DIRECTED EVOLUTION OF DNA POLYMERASES

By

STEPHANIE ANN HAVEMANN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2007

To my family and my husband. Without your constant and vigilant support, I would not be where I am today. Thank you!

ACKNOWLEDGMENTS

I would like to begin by thanking my advisor, Dr. Steven Benner for all of his wisdom and guidance; it has been an honor and a privilege to study under his tutelage. His passion for all facets of science, and how they can be intertwined, should serve as inspiration to us all.

I would like to thank the rest of my committee: Dr. Tom Lyons, for always having an open door and an receptive ear when I had questions; Dr. Nemat Keyhani, whose enthusiasm for science was contagious and whose knowledge of microbial genetics was extremely valuable; Dr. Nicole Horenstein, whose constant support and knowledge helped guide me throughout my graduate career; and Dr. Rob Ferl, whose eagerness to learn and share information about various aspects of astrobiology helped me determine the field of study I wish to pursue.

Special thanks go to Dr. Eric Gaucher, Dr. Ryan Shaw, and Dr. Nicole Leal, all of whom have worked closely with me over the past few years and who have assisted me in various experimental designs and implementations. Eric performed the rational design of the Taq mutants and was my source of knowledge for all things dealing with evolutionary biology. Ryan and I worked closely to discern the best method of creating and isolating DNA from oil-in-water emulsions; his idea of changing the composition of the oil layer drastically improved our yields. Nicole assisted me in performing some of my primer-extension assays and was a valuable source of information and never-ending support.

I am extremely grateful to Dr. Daniel Hutter for the synthesis of the 2'-deoxypseudothymidine-5'-triphosphate, to Dr. Shuichi Hoshika for the synthesis of the pseudothymidine precursor, and to Dr. Ajit Kamath for the synthesis and purification of the pseudouridine-containing oligonucleotides. Special appreciation also goes to Dr. Michael Thompson for providing the *wt taq* gene and his suggestions for the purification of the polymerase, and to Gillian Robbins for assisting on the growth curve studies.

I am also thankful for the assistance of Dr. Art Edison and Omjoy Ganesh for their assistance in the circular dichroism experiments. Finally, I would like to thank all the members of the Benner group for their advice and discussions over the years, and Romaine Hughes, without whom, our group would be in total chaos.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	adenosine
AEGIS	artificially expanded genetic information system
Amp	ampicillin
APS	ammonium persulfate
ATP	adenosine triphosphate
bp	base pair
Bst	*Bacillus stearothermophilus*
C	cytosine
Ci	Curie (1 Ci = 3.7 x 10^7 Bequerel)
CD	circular dichroism
Cfe	cell-free extract
cfu	colony forming unit
CPM	counts per minute
CNT	counts
CSR	compartmentalized self-replication
DMSO	dimethyl sulfoxide
dN	deoxyribonucleoside (dA, dG, dC, T, ψT, ψU, etc.)
DNA	deoxyribonucleic acid
DNase I	deoxyribonucleic acid specific endonuclease
ds	double-stranded nucleic acid chain
DTT	1,4-dithio-DL-threitol
E. coli	*Escherichia coli*

EDTA	ethylendiamino tetraacetate
exo-	lacking 3'→5' exonuclease activity
FLP	full-length product
G	guanosine
HIV	human immunodeficiency virus type-1
hr	hours
HPLC	high performance liquid chromatography
isoC	deoxyisocytidine
isoG	deoxyisoguanosine
LB	Luria-Bertani medium
min	minutes
M-MuLV	moloney murine leukemia virus
mRNA	messenger ribonucleic acid
MWCO	molecular weight cut-off
NMR	nuclear magnetic resonance
NSB	non-standard nucleobase
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
Pfu	*Pyrococcus furiosus*
PMSF	phenylmethylsulfonyl fluoride
PNK	polynucleotide kinase
REAP	reconstructing evolutionary adaptive paths

RNA	ribonucleic acid
RNase A	ribonucleic acid specific endonuclease
rRNA	ribosomal ribonucleic acid
RT	reverse transcriptase
SDS	sodium dodecylsulfate
s	seconds
StEP	staggered extension processes
T	thymidine
ψT	pseudothymidine
Taq	*Thermus aquaticus* DNA Polymerase I
TBE	Tris / borate / EDTA buffer
TEMED	N,N,N,N-tetramethylethylenediamine
Tet	tetracycline
Tris	tris(hydroxymethyl)aminomethane
Triton X-100	octyl phenol ethoxylate
tRNA	transfer ribonucleic acid
Tth	*Thermus thermophilus*
U	uracil
ψU	pseudouridine
UV	ultraviolet
wt	*wild type*

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

DIRECTED EVOLUTION OF DNA POLYMERASES

By

Stephanie Ann Havemann

May 2007

Chair: Steven A. Benner
Major Department: Chemistry

To achieve the long-term goal of the Benner research group to create a synthetic biology based on an Artificially Expanded Genetic Information System (AEGIS), polymerases that are able to incorporate non-standard bases (NSBs) into DNA must be identified. In this dissertation, a polymerase from *Thermus aquaticus* (*Taq* Polymerase) was identified that was able to incorporate non-standard nucleotide analogs that contain a C-glycosidic linkage. This activity was limited, meaning that the polymerase needed modification to support this goal. Further, we asked whether sequential C-glycosides destabilized the duplex and altered its structure, to better understand whether a synthetic biology based on C-glycoside nucleotides was possible.

To this end, two libraries of polymerases were created to identify mutations necessary to alter the polymerases' ability to withstand high temperatures. One library was created by the random mutagenesis of the *taq* gene, the other was rationally designed based on previous studies. Seventy-four mutants from each library were screened for their ability to generate a full-length polymerase chain reaction (PCR) product using standard nucleoside triphosphates at various temperatures; the library of random mutants contained more thermostable polymerases than the library obtained by rational design. Water-in-oil emulsions were then tested to determine whether these, as artificial cells, might deliver thermostable polymerase variants from those used

in the screen. This identified difficulties in tools used to analyze the output of the library, suggesting solutions that will guide future work. We also tested the individual components of the rationally designed library for their ability to incorporate C-glycoside triphosphates in a PCR. Structural studies with synthetic DNA containing multiple, consecutive C-glycosides showed no change in conformation, at least not one that is detectible by circular dichroism.

These results represent a step towards the goal of creating an AEGIS-based synthetic biology, an artificial chemical system that mimics emergent biological behaviors such as replication, evolution, and adaptation. In addition, the mutant polymerases created in these experiments are an inventory of polymerases useful in biotechnology, possibly allowing the development of new, as well as improving on existing, clinical diagnostic techniques and helping to facilitate a better understanding of polymerase-DNA interactions.

CHAPTER 1
INTRODUCTION

What are Nucleic Acids?

Deoxyribonucleic acid (DNA), one of the fundamental constituents of life, serves as a key component for the storage and transfer of genetic information. It is built from four building blocks, adenosine, guanosine, cytidine, and thymidine, all of which are comprised of a nucleobase attached to a 2'-deoxyribose molecule (Fig. 1-1). Similarly, ribonucleic acid (RNA) is also built from four building blocks, except that thymidine is replaced by uridine and the sugar moiety is a ribose. When a phosphate group replaces the 5'-hydroxyl group of these molecules, they become acids that can be linked by their phosphate groups, resulting in the formation of the backbone of a nucleic acid strand. Genetic information is commonly stored in a double stranded (ds) helix, which is formed when the nucleobases are paired by hydrogen bonds. These helical duplex strands are aligned so that the chains are anti-parallel to one another; in other words, one strand lies in the 5'→3' direction and the complement is in the 3'→5' orientation.

Rules of Complementarity

Watson and Crick proposed that the interactions between nucleobases are governed by two rules of complementarity: size complementarity and hydrogen-bonding complementarity (Watson and Crick, 1953a, Watson and Crick, 1953b). Size complementarity means that a large purine, such as adenosine or guanosine, pairs with a small pyrimidine, like cytosine, thymidine, or uridine. Hydrogen-bonding complementarity means that hydrogen bond donors from one nucleobase pair with the hydrogen bond acceptors from another. With these rules, it is expected that in the formation of nucleic acid duplexes, guanosine must pair with cytosine and adenosine must pair with either thymidine or uridine.

DNA Helical Conformations

The conformation of a DNA duplex is often assumed to be described using one of three abstract models: A-DNA, B-DNA, or Z-DNA (Saenger, 1984). The most common form of DNA found in living organisms is presumed to be the B-DNA helix. A-DNA is the common helical structure whose geometries are described in Table 1-1. It is also interesting to note that many other minor helical conformations of dsRNA or dehydrated DNA, but it can also be found when certain DNA sequences repeat (Ghosh and Bansal, 2003). The only left-handed helix known is the Z-DNA conformation, which appears to be a characteristic of alternating GC-rich sequences that may help stabilize DNA during transcription (Rich and Zhang, 2003). Many other helical conformations of DNA are possible, of course. Indeed, over twenty-six different forms have been described in the literature to date (Egli, 2004, Ghosh and Bansal, 2003, Saenger, 1984). Nevertheless, for this work, we will reference the A-, B-, and Z-DNA models.

In actuality, the conformation of a DNA molecule must be described by examining the structure atom by atom. Terms used to abstract the results of such an examination are described in Table 1-1. Thus, the different types of helices are characterized by different geometries, such as the number of base pairs per turn, the height of a turn, the rotation per base pair, the size and depth of the major and minor grooves, and the type of sugar pucker. The sugar pucker refers to the conformation of the sugar, which can exist in one of four envelope forms: $C_{2'}$-endo, $C_{2'}$-exo, $C_{3'}$-endo, and $C_{3'}$-exo (Fig. 1-2) (Saenger, 1984).

In some cases, helical structures can be transformed from one conformation into another simply by the modification of the humidity of the environment (for fibers) and/or the concentrations of salt in the solution (Saenger, 1984). Helical structures can also be changed by altering the chemical structure of the constituents. The conformation of the sugar pucker can alter the helical form of the DNA by increasing or decreasing the distances between the

phosphate groups, thereby changing the number of base pairs per turn and the size of the grooves. The $C_{2'}$-endo conformation is usually found in B-DNA, while the A-DNA prefers the $C_{3'}$-endo pucker. The major and minor grooves found in B-DNA can act as binding pockets for polymerases, since they allow for the presentation of nucleobase hydrogen bond donors and acceptors (Garrett and Grisham, 1999). The grooves presented by A-DNA are more symmetrical, making it difficult for polymerases to gain access to these potential hydrogen-bonding sites (Garrett and Grisham, 1999).

The conformation of a DNA helix can be assessed in several ways. X-ray crystallography is, of course, the best way to identify the position of individual atoms, with nuclear magnetic resonance (NMR) emerging as a preferred choice in solution. The general overall conformation can be extimated, however, by circular dichroism (CD) (Ghosh and Bansal, 2003).

Central Dogma of Molecular Biology

Nucleic acids maintain genetic information inside a cell by means of replication and transcription; translation uses this genetic information to create proteins. This sequence has been called the central dogma of molecular biology by Crick (Fig. 1-3) (Crick, 1970). DNA is transcribed into messenger RNA (mRNA) using RNA polymerases, which is then translated into proteins. The translation of the mRNA uses a combination of ribosomes, which are composed of ribosomal RNA (rRNA) and proteins, and transfer RNA (tRNA), which carry amino acids to the ribosomes. In situations where the genetic material is stored as RNA, such as in viruses, the information is first converted back into DNA by enzymes known as reverse transcriptases prior to being translated. DNA can replicate itself by employing enzymes known as DNA polymerases, and RNA replicates itself using RNA polymerases.

This feature of life raises an obvious question: Which came first, nucleic acids or proteins? At first glance, the answer appears to be nucleic acids, since proteins cannot store genetic

information. Upon further study, one realizes that without proteins, the genetic material could not be replicated. One possible answer to this question is that the nucleic acids were once able to act as both storage molecules and as proteins that could catalyze their own replication.

The discovery of ribozymes and deoxyribozymes lends support to this theory by showing that nucleic acid molecules are not limited to the ability to store genetic information, they can catalyze reactions both within their own structure or upon other structures (Muller, 2006, Emilsson and Breaker, 2002, Paul and Joyce, 2004). Many of these nucleic acid catalysts have been created using non-standard nucleobases (NSBs) to add additional functionality to the nucleic acid molecules (Muller, 2006).

What is AEGIS?

Using Watson and Crick's rules of complementarity and the requirement that the nucleobases be joined with three hydrogen bonds, it is feasible to create an artificially expanded genetic information system (AEGIS) containing eight additional base pairs (Fig. 1-4), thereby expanding the genetic alphabet from four to twelve letters (Switzer et al., 1989, Piccirilli et al., 1990, Geyer et al., 2003). Since these bases retain the Watson and Crick geometry, they can be incorporated into growing DNA strands via synthesis, primer-extension experiments, or by the polymerase chain reaction (PCR), which can subsequently be used in a variety of different techniques.

Use of AEGIS Components

The importance of AEGIS components has already been illustrated in many ways. It has been used in clinical diagnostics, to expand the genetic code, to understand DNA and polymerase interactions, and has even been implicated as a factor for evolution of life on Earth. These components have also been used in the first successful six-letter PCR reaction, lending support to the development of a synthetic biology.

The powerful Versant™ branched-DNA assay, used to monitor the viral load of patients infected with HIV, Hepatitis B, or Hepatitis C viruses, requires the use of at least two non-standard nucleobases (NSBs) (Collins et al., 1997). This assay uses 5-methyl-2'-deoxyisocytidine (isoC) and 5-methyl-2'-deoxyisoguanosine (isoG) to decrease the non-specific binding of a nucleic acid probe (Fig. 1-4), thereby increasing signal amplification relative to noise by eight-fold over previous systems used (Fig. 1-5) (Huisse, 2004, Collins et al., 1997). EraGen Biosciences (Madison, WI) is now using these AEGIS components in a similar multiplexed system to identify newborns with cystic fibrosis (Johnson et al., 2004). These assays have barely begun to scratch the surface of the potential clinical diagnostic uses of this expanded genetic alphabet.

The current genetic code uses 64 three-letter codons to encode for the incorporation of 20 canonical amino acids (Appendix C); use of all twelve AEGIS nucleotides would allow for 1728 three-letter codes, and if the AEGIS components were functionalized, the possibilities are seemingly nearly endless. AEGIS components have been already been used to encode for the incorporation of non-standard amino acids in ribosome-mediated translation. For example, in 1992 Bain *et al.* used isoC and isoG in a codon-anti-codon pair to generate peptides containing the non-standard amino acid L-iodotyrosine (Bain et al., 1992). More recently, Hirao *et al.* used the 2-amino-(2-thienyl)purine and pyridine-2-one in a codon-anti-codon pair in an *in vitro* transcription study to generate peptides containing 3-chlorotyrosine (Fig. 1-6) (Hirao et al., 2002, Hirao et al., 2006).

Some of these AEGIS components have also been used in the characterization of the kinetic parameters of polymerases (Joyce and Benkovic, 2004, Sismour and Benner, 2005), and in the first six-letter PCR, which was catalyzed by a mutant of the HIV-reverse transcriptase

(Sismour et al., 2004). AEGIS components have also been used to better understand the interactions between polymerases and DNA (Lutz et al., 1998, Joyce and Benkovic, 2004, Hendrickson et al., 2004, Delaney et al., 2003). For example, studies have been performed using variety of different NSBs, such as those lacking minor-groove electrons (Hendrickson et al., 2004) and those with a C-glycosidic linkage (Lutz et al., 1999), in order to identify characteristics of nucleobases that are essential for correct incorporation by polymerases.

Problems with AEGIS Components

Although the AEGIS components retain Watson and Crick geometry, it is possible that some of the features present on the NSBs, such as the absence of minor groove electrons or the presence of C-glycosidic linkages, may present a challenge to polymerases. The ability of polymerases to function in the absence of an unshared pair of electrons in the minor grove of dsDNA, as seen in the pyDAD-puADA base pair (Fig. 1-4), was previously examined by Hendrickson *et al* (Hendrickson et al., 2004). In those studies, Hendrickson discovered that the presence of electrons in the minor grove may only be necessary for exonuclease activity of polymerases, and not for incorporation (Hendrickson et al., 2004). This, however, presents a problem when trying to incorporate NSBs with efficiency and fidelity, since the polymerase has no proofreading ability. Lutz *et al.* examined the ability of polymerases to function in the presence of nucleosides exhibiting a C-glycosidic linkage, a carbon-carbon bond between the nucleobase and sugar as seen in the pyDAD, pyAAD, and pyADD nucleosides (Fig. 1-4) (Lutz et al., 1999). He also reported that polymerases with exonuclease activity were less likely to accept the C-glycoside than were those lacking the proofreading ability, making replication with fidelity difficult.

C-Glycosides

An N-glycoside is a nucleoside with a carbon-nitrogen bond linking the nucleobase to the sugar; all standard nucleosides are therefore N-glycosides. However, three of the AEGIS nucleosides use a carbon-carbon bond to join the nucleobase to the sugar, making these nucleosides C-glycosides by definition (Fig. 1-4). This carbon-carbon linkage can cause a structural change in the sugar pucker of the nucleoside, making it a $C_{3'}$-endo pucker instead of a $C_{2'}$-endo pucker, possibly changing the form of the DNA from B-DNA to A-DNA (Davis, 1995).

Wellington and Benner detailed strategies by which these molecules can be chemically synthesized in a current review article (Wellington and Benner, 2006). C-glycosides have also been found *in vivo* in various types of RNA, however (Charette and Gray, 2000). These C-glycosides are of great interest, not only because of their presence in the AEGIS nucleosides, but also for their clinical uses; many naturally occurring C-glycosides are antibiotics or antiviral agents (Michelet and Genet, 2005, Zhou et al., 2006). More generally, C-glycosides can be used in gene therapy (Li et al., 2003, Li et al., 2004).

Pseudouridine

Pseudouridine (ψU), the 5-ribosyl isomer of uridine (Fig. 1-7A), is present in both tRNA and rRNA and is vital to the fitness of organisms (Raychaudhuri et al., 1998, Charette and Gray, 2000). This modified nucleoside, found in all three domains of life, was the first naturally occurring NSB discovered (Charette and Gray, 2000), and is introduced into the RNA sequences by the posttranscriptional modification of uridine (Argoudelis and Mizsak, 1976, Grosjean et al., 1995). Pseudouridine has been reported to have a propensity to adopt a *syn* conformation around the glycosyl bond when in solution, although the data supporting this are questionable; it is, however, found only in the *anti* conformation when in a nucleic acid strand (Fig. 1-7B) (Lane et al., 1995, Neumann et al., 1980). The *anti* conformation allows the coordination of a water

molecule between the 5' phosphate group of the ψU residue, the 5' phosphate group of the preceding residue, and the N1-H of the ψU residue (Fig. 1-7C) (Arnez and Steitz, 1994). The coordination of this water molecule results in an enhanced base stacking ability and a reduced conformational flexibility of the RNA molecule, thus increasing the local rigidity of the RNA (Charette and Gray, 2000, Davis, 1995).

Pseudouridine is thought to play several roles in Nature, as described in the review by Charette and Gray (Charette and Gray, 2000). In tRNA, it is thought to play a critical role in the binding of the tRNA to the ribosome during translation because it stabilizes the tRNA structure, allowing tighter binding to occur, thereby increasing translational accuracy. Pseudouridine also has been implicated in alternative codon usage in tRNA, and as a player in the folding of rRNA and ribosome assembly by its contributions to RNA stability.

Pseudothymidine

Pseudothymidine (ψT), or 1-methylpseudouridine (Fig. 1-7D), was originally isolated from *Streptomyces platensis* in 1976 by Argoudelis and Mizsak (Argoudelis and Mizsak, 1976). This naturally occurring C-glycoside, found in RNA, is also thought to be created by a posttranscriptional modification of uridine (Limbach et al., 1994). The first successful *in vitro* transcription of ψT was performed by Piccirilli *et al.* using T7 RNA polymerases with a template containing ψT and standard ribonucleosides (Piccirilli et al., 1991). Further studies, conducted by Stefan Lutz, observed the ability of DNA polymerases to not only incorporate this NSB into a growing DNA strand in primer-extension assays, but also challenged a polymerase to use ψT in a PCR reaction that required the successful incorporation of up to three consecutive dψT residues. (Lutz et al., 1999). Since then, no further studies requiring the incorporation of this C-glycoside into nucleic acids have been performed.

DNA Polymerases

DNA polymerases are the enzymes that perform template directed DNA synthesis from deoxyribonucleotides and an existing DNA template. These enzymes, essential for the replication of the genetic information carried in all living organisms, were originally discovered in 1956 by Arthur Kornberg (Kornberg et al., 1956), for which he was awarded a Nobel Prize in 1959. The synthesis of the complementary DNA strand always occurs in the 5'→3' direction through the addition of incoming nucleotide's triphosphate group onto the 3'-OH group of the preceding nucleotide, releasing a pyrophosphate group in the process (Fig. 1-8) (Garrett and Grisham, 1999, Lewin, 1997). After the successful replication of a DNA strand, the new strand is complementary to the template (leading) strand, and identical to the lagging strand. Since all DNA polymerases function in this manner, it is easy to comprehend that their structures are also generally conserved.

General Structure of Polymerases

All DNA polymerases share a common structural framework that is commonly referred to as a right hand comprised of three subdomains: the fingers, the palm, and the thumb. The fingers domain is responsible for nucleotide recognition and binding, the thumb domain binds the DNA substrate, and the palm domain is the catalytic center of the protein. It appears that this framework is the same in all DNA polymerase families. It is not clear whether this represents convergent or divergent evolution; there is no sequence similarity between, for example, Family A and Family B polymerases that makes a case for their distant homology (Rothwell and Waksman, 2005). In 1985, the laboratory of Thomas Steitz first solved the crystal structure of the Klenow fragment, the C-terminal domain of the *Escherichia coli* DNA Polymerase I (Ollis et al., 1985). Since then, the crystal structure of many different polymerases have been solved, not only in their nascent states, but some with DNA or dNTPs and pyrophosphate bound to the

catalytic site (Rothwell and Waksman, 2005, Beese et al., 1993b, Beese et al., 1993a). It has also been determined that during polymerization, divalent metal cations, such as Mg^{2+}, are coordinated in polymerase active sites to help activate the 3'-OH group for attack on the incoming nucleotide (Steitz, 1999).

Features of polymerases that are not conserved throughout the families include both the 5' →3' and 3'→5' exonuclease subdomains that allow for proofreading, and other subunits used for different types of repair. The exonuclease subdomains, when present, are the proofreading centers of the polymerase. The 5'→3' exonuclease activity is usually involved in nick translation, or the synthesis of DNA at a location where there is a break in the phosphodiester bond of one strand (Perler et al., 1996). The 3'→5' exonuclease activity is the true "proofreading" activity of the polymerase, responsible for the excision of a newly synthesized mismatch (Perler et al., 1996).

The process by which a DNA polymerase adds an incoming nucleotide onto the 3'-hydroxyl group of the preceding nucleoside involves many steps, which are only now being fully understood. Figure 1-9 details the kinetic steps involved in this addition (Patel and Loeb, 2001, Rothwell and Waksman, 2005). In Step 1, the polymerase (E) binds to the DNA primer:template complex (TP); the polymerase then binds the incoming nucleotide triphosphate (dNTP) in Step 2. The polymerase then undergoes a conformational change (E') in Step 3 that brings the various components into positions that can support the chemistry of this reaction; this is the rate-limiting step of polymerization. The polymerase performs the addition of the nucleotide, remains complexed with the pyrophosphate, and undergoes another conformational change in Step 4. The pyrophosphate group is released in Step 5; in Step 6, the polymerase can dissociate from the DNA or translocate the substrate for another round of synthesis.

Polymerase Families

Based on sequence similarity, seven major families of homologous polymerases have been classified (Patel and Loeb, 2001, Rothwell and Waksman, 2005): A, B, C, D, X, Y, and RT. The most extensively studied are those of the Family A and Family B polymerases, but Table 1-2 identifies characteristics and representative polymerases of all seven families. Polymerases behave differently not only between the families, but also within the families themselves, based on their ability to repair, their processivity, and their fidelity. Processivity is defined as the ability of the polymerase to continue catalysis without dissociating from the DNA (Kelman et al., 1998); this is important when dealing with AEGIS components since it has been previously shown that polymerases tend to "pause," or fall off the DNA, after the incorporation of a NSB (Lutz et al., 1999, Sismour and Benner, 2005). Fidelity is the ability of the polymerase to select and incorporate the correct complementary nucleoside opposite the template from a pool of similar structures (Beard et al., 2002, Cline et al., 1996); this is important to AEGIS components to guarantee that the newly replicated DNA contains the correct sequence.

Family A polymerases, which contain some of the prokaryotic, eukaryotic, and viral polymerases, are best known for the *E. coli* DNA Pol I, *Thermus aquaticus* (*Taq*) Pol I, and the T7 DNA polymerases (Perler et al., 1996). The *E. coli* DNA Pol I and *Taq* polymerases are known as repair polymerases since they contain the 5'→3' exonuclease domains, while the T7 is known as a replicative polymerase since it has a strong 3'→5' exonuclease activity (Rothwell and Waksman, 2005, Kunkel and Bebenek, 2000).

Family B polymerases contain representatives from prokaryotic, eukaryotic, archaeal, and viral polymerases, this is the only family of polymerases with members from all four of these populations (Patel and Loeb, 2001). This family of polymerases is predominately involved with DNA replication, as opposed to repair, and exhibit extremely strong 3'→5' exonuclease

activities. In eukaryotes, these polymerases carry out the replication of chromosomal targets during cell division. The most well known of the archaeal polymerases from this family, *Pyrococcus furiosus* (*Pfu*) DNA Polymerase, has the lowest known error rate of all thermophilic DNA polymerases that can be used for PCR amplification (mutational frequency/bp/duplication is 1.3×10^{-6}) (Hogrefe et al., 2001, Cline et al., 1996).

Family C polymerases contain the bacterial chromosomal replicative polymerases, and Family D polymerases are suggested to act as archaeal replicative polymerases (Patel and Loeb, 2001, Rothwell and Waksman, 2005). Family X polymerases are found in eukaryotes, and are believed to play a role in the base-excision repair pathway that is important for correcting abasic sites in DNA (Patel and Loeb, 2001, Rothwell and Waksman, 2005). Family Y polymerases, found in prokaryotes, eukaryotes, and archaea, are part of a replicative complex, and function by recognizing and bypassing lesions created by UV damage so that replication of the DNA is not stalled (Zhou et al., 2001, Rothwell and Waksman, 2005). The last characterized family of polymerases, the reverse transcriptases (RT), found in eukaryotes and viruses, catalyze the conversion of RNA into DNA, but they can also replicate DNA templates as well (Najmudin et al., 2000, Goldman and Marcy, 2001, Rothwell and Waksman, 2005).

Taq Polymerase

Thermus aquaticus, an organism found in thermal springs, hydrothermal vents, and even hot tap water, was first isolated by Brock and Freeze in 1969 (Brock and Freeze, 1969). *Taq* polymerase, a 94 kDa protein, was isolated from this organism by Chien *et al.* in 1976 (Chien et al., 1976), and belongs to the Family A polymerases. This thermophilic polymerase has 5'→3' exonuclease activity, but lacks the 3'→5' exonuclease activity required for the proofreading ability, therefore giving this polymerase a low replication fidelity of about 8×10^{-6} (mutational frequency/bp/duplication) (Cline et al., 1996). However, *Taq* is fairly processive with an

average incorporation of 40 nucleotides before dissociating from the DNA, and it has a quick extension rate of about 100 nucleotides per second (Pavlov et al., 2004, Perler et al., 1996).

