 unique
sequences representing 7 different segment configurations were found as summarized in Table 6. The variability at the segment boundaries of all unique sequences for either the Pol $\beta$ or Pol X-sized complementors was analyzed and the results shown in Figure 21. Regardless of the sequence and structure homology at a particular junction, the zero position was only the most favored at one boundary. For example, the chosen zero point of fusion boundary 1 agrees well with the structure-based sequence alignment reported later, yet there is almost invariably a deletion (-2 or -3) in the chimeras of Pol X and Pol $\beta$ size. The observed bias for deleted or expanded junctions suggests the importance of engineering flexibility into the connections between segments to accommodate the “best” fit and inaccuracy in the choice of their points of fusion.
<table>
<thead>
<tr>
<th>Generation</th>
<th>Segment configuration</th>
<th>Size</th>
<th>Frequency</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>F₅</td>
<td>§</td>
<td>202</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>195</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>190</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>193</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>198</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>194</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F₄</td>
<td>§</td>
<td>191</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>185</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>§</td>
<td>189</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>194</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>F₃</td>
<td>§</td>
<td>206</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>205</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>204</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>184</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>178</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>180</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

* Size is defined as the number of amino acids.
^ Frequency refers to the number of times a particular sequence was found.
® Total is the number of clones from a particular generation that were sequenced.
Segments colored in black and white represent sequence derived from Pol X and Pol β respectively.

Table 5. Pol X-sized hybrids recovered from selection experiments.
<table>
<thead>
<tr>
<th>Generation</th>
<th>Segment configuration</th>
<th>Size*</th>
<th>Frequency^</th>
<th>Total@</th>
</tr>
</thead>
<tbody>
<tr>
<td>F₄</td>
<td></td>
<td>318</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>319</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>325</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>319</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>318</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>322</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>324</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>329</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>F₃</td>
<td></td>
<td>317</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>318</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>324</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>318</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>336</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>334</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>325</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>F₂</td>
<td></td>
<td>327</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>F₂</td>
<td></td>
<td>320</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>328</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>322</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>319</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

* Size is defined as the number of amino acids.
^ Frequency refers to the number of times a particular sequence was found.
@ Total is the number of clones from a particular generation that were sequenced.
Segments colored in black and white represent sequence derived from Pol X and Pol β respectively.

**Table 6.** Pol β-sized hybrids recovered from selection experiments.
Figure 21. The junctional preferences of complementing hybrids. The amino acid sequences derived from the DNA sequencing data were analyzed and the junctional positions at segment boundaries (as outlined in Fig. 14) was tabulated. A total of 60 complementing Pol X-sized hybrids from the F$_3$ to the F$_5$ generation and 39 complementing Pol β-sized hybrids from F$_2$ to F$_4$ generation were sequenced and the position of their junctions were tabulated. The % occurrence is calculated as the # of times a junction position occurred divided by the total # of hybrids with that segment boundary multiplied by 100.
Probability Analysis

The mathematical benefit of sampling generations individually as opposed to a complex mixture is to increase the probability that each unique member is screened. This can be formally expressed in eq. 2, assuming that the number of copies of a gene in a collection of unique members is not limiting:

\[ p(n) = 1 - \sum_{i=1}^{n-1} (-1)^{i+1} \frac{n!}{i! (n-i)!} \left[ \frac{(n-i)^k}{n} \right] \] (2)

For \( k \geq n \), where \( k \) is the sample size and \( n \) is the number of unique members defined here by \( n = f \), the probability \( p \) that a sample size of \( k \) contains at least one representative of each unique member can be calculated. If \( k \) is a fixed multiple of \( n \) (\( k = 10 \) \( n \) 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 segments and the variability in the junctions connecting them in the design of libraries as shown in Figure 22.
Figure 22. Probability as a function of generation and junctional variability. Given the relation $n = l^g$, where $n =$ complexity of the most shuffled generation, $l =$ junctional variability and $g =$ the generation, for a sample size that is 10 times the number of unique members of a generation (for $k = 10n$), the probability ($p$) that a representative of each type is screened is plotted. Probability is calculated according to equation 2, and ranges from 1 in the upper left to zero in the bottom right as the numerical complexity increases.
Conclusion

The concept of exon rearrangement as a mathematical solution to the complexity dilemma faced by nature is a valuable lesson that can be utilized in the laboratory. The possibility of shuffling exons as nature does for directed evolution of enzymes has been previously suggested (Kolkman and Stemmer, 2001). However, if the genes of interest don’t have an intron-exon organization, then assignment and combinatorial assembly of “pseudo-exons” can be performed, as presented here. In practice, structural or sequence information can be used to guide the fragmentation of genes into equivalent pieces and the IMI technique, also developed in this work, can assemble the combinatorial libraries.

Structure-based shuffling allows the generation of libraries with greater “effective” complexity than existing techniques by accessing larger regions of sequence space through shuffling non-homologous genes. The IMI approach enables their controlled synthesis resulting in partitioning of the library into discrete collections of unique members with the benefit of expanding the “numerical” complexity that can be screened. The practical benefits are especially apparent when there are screening limitations imposed by the selection system. In principle, homology between the protein sequences alone can guide library design within the desired probability of screening all variants as outlined here. Knowledge of tertiary structures of the proteins is useful but not essential. In fact, the sequence alignment used to guide this work Figure 14. (C) differs from the structure-based sequence alignment reported after the completion of this work.
The effectiveness of this design strategy has been demonstrated by successfully producing a large number of biologically active chimeras from the distantly related proteins Pol X and Pol β. The equivalent segments that are exchanged can have as low as 14.3% amino acid sequence identity, yet can be assembled to give rise to functional proteins. The preferences at certain junctions suggest the importance of elasticity in their design when combining structurally equivalent segments of low sequence identity. Further, these Pol X-sized DNA polymerases have enhanced activity in vivo relative to Pol X. It will be of great significance to characterize the detailed structural and functional properties of these new chimeric polymerases. It is likely that some of these chimeras will show novel substrate specificity. Such studies could lead to an improved understanding of the structural basis of the fidelity of DNA polymerases.

In conclusion, structure-based gene shuffling by iterative molecular inbreeding provides a controlled method for connecting distant points in sequence space through shuffling non-homologous genes. Plotting the structural and functional properties of these proteins, once obtained, as a function of sequence space will ultimately enable the construction of maps, which will guide future navigation through the vastness of sequence space.
BIBLIOGRAPHY


80