Taq polymerase, one of the most extensively studied polymerases, was the first thermostable polymerase to be used in PCR; thereby eliminating the need to add additional polymerase after every round of PCR as was necessary when *E. coli* DNA Pol I was used for thermocycling experiments (Saiki et al., 1988). In 1995, the Steitz laboratory was the first to crystallize nascent *Taq* polymerase (Kim et al., 1995), and have since crystallized the polymerase with DNA at the active site (Eom et al., 1996). These, and other studies, have allowed researchers to identify the active site of the polymerase and the specific residues which contact the DNA, the incoming nucleotides, or are involved in metal ion chelation (Eom et al., 1996, Fa et al., 2004, Li et al., 1998b, Li et al., 1998a, Kim et al., 1995, Suzuki et al., 1996).

Due to *Taq* polymerase's lack of proofreading ability, it has been identified previously as a candidate for replication of DNA containing non-standard nucleosides (Lutz et al., 1999). *Taq* has been used to incorporate and/or replicate NSBs exhibiting C-glycosidic linkages (Lutz et al., 1999), NSBs lacking an unshared pair of electrons in the minor groove (Hendrickson et al., 2004), and nonpolar nucleoside isoteres (Morales and Kool, 2000). Directed evolution has created *Taq* polymerase mutants that have been used to incorporate an even larger repertoire of NSBs (Henry and Romesberg, 2005).

Directed Evolution

A recent review by Griffiths and Tawfik discussed the application of techniques developed for the *in vitro* evolution of various proteins to increase their rate of catalysis, perform different functions, and accept new substrates (Griffiths and Tawfik, 2006). These procedures all select for desired enzyme characteristics from pools of millions of genes with schemes designed to link genotype to phenotype. This provides a great advantage over the older methods of screening

mutant library members individually, because these approaches use a "one-pot" technique that allows for the testing of a large number of variants (2×10^8 or more) at once (Griffiths and Tawfik, 2006)

Other common features of these directed evolution systems include the development of a mutagenic library, expression of this library, a high-throughput assay designed to identify individuals with the desired characteristics, and a means for reshuffling mutants between rounds of selection (Brakmann, 2005, Lutz and Patrick, 2004, Arnold and Georgiou, 2003a). The most challenging part of any selection experiment is the design of the technique that will be used to isolate variants with the desired characteristics (Brakmann, 2005), because "you get what you select for." In other words, scientists may want to select for a specific characteristic of an enzyme, but if the technique is not designed correctly, they may end up selecting for an enzyme with a different characteristic.

Mutagenic Libraries

The first step in any directed evolution experiment is to create a large library of mutant enzymes. There are many ways to accomplish this task, varying from the rational design of mutations at selected sites to the random mutagenesis of residues along the length of the sequence. Francis Arnold co-authored a book with George Georgiou that gave detailed instructions on how to perform nineteen different techniques to generate libraries for directed evolution (Arnold and Georgiou, 2003b). This book gave attention to standard error-prone PCR techniques that use $MnCl_2$ instead of $MgCl_2$ in PCR reactions catalyzed by a polymerase with low fidelity, such as *Taq*, and to methods that could be used for the rediversification of libraries between rounds of selection, such as the staggered extension process (Fig. 1-11).

An important consideration when creating a true random library of mutants is the bias of some techniques to create certain transitional or transversional mutations preferentially.

Transitional mutations occur when one purine-pyrimidine pair is replaced with another purine-pyrimidine pair; this creates four possible transition mutations with the standard nucleotides. Transversional mutations occur when a purine-pyrimidine pair is replaced by a pyrimidine-purine pair, creating eight possible transition mutations when using standard dNTPs. When creating an unbiased library, sometimes it is necessary to use two or more methods in order to allow for the same approximate percentage of transitional and transversional mutations to occur.

The use of the $MnCl_2$ and *Taq* polymerase in an error-prone PCR allows for all four transitions and all eight transversions to occur, however the A-T to T-A transition and A-T to G-C transversion tend to be more prevalent when using this technique (Vartanian et al., 1996, Lin-Goerke et al., 1997, Arnold and Georgiou, 2003b). Biases such as this can be altered by increasing or decreasing the concentrations of some of the nucleotides in the reaction. This technique can be performed on a low budget, and can be easily modified to increase or decrease the frequency of mutagenesis by altering the concentration of dNTPs or the number of PCR cycles (Arnold and Georgiou, 2003b).

Another method of creating mutagenic libraries is by rational design. The random library approach generates a large, diverse repertoire of polymerases, but a low number of active clones. Guo *et al.* has shown that at least one-third of all random amino acid changes will result in the inactivation of a protein (Guo et al., 2004), so it is likely that a protein with more than a few random amino acid changes will be inactive. Furthermore, Guo *et al.* also calculated that approximately 70% of random mutations in the active sites of polymerases will result in an inactive polymerase variant (Guo et al., 2004). A desirable library for directed evolution experiments would optimally have a large, diverse number of proteins with a high number of active clones (Hibbert and Dalby, 2005). To generate a library such as this, the reconstructing

evolutionary adaptive paths (REAP) approach can be used (Gaucher, 2006); this approach allows researches to modify only the sites where functional divergence occurred within a family of polymerases. In other words, sites that, in the historical evolution of the polymerase, had a split "conserved but different" pattern of evolutionary variation, are chosen for modification. In theory, this technique has a high probability to generate new activities and functions (Gaucher, 2006).

Systems of Directed Evolution

Some of the more common methods used in directed evolution experiments include phage display (Fa et al., 2004), ribosome display (Yan and Xu, 2006), complementation (Arnold and Georgiou, 2003a), and compartmentalized self-replication (CSR) (Ghadessy et al., 2001, Tawfik and Griffiths, 1998). Two of these techniques, phage display and CSR (Henry et al., 2004), were applied to the evolution of polymerases to increase thermostability (Ghadessy et al., 2001), permit activity in the presence of an inhibitor (Ghadessy et al., 2001), and allow incorporation of non-standard bases (Ghadessy et al., 2004, Fa et al., 2004, Xia et al., 2002). Both phage display and CSR systems have been successfully used to evolve *Taq* polymerase *in vitro* (Ghadessy et al., 2001, Ghadessy et al., 2004, Fa et al., 2004).

Phage display

The phage display directed evolution system was developed by attaching a fragment of *Taq* polymerase and an oligonucleotide primer substrate to the exterior of a phage particle via its minor phage coat protein pIII (Fa et al., 2004). Since there are approximately five of these coat proteins per phage, all localized to one area on the phage coat, researchers were able to successfully link phenotype to genotype (Fig. 1-12). The mutant polymerases were challenged to add non-standard nucleosides and one biotinylated nucleoside onto the oligonucleotide primer by template directed synthesis; those polymerases with the ability to do so were immobilized on

streptavidin beads, and were recovered. The genes encoding the active polymerases were identified by sequencing, or rediversified and shuttled into another round of selection. This technique, while excellent for identifying polymerase mutants able to incorporate a small number of non-standard bases, does not require the polymerase to perform a PCR; this would not be conducive to the design of an AEGIS based synthetic biology that requires the polymerase to replicate its own gene.

Compartmentalized self-replication

Compartmentalized self-replication makes use of water-in-oil emulsions as a way to link genotype to phenotype, and requires polymerase mutants to replicate their encoding gene in a PCR reaction (Tawfik and Griffiths, 1998, Ghadessy et al., 2004, Ghadessy et al., 2001, Williams et al., 2006), theoretically an excellent technique for developing polymerases for a synthetic biology. A library of polymerase gene variants is cloned and expressed in cells (Fig.1-13A); the bacterial cells containing the polymerases and their encoding genes are then suspended in aqueous droplets in an oil emulsion. Each of these droplets, on average, contains one cell as well as the primers and dNTPs/NSBs required for PCR (Fig. 1-13B). The thermostable polymerase is released from the cell during the first denaturing cycle of PCR, allowing replication of its encoding gene to proceed. Poorly adapted polymerases fail to replicate their encoding gene, while better-adapted polymerases succeed in replication (Fig. 1-13C). The resulting polymerase genes are then released from emulsions by extraction with ether; those encoding the most active polymerases dominate these clones. A run-off PCR using standard nucleotides prepares the DNA for recloning, which can then be subjected to another cycle of selection (Fig.1-13E).

CSR has been previously used to generate *Taq* polymerase variants that are more thermostable (Ghadessy et al., 2001), have an increased resistance to inhibitors (Ghadessy et al.,

2001), and are able to incorporate various non-standard bases (Ghadessy et al., 2004). More recently, Philipp Holliger and co-workers, who originally performed the aforementioned selections, have modified this technique to change a selected region of the polymerase sequence, and replicate that region in CSR reactions (Ong et al., 2006). This short-patch compartmentalized self-replication reaction (spCSR) has already been used to develop *Taq* polymerase variants able to function with both NTPs and dNTPs, and variants that are able to incorporate NSBs with 2'-substitutions. This technique allows the researcher to mutate only the active site of the polymerase, and then challenges the polymerase to amplify the region encoding the active site; this makes it easier for polymerases with the ability to incorporate NSBs, but who lack the catalytic efficiency and processivity, to be isolated from a pool of mutants. By reducing the stringency of the initial selections, more clones can be isolated with the desired traits; catalytic efficiency and processivity of the polymerase can be selected for later using the polymerase sequence of the desired variant under normal CSR conditions.

Research Overview

To create an AEGIS, the first step should be to create or identify polymerases with the ability to incorporate multiple, consecutive NSBs into a growing strand of dsDNA, efficiently and faithfully. Rather than challenging a polymerase with a gamut of NSBs containing different unique features, we decided to focus on one unique characteristic of AEGIS nucleosides, the C-glycosidic linkage. Previous studies have shown that polymerases have a difficult time incorporating the non-standard base pairs containing a C-glycosidic linkage (Switzer et al., 1993, Sismour et al., 2004), therefore representative C-glycosides, 2'-deoxypseudouridine (dψU) and 2'-deoxypseudothymidine (dψT), that could base pair with a canonical nucleotide, in order to decrease the strain on the polymerase, were selected for study (Lutz et al., 1999).

The research presented here began with the determination of the effect of multiple, sequential C-glycosides on duplex DNA structure, to better understand the obstacles a polymerase would have to overcome in order to incorporate bases exhibiting C-glycosides. Next, a screening of a variety of Family A and Family B polymerases, identified *Taq* as a polymerase that exhibited a limited ability to incorporate non-standard bases that contain a C-glycosidic linkage. However, further modification of the protein sequence of this enzyme was needed to identify a mutant *Taq* polymerase with an increased ability to incorporate multiple, sequential C-glycosides NSBs more efficiently.

To achieve this, the second part of this dissertation focused on the creation of a rationally designed (RD) library of 74 mutant *Taq* polymerases. Variants were screened for the ability to incorporate dψU in a PCR amplification of their encoding gene. None of these variants were shown to produce more full-length PCR product than the *wild type Taq* polymerase. Only 18 variants showed any activity at all in this first test, even with standard dNTPs, under these reaction conditions. A rationally designed library was then used to perform an initial selection, by using water-in-oil emulsions to select for the active mutant polymerases we identified in our initial screen.

It was postulated that the low number of active variants in our RD library was due to a decrease in the thermostability of the enzyme. After altering the PCR reaction conditions to test this hypothesis, we were able to identify 33 active mutant polymerases in this library. Since this library was rationally designed, it was interesting to speculate as to whether a randomly created library of polymerase clones would tend to have increased or decreased thermostability when compared to the number of active clones in our RD library. A random library (L4) was created for this purpose, and was screened for activity at various temperatures in PCR reactions; 39

clones were found to be active. This comparison of the thermostability of the two libraries shows that the randomly created library has an enhanced ability to retain polymerase thermostability when compared to our rationally designed library.

The RD library was designed to identify mutants able to incorporate non-standard bases, and not to have a high degree of thermostability. Optimal temperatures for function in a PCR were determined for each of the RD variants, and the mutants were then screened for their ability to incorporate various concentrations of dψU at that optimal temperature. One mutant in the pSW27 plasmid, containing the A597S, A740R, and E742V residue changes, was identified with the ability to generate, on average, 72% more product at all dψU concentrations tested, than *wt Taq* polymerase at a temperature of 86.3 °C.

While dψU is a C-glycoside with the ability to pair with 2'-deoxyadenosine, it has been shown to epimerize (Wellington and Benner, 2006, Cohn, 1960, Chambers et al., 1963). Since dψT cannot epimerize, due to the presence of the extra methyl group, we performed a comparative analysis between *wt Taq* polymerases' ability to cope with dψU and dψT in various concentrations and at different temperatures in a PCR. Results indicated that it may be the epimerization of the nucleotide hindering the incorporation of dψU, and therefore it should not be used as a model C-glycoside for directed evolution studies.

These results presented in this work represent a significant step towards the long-term goal of creating an AEGIS-based synthetic biology. In addition, the repertoire of mutant polymerases designed and created in these experiments will assist in creating an inventory of polymerases useful in biotechnology, possibly allowing the development of new, as well as improving on existing diagnostic techniques and helping to facilitate a better understanding of polymerase-DNA interactions.

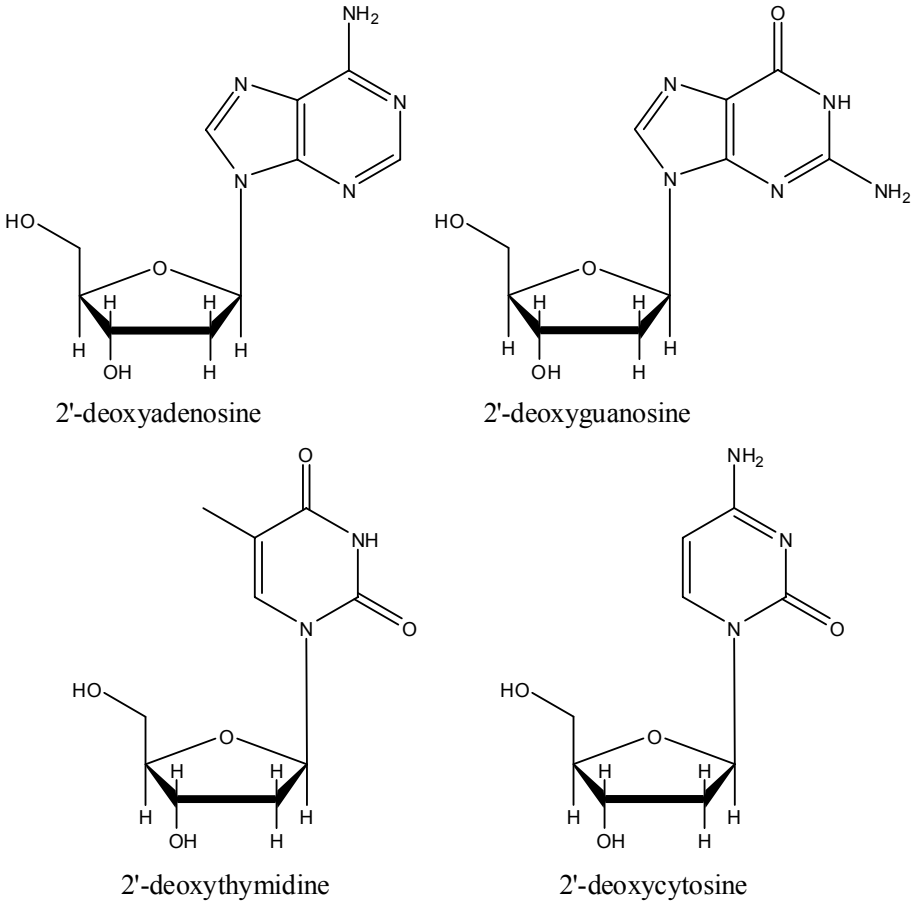


Figure 1-1. The standard deoxyribonucleotides. The nucleobases pair based on the two rules of complementarity: hydrogen-bonding complementarity, when the hydrogen bond donor from one nucleobase pairs with the hydrogen bond acceptor from another, and size complementarity, when a large purine (top row) pairs with small pyrimidine (bottom row) (Watson and Crick, 1953a, Watson and Crick, 1953b). Therefore, 2'-deoxyadenosine joins with 2'-deoxythymidine and 2'-deoxyguanosine joins with 2'-deoxycytosine. When a phosphate group replaces the 5'-hydroxyl group of these molecules, they become acids and can be linked by their phosphate groups to create a DNA strand.

Table 1-1. Comparison of the structural geometries of A, B, and Z-DNA forms.

Geometry	A-DNA	B-DNA	Z-DNA
Helical Sense	Right-handed	Right-handed	Left-handed
Helix diameter	2.6 nm	2.0 nm	1.8 nm
Repeating unit	1 base pair	1 base pair	2 base pairs
Rotation per base pair	34°	36°	60°/2
Rise per base pair	0.256 nm	0.338 nm	0.38 nm
Base pairs per turn	11	10	12
Pitch per turn of helix	2.82 nm	3.38 nm	4.56 nm
Major Groove	Very narrow and very deep	Very wide and deep	Flat
Minor Groove	Very broad and very shallow	Narrow and deep	Very narrow and deep
Sugar Pucker	$C_{3'}$-endo	$C_{2'}$-endo	C: $C_{2'}$-endo & G: $C_{2'}$-exo

*Data adapted from Saenger and Garrett & Grisham (Saenger, 1984, Garrett and Grisham, 1999).

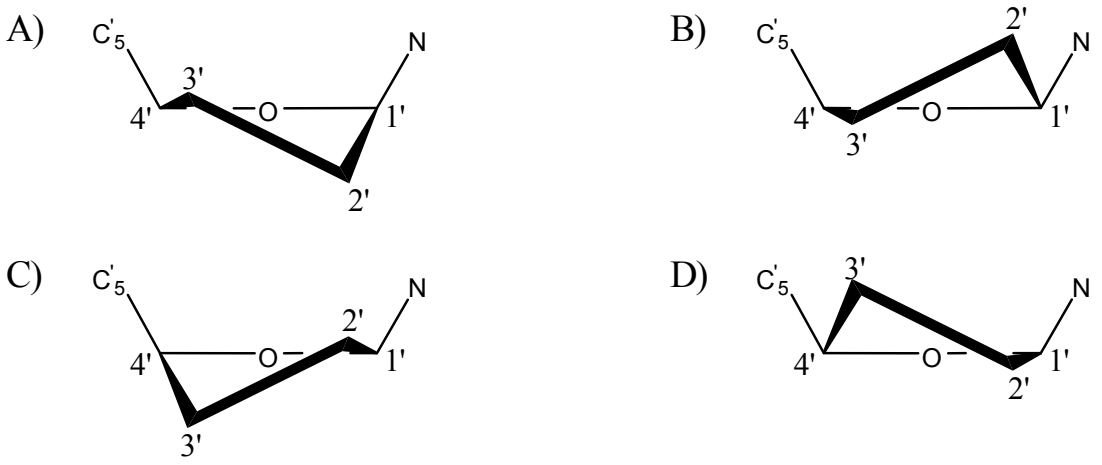


Figure 1-2. Puckering of the furanose ring of nucleosides into various envelope forms. In the envelope form, four of the five atoms are coplanar, the remaining atom departs this plane: A) a $C_{2'}$-exo sugar pucker, B) a $C_{2'}$-endo sugar pucker, C) a $C_{3'}$-exo sugar pucker, and D) a $C_{3'}$-endo sugar pucker. B-DNA has a $C_{2'}$-endo pucker, while A-DNA exhibits a $C_{3'}$-endo pucker (Saenger, 1984).

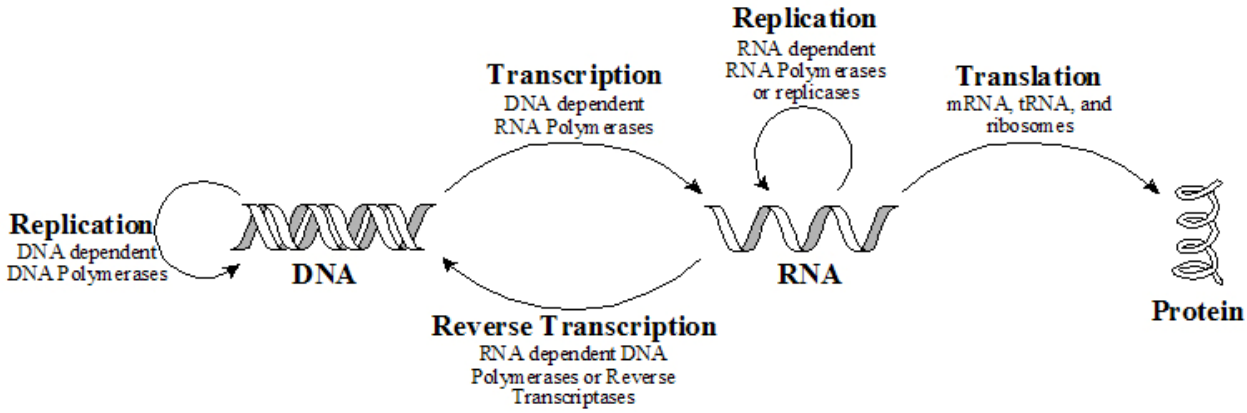


Figure 1-3. The central dogma of molecular biology (Lewin, 1997, Crick, 1970). Genetic material, in the form of DNA, is first transcribed into RNA and then is translated into proteins. On the occasion that genetic material is stored as RNA, it first undergoes reverse transcription to create DNA before it is shuttled back into the system.

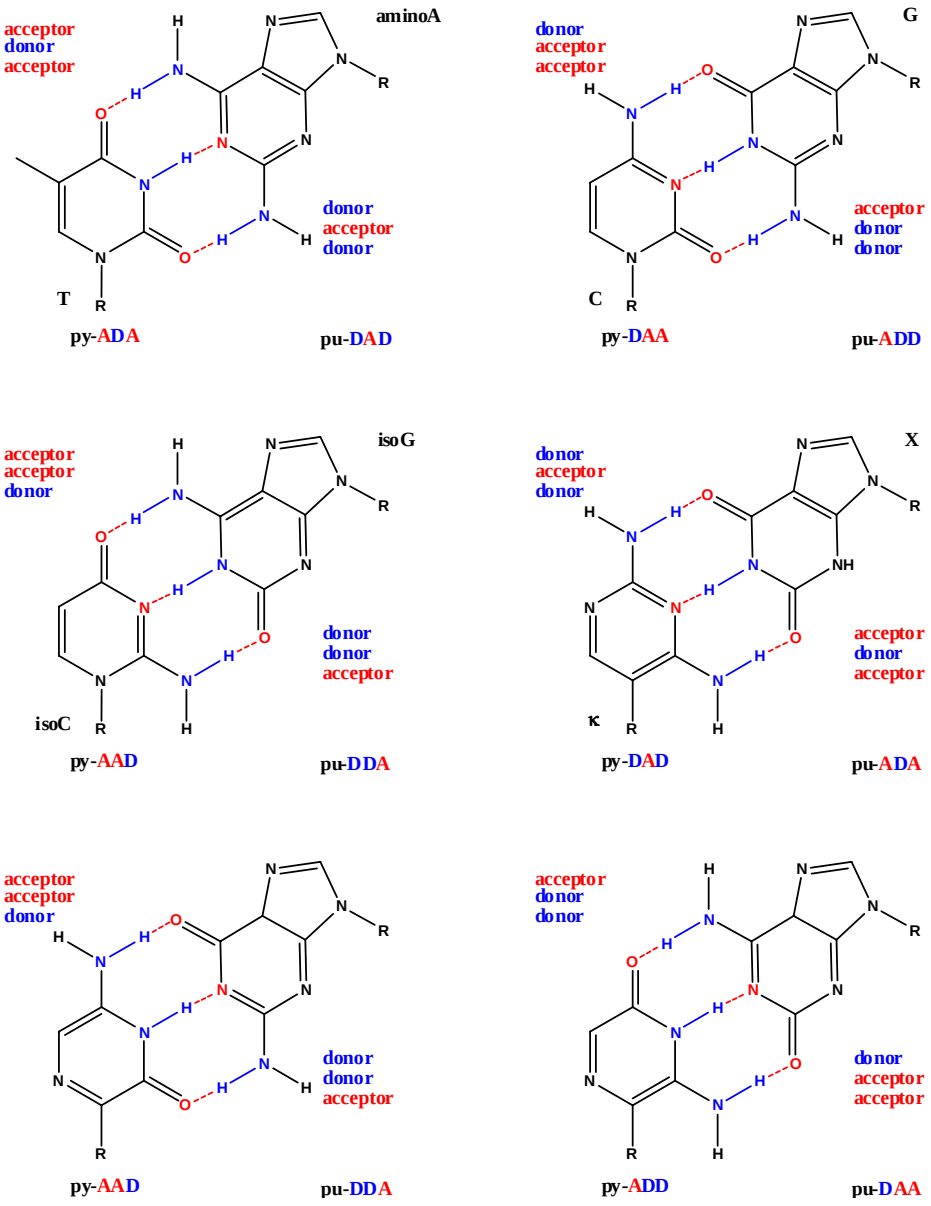


Figure 1-4. The six hydrogen bond patterns in an artificially expanded genetic information system (AEGIS). These patterns are constrained by Watson and Crick's rules of complementarity and by the requirement that the nucleobases be joined by three hydrogen bonds (Switzer et al., 1989, Piccirilli et al., 1990, Geyer et al., 2003, Benner, 2004, Watson and Crick, 1953a, Watson and Crick, 1953b). Purines are denoted by "pu," pyrimidines by "py," hydrogen-bond acceptors by "A," hydrogen bond donors by "D," and R indicates the point of attachment of the backbone. Note the presence of a C-glycosidic linkage in the pyDAD, pyADD, and pyDDA nucleotides.

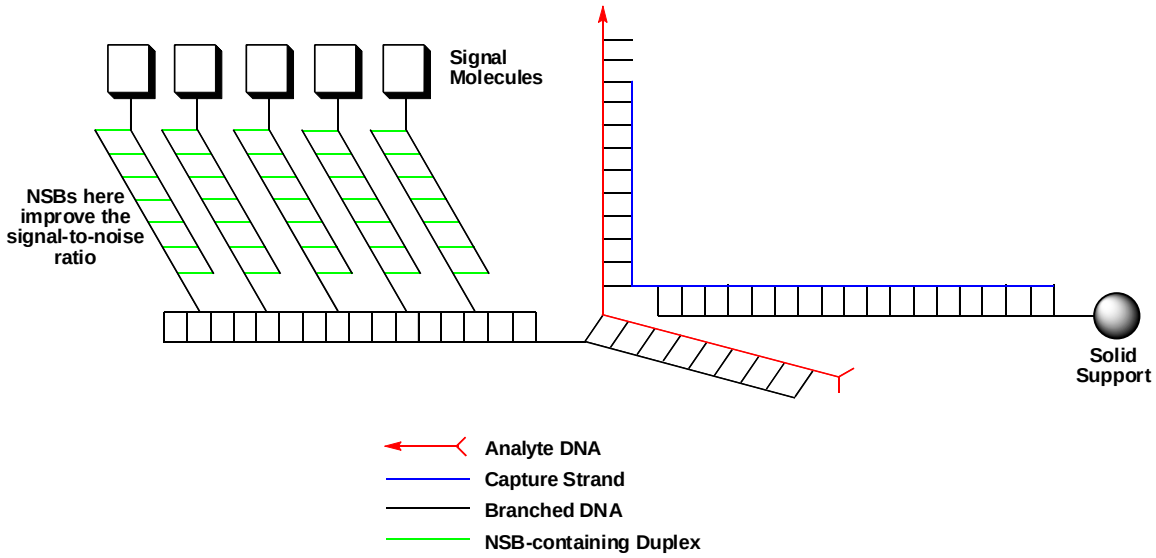


Figure 1-5. The Versant™ branched DNA assay. This assay exploits the pairing of non-standard bases (NSBs) to reduce the signal to noise ratio 8-fold over a previous version of the assay that did not use NSBs (Huisse, 2004, Collins et al., 1997). The branched DNA assay is used to monitor the viral load counts of patients with the HIV, Hepatitis B, or Hepatitis C viruses (Collins et al., 1997).

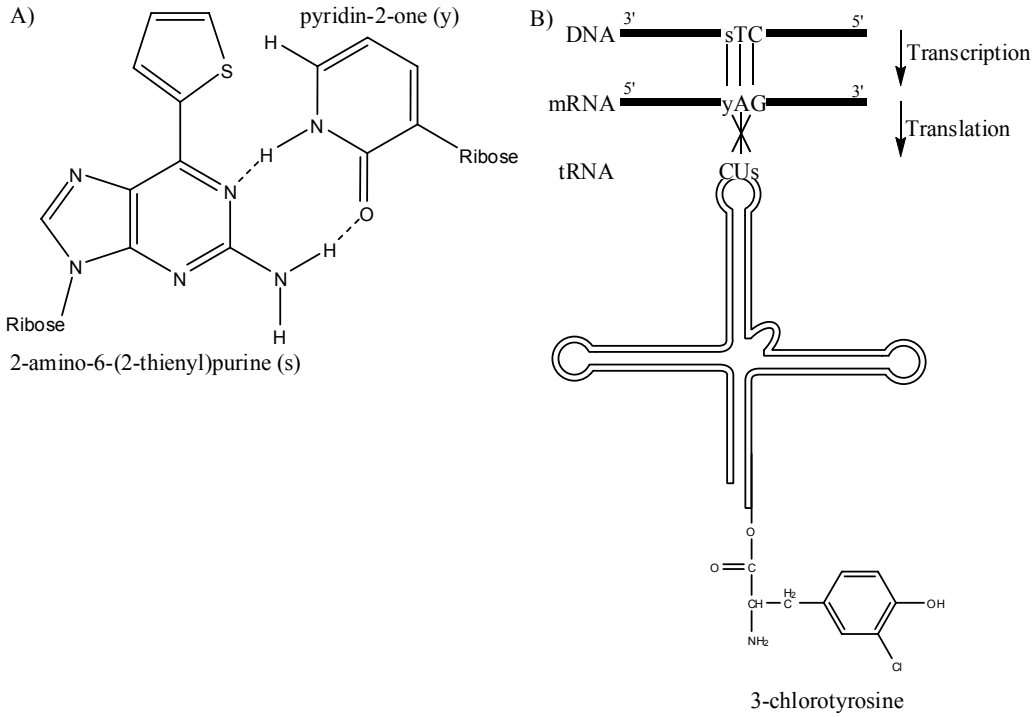


Figure 1-6. An example of non-standard nucleobases coding for a non-standard amino acid. This shows the transcription and translation (seen in B) of the non-standard base pair (seen in A and denoted as s and y) to generate a protein containing the non-standard amino acid 3-chlorotyrosine. This picture is adapted from Hirao *et al* (Hirao et al., 2002, Hirao et al., 2006).

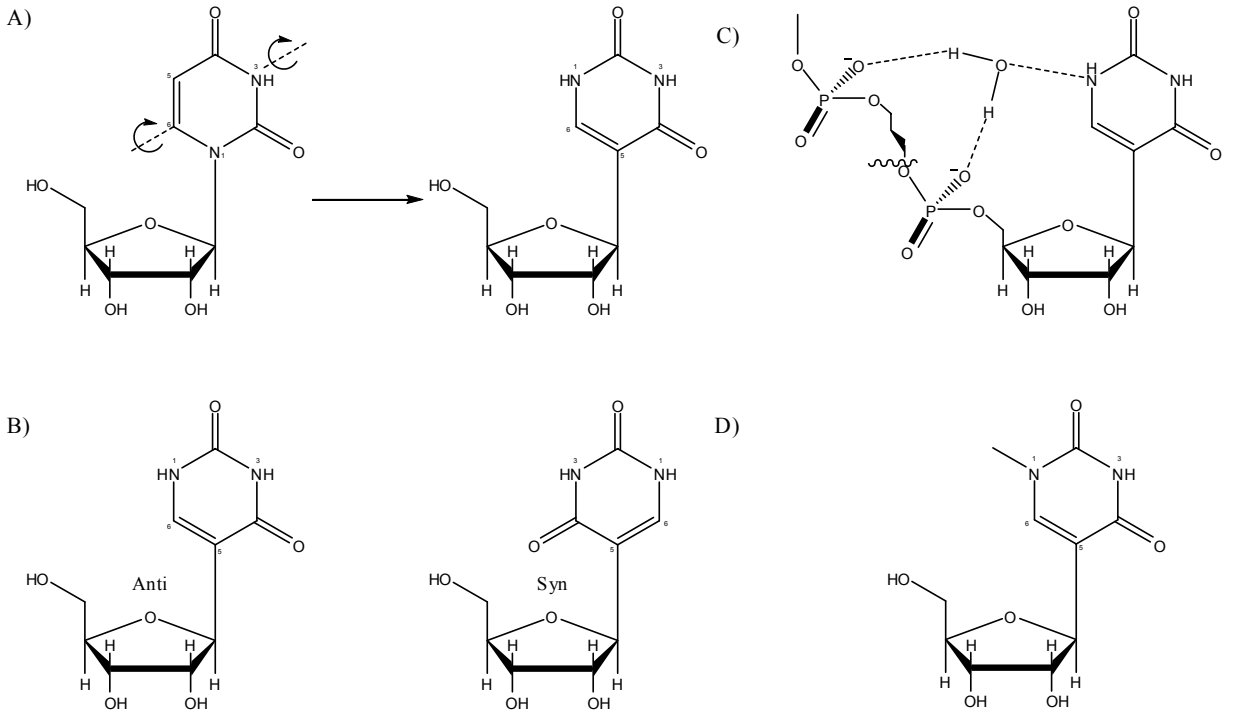


Figure 1-7. Pseudouridine and pseudothymidine. A) This naturally occurring C-glycoside, found in RNA, is thought to be created by a posttranscriptional isomerization of uridine (Argoudelis and Mizsak, 1976, Grosjean et al., 1995). B) Pseudouridine has a propensity to adopt a *syn* conformation around the glycosyl bond when in solution, but it is only found in the *anti* conformation when in a nucleic acid strand (Lane et al., 1995, Neumann et al., 1980). C) The *anti* conformation allows for the coordination of a water molecule between the 5' phosphate group of the ψU residue, the 5' phosphate group of the preceding residue, and the N1-H of the ψU residue (Arnez and Steitz, 1994). The coordination of this water molecule results in an enhanced base stacking ability and a reduced conformational flexibility of the RNA molecule, thus increasing the local rigidity of the RNA (Charette and Gray, 2000, Davis, 1995). D) The structure of pseudothymidine (1-methylpseudouridine). This naturally occurring C-glycoside, found in RNA, is also thought to be created by a posttranscriptional modification of uridine (Limbach et al., 1994).

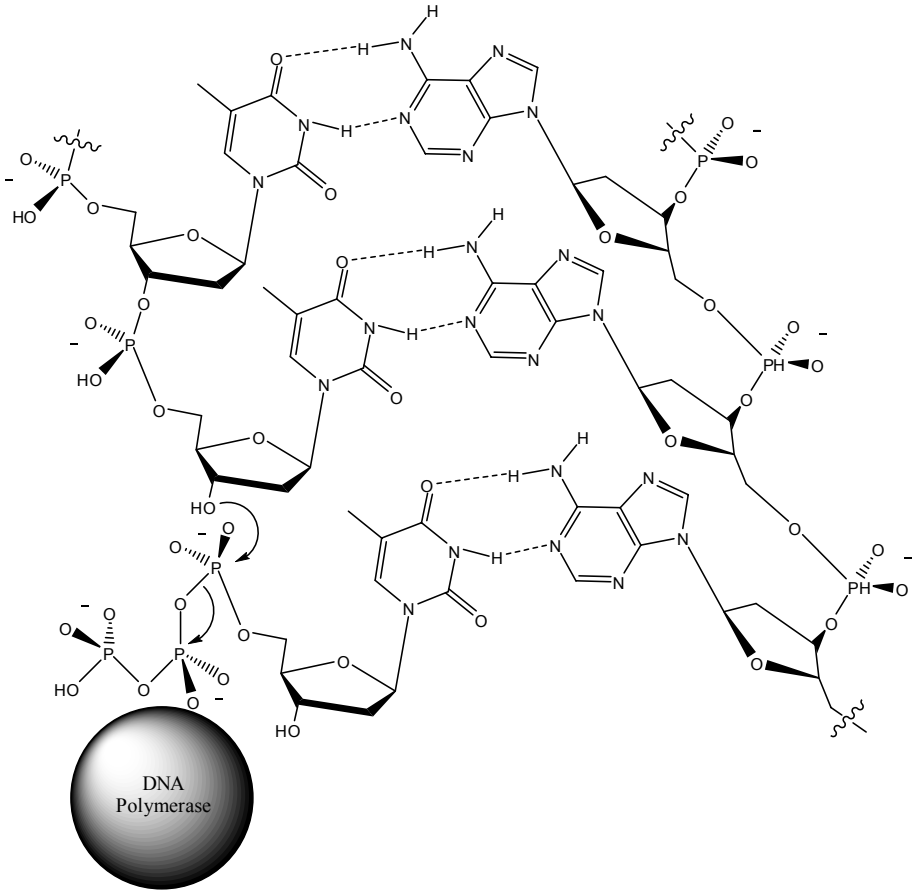


Figure 1-8. The polymerization reaction of deoxyribonucleotides triphosphates catalyzed by DNA polymerases. The triphosphate of the incoming group is linked to the 3'-hydroxyl group of the preceding nucleoside, releasing a pyrophosphate in the process; therefore DNA synthesis requires synthesis of new molecules in the 5'→3' direction (Garrett and Grisham, 1999).

$$E + TP \overset{1}{\rightleftharpoons} E\text{-}TP \overset{2}{\rightleftharpoons} E\text{-}TP\text{-}dNTP \overset{3}{\rightleftharpoons} E'\text{-}TP\text{-}dNTP \overset{4}{\rightleftharpoons} E\text{-}TP_{+1}\text{-}PP_i \overset{5}{\rightleftharpoons} E\text{-}TP_{+1}+PPi \overset{6}{\rightleftharpoons} E\text{-}TP_{+1}$$

Figure 1-9. Kinetic steps involved in the nucleotide incorporation pathway. The kinetic steps involved in the addition of a nucleotide onto a growing DNA strand (Patel and Loeb, 2001, Rothwell and Waksman, 2005). In Step 1, the polymerase (E) binds to the DNA primer:template complex (TP); the polymerase then binds the incoming nucleotide triphosphate (dNTP) in Step 2. The polymerase then undergoes a conformational change (E') in Step 3 that brings the various components into positions that can support the chemistry of this reaction; this is the rate-limiting step of polymerization. The polymerase performs the addition of the nucleotide, remains complexed with the pyrophosphate, and undergoes another conformational change in Step 4. The pyrophosphate group is released in Step 5; in Step 6, the polymerase can dissociate from the DNA or translocate the substrate for another round of synthesis.

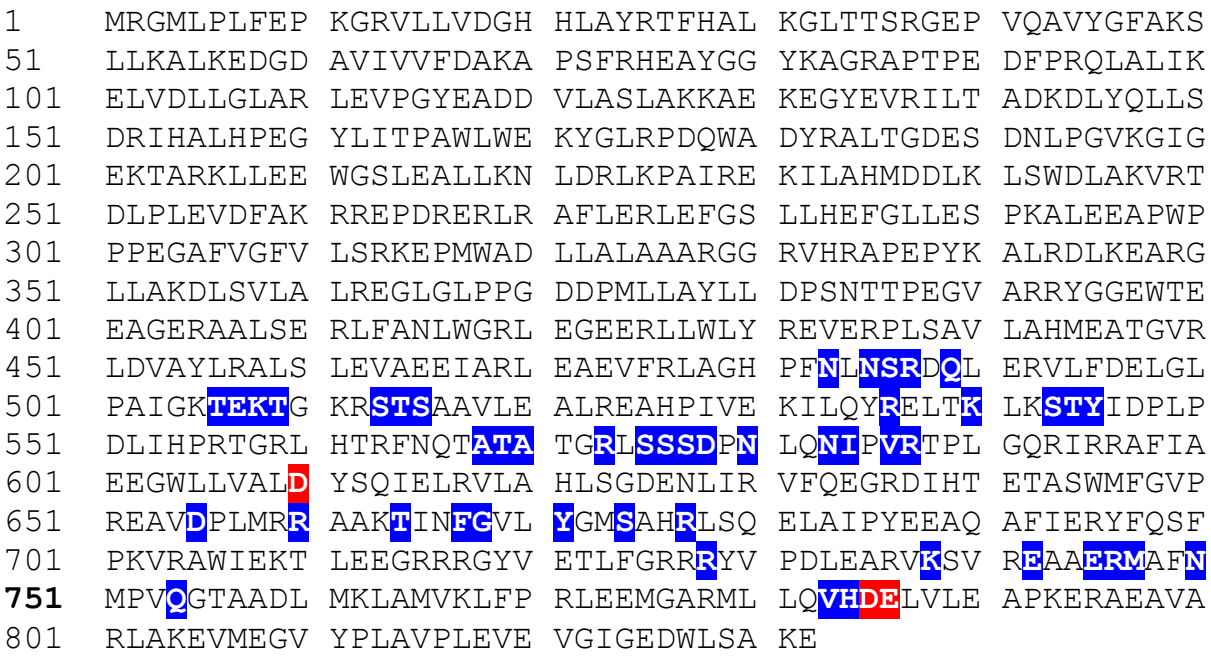


Figure 1-10. Locations of active site residues in *Taq* polymerase. Residues shown in blue are involved in contacting the DNA during polymerization; those shown in red indicate residues involved in metal ion coordination (Eom et al., 1996, Fa et al., 2004, Li et al., 1998b, Li et al., 1998a, Kim et al., 1995, Suzuki et al., 1996).

Table 1-2. Characteristics of the various polymerase families.

Feature	**A**	**B**	**C**	**D**	**X**	**Y**	**RT**
Domains Containing Polymerase	Prokaryotes, Eukaryotes, Viruses	Prokaryotes, Eukaryotes, Archaea, Viruses	Prokaryotes	Archaea	Eukaryotes	Prokaryotes, Eukaryotes, Archaea	Eukaryotes, Viruses
Representative Polymerases	*E. coli* DNA Pol I; *Taq* Pol I; T7 DNA Pol	*Pfu* DNA Pol I; Eukaryotic DNA Pol a	*E. coli* Pol III(a)	*Pfu* DNA Pol II	Eukaryotic DNA Pol b	*E. coli* DNA Pol IV; *E. coli* DNA Pol V	HIV-RT; M-MuLV-RT; Eukaryotic telomerases
General Use	Repair	Replicative	Replicative	Replicative	Repair	Replicative/Repair	Replicative
Fidelity	Good	Excellent	Excellent	Excellent	N/A	Poor	Good

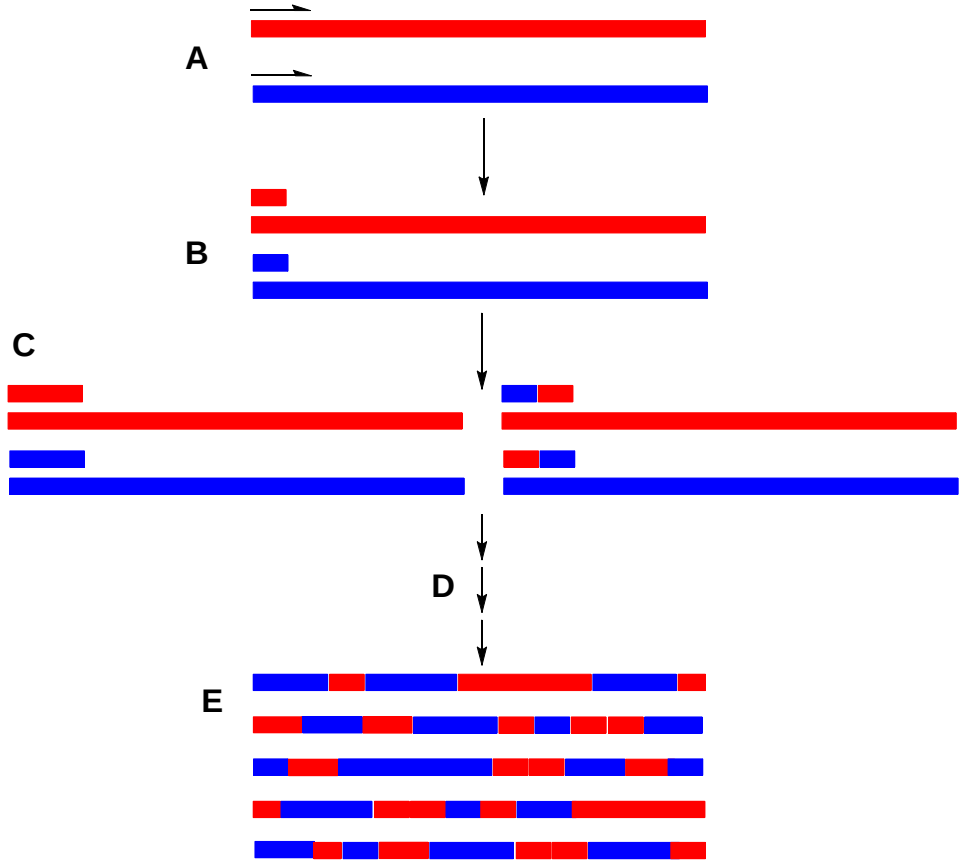


Figure 1-11. The staggered extension process (StEP) for rediversification of mutant libraries. This process has already been successfully used to rediversify libraries between rounds of selection in CSR reactions (Arnold and Georgiou, 2003b, Zhao et al., 1998, Ghadessy et al., 2001). A) Denatured template genes are primed with the same primer. B) Short fragments are produced by brief primer-extension. C) In the next cycle, fragments randomly prime the templates and extend further. D) This process is repeated until full-length genes are produced. E) Full-length genes are then purified, amplified, and recloned into a vector for another round of selection.

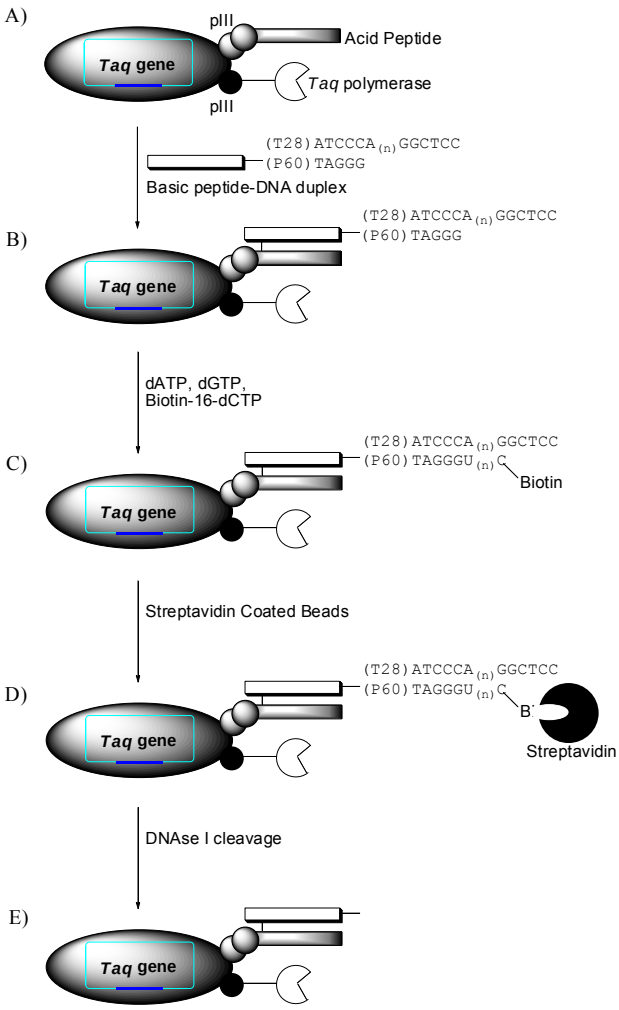


Figure 1-12. Phage display selection scheme. This details the scheme used in the directed evolution of a *Taq* polymerase fragment to incorporate non-standard nucleosides into a growing DNA strand (Fa et al., 2004). A) A phage particle is displaying an acidic peptide and a mutant polymerase on the pIII minor coat protein of the phage. These coat proteins are localized to one area on the phage molecule, allowing genotype to be linked to phenotype. B) The primer-template complex is attached to the phage particle via a basic peptide, which links with the acidic peptide displayed on the coat protein. C) The polymerase incorporates modified nucleotides in a primer-extension assay, which terminates with the addition of a biotinylated standard nucleotide. D) The biotin tag is captured by streptavidin and the entire complex is immobilized on magnetic beads, allowing those phage particles displaying inactive polymerases to be washed away. E) DNase I is used to dissociate the phage complex from the DNA strands, allowing the phage displaying the active polymerase to be captured in an elution. The genes encoding the active polymerases can then be identified by sequencing and/or rediversified and shuttled into another round of selection.

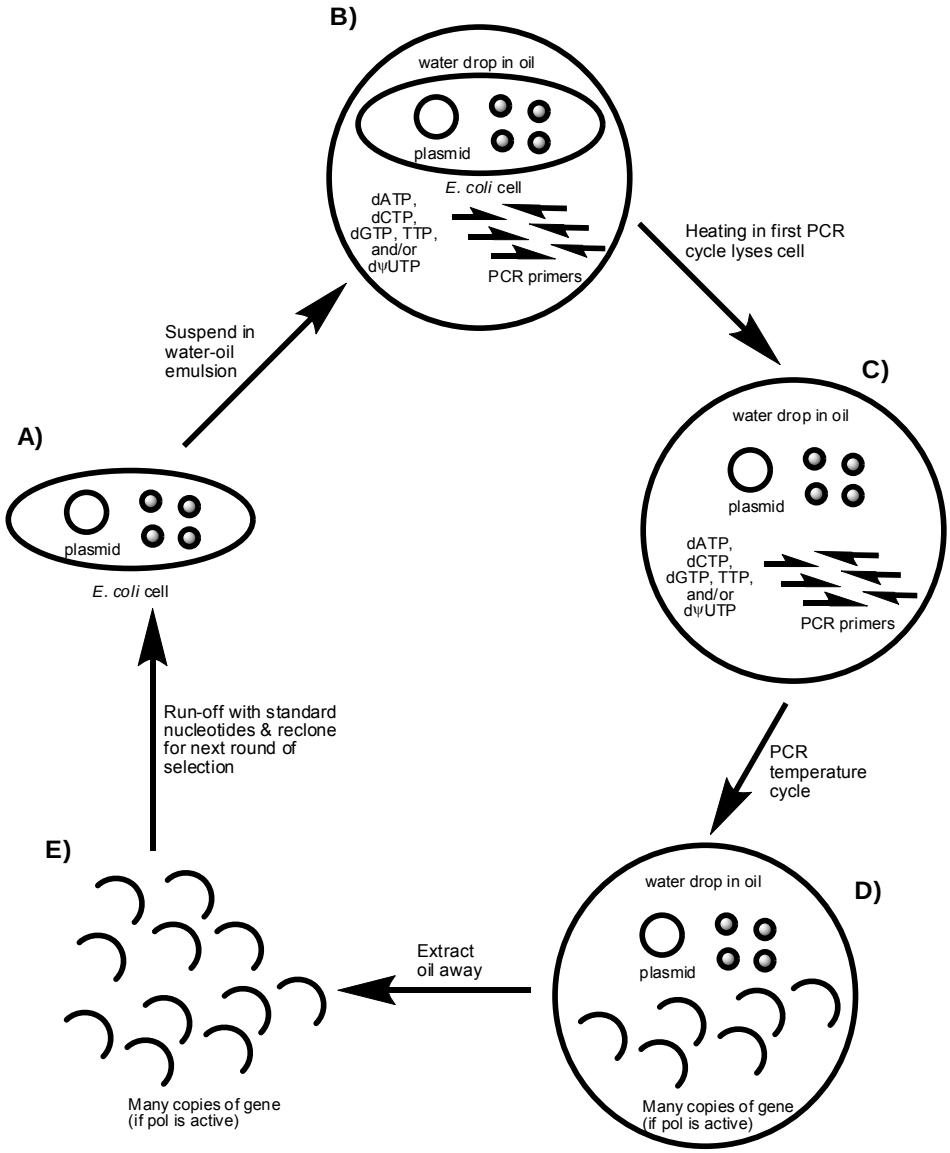


Figure 1-13. General scheme for CSR. CSR allows for the selection of polymerases with an ability to incorporate an unnatural nucleotide using water-in-oil emulsions.) A library of polymerase gene variants is cloned and expressed in *E. coli*. Spheres represent active polymerase molecules inside of a bacterial cell. B) The bacterial cells containing the polymerases and their encoding genes are suspended in aqueous droplets in an oil emulsion. C) The thermostable polymerase enzyme and encoding gene are released from the cell during the first denaturing cycle of PCR, allowing self-replication to proceed. D) The resulting mixture of polymerase genes is released by extraction with ether. E) A single run-off PCR with standard nucleotides prepares the DNA for recloning and another cycle of selection.

CHAPTER 2
POLYMERASE INCORPORATION OF MULTIPLE C-GLYCOSIDES INTO DNA: PSEUDOTHYMIDINE AS A COMPONENT OF AN ALTERNATIVE GENETIC SYSTEM

Introduction

Each of the four standard nucleobases found in natural DNA (adenine, guanine, cytosine, and thymine) is joined to their sugar via a carbon-nitrogen bond. This, by definition, makes standard nucleotides N-glycosides. The nature of the glycosidic linkage is believed to have consequences on the detailed conformation of the nucleoside, including through the operation of the anomeric effect. In particular, the nature of the glycosidic bond may influence the puckering of the sugar.

Unlike the standard nucleotides, the nucleotides that allow artificially expanded genetic information systems (AEGIS) to be created are frequently C-glycosides, which have a carbon-carbon bond between the nucleobase and the sugar. This is exemplified in the case of non-standard pyrimidines that present Donor- Donor-Acceptor, Donor-Acceptor-Donor and Acceptor-Donor-Donor hydrogen bonding patterns seen in Figure 1-4. If replacing the N-glycosidic linkage by a C-glycosidic linkage changes features of the nucleoside that are important specificity determinants for polymerases, problems are created for those seeking to expand the genetic alphabet artificially and develop a synthetic biology from an expanded genetic alphabet.

Reverse transcriptases have an ability to process both DNA and RNA, whose sugars have different conformations. Reverse transcriptases, therefore, should be able to accept components of an artificially expanded genetic information system that incorporate C-glycosides. Perhaps it is not surprising that the first reported example of PCR amplification of a six letter genetic alphabet, where one the extra two letters was a C-glycoside, exploited HIV-RT (Sismour et al., 2004).

When attempting to develop a synthetic biology using C-glycosides, the physical structure of the DNA must be considered, especially since the presence of multiple, sequential C-glycosides can possibly alter the structure and stability of duplex DNA. Previous studies have shown that poly(U)●poly(A) helices favor the A-DNA form while poly(T)●poly(A) helices display perfect B-DNA structure (Ivanov et al., 1973, Saenger, 1984, Chandrasekaran and Radha, 1992). Circular dichroism was employed to infer the secondary structure of our DNA, since the spectra generated by A-DNA and B-DNA are quite different (Fig. 2-1) (Ivanov et al., 1973). Duplex DNA containing one to twelve consecutive dA-dψU base pairs was studied and it was determined that all remained in the B-DNA form.

To take the next step towards a synthetic biology with an expanded genetic alphabet, it would be desirable to have DNA polymerases that accept multiple C-glycoside nucleotides. To determine whether natural DNA polymerases have this capability and the extent to which this capability is conserved, four Family A DNA polymerases and four Family B DNA polymerases were screened for their ability to incorporate multiple 2'-deoxypseudothymidine-5'-triphosphate (dψTTP) and 2'-deoxypseudouridine-5'-triphosphate (dψUTP) across from template dA. These C-glycosides are steric analogs of thymidine-5'-triphosphate (TTP) and present the same hydrogen bonding pattern to a complementary strand as TTP (Fig. 2-2). Consequently, they should serve as a relatively specific probe for this non-standard structural feature.

In these experiments, all of the polymerases tested were able to incorporate both C-glycosides to an extent; but there was room for improvement in some, such as *Taq*. To determine the extent of *Taq* polymerase's ability to incorporate the C-glycosides, it was screened for its ability to incorporate anywhere from one to twelve consecutive dψTTP or dψUTP across from template dA.

Materials and Methods

Synthesis of Triphosphates and Oligonucleotides

Dr. Shuichi Hoshika, from the Foundation for Applied Molecular Evolution (FfAME, Gainesville, Florida), synthesized the pseudothymidine precursor as described in Appendix A. Dr. Daniel Hutter (FfAME) synthesized 2'-deoxypseudothymidine-5'-triphosphate (dψTTP) as described in Appendix A. 2'-Deoxypseudouridine-5'-triphosphate (dψUTP) was purchased from TriLink BioTechnologies (San Diego, California). Standard deoxynucleotide triphosphates (dNTPs) of 2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxycytidine-5'-triphosphate (dCTP), 2'-deoxyganosine-5'-triphosphate (dGTP), and thymidine-5'-triphosphate (TTP) and were purchased from Promega Corporation (Madison, Wisconsin). Triphosphate solutions identified as dψTNTPs were comprised of dATP, dCTP, dGTP, and dψTTP, while those acknowledged as dψUNTPs were contained dATP, dCTP, d GTP, and dψUTP.

The oligonucleotides used for these experiments are listed in Table 2-1. Those sequences containing only standard nucleotides were commercially obtained from Integrated DNA Technologies (Coralville, Iowa) as desalted or PAGE (Polyacrylamide Gel Electrophoresis) purified oligonucleotides. Those oligonucleotides containing dψU were synthesized by Dr. Ajit Kamath (University of Florida, Gainesville, Florida) and were prepared using standard monomers and reagents (Glen Research, Sterling, Virginia) on an Expedite 8909 DNA Synthesizer (PerSeptive Biosystems, Inc., Framingham, Massachusetts). The crude products were digested, with agitation, in 1 mL of concentrated ammonium hydroxide at 55 °C for 16 hrs to release and deprotect the oligonucleotide (Sambrook et al., 1989). The mixtures were briefly centrifuged and the supernatants were passed through 2 μm cellulose acetate syringe filters. The residual products were washed three times with 1 mL portions of sterile water. The combined

filtrates were lyophilized to dryness and were purified by polyacrylamide gel electrophoresis (PAGE) and isolated by reversed-phase chromatography on a silica gel as described previously (Sambrook et al., 1989).

Circular Dichroism

Each template, containing one through twelve consecutive dA or dψU residues (T-13 through T-22 or T-23 through T-34, respectively), was annealed to its complement template, containing consecutive dT or dA residues (T-35 through T-46 or T-47 through T-58, respectively). Reactions contained 5 nmol of each template and 290 μL of CD buffer (1 M NaCl, 10 mM Na_2HPO_4, 1 mM Na_2EDTA at pH 7.0) for a total volume of 300 μL. The mixtures were incubated for 5 min at 96 °C and allowed to cool to room temperature over the course of 1 hr.

The CD spectra from 200 to 320 nm, using a wavelength step of 1 nm, were measured in a nitrogen atmosphere at 25 °C in a 0.1 cm pathlength cuvette, using an Aviv Model 215 Circular Dichroism Spectrometer (Proterion Corporation, Inc., Piscataway, NJ). Scans were performed in triplicate for each sample mixture and the data was averaged.

Standing Start Primer-Extension Assays

Radiolabeled primer was prepared by incubating 0.5 nmol P-1, 100 μCi γ-^{32}P-ATP, 1X T4 Polynucleotide Kinase (PNK) Buffer, 50 U T4 PNK (New England BioLabs, Beverly, Massachusetts), and sterile dH_2O in a final volume of 100 μL, for 1 hr at 37 °C. The radiolabeled primer was purified using the QIAquick Nucleotide Removal Kit (Qiagen, Valencia, California) and eluted from the column in 100 μL Buffer EB (10 mM Tris-HCl, pH 8.5).

Radiolabeled template, to depict the location of full-length product (FLP), was prepared by incubating 50 pmol T-4, 10 μCi γ-^{32}P-ATP, 1X T4 PNK Buffer, 25 U T4 PNK, and sterile dH_2O in a final volume of 50 μL, for 1 hr at 37 °C. The radiolabeled T-4 was purified using the QIAquick Nucleotide Removal Kit, and eluted from the column in 50 μL Buffer EB. 200 μL DNA PAGE Loading Dye (98% formamide, 10 mM EDTA, 1 mg/mL xylene cyanol, and 1 mg/mL bromophenol blue) was added to the 1 μM radiolabeled T-4 for a final concentration of 0.2 μM radiolabeled T-4.

Radiolabeled 10 base-pair (bp) ladder was prepared by first incubating 1.95 μg 10 bp DNA Step Ladder (Promega Corporation), 30 μCi γ-^{32}P-ATP, 1X T4 PNK Buffer, and sterile dH_2O in a final volume of 27 μL, for 1 min at 90 °C. Immediately following, 30 U T4 PNK was added and the mixture was incubated for 30 min at 37 °C. The radiolabeled 10 bp ladder was purified using the QIAquick Nucleotide Removal Kit, and eluted from the column in 30 μL Buffer EB. 120 μL DNA PAGE Loading Dye was added to the 65 ng/μL radiolabeled 10 bp DNA Ladder for a final concentration of 13 ng/μL radiolabeled 10 bp DNA Ladder.

Polymerase screen primer-extension assays

Klenow Fragment (3'→5' exo-), *Bst* DNA Polymerase (Large Fragment), *Taq* DNA Polymerase, Vent$_R$® (exo-) DNA Polymerase, Deep Vent$_R$® (exo-) DNA Polymerase, and Therminator™ DNA Polymerase were purchased from New England BioLabs. *Tth* DNA Polymerase was purchased from Promega Corporation. *Pfu* (exo-) DNA Polymerase was purchased from Stratagene (La Jolla, California). Buffers used in these experiments were supplied by the manufacturer as follows: reactions using *Bst*, *Taq*, *Tth*, Vent (exo-), Deep Vent (exo-), and Therminator were performed in 1X ThermoPol Buffer (20 mM Tris-HCl (pH 8.8), 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton X-100); Klenow (exo-) reactions were

performed in 1X NEBuffer 2 (10 mM Tris-HCl (pH 7.9), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM dithiothreitol); and reactions using *Pfu* (exo-) were performed in 1X Cloned *Pfu* Buffer (20 mM Tris-HCl (pH 8.8), 2 mM $MgSO_4$, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton X-100, 0.1 mg/mL nuclease-free Bovine Serum Albumin). Optimal temperatures for polymerase function were 37 °C for Klenow (exo-), 65 °C for *Bst*, and 72 °C for *Taq*, *Tth*, Vent (exo-), Deep Vent (exo-), *Pfu* (exo-), and Therminator.

T-4 Primer-Template complex was prepared by mixing 25 pmol radiolabeled P-1, 200 pmol non-radiolabeled P-1, and 300 pmol non-radiolabeled T-4, in a final volume of 15 μL. The mixture was incubated for 5 min at 96 °C and allowed to cool to room temperature over the course of 1 hr.

For primer-extension assays, 1.5 μL of the primer-template complex, 1X of the appropriate manufacturer's supplied buffer, 1 U/μL of the appropriate polymerase, and sterile dH_2O were used in a final volume of 9 μL. Reactions were then incubated at the appropriate temperature for 30 s. Each reaction was initiated by adding 1 μL of one of the following: 1 mM dTTP, 1 mM dψTTP, 1 mM dψUTP, 1 mM dNTPs, 1 mM dψTNTPs, or 1 mM dψUNTPs, and incubated for two more minutes at the appropriate temperature. Reactions were immediately quenched with 5 μL of DNA PAGE Loading Dye. Samples (1 μL) were resolved on denaturing PAGE gels (7 M Urea and 20% 40:1 acrylamide: bisacrylamide) and analyzed on a Molecular Imager FX System (Bio-Rad, Hercules, California).

Taq polymerase primer-extension assays

Primer-Template complexes were prepared by mixing 25 pmol radiolabeled P-1, 200 pmol non-radiolabeled P-1, and 300 pmol of non-radiolabeled template (T-1 through T-12), in a final

volume of 15 μL. The mixtures were incubated for 5 min at 96 °C and allowed to cool to room temperature over the course of 1 hr.

For primer-extension assays, 1.5 μL of the appropriate primer-template complex, 1X ThermoPol buffer, 1 U/μL *Taq* Polymerase, and sterile dH_2O were used in a final volume of 9 μL. Reactions were then incubated at 72 °C for 30 s. Each reaction was initiated by adding 1 μL of one of the following: 1 mM dNTPs, 1 mM dψTNTPs, or 1 mM dψUNTPs, and incubated for two more minutes at 72 °C. Reactions were immediately quenched with 5 μL of DNA PAGE Loading Dye. Samples (1 μL) were resolved on denaturing PAGE gels (7 M Urea and 20% 40:1 acrylamide: bisacrylamide) and analyzed on a Molecular Imager FX System (Bio-Rad).

Results

Circular Dichroism

Duplexes were formed by annealing each template (T-13 through T-34) to its complement sequence (T-35 through T-58) creating twelve control helices containing only thymidine and twelve helices containing pseudouridine. Figure 2-3[A-E] shows a representative set of these spectra, specifically the spectra of duplexes containing 1, 3, 6, 9, or 12 A-ψU base pairs. When compared to the spectra seen in Figure 2-1, all spectra are consistent with B-DNA being the overall conformation of all duplexes. In addition, the spectra representing the oligonucleotides containing the dA-dψU base pairs are similar to the patterns of the spectra containing the dA-dT base pairs.

Polymerase Screen Primer-Extension Assays

Four Family A and four Family B polymerases were screened for their ability to incorporate non-standard bases exhibiting a C-glycosidic linkage with efficiency. Polymerases were tested in both 4-base and 13-base extension assays, and were challenged to incorporate (4-

bases) or incorporate and extend beyond (13-bases) four consecutive dT, dψT, or dψU residues across from template dA under the polymerases' optimal conditions (Fig. 2-4). Reactions used TTP, dψTTP, or dψUTP in the 4-base extensions and either dNTPs, dψTNTPs, or dψUNTPs for the 13-base extension reactions. Family A polymerases (Fig. 2-5[A-B]) were represented by Klenow (exo-), *Bst*, *Taq*, and *Tth*; Family B polymerases (Fig. 2-6[A-B]) were represented by Vent (exo-), Deep Vent Exo-, *Pfu* (exo-), and Therminator.

Pfu (exo-) was the only polymerase that was not able to generate FLP when challenged to incorporate and extend beyond both of the non-standard bases. All other Family A and Family B polymerases were able to incorporate the four consecutive non-standard bases (NSBs) and extend beyond them, to some measure, to generate FLP. *Bst* and Therminator polymerases appeared to have consumed almost all of the primer in the course of their reactions, generating large amounts of FLP, with all of the different NTPs tested. Klenow (exo-) and Vent (exo-) also did an exceptional job at incorporating the NSBs, but the remainder of the polymerases did appear to have difficulty given the intensity of the pause sites relative to the intensity of the FLP bands.

Taq Polymerase Primer-Extension Assays

To replicate its own encoding polymerase gene, *Taq* polymerase would be required to incorporate and extend beyond four consecutive dT, dψT, or dψU residues. In these experiments, *Taq* polymerase was challenged to incorporate and extend beyond twelve consecutive dT/dψT/dψU residues opposite template dA. From these results (Fig. 2-7[A-B]), it was determined that *Taq* appears to have some difficulty incorporating twelve consecutive dT residues, as evidenced by the pausing in those lanes, but it is still able to generate FLP (N+13). It is also apparent that *Taq* has difficulty incorporating multiple consecutive residues of C-

glycosides, since it was not able to generate FLP when forced to incorporate five or more dψT or dψU residues. However, it does, generate a small amount of FLP when challenged to insert four consecutive dT, dψT, or dψU residues, and therefore should be able to replicate its own gene using a C-glycoside substitute for TTP.

Discussion

It was first necessary to determine if the presence of multiple dψU residues in double-stranded DNA would perturb the helical structure to a point where there is a phase transition from B-DNA to A-DNA, perhaps making it difficult for polymerases to replicate the DNA. It is well known that poly(U)●poly(A) favors the A-helicies, while poly(T)●poly(A) favors B-DNA helicies (Ivanov et al., 1973, Saenger, 1984, Chandrasekaran and Radha, 1992).

The distinctive differences in the CD between the canonical A-duplex and the canonical B-duplex structures involves a shift of the positive potion of the spectrum to shorter wavelengths, to 267 nm for the A-form compared to 275 nm for the B-form (Ivanov et al., 1973). A similar shift with a similar magnitude is seen in the negative portion. Further, the Q-DNA shows a stronger Cotton effect than the B-DNA. Therefore, to determine whether the addition of C-glycosidic units tends to drive the conformation of the duplex from B towards A, we look for an increase in the Cotton effect and a shift towards shorter wavelengths.

Circular dichroism was performed on 24 duplex DNA molecules containing anywhere from one to twelve consecutive dψU●dA or dT●dA base pairs. The observed spectra (Fig. 2-3[A-E]) were compared to those in Figure 2-1, the reference spectra for canonical A and B duplexes. In all spectra containing dψU, the wavelength shifted marginally (ca. 4 nm) towards longer wavelengths. This shift does not display a trend, however. The shift is the same no matter how many dψU units are incorporated into the strand.

The only possible trend is a change in the relative intensity of the positive (at 275 nm) and negative (at 264 nm) band intensities (Ivanov et al., 1973). Here the intensity of the 246 nm band and the 275 nm band both decrease. As concentrations were carefully controlled, we do not believe that this reflects a change in the concentration of the oligonucleotides. This is also suggested by the intensity of signals at lower wavelengths, although these are notoriously compromised by any trace of impurity. Disregarding this detail, the trend is the opposite of what one expects for the conversion of the duplex structure from canonical B to canonical A.

These results provide no evidence that addition of dψU units causes the duplex structure to change from a B-DNA to an A-DNA conformation. Thus, there was no evidence to suggest that there would be a conformational problem with the duplex structure when incorporating multiple, sequential C-glycosides. It should be mentioned, however, that CD is indicative only of the gross properties of the system; it does not provide information about detailed structure. It is conceivable that the conformation is changed in a different way, or some subtly.

Nevertheless, these results encouraged us to test polymerases for their ability to work with C-glycosides. Polymerases that already display some of the desired catalytic activity, in this case the incorporation of the C-glycosides, should facilitate in the evolution and/or creation of an AEGIS. Previous studies have shown that polymerases are able to incorporate up to three C-glycosides, but have not tested their ability to incorporate more than three multiple, sequential C-glycosides that would be required for an AEGIS (Lutz et al., 1999, Sismour et al., 2004, Piccirilli et al., 1991). Accordingly, four Family A polymerases, Klenow (exo-), *Bst* , *Taq*, and *Tth*, and four Family B polymerases, Deep Vent (exo-), Vent (exo-), *Pfu* (exo-), and Therminator, were screened for the ability to incorporate TTP, d ψTTP, and dψUTP across from template dA in both 4-base and 13-base primer extension assays (Fig. 2-5[A-B] and Fig. 2-6[A-B]). In the 4-

base extension assay, polymerases were challenged to incorporate four consecutive TTP, dψTTP, or dψUTP across from template dA during two-minute incubations at the optimal temperature for each enzyme. The 13-base assay, incubated as described above, took place in the presence of dCTP, dGTP, dATP, and TTP, dψTTP, or dψUTP, and required incorporation and extension beyond the four consecutive TTP, dψTTP, or dψUTP.

The *Bst* and Therminator polymerases appeared to have worked extremely well and consumed almost all of the primer in the course of all of their reactions, while *Pfu* (exo-) did not appear to generate any 13-base FLP when presented with either of the two NSBs. All other polymerases generated varying amounts of FLP with both of the NSBs, suggesting that any of the aforementioned polymerases could be potential candidates for adaptation to an AEGIS, based on the qualification that the polymerase must already be able to incorporate C-glycosides. However, two of these polymerases, Klenow (exo-) and *Bst* are not thermostable, and thus could not undergo PCR and, according to the manufacturer, Therminator is not recommended for any applications except DNA sequencing and primer-extension reactions, thereby making these three polymerases unlikely candidates for future studies. As in previous studies, *Taq* was selected as the best polymerase candidate to undergo further testing since it so readily accepted the consecutive non-standard bases (Lutz et al., 1999).

In an AEGIS system, a polymerase would be required to replicate its own encoding gene with efficiency and fidelity. In order for *Taq* to replicate its encoding polymerase gene, it would be required to incorporate four consecutive dψT or dψU across from template dA. Since we have already shown that *Taq* can in fact incorporate and extend beyond four consecutive C-glycosides, we next tested its ability to incorporate and extend beyond up to twelve consecutive dψT-dA or dψU-dA base pairs. Primer extension experiments were performed under optimal

polymerase conditions using templates T-1 through T-12. Based on the results of the study (Fig. 2-7[A-B]), *Taq* polymerase will not readily incorporate and extend beyond more than five consecutive C-glycosides to generate FLP. If this polymerase is to be used as a potential candidate for an AEGIS system, it must be modified, possibly by directed evolution experiments, so that it can incorporate more of these non-standard bases.

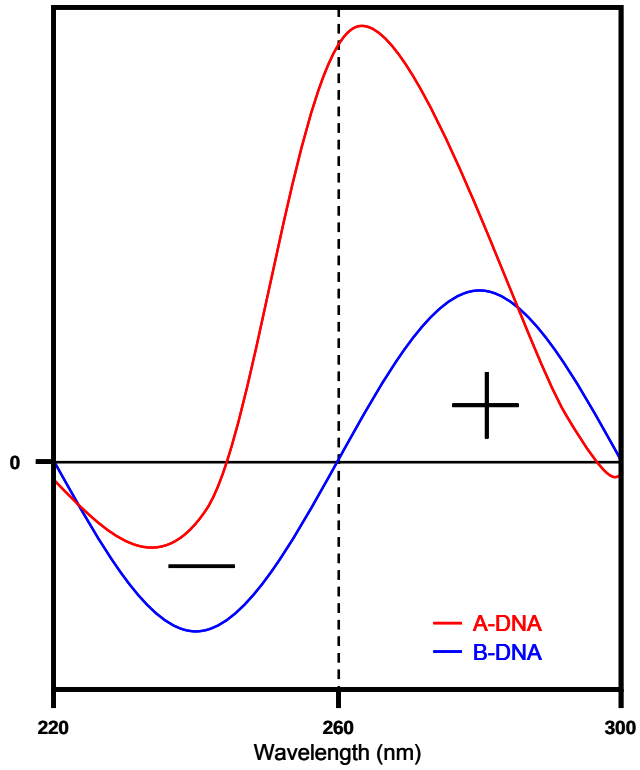


Figure 2-1. A schematic representation of the CD spectra of A- and B-DNA forms. The dotted line indicates the position of the absorption maxima (adapted from Ivanov *et al.*, 1973 (Ivanov et al., 1973)).

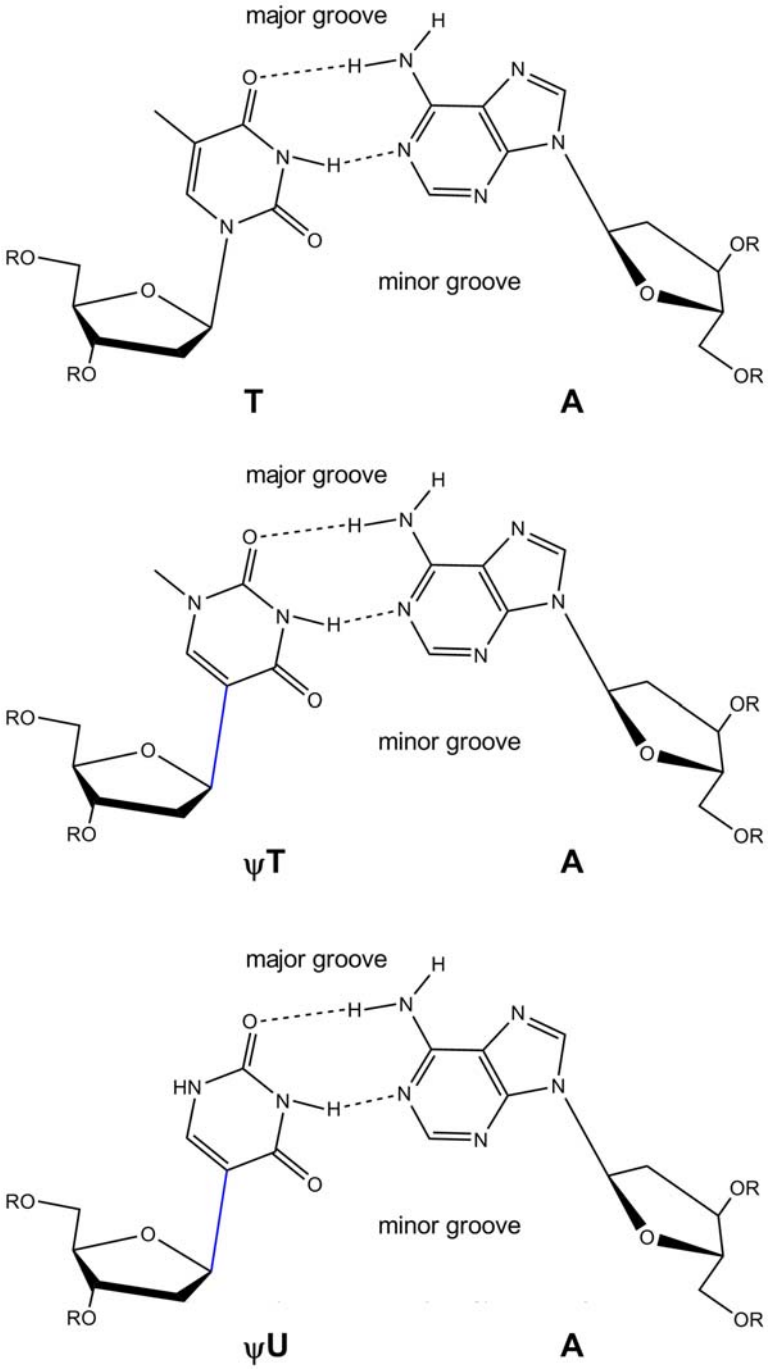


Figure 2-2. The base pairing interactions between a standard A-T base pair and the non-standard ψT-A and ψU-A base pairs. Note the C-glycosidic bond (shown in blue) between the base and the sugar in both ψT and ψU.

Table 2-1. Oligonucleotides used in this study.

Oligo	Sequence (5'→3' Direction)	Purification
P-1	GCG TAA TAC GAC TCA CTA TAG	PAGE
T-1	GTT CCT GTG TCG ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-2	TTC CTG TGT CGA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-3	TCC TGT GTC GAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-4	CCT GTG TCG AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-5	CTG TGT CGA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-6	TGT GTC GAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-7	GTG TCG AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-8	TGT CGA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-9	GTC GAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-10	TCG AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-11	CGA AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-12	GAA AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-13	CGG CGT AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-14	GGC GTA AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-15	GCG TAA AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-16	CGT AAA AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-17	GTA AAA AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-18	GAA AAA AAA CTA TAG TGA GTC GTA TTA CGC	Desalted
T-19	GTT CAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-20	GTC AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-21	GCA AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-22	GAA AAA AAA AAA ACT ATA GTG AGT CGT ATT ACG C	Desalted
T-23	CAG AGA CGψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-24	CGG ACG Aψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-25	CGG CGA ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-26	GGC GAψ ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-27	GCG Aψψ ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-28	CGA ψψψ ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-29	GAψ ψψψ ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-30	Gψψ ψψψ ψψψ CTA TAG TGA GTC GTA TTA CGC	PAGE
T-31	GAA Cψψ ψψψ ψψψ ψCT ATA GTG AGT CGT ATT ACG C	PAGE
T-32	GAC ψψψ ψψψ ψψψ ψCT ATA GTG AGT CGT ATT ACG C	PAGE
T-33	GCψ ψψψ ψψψ ψψψ ψCT ATA GTG AGT CGT ATT ACG C	PAGE
T-34	Gψψ ψψψ ψψψ ψψψ ψCT ATA GTG AGT CGT ATT ACG C	PAGE
T-35	GCG TAA TAC GAC TCA CTA TAG TCG ACA CAG	Desalted
T-36	GCG TAA TAC GAC TCA CTA TAG TTA CGA CCG	Desalted
T-37	GCG TAA TAC GAC TCA CTA TAG TTT ACG CCG	Desalted
T-38	GCG TAA TAC GAC TCA CTA TAG TTT TAC GCC	Desalted
T-39	GCG TAA TAC GAC TCA CTA TAG TTT TTA CGC	Desalted
T-40	GCG TAA TAC GAC TCA CTA TAG TTT TTT ACG	Desalted
T-41	GCG TAA TAC GAC TCA CTA TAG TTT TTT TAC	Desalted
T-42	GCG TAA TAC GAC TCA CTA TAG TTT TTT TTC	Desalted
T-43	GCG TAA TAC GAC TCA CTA TAG TTT TTT TTT GAA C	Desalted
T-44	GCG TAA TAC GAC TCA CTA TAG TTT TTT TTT TGA C	Desalted
T-45	GCG TAA TAC GAC TCA CTA TAG TTT TTT TTT TTG C	Desalted
T-46	GCG TAA TAC GAC TCA CTA TAG TTT TTT TTT TTT C	Desalted
T-47	GCG TAA TAC GAC TCA CTA TAG ACG TCT CTG	Desalted
T-48	GCG TAA TAC GAC TCA CTA TAG AAT CGT CCG	Desalted
T-49	GCG TAA TAC GAC TCA CTA TAG AAA TCG CCG	Desalted
T-50	GCG TAA TAC GAC TCA CTA TAG AAA ATC GCC	Desalted
T-51	GCG TAA TAC GAC TCA CTA TAG AAA AAT CGC	Desalted
T-52	GCG TAA TAC GAC TCA CTA TAG AAA AAA TCG	Desalted
T-53	GCG TAA TAC GAC TCA CTA TAG AAA AAA ATC	Desalted
T-54	GCG TAA TAC GAC TCA CTA TAG AAA AAA AAC	Desalted
T-55	GCG TAA TAC GAC TCA CTA TAG AAA AAA AAA GTT C	Desalted
T-56	GCG TAA TAC GAC TCA CTA TAG AAA AAA AAA AGT C	Desalted
T-57	GCG TAA TAC GAC TCA CTA TAG AAA AAA AAA AAG C	Desalted
T-58	GCG TAA TAC GAC TCA CTA TAG AAA AAA AAA AAA C	Desalted

*The ψ represent the incorporation of a pseudouridine residue.

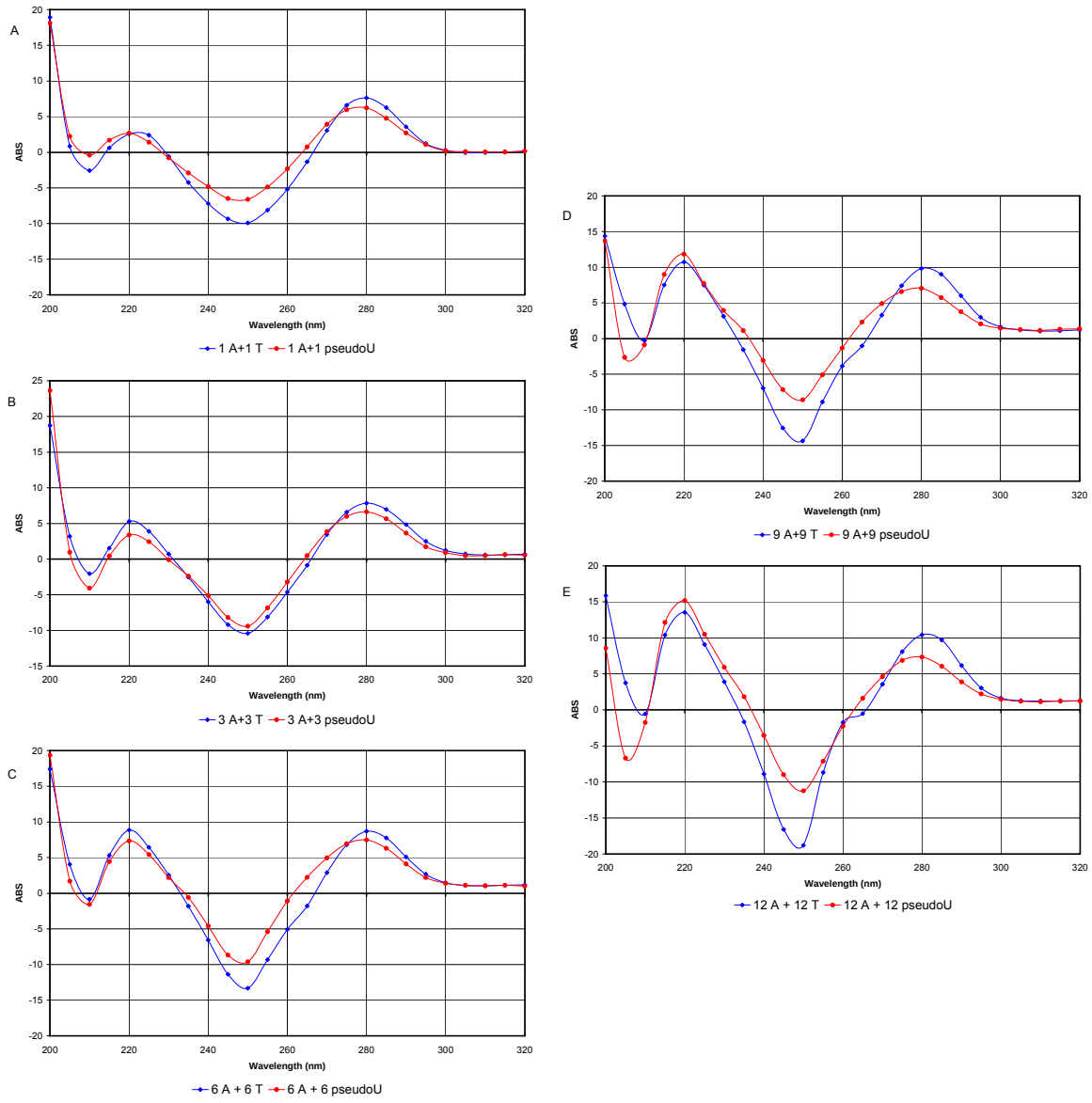


Figure 2-3. Representative CD Spectra. Circular dichroism spectra of select double stranded templates with their complements containing varying amounts of dA-dT or dA-dψU base pairs at 25 °C. All of the spectra above are indicative of B-DNA (Ivanov et al., 1973). Note that the conformation does not dramatically change as the amount of ψU is increased. (A) The spectra of duplexes containing 1 dA-dT base pair vs. 1 dA-dψU base pair. (B) The spectra of duplexes containing 3 dA-dT base pairs vs. 3 dA-dψU base pairs. (C) The spectra of duplexes containing 6 dA-dT base pairs vs. 6 dA-dψU base pairs. (D) The spectra of duplexes containing 9 dA-dT base pairs vs. 9 dA-dψU base pairs. (E) The spectra of duplexes containing 12 dA-dT base pairs vs. 12 dA-dψU base pairs.

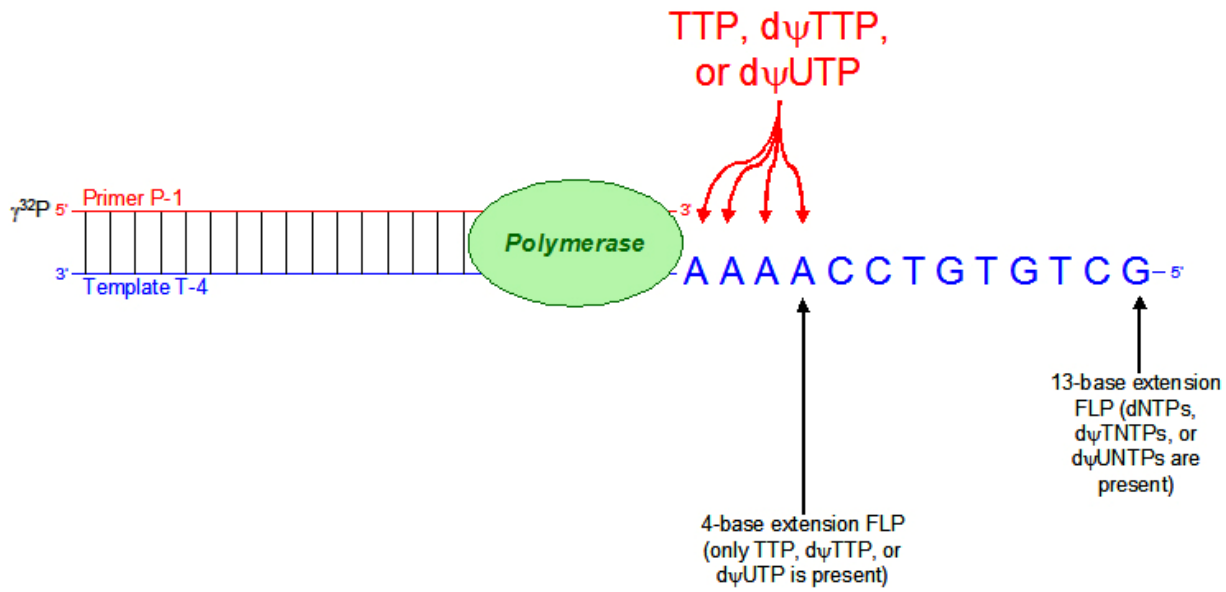


Figure 2-4. Depiction of primer-extension assays used in the polymerase screen. In the 4-base extension assays, polymerases were challenged to incorporate up to four consecutive dT, dψT, or dψU residues across from template dA. In the 13-base extension assays, the polymerases were forced to incorporate and extend beyond those first four residues.

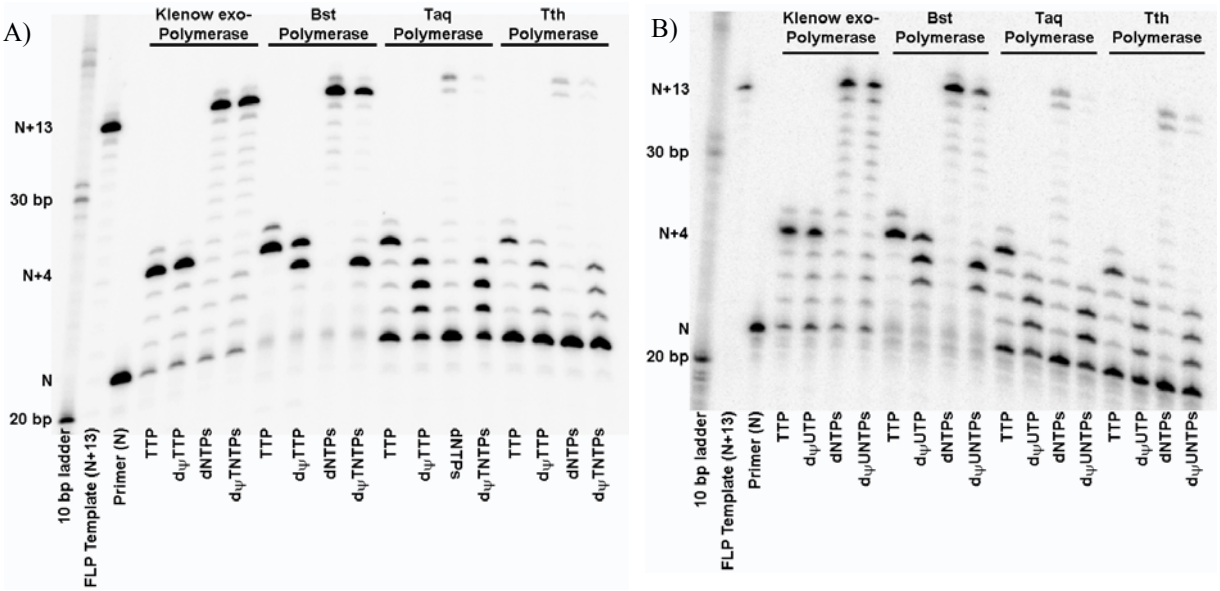


Figure 2-5. Family A polymerase screen. Unextended primer is at position N; N+4 is the full-length product (FLP) for the 4-base extension assays; N+13 is the FLP for the 13-base extension assays. Final concentrations: TTPs/dψTTPs/dψUTPs/dNTPs/ dψTNTPs/dψUNTPs (100 μM), radiolabeled P-1 (2.5 pmol), non-radiolabeled P-1 (20 pmol), non-radiolabeled template T-4 (30 pmol), and appropriate polymerase (1 U). The mixtures were prewarmed to the polymerase's optimal temperature for 30 s and initiated with the appropriate NTP mixture. The mixtures were incubated at the polymerase's optimal temperature for 2 min and immediately terminated with DNA PAGE Loading Dye (formamide, EDTA, and dyes). An aliquot (1 μL) was loaded onto denaturing polyacrylamide gels (20%, 7 M urea) and resolved. A) The incorporation and extension of dT and dψT by various Family A polymerases. All polymerases were able to incorporate and extend beyond the four consecutive A-T or A-ψT base pairs to generate some FLP in both the 4-base and 13-base extension assays. Klenow (exo-) and *Bst* most likely generated higher amounts of ψT containing FLP since their optimal temperatures are lower than that of *Taq* and *Tth*. B) The incorporation and extension of dT and dψU by various Family A polymerases. All polymerases were able to incorporate and extend beyond the four consecutive A-T or A-ψU base pairs to generate some FLP in both the 4-base and 13-base extension assays. Klenow (exo-) and *Bst* most likely generated higher amounts of ψU containing FLP since their optimal temperatures are lower than that of *Taq* and *Tth*.

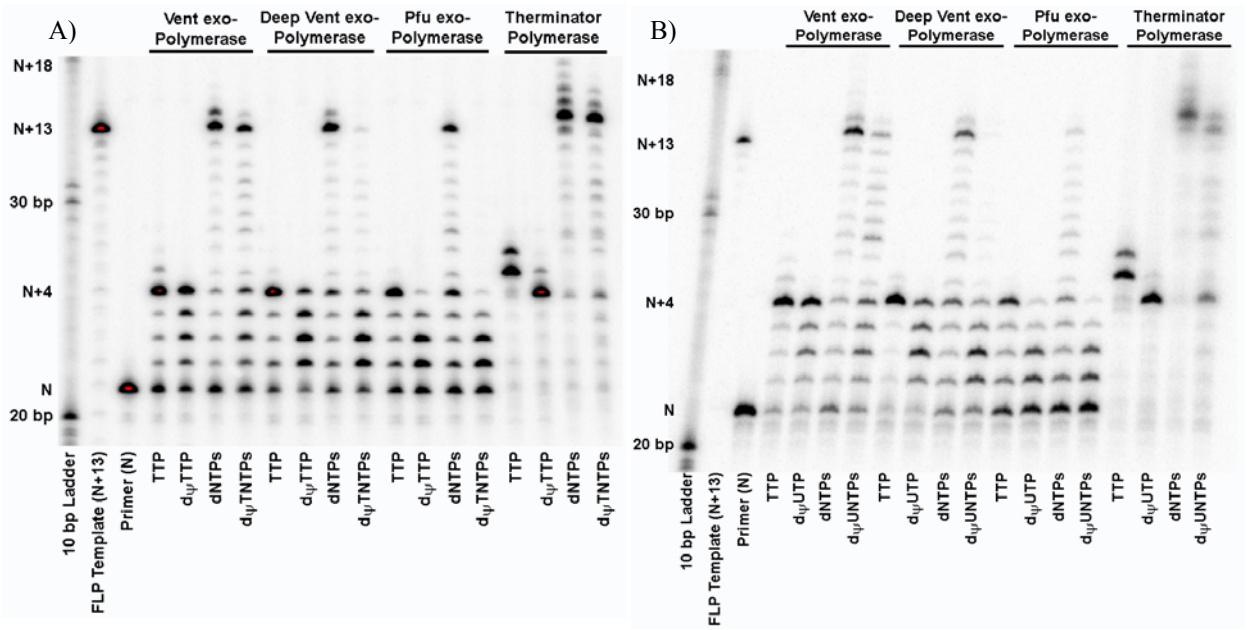


Figure 2-6. Family B polymerase screen. Unextended primer is at position N; N+4 is the full-length product (FLP) for the 4-base extension assays; N+13 is the FLP for the 13-base extension assays. Final concentrations: TTPs/dψTTPs/dψUTPs/dNTPs/ dψTNTPs/dψUNTPs (100 μM), radiolabeled P-1 (2.5 pmol), non-radiolabeled P-1 (20 pmol), non-radiolabeled template T-4 (30 pmol), and appropriate polymerase (1 U). The mixtures were prewarmed to the polymerase's optimal temperature for 30 s and initiated with the appropriate triphosphate mixture. The mixtures were incubated at the polymerase's optimal temperature for 2 min and immediately terminated with DNA PAGE Loading Dye (formamide, EDTA, and dyes). An aliquot (1 μL) was loaded onto denaturing polyacrylamide gels (20%, 7 M urea) and resolved. A) The incorporation and extension of dT and dψT by various Family B polymerases. All polymerases, except *Pfu* (exo-), were able to incorporate and extend beyond the four consecutive A-T or A-ψT base pairs to generate some FLP in both the 4-base and 13-base extension assays. *Pfu* (exo-) was able to generate FLP in the 4-base assay, but not the 13-base assay. Therminator was extremely adept at incorporating the dψT residues, as depicted by the low levels of unextended primer remaining in those lanes. B) The incorporation and extension of dT and dψU by various Family B polymerases. All polymerases, except *Pfu* (exo-), were able to incorporate and extend beyond the four consecutive A-T or A-ψT base pairs to generate some FLP in both the 4-base and 13-base extension assays. *Pfu* (exo-) was able to generate FLP in the 4-base assay, but not the 13-base assay. Therminator was extremely adept at incorporating the dψU residues, as depicted by the low levels of unextended primer remaining in those lanes.

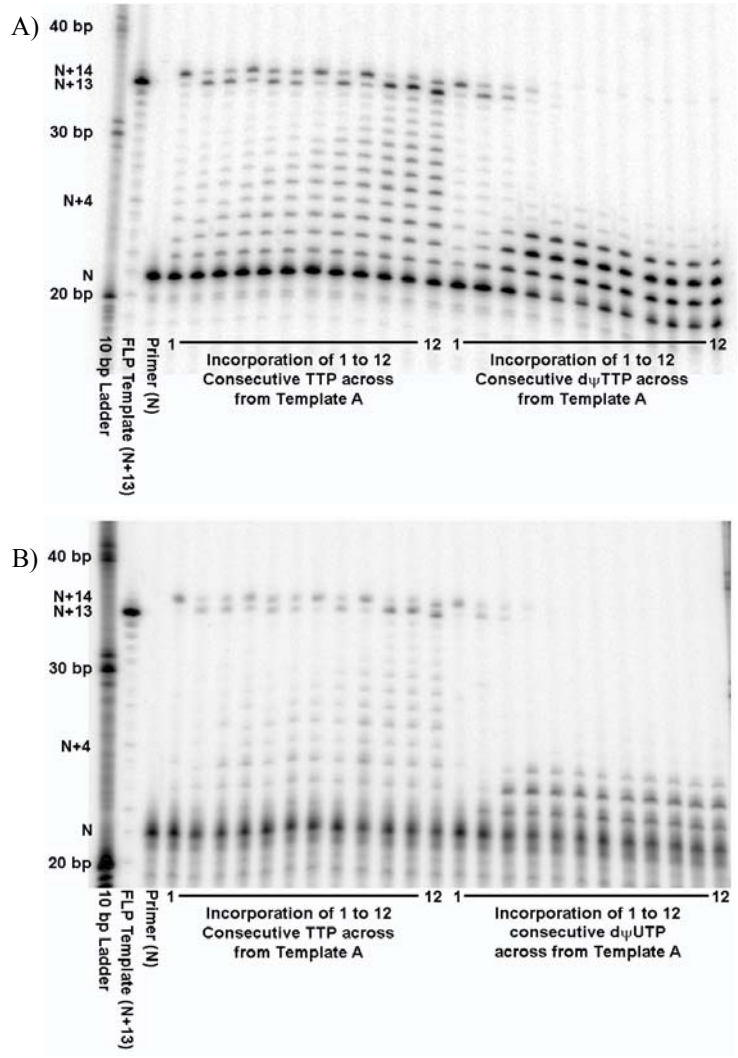


Figure 2-7. Incorporation of one to twelve consecutive dT, dψT, or dψU residues by *Taq* polymerase. Unextended primer is at position N; FLP is denoted by N+13 in all of these assays (see Table 2-1 for oligonucleotides used). Final concentrations: dNTPs/dψTNTPs/dψUNTPs (100 μM), radiolabeled P-1 (2.5 pmol), non-radiolabeled P-1 (20 pmol), non-radiolabeled templates T-1 through T-12 (30 pmol), and *Taq* polymerase (1 U). The mixtures were prewarmed to 72 °C for 30 s and initiated with the appropriate NTP mixture. The mixtures were incubated at 72 °C for 2 min and immediately terminated with DNA PAGE Loading Dye (formamide, EDTA, and dyes). An aliquot (1 μL) was loaded onto denaturing polyacrylamide gels (20%, 7 M urea) and resolved. A) The incorporation and extension of 1 to 12 dT or dψT residues across from template A by *Taq* polymerase. It appears that very little to no FLP is generated after the incorporation of five or more consecutive dψTs. B) The incorporation and extension of 1 to 12 dT or dψU residues across from template A by *Taq* polymerase. It appears that very little to no FLP is generated after the incorporation of five or more consecutive dψUs.

CHAPTER 3
CREATION OF A RATIONALLY DESIGNED MUTAGENIC LIBRARY AND SELECTION OF THERMOSTABLE POLYMERASES USING WATER-IN-OIL EMULSIONS

Introduction

To create synthetic biology using an artificially expanded genetic information system (AEGIS), a polymerase that is capable of incorporating non-standard nucleotides (NSBs) is needed. Unfortunately, studies have not found an extant thermostable polymerase able to incorporate a variety of NSBs with efficiency and fidelity. Polymerases usually perform more efficiently with one type of NSB, than they do with another (Hendrickson et al., 2004, Leal et al., 2006, Roychowdhury et al., 2004).

Directed evolution may help to rectify this situation and allow us to mutate an existing polymerase to generate one with an increased ability to incorporate a variety of NSBs (Ghadessy et al., 2001, Ghadessy et al., 2004). Therefore, we became interested in directed evolution as a way to modify *Taq* polymerase to better incorporate NSBs, specifically ones exhibiting a C-glycosidic linkage.

Taq polymerase, a member of the Family A polymerases, has already been successfully evolved under direction to incorporate various other NSBs using directed evolution (Ghadessy et al., 2001, Ghadessy et al., 2004, Fa et al., 2004). Ghadessy *et al.* provided a procedure for doing so using water droplets in oil (Ghadessy et al., 2004, Ghadessy et al., 2001); these served as artificial cells. They began with large, diverse random libraries of the *Taq* polymerase, with approximately 7 amino acid residue replacements. Ghadessy *et al.* found that three to four rounds of selection was sufficient to identify a polymerase able to incorporate various NSBs using these random libraries.

This result was initially surprising, as Guo *et al.* has shown that approximately one-third of all random multiple amino acid changes will result in the inactivation of a protein, and that 70%

of random changes in the active site of a polymerase will also result in inactivation (Guo et al., 2004). This implies that a protein having more than a few random amino acid changes has a high likelihood of being inactive. One might have expected that a very large fraction of the variants created by Ghadessy *et al.* would have been inactive, especially at high temperatures, and this expectation is consistent with results reported below.

This raises a general question: What is the likelihood that a library contains a protein having a novel but desirable property? A desirable library for directed would optimally have a large, diverse number of proteins with a high number of active clones (Hibbert and Dalby, 2005). One approach to achieving this goal involves the selection of sites to introduce replacements. For example, if replacements throughout the protein are equally likely to lower thermal stability, while replacements in sites near the active site are more likely to change catalytic behavior, it makes sense to focus randomization in residues near the active site (Arnold and Georgiou, 2003b, Arnold and Georgiou, 2003a, Fa et al., 2004, Miller et al., 2006, Ghadessy et al., 2004, Ghadessy et al., 2001).

An alternative approach recognizes that natural history has already explored polymerase "sequence space." Much of this natural history is available to us in genomic sequence databases. This permits an approach, originally called "evolutionary guidance," that extracts information from that history to identify sites that are more likely to influence behavior in a way that is desired, and less likely to damage the enzyme (Allemann et al., 1991, Presnell and Benner, 1988).

Eric Gaucher, at the Foundation for Applied Molecular Evolution (FfAME), recently developed this approach a step further under the reconstructing evolutionary adaptive paths (REAP) rubric (Gaucher, 2006). He identified sites where functional divergence occurred within

a family of polymerases, but where natural history suggested that the site was under strong selective pressure. In theory, this has the highest probability to generate new activities and functions.

Using the sites identified by the REAP approach, the Type II sequence divergence of the Family A polymerases was studied (Gu, 2002, Gu, 1999). In this approach, sites were identified that had a split "conserved but different" pattern of historical evolutionary variation, and had been previously suggested to lead to a change in the function or behavior of the polymerase. Using Pfam (Fig. 3-2), a total of 57 amino acid changes across 35 sites within the 719 members of Family A polymerases that were available were identified (Bateman, 2006, Finn et al., 2006). The 35 sites for mutational studies, distributed as seen in Figure 3-2, were derived from these analyses, and from sequences discussed in a recent review by Henry and Romesberg on the evolution of novel polymerase activities (Henry and Romesberg, 2005). The 57 replacement amino acid residues were selected based on the Family A viral polymerase sequences at the 35 mutational sites. The viral sequences were exploited since literature has told us that viral polymerases are more adept at incorporating NSBs than other polymerases (Sismour et al., 2004, Leal et al., 2006, Horlacher et al., 1995), and ancient viruses have also been implicated in the origins of cellular DNA replication machinery (Forterre, 2006).

The company DNA 2.0 created and synthesized the rationally designed (RD) library containing 74 different mutants using the 57 amino acid changes identified by REAP, in various combinations to yield three or four amino acid mutations per sequence. In addition to creating the library, DNA 2.0 also designed and generated a version of the *wt taq* polymerase gene that was optimized for codon usage in *E. coli* cells (co-*Taq* polymerase). The optimization of codon usage results in higher expression levels of the protein within the cell (Gustafsson et al., 2004).

Each of these 75 polymerases (co-*Taq* and the 74 mutants) were tested for their ability to incorporate increasing concentrations of a representative C-glycoside (Fig. 2-3), 2'-deoxypesuouridine-5'triphosphate (dψUTP). None were able to incorporate dψUTP more efficiently than the co-*Taq* polymerase, and only eighteen of the 74 mutants of the RD Library showed activity with the canonical dNTPs under the conditions with which they were presented.

Selections require that some members of the library perform differently than the original protein of interest (Arnold and Georgiou, 2003a, Lutz and Patrick, 2004). We did not perform a selection to identify a polymerase with an increased ability to incorporate dψUTP, since we determined there were no clones in the RD Library that functioned with the NSB better than co-*Taq* polymerase. In order to demonstrate our laboratories ability to perform *in vitro* selections, we decided select for the eighteen mutant polymerases that exhibited activity with dNTPs from the pool of 74 mutants.

To perform our selection experiments, we used a variation of the compartmentalized self-replication (CSR) method developed in the laboratories of Griffiths and Holliger to create water-in-oil emulsions as a way to link genotype to phenotype (Miller et al., 2006, Tawfik and Griffiths, 1998, Ghadessy et al., 2001, Ghadessy et al., 2004). This method (Fig. 1-13) uses cells expressing the polymerase as the sole source of polymerase and plasmid template in a PCR reaction, which takes place inside the aqueous phase of the emulsion. Inactive polymerases fail to replicate their encoding gene, so they are effectively selected against after the extraction of products from the emulsion.

After our selection, products were recloned into the expression vector using a version of the megaprimer PCR method (Miyazaki and Takenouchi, 2002). As this protocol generated products that were crossover mutations, sequencing of the products provided a list of the

mutations that survived the selection, without providing information about which mutations were associated with each other. The megaprimer PCR is, nevertheless, an effective method for library rediversification between rounds of selection.

Materials and Methods

DNA Sequencing and Analysis

DNA sequencing was carried out by the University of Florida Interdisciplinary Center for Biotechnology Research, DNA Sequencing Core Facility using an ABI 3130*xl* Genetic Analyzer (Applied Biosystems, Foster City, California) and primers P-6 through P-9 (Table 3-1). BLAST 2 software was used for sequence similarity searching (Tatusova and Madden, 1999); Derti's Reverse and/or complement DNA sequences website was used to find the reverse complement of various DNA strands (Derti, 2003); and ExPASy's translate tool was used to translate DNA sequences into their amino acid counterparts (Swiss Institute of Bioinformatics, 1999).

Construction of Plasmids

Construction of pSW1

The gene (*wt taq*) encoding *wt Taq* polymerase was cloned from a vector generously donated by Dr. Michael Thompson (UNC, Chapel Hill, North Carolina) using primers P-2 and P-3. The product was digested with the *Sac*II and *Nco*I restriction enzymes (New England BioLabs, Beverly, Massachusetts) according to manufacturer's protocol. The restricted *wt taq* was then ligated into the identically digested pASK-IBA43plus vector (IBA GmbH, St. Louis, Missouri)(Fig. 3-3), using T4 DNA ligase (New England BioLabs) according to manufacturer's protocol (16 °C overnight with a 4:1 insert:vector ratio) to make the new plasmid pSW1 (Fig. 3-4), and adding an N-terminal hexahistidine tag onto the *wt taq* gene ($His_{(6)}$-*wt Taq*). Plasmid constructs were verified by restriction digest analysis, using the enzymes *Bam*HI and *Nco*I according to the manufacturer's protocol (New England BioLabs), as well as sequencing.

Rationally designed mutagenic library (RD Library) creation

DNA 2.0 (Menlo Park, California) synthesized a variant of the *wt taq* polymerase gene (co-*taq*) that was optimized for the codon-usage of *E. coli*, which was then used to construct the pSW2 plasmid (Fig. 3-5). Plasmids pSW3 – pSW76 (Table 3-2) were designed by Dr. Eric Gaucher (Foundation for Applied Molecular Evolution, Gainesville, Florida) and DNA 2.0 using the REAP approach. Sequence alignments and phylogenetic tree construction of 719 Family A polymerase protein sequences were generated using the Pfam website (Bateman, 2006, Finn et al., 2006). Type II functional divergence between the bacterial/eukaryotic Family A polymerases and the viral Family A polymerases was estimated with DIVERGE 2.0 software (Gu, 2002, Gu, 1999). The 35 sites for mutational studies were derived from these analyses, as well as sequences discussed in Henry and Romesberg (Henry and Romesberg, 2005); the replacement amino acid residues were selected based on the viral sequences at those sites. The sites chosen are all located in or near the active site of the polymerase.

DNA 2.0 randomized the mutations throughout the 74 sequences so they were equally distributed (3 to 4 amino acid changes per gene). In addition to the synthesis of the genes, DNA 2.0 cloned all 75 of these plasmids (co-*taq* and 74 mutants) into the pASK-IBA43plus vector using the *Sac*II and *Nco*I restriction sites. Plasmid constructs were verified both by restriction digest analysis, using the enzymes *Bam*HI and *Nco*I according to the manufacturer's protocol (New England BioLabs), and by sequencing.

Growth Curves and Cell Counts

The bacterial strains used in this study are listed in Table 3-3. The rich media used in these studies was Luria-Bertani (LB) medium (Difco Laboratories, Detroit, Michigan) (Miller, 1972). Ampicillin was provided in liquid or solid medium at a final concentration of 100 μg/mL. Plasmids were transformed into the *E. coli* TG-1 cell line according to manufacturer's protocol

(Zymo Research, Orange, California). Cell growth was determined by measuring optical density at 550 nm using a SmartSpec Plus Spectrophotometer (Bio-Rad, Hercules, California). Anhydrotetracycline (2 mg/mL stock in *N*,*N*-dimethylformamide) was used at a final concentration of final concentration of 0.2 ng/μL to induce expression.

Inocula for the growth experiments were prepared as follows: bacterial strains were grown overnight (14.25 hrs) at 37 °C and 250 rpm in LB medium (supplemented with ampicillin, if applicable) in 14 mL 2059 Falcon Tubes (BD Biosciences, San Jose, California). Cells (1 mL) from the 5 mL overnight culture were used to inoculate 100 mL LB or LB-Amp cultures in 500 mL baffled flasks. Cultures were grown at 37 °C and 250 rpm for 8.75 hrs. Cell counts were measured by performing a dilution series using 10-fold dilutions of the cells in 0.85% NaCl. Dilutions were plated onto LB plates (supplemented with ampicillin, if applicable), grown overnight at 37 °C, and colonies were counted the next morning to determine the number of colony-forming units per milliliter of culture (cfu/mL).

Samples of cells were taken at various time points to determine the levels of protein expression, before and after induction. 2X SDS-PAGE (62.5 mM, pH 6.8, 25% glycerol, 2% SDS, 0.01% bromophenol blue, 5% β-mercaptoethanol (Laemmli, 1970))loading dye was added to the samples, and to 50 U *Taq* Polymerase (New England BioLabs). Samples were boiled for 8 minutes, then loaded onto a Tris-HCl Ready Gel (7.5%, Bio-Rad) and resolved for 45 min at 200 V. Gels were stained via the Fairbanks Method (Fairbanks et al., 1971).

Purification of $His_{(6)}$-*wt* Taq Polymerase

The SW3 cell line was grown overnight in 5 mL of LB-Amp broth for 14.25 hr at 37 °C and 250 rpm in 14 mL 2059 Falcon Tubes (BD Biosciences). Approximately 2 x 10^8 colony-forming units (cfu), roughly equal to 500 μL of a culture with an OD_{550nm} of 4.0, were used to

inoculate two 100 mL cultures of LB-Amp in 500 mL baffled flasks. These cultures were grown at 37 °C and 250 rpm for 3.75 hrs to an approximate OD_{550nm} of 1.8, and were then induced by addition of anhydrotetracycline (0.2 ng/μL final concentration). The cells were allowed to grow for an additional 5 hrs to an approximate OD_{550nm} of 3.5. Samples of the undinduced and induced cells were taken and stored at -20 °C for further analysis.

Cultures were then combined and the cells harvested by centrifugation (9000 rpm, 10 min, 4 °C). The SW3 cells were washed in 40 mL of Cell Harvest Buffer (50 mM Tris-HCl, pH 7.9, 50 mM dextrose, 1 mM EDTA, 4 °C) and centrifuged again (8000 rpm, 10 min, 4 °C). The cell pellet was then resuspended in Cell Lysis Buffer (20 mM Tris-HCl, pH 7.9, 50 mM NaCl, 5 mM imidazole, 1 mg/mL lysozyme, 5 μg/mL DNaseI, and 10 μg/mL RNaseI) at a concentration of 2 mL/gram of cells.

The cells were gently lysed by rocking (GyroMini Nutating Mixer) at ambient temperature for 15 min, the proteins were then denatured by heating to 75 °C for 20 min. The lysed cells were centrifuged (39,000 x g, 10 min, 4 °C) and the cell-free extract (cfe) removed and placed into a clean tube. The cfe was then sonicated with six 10 s bursts at 71% output with a 10 s cooling periods at 4 °C between each burst (Model 500 Sonic Dismembrator with a 1/2 inch tapped horn with flat tip, Fisher Scientific, Suwannee, Georgia). The cfe was centrifuged (39,000 x g, 10 min, 4 °C) and the supernatant (cleared cfe) was removed.

The cleared cfe was added to 1 mL of a 50% Ni-NTA slurry (Qiagen, Valencia, California) and incubated at 4 °C for 60 min with gentle mixing (GyroMini Nutating Mixer). The lysate-Ni-NTA mixture was loaded onto a Poly-Prep Column (Bio-Rad, Hercules, California) and allowed to settle for 10 min at 4 °C. A portion of the flow-through (10 μL) was then collected and saved for analysis. The column was washed twice with 4 mL of Ni-NTA Wash Buffer (20 mM Tris-

HCl, pH 7.9), 50 mM NaCl, 60 mM imidazole) and a portion of the flow-through (10 μL) was saved for future analysis. The protein was eluted four times (0.5 mL each) with Ni-NTA Elution Buffer (10 mM Tris-HCL, pH 7.9, 250 mM NaCl, 500 mM imidazole) and portions of each (10 μL) were saved for future analysis at -20 °C. 2X SDS-PAGE loading dye was added to each of the samples mentioned above. Samples were prepared, resolved, stained, as described in the previous section, and the elutions containing the majority of the purified $His_{(6)}$-*wt Taq* polymerase were identified.

Elution fractions 2 – 4 were combined and loaded into a Slide-A-Lyzer 10K MWCO 0.5 – 3 mL Dialysis Cassette (Pierce, Rockford, Illinois) that was pre-hydrated in Taq Dialysis Buffer A (50 mM Tris-HCl, pH 8.0, 50 mM KCl, 0.1 mM EDTA, 0.5 mM PMSF, 0.5% Nonidet-P40, 0.5% Triton X-100). The sample was dialyzed at 4 °C for 4 hrs against 500 mL of Dialysis Buffer A. It was then dialyzed for another 4 hrs at 4 °C against 500 mL of Taq Dialysis Buffer B (50 mM Tris-HCl, pH 8.0, 50 mM KCl, 0.1 mM EDTA, 0.5 mM PMSF, 0.5% Nonidet-P40, 0.5% Triton X-100, 1 mM DTT). Finally, it was dialyzed for 8 hrs at 4 °C against 1 L of Taq Storage Buffer (50 mM Tris-HCl, pH 8.0, 50 mM KCl, 1 mM DTT, 0.1 mM EDTA, 0.5 mM PMSF, 0.5% Nonidet-P40, 0.5% Triton X-100, 1 mM DTT, 50% glycerol). The sample was removed, quantitated, and the protein concentration determined using the Bio-Rad Protein Assay Dye Reagent according to manufacturer's instructions.

The purified $His_{(6)}$-*wt Taq* polymerase and *Taq* polymerase (New England BioLabs) were used in separate PCR reactions. The same concentration of each polymerase (enough protein to equate to 3 U of *Taq* polymerase from New England BioLabs) were added to PCR reactions containing: 1X Modified ThermoPol Buffer (2 mM Tris-HCl, pH 9, 10 mM KCl, 1 mM $(NH_4)_2SO_4$, 2.5 mM $MgCl_2$, 0.2% Tween 20), 250 μM dNTPs, 1.0 μM P-4, 1.0 μM P-5, and 1

ng/μL pSW1. The PCRs (50 μL) were run under the following conditions: 5 min, 94 °C; (1 min, 94.0 °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C. Products were analyzed by agarose gel electrophoresis and quantitated using the Molecular Imager Software (Bio-Rad).

Incorporation of dψUTP by RD Library

2'-deoxypseudouridine-5'-triphosphate (dψUTP) was purchased from TriLink BioTechnologies (San Diego, California). Standard deoxynucleotide triphosphates (dNTPs) were comprised of 2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxycytidine-5'-triphosphate (dCTP), 2'-deoxyganosine-5'-triphosphate (dGTP), and thymidine-5'-triphosphate (TTP) and were purchased from Promega Corporation (Madison, Wisconsin). dψUNTPs were comprised of dATP, dCTP, d GTP, and dψUTP.

Individual cultures (5 mL LB-Amp) of the SW5 – SW78 cell lines were grown for 14.25 hrs at 250 rpm and 37 °C in 14 mL 2059 Falcon Tubes (BD Biosciences). Approximately 2 x 10^8 colony-forming units (cfu), roughly equal to 500 μL of a culture with an OD_{550nm} of 4.0, were used to inoculate individual 100 mL cultures of LB-Amp in 500 mL baffled flasks. These cultures were grown at 37 °C and 250 rpm for 3.75 hrs to an approximate OD_{550nm} of 1.8, and were then induced with anhydrotetracycline. The cells were allowed to grow for 1 hr longer to an approximate OD_{550nm} of 3.0.

Approximately 1 x 10^6 cfu (~2 μL cells) were used as the sole source of polymerase and template in separate PCR reactions containing final concentrations of these constituents: 1X Modified ThermoPol Buffer, 1.4 μM P-4, 1.4 μM P-5, 1.1 ng/μL RNaseA, and 6% DMSO. The final concentration of nucleotide triphosphates added to the reactions were one of the following: 500 μM dNTPs; 500 μM dATP/dGTP/dCTP; 500 μM dATP/dGTP/dCTP + 450 μM TTP + 50 μM dψUTP; 10 μM dATP/dGTP/dCTP + 400 μM TTP + 100 μM dψUTP; 10 μM

dATP/dGTP/dCTP + 350 μM TTP + 150 μM dψUTP; 10 μM dATP/dGTP/dCTP + 300 μM TTP + 200 μM dψUTP; 10 μM dATP/dGTP/dCTP + 250 μM TTP + 250 μM dψUTP; 10 μM dATP/dGTP/dCTP + 200 μM TTP + 300 μM dψUTP; 10 μM dATP/dGTP/dCTP + 150 μM TTP + 350 μM dψUTP; 10 μM dATP/dGTP/dCTP + 100 μM TTP + 400 μM dψUTP; 10 μM dATP/dGTP/dCTP + 50 μM TTP + 450 μM dψUTP; 500 μM dψUTPs. The PCRs (50 μL) were run under the following conditions: 5 min, 94 °C; (1 min, 94.0 °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C. Products were analyzed by agarose gel electrophoresis and quantitated using the GeneTools Software, version 3.07 (SynGene, Cambridge, England).

Selection of Thermostable Mutants Using Water-In-Oil Emulsions

Water-in-oil emulsions

The appropriate cell line was grown overnight in LB-Amp broth (5 mL) for 14.25 hr at 37 °C and 250 rpm in 14 mL 2059 Falcon Tubes (BD Biosciences). Approximately 2 x 10^8 colony-forming units (cfu), roughly equal to 500 μL of a culture with an OD_{550nm} of 4.0, were used to inoculate a 100 mL culture of LB-Amp in 500 mL baffled flasks. These cultures were grown at 37 °C and 250 rpm for 3.75 hrs to an approximate OD_{550nm} of 1.8, induced with anhydrotetracycline, and allowed to grow for 1 hr longer to an approximate OD_{550nm} of 3.0. The amount of culture containing 2 x 10^8 cfu was determined; that amount was centrifuged (13,000 rpm, 2 min), the supernatant removed, and the remaining pellet was stored on ice.

The aqueous phase of the emulsions was prepared by resuspending the cell pellet in a 200 μL solution containing: 1X Modified ThermoPol Buffer, 500 μM dNTPs, 1.4 μM P-4, 1.4 μM P-5, 1.1 ng/μL RNaseA, and 6% DMSO. For control reactions, without cells, 1 ng/μL of pSW2 and 10 U *Taq* Polymerase were added to the aqueous phase. Reactions were stored on ice until further use.

To prepare the oil-phase of the emulsions, Arlacel P135 (Uniqema, New Castle, Delaware) was heated to 75 °C, as was mineral oil (Sigma-Aldrich, St. Louis, Missouri). The mineral oil was mixed with the Arlacel P135 (1.5% v/v) in a 5 mL Corning Externally Threaded Cryogenic Vial (Corning, Acton, Massachusetts) containing an 8 x 3 mm stir bar with pivot ring. The oil-phase was stirred at 1000 rpm on ice while the 200 μL aqueous phase was added drop-wise over a period of 2 minutes. The emulsion was stirred for 5 min longer, then subjected to PCR [5 min, 94 °C; (1 min, 94.0 °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C].

Products were extracted from the emulsions with the addition of two volumes of water-saturated ether. The ether and emulsions were mixed by vortexing, centrifuged (5 min, 8000 rpm), and the aqueous phases extracted. To rid the aqueous phases of contaminating enzyme, the products were subjected to a QIAquick PCR Purification Kit (Qiagen), and products were eluted from the column in Qiagen Buffer EB (50 μL). Products were separated using agarose gel electrophoresis; the product band was extracted and then purified using a QIAquick Gel Extraction Kit (Qiagen). Samples were eluted in Qiagen Buffer EB (50 μL), and product concentration was determined by measuring absorption at 260 nm.

Re-cloning of selected mutants

The final products of the emulsions were used in an adaptation of the Miyazaki and Takenouchi megaprimer PCR protocol (Miyazaki and Takenouchi, 2002). CSR products were digested with *Nco*I and *Sac*II according to manufacturer's protocol (New England BioLabs). Digested samples (10 ng in 1 μL) were added to a 49 μL PCR mixture (1X Native *Pfu* Buffer, 100 ng pSW2, 500 μM dNTPs, 6% DMSO). Mixture was heated to 96 °C for 30 s prior to the addition of 0.05 U/μL Native *Pfu* Polymerase (Stratagene, La Jolla, California). Samples were

then subjected to PCR [2 min, 96 °C; (30 s, 96.0 °C; 10 min, 68.0 °C)x25 cycles; 30 min, 72.0 °C].

The template strands of DNA (pSW2 plasmid in the PCR) were digested with 2 U *Dpn*I (New England BioLabs) at 37 °C for 2.5 hrs. Reactions were cooled to room temperature, purified using a Qiagen PCR Purification Kit, and eluted with Qiagen Buffer EB (30 μL). Purified products were transformed into the *E. coli* DH5α cell line according to manufacturer's protocol (Invitrogen, Carlsbad, California). Fifty isolated colonies were selected after the transformation (cell lines SW79 through SW128). Overnight 5 mL LB-Amp cultures (250 rpm, 37 °C) were grown for each colony, and their plasmids isolated using the QIAprep Spin Miniprep Kit (Qiagen). Plasmid constructs were verified by restriction digest analysis, using the enzymes *Bam*HI and *Nco*I according to the manufacturer's protocol (New England BioLabs), and mutations were determined by sequencing.

Results

Growth Curves and Cell Counts

Growth curves, cfu counts, and protein expression studies were performed on the SW1 – SW4 cell lines to determine the optimal times for induction (Fig. 3-6[A-C]). The optimal time (1 hr) for induction for both the SW3 and SW4 cell lines was found to be during late log phase at an optical density of approximately 1.8 at 550 nm. The optimal length of induction was 1 hr, due to the rapid death of the cells after the induction of the *taq* gene, as is evidenced by a drop in the cfu/mL counts (Fig. 3-6B). Inductions longer than 1 hr, or induction at early to mid-log phases caused the cells to perish due to toxicity because of the over-expression of a polymerase *in vivo* (data not shown) (Moreno et al., 2005, Andraos et al., 2004). When the migration of the recombinant *Taq* polymerases ($His_{(6)}$-*wt Taq* and co-*Taq*) are compared to that of the *Taq*

Polymerase purchased from New England BioLabs, they all appear to have the same observed molecular weight of 94 kDa on a Coomassie Blue stained SDS-PAGE (7.5%) gel (Fig. 3-6C).

Purification of $His_{(6)}$-*wt* Taq Polymerase

The $His_{(6)}$-*wt Taq* polymerase was purified from SW3 cells that were over-expressing the $His_{(6)}$-*wt taq* gene using nickel affinity chromatography. The polymerase was purified to a single band on a Coomassie Blue stained SDS-PAGE (7.5%) gel (Fig. 3-7A), and elution fractions 2 – 4 were combined and concentrated via dialysis to generate a working stock of $His_{(6)}$-*wt Taq* polymerase. The protein concentration was determined to be 0.744 μg/μL, using the Bio-Rad Protein Assay Dye Reagent. To verify the ability of the purified $His_{(6)}$-*wt Taq* polymerase to amplify DNA in a PCR reaction, similar to that of *Taq* polymerase (New England BioLabs), each of these polymerases were used in separate, identical PCRs. The final concentration of polymerase (5.5 μg/mL) in each reaction was kept constant. Figure 3-7B shows the products of these PCRs, and after analysis it was determined that the densities of these two bands were almost identical.

Incorporation of dψUTP by RD Library

In efforts to find a polymerase that can incorporate and extend beyond dψUs with higher efficiency than the co-*Taq* polymerase, each of the mutant *Taq* polymerases in the RD Library were tested for their ability to incorporate dψUTP across from template dA in PCR reactions containing varying ratios of TTP to dψUTP. Reactions contained induced cells as the sole source of polymerase and template plasmid, so active polymerases were forced to replicate their own encoding gene (2603 bp).

Figure 3-8[A-B] shows the difference between the PCR products from the co-*Taq* polymerase screen (Fig. 3-8A) and a representative (SW21) of the RD Library (Fig. 3-8B). In

both of these reactions, the polymerase could not produce full-length product (FLP) with concentrations of dψUTP higher than 400 μM (final concentration). Based on the product band densities, it was found that none of the active RD Library polymerases displayed a higher propensity for the incorporation of dψUTP than the co-*Taq* polymerase (Table 3-4). It was also noted that only 18 of the 74 mutant polymerases tested showed activity with only dNTPs under these assay conditions (Table 3-2 and Table 3-4).

Selection and Identification of Thermostable Mutants Using Water-In-Oil Emulsions

We pooled all 74 RD Library strains to perform a selection in water-in-oil emulsions to isolate those 18 mutants that showed activity. After the products were isolated, they were used in a modified version of the Miyazaki and Takenouchi megaprimer PCR protocol (Miyazaki and Takenouchi, 2002), creating the full-length plasmid (pASK-IBA43plus with insert). Purified products were transformed into the *E. coli* DH5α cell line; fifty clones were isolated, sequenced, and compared to the co-*Taq* amino acid sequence (Table 3-5). Of these fifty clones, 22 showed no changes relative to the co-*Taq* sequence, and the remaining 28 had at least one residue modified. Table 3-6 shows a breakdown of these mutations, and states whether they are random mutations or RD Library mutations. In the case of the RD Library mutations, it is indicated if they are true RD Library sequences, RD Library sequences with additional mutations, RD Library sequences with reversions to the co-*Taq* sequence, and/or crossovers between two or more RD Library sequences. In addition, only 5% of the mutations found in these sequences encode silent mutations (Table 3-6).

As a control, the selection was also performed using only cells expressing the co-*Taq* polymerase. Five clones were submitted for sequencing following the megaprimer PCR protocol. Of these five, four were the correct co-*Taq* polymerase sequence found in SW4, and

the fifth contained only two amino acid mutations in relation to the co-*Taq* sequence (data not shown).

Discussion

Previously, directed evolution experiments have defined mutations that allow *Taq* polymerase, and other Family A polymerases, to be used in different situations; for example, a few allow for the incorporation of non-standard bases, others are more thermostable, and some are resistant to inhibitors (Ghadessy et al., 2001, Ghadessy et al., 2004, Henry and Romesberg, 2005). The design of our RD Library was based off mutations discussed in the review by Henry and Romesberg (Henry and Romesberg, 2005), and were carried out by using the REAP approach with the Family A polymerases. A library of 74 polymerases was designed, which contained three to four amino acid mutations per polymerase out of a pool of thirty-five possible mutations, in an attempt to identify a polymerase with the ability to incorporate non-standard bases, exhibiting a C-glycosidic linkage, with efficiency and fidelity.

It has been demonstrated previously that the over-expression of a polymerase in a cell can cause toxicity problems and cause premature cell death (Moreno et al., 2005, Andraos et al., 2004). To circumvent this problem, the gene encoding $His_{(6)}$-*wt Taq* polymerase was optimized for codon-usage in *E. coli*, and cloned into a tightly-regulated plasmid (Skerra, 1994) in an attempt to express the polymerase at higher levels only after induction. After appropriate expression conditions were found, the members of the RD Library were individually tested for their ability to incorporate dψUTP, a representative non-standard nucleotide exhibiting a C-glycosidic linkage. The polymerases were challenged with increasing concentrations of the dψUTP as the concentration of TTP presented was decreased. None of the RD Library polymerases were able to incorporate dψUTP more efficiently than the codon-optimized *Taq*

sequence. In the future, other possible mutation sites and combinations of mutations may need to be made and tested to find a polymerase that can accomplish this task. Interestingly, only eighteen of the 74 mutant polymerases tested showed activity with standard dNTPs under these assay conditions.

Ideally, a selection would have been performed using the RD Library to identify polymerases able to incorporate dψUTP with efficiency. Since none were able to incorporate the NSB more efficiently than the co-*Taq* polymerase, as evidenced by the densities of the FLP bands, a selection was performed to identify those polymerases that showed activity with the dNTPs under these assay conditions. A water-in-oil emulsion system, similar to that Ghadessy *et al.* described (Ghadessy et al., 2001), was used as a means to link geneotype to phenotype, forcing active polymerases to replicate their own genes in a PCR reaction. All 74 cell lines containing the RD-Library were used in equal proportions to perform such a selection. After products were extracted from the emulsion system, they were recloned into a plasmid using a version of the megaprimer PCR (Miyazaki and Takenouchi, 2002).

The megaprimer PCR method was chosen as the method for recombining the polymerase genes with the plasmid based on its "one pot" approach. After extracting the final products from the emulsions, all further recloning can take place in one reaction vessel, and undergoes only one purification step prior to transformation into a cell line. Other methods, using digestions and ligations, require several purification steps between the various procedures, resulting in low yields of final product.

After sequencing, it was noted that 22 out of the 50 clones sequenced contained the original co-*Taq* polymerase sequence; 15 carried partial forms of the original RD Library sequences, and only four were true RD library sequences. The remaining nine sequences were

random mutations most likely created during the PCR in the emulsions. This could be due to the fact that *Taq* polymerase has an error rate of approximately 8 x 10^{-6} (mutational frequency/bp/duplication) (Cline et al., 1996). It is also noteworthy that two of 50 sequences (SW119 and SW122) contained frameshift mutations, which tend to occur once every 2.4 x 10^{-5} base pairs when using *Taq* polymerase (Tindall and Kunkel, 1988).

Since the plasmid carrying the co-*taq* gene was only introduced during the megaprimer PCR, and the plasmid used as template was digested with *Dpn*I, it was determined that during the course of the megaprimer PCR reaction, recombinations and reversions of the various sequences most likely occurred during this procedure. This would explain the high number of co-*Taq* sequences and the large number that contain various additions, reversions, and crossovers relative to the original RD Library mutations. This also accounts for the presence of the numerous co-*Taq* polymerase clones identified after sequencing.

Out of the four exact RD library sequences that were recovered, only one coded for a mutant that was previously shown to have activity in the assay using dψUTP. This could indicate that the emulsions are breaking, allowing active polymerases to replicate the genes of inactive polymerases. Further tests could be performed to confirm or deny this conclusion; an example would be using two different cell lines in an emulsion, one expressing active polymerase and one expressing inactive polymerase. Identification of the final product would allow us to determine if indeed these emulsions are rupturing. If this is the case, modifications could be made to the oil phase of the emulsions, such as increasing the percentage of Arlacel P135, to prevent this from occurring.

We have determined that the megaprimer PCR method would be an efficient way of introducing diversity into a library between rounds of selection, but it is not an effective means

for recloning if trying to identify specific products. Once the stability of the emulsion system is verified, and the recloning of the CSR products is performed using the standard digestion/ligation/transformation protocol (Sambrook et al., 1989), it is likely that we will be able to identify thermostable polymerases using this technique. The next step would be using this method with a random library, instead of a rationally designed library, to identify thermostable polymerases and/or polymerases that can incorporate C-glycosides with efficiency and fidelity. After several rounds of evolution, we may be able to identify a polymerase capable of functioning with an AEGIS.

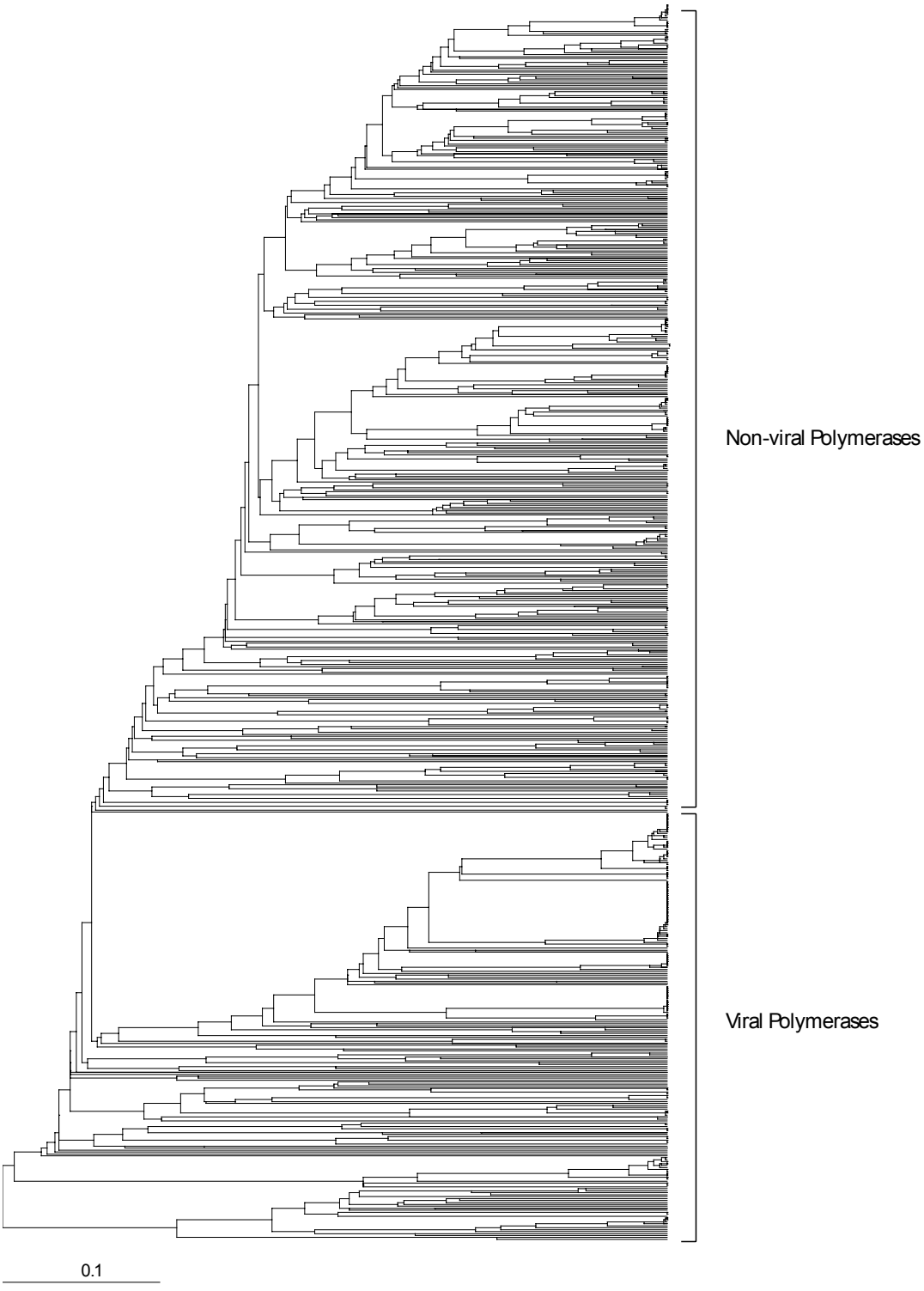


Figure 3-1. A phylogenetic tree of the Family A polymerases. This tree was generated using Pfam (Bateman, 2006, Finn et al., 2006), and analyzed for sites that underwent Type II functional divergence. Appendix B has parts of this tree expanded so that it is readable.

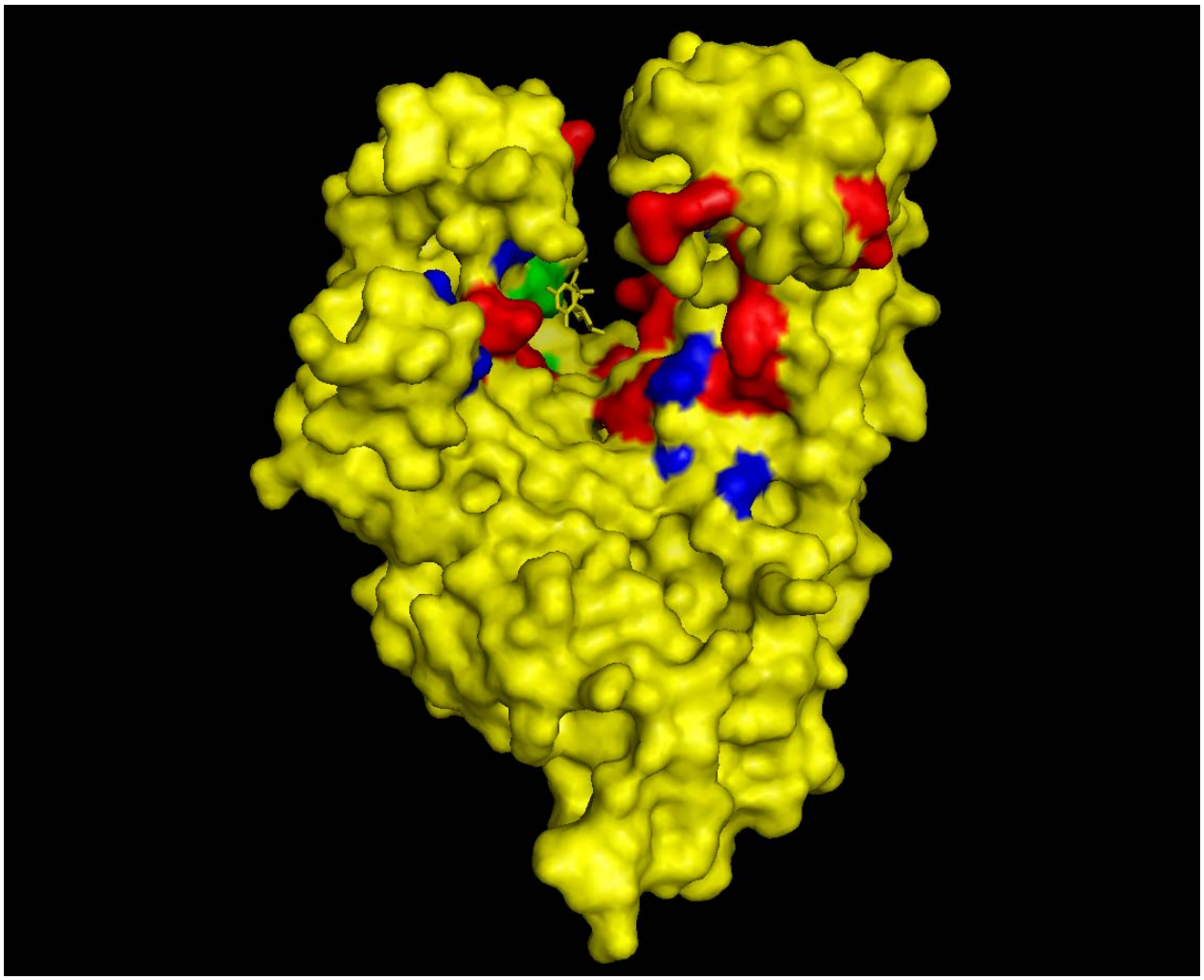


Figure 3-2. Locations of the 35 rationally designed (RD) sites in the *Taq* polymerase structure. These held the mutations in the RD Library. There were 57 mutations made at these sites: sites in red were sites where the natural amino acid was replaced by one different amino acid. Amino acids in blue indicate sites that were replaced by two different amino acids. Sites in green represent sites where three residues were substituted for the original amino acid. Image created by Dr. Eric Gaucher using the PyMOL Molecular Graphic System (DeLano, 2002).

Table 3-1. Oligonucleotides used in this study.

Oligo	Sequence (5′→3′ Direction)	Purification
P-2	GAT GAC CGC GGT ATG CTG CCC CTC	Desalted
P-3	CAT TAC AGA CCA TGG TCA CTC CTT GGC GGA G	Desalted
P-4	CAA ATG GCT AGC AGA GGA TCG CAT CAC CAT CAC	Desalted
P-5	CAG GTC AAG CTT ATT ATT TTT CGA ACT GCG GGT GGC	Desalted
P-6	GAG TTA TTT TAC CAC TCC CT	Desalted
P-7	CGC AGT AGC GGT AAA CG	Desalted
P-8	GAA AAC CGC GCG TAA ACT GC	Desalted
P-9	CCT GGA ACA CGC GAA TCA GG	Desalted

*All oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, Iowa).

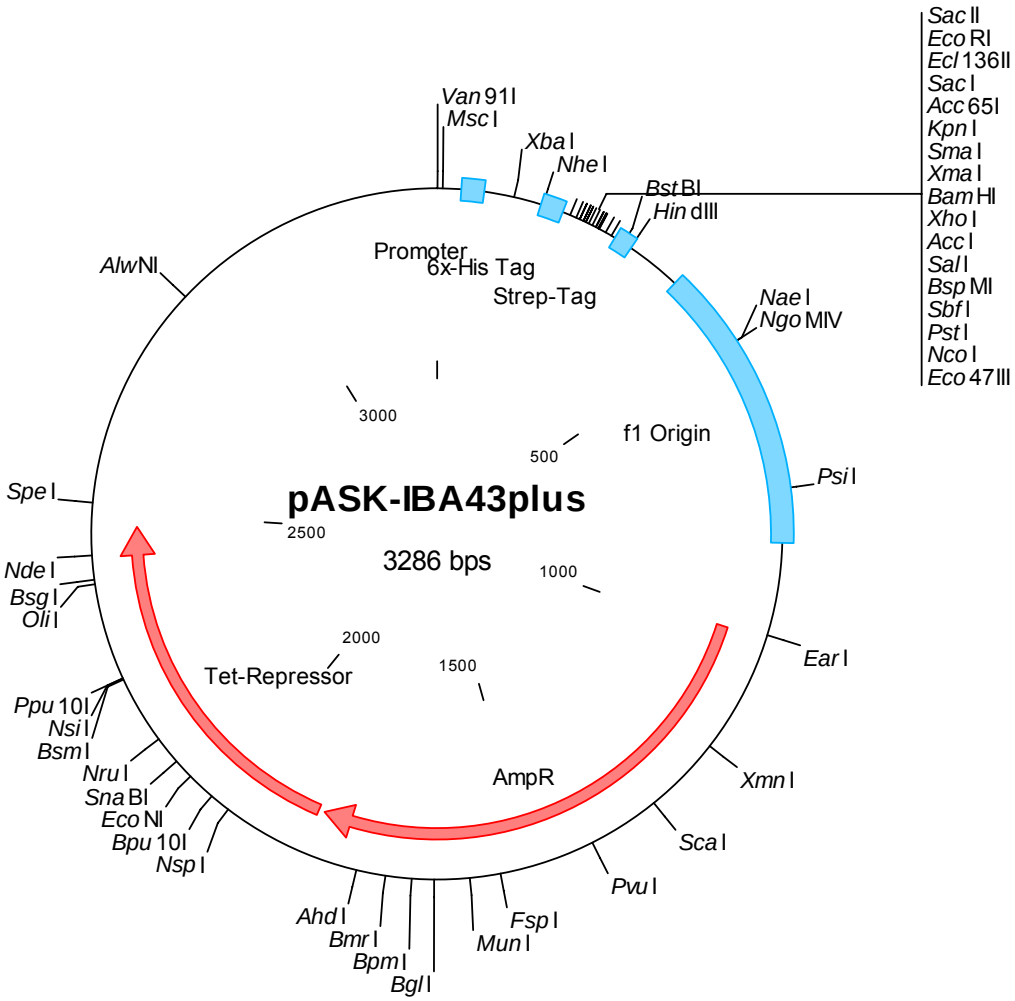


Figure 3-3. View of the pASK-IBA43plus plasmid. This plasmid was purchased from IBA GmbH (St. Louis, Missouri) and it can generate an N-terminal hexahistidine and a C-terminal *Strep*-tag®. This high copy number plasmid is a tightly controlled tetracycline expression system conferring ampicillin resistance.

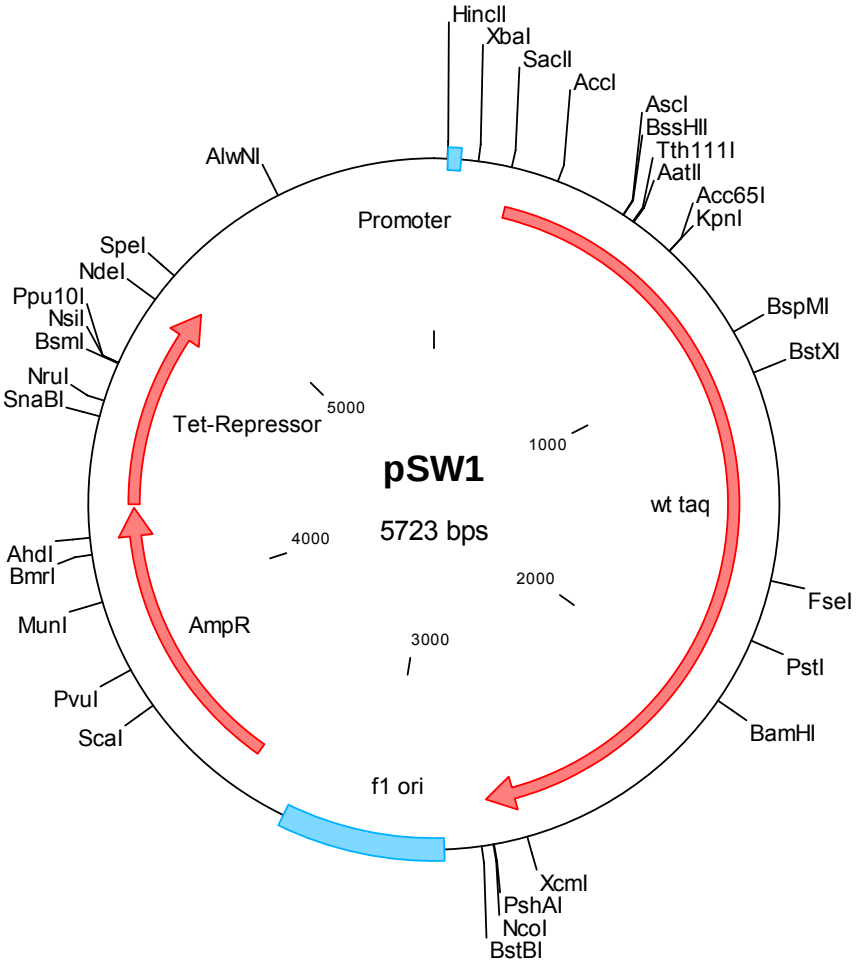


Figure 3-4. View of the pSW1 plasmid. This is a ligation of the pASK-IBA43plus plasmid with the $His_{(6)}$-*wt taq* polymerase gene using the *Sac*II and *Nco*I restriction sites. This plasmid generates an N-terminal hexahistidine translated with the $His_{(6)}$-*wt taq* gene. This high copy number plasmid is a tightly controlled tetracycline expression system conferring ampicillin resistance.

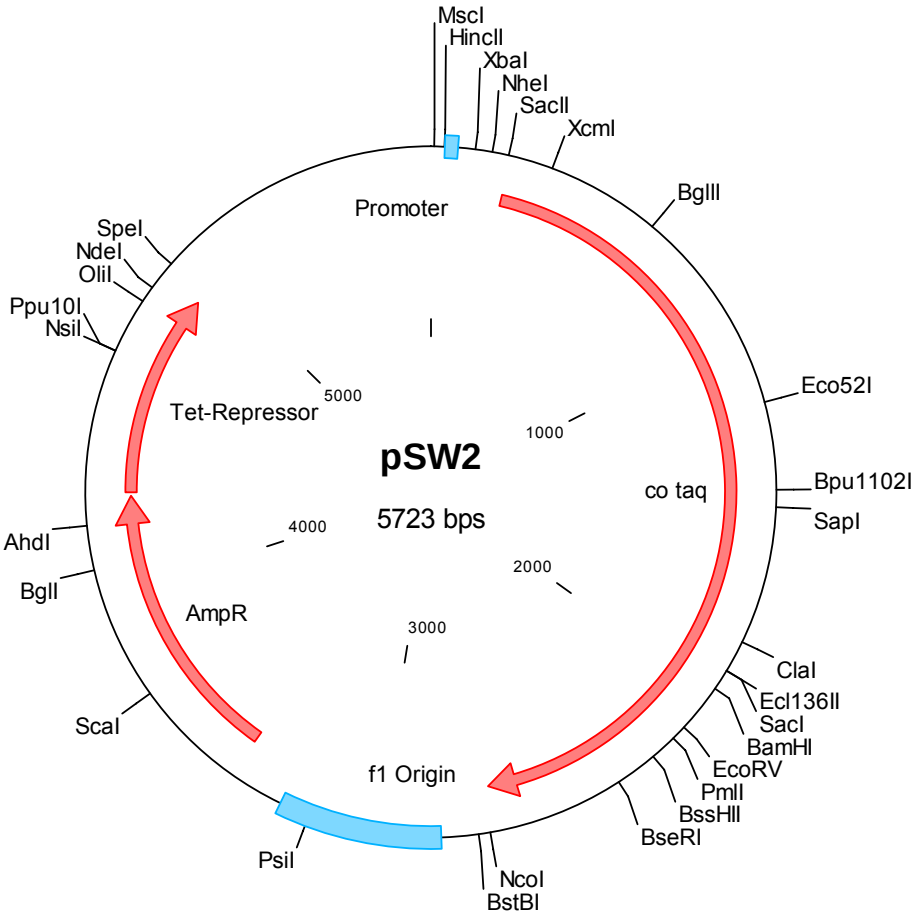


Figure 3-5. View of the pSW2 plasmid. This is a ligation of the pASK-IBA43plus plasmid with the codon-optimized *taq* polymerase gene using the *Sac*II and *Nco*I restriction sites. This plasmid generates an N-terminal hexahistidine translated with the co-*taq* gene. This high copy number plasmid is a tightly controlled tetracycline expression system conferring ampicillin resistance.

Table 3-2. Rationally Designed (RD) Mutant Library.

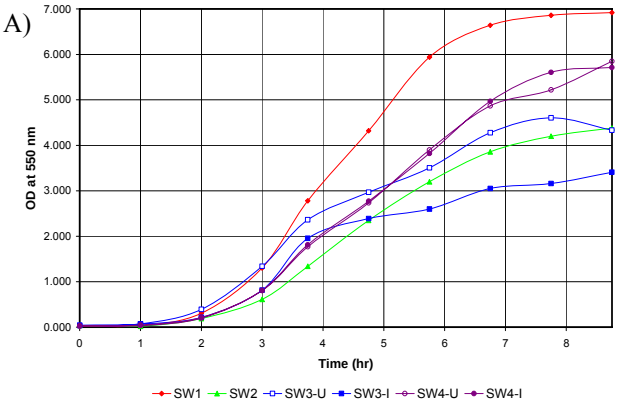
Plasmid Name	DNA 2.0 Gene ID #	Mutations Present in RD *Taq* Library	Plasmid Name	DNA 2.0 Gene ID #	Mutations Present in RD *Taq* Library
pSW3	5339	S573E,Y668F,A740S	pSW40	5383	S573H,D575T,L613I
pSW4	5340	Q486H,K537I,M670G	pSW41	5384	T541A,L606P,L613D
pSW5	5342	A605G,L613A,E739P	pSW42	5385	Y542E,V583K,A605E
pSW6	5343	D575F,L606C,A740S	pSW43	5387	E517I,F595W,A605E,I611E
pSW7	5344	T511V,R584V,I611E	pSW44	5388	T541A,D575F,L613A,D622A
pSW8	5345	N480R,F595V,E742H	pSW45	5389	T511V,A594C,L606S,A740R
pSW9	5346	E517I,V583K,A597S	pSW46	5390	Q486H,R533I,L606C,L613A
pSW10	5347	D575F,V583K,M670A	pSW47	5391	Q486H,F595V,D622A,F664Y
pSW11	5348	E517I,D607W,D622S	pSW48	5393	E517I,S573H,A605G,E612I
pSW12	5349	A594C,F664Y,A774H	pSW49	5395	Y542E,R584V,A605K,E612I
pSW13	5350	F595W,L606P,D622S	pSW50	5396	D575T,A605E,L606C,D622A
pSW14	5351	S573E,D575F,F595V	pSW51	5397	A594T,L613A,F664Y,E742H
pSW15	5352	S510I,A605K,L606S	pSW52	5398	D575F,N580Q,W601G,D622S
pSW16	5353	S573E,D622L,E742H	pSW53	5399	K537I,L606P,A740S,E742H
pSW17	5356	N480R,T511V,Y542E	pSW54	5400	A597S,W601G,L606S,F664H
pSW18	5357	A594C,F664H,M670G	pSW55	5401	S510I,E517G,D607W,I611E
pSW19	5358	Q486H,D575T,N580S	pSW56	5402	S510I,V583K,R584V,L606P
pSW20	5359	S510I,A605E,E612I	pSW57	5405	N480R,R533I,A597S,M670G
pSW21	5360	A594C,E612I,M670A	pSW58	5408	E612I,D622L,F664L,E739P
pSW22	5361	S510I,Q579A,I611Q	pSW59	5409	I611Q,M670G,E739P,E742H
pSW23	5363	A594T,L606C,R657D	pSW60	5410	F595W,F664H,Y668F,E739P
pSW24	5364	T541A,A605G,L606S	pSW61	5411	A597S,A605G,D622A,F664L
pSW25	5365	E517G,K537I,L613A	pSW62	5413	L606P,I611E,E739R,R743A
pSW26	5366	K537I,Q579A,E742V	pSW63	5414	D607W,I611Q,R657D,E742V
pSW27	5367	A597S,A740R,E742V	pSW64	5417	T541A,I611Q,L613I,D622L
pSW28	5368	N580Q,A605E,L613I	pSW65	5418	K537I,S573H,N580S,D622S
pSW29	5369	N580S,F595V,A605G	pSW66	5419	N480R,S573E,D607W,A740R
pSW30	5370	N580S,D622L,A774H	pSW67	5420	D575T,L613D,E739R,A774H
pSW31	5371	R533I,R584V,F664L	pSW68	5421	Q579A,R657D,F664Y,A740R
pSW32	5372	Q486H,E517G,A605K	pSW69	5422	R533I,K537I,A605K,L613I
pSW33	5375	S573H,F664Y,R743A	pSW70	5423	T511V,E517G,L606C,F664Y
pSW34	5376	D575T,N580Q,R584V	pSW71	5425	D575T,F664H,E742V,R743A
pSW35	5377	T541A,F664L,R743A	pSW72	5426	A594C,I611E,F664L,A740S
pSW36	5378	T511V,R533I,D622A	pSW73	5427	N580S,L613A,A740S,R743A
pSW37	5379	A597S,I611E,Y668F	pSW74	5428	S510I,T511V,L613I,E739R
pSW38	5381	Y542E,F595W,L606C	pSW75	5429	V583K,E612I,L613D,Y668F
pSW39	5382	L606S,R657D,E739R	pSW76	5430	S573E,R584V,A594C,D622S

*The pink cells denote the sequences of polymerases showing activity. The blue cells signify the sequences of polymerases that lack evidence of activity under these assay conditions. All are derivatives of the co-*taq* gene and inserted into the pASK-IBA43plus vector. Mutations were designed by Dr. Eric Gaucher (Foundation for Applied Molecular Evolution) and were synthesized and assembled by DNA 2.0.

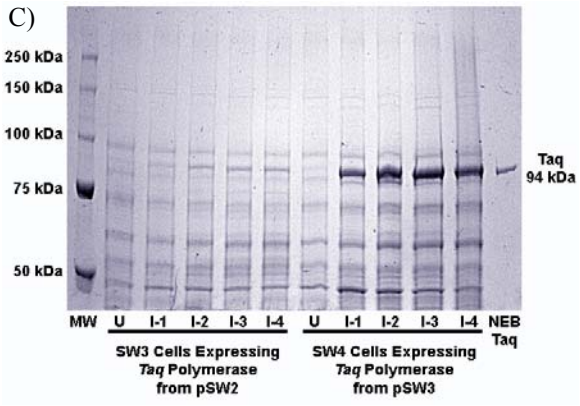
Table 3-3. Bacterial strains used in this study.

Name	Strain	Genotype
SW1	*E. coli* TG-1	F'traD36 lacI^{q}Δ(lacZ) M15 proA$^+$B$^+$ /supE Δ(hsdM-mcrB)5 (r_k - m_k - McrB$^-$) thi Δ(lac-proAB)
SW2	*E. coli* TG-1	SW1/pASK-IBA43+
SW3	*E. coli* TG-1	SW1/pSW1 (pASK-IBA43+ with *wt taq* insert, Ap^r)
SW4	*E. coli* TG-1	SW1/pSW2 (pASK-IBA43+ with co *taq* insert, Ap^r)
SW5	*E. coli* TG-1	SW1/pSW3 (pASK-IBA43+ with co *taq* mutant 5339, Ap^r)
SW6	*E. coli* TG-1	SW1/pSW4 (pASK-IBA43+ with co *taq* mutant 5340, Ap^r)
SW7	*E. coli* TG-1	SW1/pSW5 (pASK-IBA43+ with co *taq* mutant 5342, Ap^r)
SW8	*E. coli* TG-1	SW1/pSW6 (pASK-IBA43+ with co *taq* mutant 5343, Ap^r)
SW9	*E. coli* TG-1	SW1/pSW7 (pASK-IBA43+ with co *taq* mutant 5344, Ap^r)
SW10	*E. coli* TG-1	SW1/pSW8 (pASK-IBA43+ with co *taq* mutant 5345, Ap^r)
SW11	*E. coli* TG-1	SW1/pSW9 (pASK-IBA43+ with co *taq* mutant 5346, Ap^r)
SW12	*E. coli* TG-1	SW1/pSW10 (pASK-IBA43+ with co *taq* mutant 5347, Ap^r)
SW13	*E. coli* TG-1	SW1/pSW11 (pASK-IBA43+ with co *taq* mutant 5348, Ap^r)
SW14	*E. coli* TG-1	SW1/pSW12 (pASK-IBA43+ with co *taq* mutant 5349, Ap^r)
SW15	*E. coli* TG-1	SW1/pSW13 (pASK-IBA43+ with co *taq* mutant 5350, Ap^r)
SW16	*E. coli* TG-1	SW1/pSW14 (pASK-IBA43+ with co *taq* mutant 5351, Ap^r)
SW17	*E. coli* TG-1	SW1/pSW15 (pASK-IBA43+ with co *taq* mutant 5352, Ap^r)
SW18	*E. coli* TG-1	SW1/pSW16 (pASK-IBA43+ with co *taq* mutant 5353, Ap^r)
SW19	*E. coli* TG-1	SW1/pSW17 (pASK-IBA43+ with co *taq* mutant 5356, Ap^r)
SW20	*E. coli* TG-1	SW1/pSW18 (pASK-IBA43+ with co *taq* mutant 5357, Ap^r)
SW21	*E. coli* TG-1	SW1/pSW19 (pASK-IBA43+ with co *taq* mutant 5358, Ap^r)
SW22	*E. coli* TG-1	SW1/pSW20 (pASK-IBA43+ with co *taq* mutant 5359, Ap^r)
SW23	*E. coli* TG-1	SW1/pSW21 (pASK-IBA43+ with co *taq* mutant 5360, Ap^r)
SW24	*E. coli* TG-1	SW1/pSW22 (pASK-IBA43+ with co *taq* mutant 5361, Ap^r)
SW25	*E. coli* TG-1	SW1/pSW23 (pASK-IBA43+ with co *taq* mutant 5363, Ap^r)
SW26	*E. coli* TG-1	SW1/pSW24 (pASK-IBA43+ with co *taq* mutant 5364, Ap^r)
SW27	*E. coli* TG-1	SW1/pSW25 (pASK-IBA43+ with co *taq* mutant 5365, Ap^r)
SW28	*E. coli* TG-1	SW1/pSW26 (pASK-IBA43+ with co *taq* mutant 5366, Ap^r)
SW29	*E. coli* TG-1	SW1/pSW27 (pASK-IBA43+ with co *taq* mutant 5367, Ap^r)
SW30	*E. coli* TG-1	SW1/pSW28 (pASK-IBA43+ with co *taq* mutant 5368, Ap^r)
SW31	*E. coli* TG-1	SW1/pSW29 (pASK-IBA43+ with co *taq* mutant 5369, Ap^r)
SW32	*E. coli* TG-1	SW1/pSW30 (pASK-IBA43+ with co *taq* mutant 5370, Ap^r)
SW33	*E. coli* TG-1	SW1/pSW31 (pASK-IBA43+ with co *taq* mutant 5371, Ap^r)
SW34	*E. coli* TG-1	SW1/pSW32 (pASK-IBA43+ with co *taq* mutant 5372, Ap^r)
SW35	*E. coli* TG-1	SW1/pSW33 (pASK-IBA43+ with co *taq* mutant 5375, Ap^r)
SW36	*E. coli* TG-1	SW1/pSW34 (pASK-IBA43+ with co *taq* mutant 5376, Ap^r)
SW37	*E. coli* TG-1	SW1/pSW35 (pASK-IBA43+ with co *taq* mutant 5377, Ap^r)
SW38	*E. coli* TG-1	SW1/pSW36 (pASK-IBA43+ with co *taq* mutant 5378, Ap^r)
SW39	*E. coli* TG-1	SW1/pSW37 (pASK-IBA43+ with co *taq* mutant 5379, Ap^r)
SW40	*E. coli* TG-1	SW1/pSW38 (pASK-IBA43+ with co *taq* mutant 5381, Ap^r)
SW41	*E. coli* TG-1	SW1/pSW39 (pASK-IBA43+ with co *taq* mutant 5382, Ap^r)
SW42	*E. coli* TG-1	SW1/pSW40 (pASK-IBA43+ with co *taq* mutant 5383, Ap^r)
SW43	*E. coli* TG-1	SW1/pSW41 (pASK-IBA43+ with co *taq* mutant 5384, Ap^r)
SW44	*E. coli* TG-1	SW1/pSW42 (pASK-IBA43+ with co *taq* mutant 5385, Ap^r)
SW45	*E. coli* TG-1	SW1/pSW43 (pASK-IBA43+ with co *taq* mutant 5387, Ap^r)
SW46	*E. coli* TG-1	SW1/pSW44 (pASK-IBA43+ with co *taq* mutant 5388, Ap^r)
SW47	*E. coli* TG-1	SW1/pSW45 (pASK-IBA43+ with co *taq* mutant 5389, Ap^r)
SW48	*E. coli* TG-1	SW1/pSW46 (pASK-IBA43+ with co *taq* mutant 5390, Ap^r)
SW49	*E. coli* TG-1	SW1/pSW47 (pASK-IBA43+ with co *taq* mutant 5391, Ap^r)
SW50	*E. coli* TG-1	SW1/pSW48 (pASK-IBA43+ with co *taq* mutant 5393, Ap^r)
SW51	*E. coli* TG-1	SW1/pSW49 (pASK-IBA43+ with co *taq* mutant 5395, Ap^r)
SW52	*E. coli* TG-1	SW1/pSW50 (pASK-IBA43+ with co *taq* mutant 5396, Ap^r)
SW53	*E. coli* TG-1	SW1/pSW51 (pASK-IBA43+ with co *taq* mutant 5397, Ap^r)
SW54	*E. coli* TG-1	SW1/pSW52 (pASK-IBA43+ with co *taq* mutant 5398, Ap^r)
SW55	*E. coli* TG-1	SW1/pSW53 (pASK-IBA43+ with co *taq* mutant 5399, Ap^r)
SW56	*E. coli* TG-1	SW1/pSW54 (pASK-IBA43+ with co *taq* mutant 5400, Ap^r)
SW57	*E. coli* TG-1	SW1/pSW55 (pASK-IBA43+ with co *taq* mutant 5401, Ap^r)
SW58	*E. coli* TG-1	SW1/pSW56 (pASK-IBA43+ with co *taq* mutant 5402, Ap^r)
SW59	*E. coli* TG-1	SW1/pSW57 (pASK-IBA43+ with co *taq* mutant 5405, Ap^r)
SW60	*E. coli* TG-1	SW1/pSW58 (pASK-IBA43+ with co *taq* mutant 5408, Ap^r)
SW61	*E. coli* TG-1	SW1/pSW59 (pASK-IBA43+ with co *taq* mutant 5409, Ap^r)
SW62	*E. coli* TG-1	SW1/pSW60 (pASK-IBA43+ with co *taq* mutant 5410, Ap^r)
SW63	*E. coli* TG-1	SW1/pSW61 (pASK-IBA43+ with co *taq* mutant 5411, Ap^r)
SW64	*E. coli* TG-1	SW1/pSW62 (pASK-IBA43+ with co *taq* mutant 5413, Ap^r)
SW65	*E. coli* TG-1	SW1/pSW63 (pASK-IBA43+ with co *taq* mutant 5414, Ap^r)
SW66	*E. coli* TG-1	SW1/pSW64 (pASK-IBA43+ with co *taq* mutant 5417, Ap^r)
SW67	*E. coli* TG-1	SW1/pSW65 (pASK-IBA43+ with co *taq* mutant 5418, Ap^r)
SW68	*E. coli* TG-1	SW1/pSW66 (pASK-IBA43+ with co *taq* mutant 5419, Ap^r)
SW69	*E. coli* TG-1	SW1/pSW67 (pASK-IBA43+ with co *taq* mutant 5420, Ap^r)
SW70	*E. coli* TG-1	SW1/pSW68 (pASK-IBA43+ with co *taq* mutant 5421, Ap^r)
SW71	*E. coli* TG-1	SW1/pSW69 (pASK-IBA43+ with co *taq* mutant 5422, Ap^r)
SW72	*E. coli* TG-1	SW1/pSW70 (pASK-IBA43+ with co *taq* mutant 5423, Ap^r)
SW73	*E. coli* TG-1	SW1/pSW71 (pASK-IBA43+ with co *taq* mutant 5425, Ap^r)
SW74	*E. coli* TG-1	SW1/pSW72 (pASK-IBA43+ with co *taq* mutant 5426, Ap^r)
SW75	*E. coli* TG-1	SW1/pSW73 (pASK-IBA43+ with co *taq* mutant 5427, Ap^r)
SW76	*E. coli* TG-1	SW1/pSW74 (pASK-IBA43+ with co *taq* mutant 5428, Ap^r)
SW77	*E. coli* TG-1	SW1/pSW75 (pASK-IBA43+ with co *taq* mutant 5429, Ap^r)
SW78	*E. coli* TG-1	SW1/pSW76 (pASK-IBA43+ with co *taq* mutant 5430, Ap^r)
SW79	*E. coli* DH5α	SW1/pCSRMut1 (pASK-IBA43+ with co *taq* CSR mut 1, Ap$_r$)
SW80	*E. coli* DH5α	SW1/pCSRMut2 (pASK-IBA43+ with co *taq* CSR mut 2, Ap$_r$)
SW81	*E. coli* DH5α	SW1/pCSRMut3 (pASK-IBA43+ with co *taq* CSR mut 3, Ap$_r$)
SW82	*E. coli* DH5α	SW1/pCSRMut4 (pASK-IBA43+ with co *taq* CSR mut 4, Ap$_r$)
SW83	*E. coli* DH5α	SW1/pCSRMut5 (pASK-IBA43+ with co *taq* CSR mut 5, Ap$_r$)
SW84	*E. coli* DH5α	SW1/pCSRMut6 (pASK-IBA43+ with co *taq* CSR mut 6, Ap$_r$)
SW85	*E. coli* DH5α	SW1/pCSRMut7 (pASK-IBA43+ with co *taq* CSR mut 7, Ap$_r$)
SW86	*E. coli* DH5α	SW1/pCSRMut8 (pASK-IBA43+ with co *taq* CSR mut 8, Ap$_r$)
SW87	*E. coli* DH5α	SW1/pCSRMut9 (pASK-IBA43+ with co *taq* CSR mut 9, Ap$_r$)
SW88	*E. coli* DH5α	SW1/pCSRMut10 (pASK-IBA43+ with co *taq* CSR mut 10, Ap$_r$)
SW89	*E. coli* DH5α	SW1/pCSRMut11 (pASK-IBA43+ with co *taq* CSR mut 11, Ap$_r$)
SW90	*E. coli* DH5α	SW1/pCSRMut12 (pASK-IBA43+ with co *taq* CSR mut 12, Ap$_r$)
SW91	*E. coli* DH5α	SW1/pCSRMut13 (pASK-IBA43+ with co *taq* CSR mut 13, Ap$_r$)
SW92	*E. coli* DH5α	SW1/pCSRMut14 (pASK-IBA43+ with co *taq* CSR mut 14, Ap$_r$)
SW93	*E. coli* DH5α	SW1/pCSRMut15 (pASK-IBA43+ with co *taq* CSR mut 15, Ap$_r$)
SW94	*E. coli* DH5α	SW1/pCSRMut16 (pASK-IBA43+ with co *taq* CSR mut 16, Ap$_r$)
SW95	*E. coli* DH5α	SW1/pCSRMut17 (pASK-IBA43+ with co *taq* CSR mut 17, Ap$_r$)
SW96	*E. coli* DH5α	SW1/pCSRMut18 (pASK-IBA43+ with co *taq* CSR mut 18, Ap$_r$)
SW97	*E. coli* DH5α	SW1/pCSRMut19 (pASK-IBA43+ with co *taq* CSR mut 19, Ap$_r$)
SW98	*E. coli* DH5α	SW1/pCSRMut20 (pASK-IBA43+ with co *taq* CSR mut 20, Ap$_r$)
SW99	*E. coli* DH5α	SW1/pCSRMut21 (pASK-IBA43+ with co *taq* CSR mut 21, Ap$_r$)
SW100	*E. coli* DH5α	SW1/pCSRMut22 (pASK-IBA43+ with co *taq* CSR mut 22, Ap$_r$)
SW101	*E. coli* DH5α	SW1/pCSRMut23 (pASK-IBA43+ with co *taq* CSR mut 23, Ap$_r$)
SW102	*E. coli* DH5α	SW1/pCSRMut24 (pASK-IBA43+ with co *taq* CSR mut 24, Ap$_r$)
SW103	*E. coli* DH5α	SW1/pCSRMut25 (pASK-IBA43+ with co *taq* CSR mut 25, Ap$_r$)
SW104	*E. coli* DH5α	SW1/pCSRMut26 (pASK-IBA43+ with co *taq* CSR mut 26, Ap$_r$)
SW105	*E. coli* DH5α	SW1/pCSRMut27 (pASK-IBA43+ with co *taq* CSR mut 27, Ap$_r$)
SW106	*E. coli* DH5α	SW1/pCSRMut28 (pASK-IBA43+ with co *taq* CSR mut 28, Ap$_r$)
SW107	*E. coli* DH5α	SW1/pCSRMut29 (pASK-IBA43+ with co *taq* CSR mut 29, Ap$_r$)
SW108	*E. coli* DH5α	SW1/pCSRMut30 (pASK-IBA43+ with co *taq* CSR mut 30, Ap$_r$)
SW109	*E. coli* DH5α	SW1/pCSRMut31 (pASK-IBA43+ with co *taq* CSR mut 31, Ap$_r$)
SW110	*E. coli* DH5α	SW1/pCSRMut32 (pASK-IBA43+ with co *taq* CSR mut 32, Ap$_r$)
SW111	*E. coli* DH5α	SW1/pCSRMut33 (pASK-IBA43+ with co *taq* CSR mut 33, Ap$_r$)
SW112	*E. coli* DH5α	SW1/pCSRMut34 (pASK-IBA43+ with co *taq* CSR mut 34, Ap$_r$)
SW113	*E. coli* DH5α	SW1/pCSRMut35 (pASK-IBA43+ with co *taq* CSR mut 35, Ap$_r$)
SW114	*E. coli* DH5α	SW1/pCSRMut36 (pASK-IBA43+ with co *taq* CSR mut 36, Ap$_r$)
SW115	*E. coli* DH5α	SW1/pCSRMut37 (pASK-IBA43+ with co *taq* CSR mut 37, Ap$_r$)
SW116	*E. coli* DH5α	SW1/pCSRMut38 (pASK-IBA43+ with co *taq* CSR mut 38, Ap$_r$)
SW117	*E. coli* DH5α	SW1/pCSRMut39 (pASK-IBA43+ with co *taq* CSR mut 39, Ap$_r$)
SW118	*E. coli* DH5α	SW1/pCSRMut40 (pASK-IBA43+ with co *taq* CSR mut 40, Ap$_r$)
SW119	*E. coli* DH5α	SW1/pCSRMut41 (pASK-IBA43+ with co *taq* CSR mut 41, Ap$_r$)
SW120	*E. coli* DH5α	SW1/pCSRMut42 (pASK-IBA43+ with co *taq* CSR mut 42, Ap$_r$)
SW121	*E. coli* DH5α	SW1/pCSRMut43 (pASK-IBA43+ with co *taq* CSR mut 43, Ap$_r$)
SW122	*E. coli* DH5α	SW1/pCSRMut44 (pASK-IBA43+ with co *taq* CSR mut 44, Ap$_r$)
SW123	*E. coli* DH5α	SW1/pCSRMut45 (pASK-IBA43+ with co *taq* CSR mut 45, Ap$_r$)
SW124	*E. coli* DH5α	SW1/pCSRMut46 (pASK-IBA43+ with co *taq* CSR mut 46, Ap$_r$)
SW125	*E. coli* DH5α	SW1/pCSRMut47 (pASK-IBA43+ with co *taq* CSR mut 47, Ap$_r$)
SW126	*E. coli* DH5α	SW1/pCSRMut48 (pASK-IBA43+ with co *taq* CSR mut 48, Ap$_r$)
SW127	*E. coli* DH5α	SW1/pCSRMut49 (pASK-IBA43+ with co *taq* CSR mut 49, Ap$_r$)
SW128	*E. coli* DH5α	SW1/pCSRMut50 (pASK-IBA43+ with co *taq* CSR mut 50, Ap$_r$)
SW129	*E. coli* DH5α	SW1/pCSRwt1 (pASK-IBA43+ with co *taq wt* mut 1, Ap$_r$)
SW130	*E. coli* DH5α	SW1/pCSRwt2 (pASK-IBA43+ with co *taq wt* mut 2, Ap$_r$)
SW131	*E. coli* DH5α	SW1/pCSRwt3 (pASK-IBA43+ with co *taq wt* mut 3, Ap$_r$)
SW132	*E. coli* DH5α	SW1/pCSRwt4 (pASK-IBA43+ with co *taq wt* mut 4, Ap$_r$)
SW133	*E. coli* DH5α	SW1/pCSRwt5 (pASK-IBA43+ with co *taq wt* mut 5, Ap$_r$)
SW134	*E. coli* DH5α	F$^-$ 80d*lac* ZΔM15 Δ(*lac* ZYA-*arg* F) U169 *rec* A1 *end* A1 *hsd* R17(r_{k-}, m_{k+}) *gal*$^-$ *pho* A *sup* E44 λ$^-$ *thi*$^-$1 *gyr* A96 *rel* A1

A)



C)



B)

Time (hr)	Colony Counts (cfu/mL)					
	SW1	**SW2**	**SW3-U**	**SW3-I**	**SW4-U**	**SW4-I**
0	1.36E+07	2.01E+07	6.10E+06	7.95E+06	6.00E+06	7.22E+06
1	5.40E+07	1.31E+07	1.22E+07	8.73E+06	9.53E+06	1.04E+07
2	1.80E+08	3.04E+07	8.14E+07	5.21E+07	2.65E+07	3.09E+07
3	6.48E+08	8.63E+07	2.87E+08	2.79E+08	3.88E+08	1.90E+08
3.75	1.47E+09	3.22E+08	6.83E+08	4.93E+08	4.09E+08	4.36E+08
4.75	2.73E+09	7.95E+08	1.08E+09	2.36E+08	8.26E+08	5.66E+08
5.75	2.94E+09	1.01E+09	1.08E+09	5.07E+07	1.10E+09	6.29E+08
6.75	6.06E+09	1.39E+09	1.88E+09	1.22E+07	1.53E+09	5.01E+08
7.75	3.08E+09	1.11E+09	1.77E+09	7.39E+06	2.23E+09	7.63E+08
8.75	3.14E+09	1.14E+09	1.55E+09	3.90E+06	2.15E+09	5.69E+08

Figure 3-6. Growth curves, cell counts, and expression of various *E. coli* TG-1 cell lines. The SW3 (denoted SW3-I) and SW4 (denoted SW4-I) cell lines were induced after 3.75 hrs with a final concentration of 0.2 ng/µL anhydrotetracycline. A) Growth curves of various cell lines. Samples were grown in LB media, cultures SW2 – SW4 were supplemented with ampicillin (100 µg/mL final concentration), at 250 rpm and 37 °C for 8.75 hrs. B) Colony counts (cfu/mL) of each of the cell lines in part A at the various time points. Cells were grown on LB or LB-Amp agar overnight at 37 °C. C) Coomassie Blue stained SDS-PAGE (7.5%) gel showing protein expression of induced cells at various time points. U stands for uninduced cells, I-1 through I-4 indicate time-points at hours one through four after induction (t = 4.75 through t = 7.75 hrs), and NEB *Taq* depicts the migration of the 94 kDa *Taq* polymerase purchased from New England BioLabs. Since their genetic code has been optimized for use in *E. coli* cells, the SW4 strain, containing the co-*taq* gene, appear to grow to a higher OD550nm than the SW3 strain containing the $His_{(6)}$-*wt taq* gene.

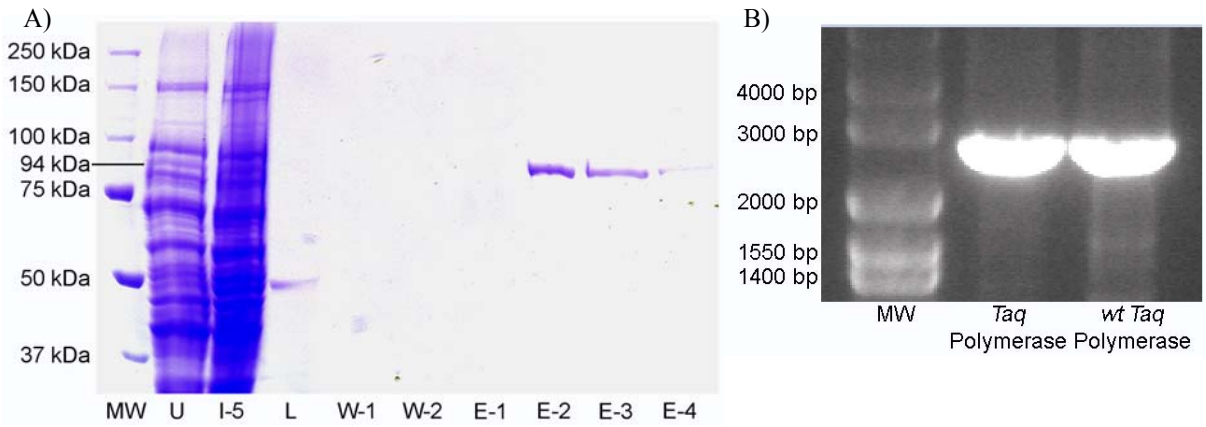


Figure 3-7. Purification and activity of $His_{(6)}$-*wt Taq* polymerase. A) The purification of $His_{(6)}$-*wt Taq* polymerase from SW3 cells after five hours of induction. U – uninduced cells, I-5 – cells after 5 hrs of induction, L – load from the Ni^{2+} column, W-1 and W-2 – wash fractions from the column, E-1 through E-4 – elution fractions from the column. Elution fractions 2 – 4 were combined and subjected to dialysis. B) Products of PCRs comparing identical concentrations of *Taq* polymerase (New England BioLabs) and $His_{(6)}$-*wt Taq* polymerase. The amount of product generated with each polymerase was almost identical considering the density of the product band using *Taq* polymerase was 1980 CNT/mm^2 and the density of the product band using $His_{(6)}$-*wt Taq* polymerase was 1925 CNT/mm^2.

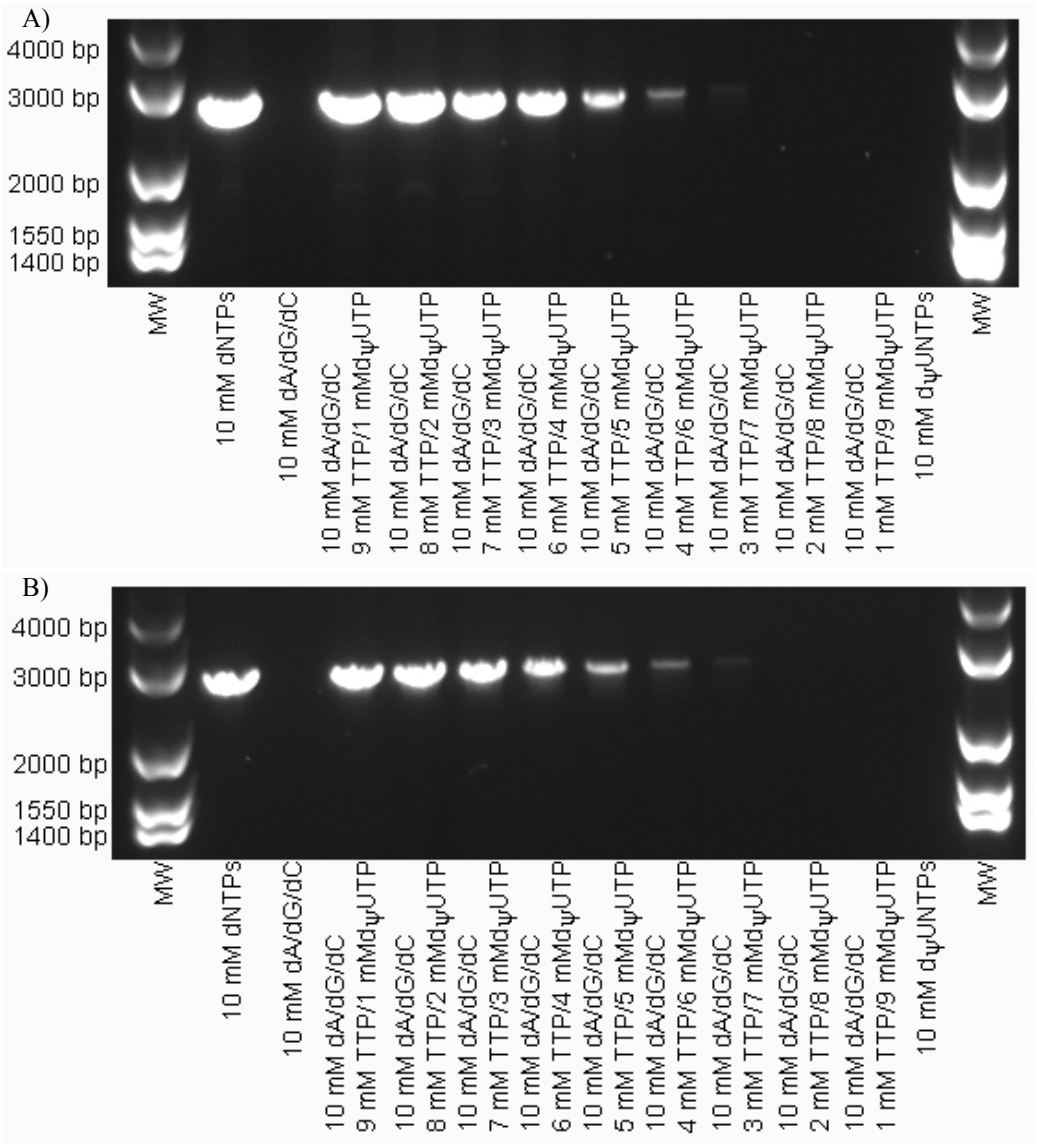


Figure 3-8. Representative gels showing the amount of full-length PCR products generated with different dNTP/dψUNTP ratios and the indicated polymerases. Concentrations of dNTPs/dψUNTPs listed are the starting concentrations (see Materials and Methods for listing of final concentrations). All PCRs used 1 x 10^6 cfu of cells expressing polymerase as the sole source of polymerase and template plasmid for the reaction. Polymerases were forced to replicate their own encoding gene (2603 bp). A) Incorporation of various dNTP/dψUNTP ratios by co-*Taq* polymerase. FLP is not generated beyond the ratio of 3 mM TTP/7 mM dψUTP. B) Incorporation of various dNTP/dψUNTP ratios by a representative of the RD Library (SW21). FLP is not generated beyond the ratio of 3 mM TTP/7 mM dψUTP.

Table 3-4. Incorporation of dψUTP at 94.0 °C by RD Library.

Cell Line	Substitutions	Raw Densities (CNT/mm^2)										
		All dNTPs	9 mM dT/ 1 mM dψU	8 mM dT/ 2 mM dψU	7 mM dT/ 3 mM dψU	6 mM dT/ 4 mM dψU	5 mM dT/ 5 mM dψU	4 mM dT/ 6 mM dψU	3 mM dT/ 7 mM dψU	2 mM dT/ 8 mM dψU	1 mM dT/ 9 mM dψU	All dψUNTPs
SW4	Codon-Optimized (co) *wt* Taq	2244256	2005371	1995649	1535822	1255379	589637	188752	64360	0	0	0
SW8	D575F,L606C,A740S	585777	540589	407327	307603	202787	118788	52491	19033	0	0	0
SW10	N480R,F595V,E742H	547281	164779	170880	221286	265684	97901	49321	0	0	0	0
SW11	E517I,V583K,A597S	919340	836722	747138	620995	412059	162453	62919	19455	0	0	0
SW12	D575F,V583K,M670A	24794	26679	15530	20064	0	0	0	0	0	0	0
SW14	A594C,F664Y,A774H	669933	129967	142533	99426	59152	31362	0	0	0	0	0
SW17	S510I,A605K,L606S	32536	0	0	0	0	0	0	0	0	0	0
SW21	Q486H,D575T,N580S	1344877	1310961	1174649	999650	640497	333511	112569	48284	0	0	0
SW25	A594T,L606C,R657D	509241	201540	119721	69894	46402	0	0	0	0	0	0
SW27	E517G,K537I,L613A	137180	53372	32796	0	22605	0	0	0	0	0	0
SW29	A597S,A740R,E742V	766112	500823	350791	263933	233934	77446	38731	22923	0	0	0
SW30	N580Q,A605E,L613I	28402	0	0	0	0	0	0	0	0	0	0
SW31	N580S,F595V,A605G	43184	0	0	0	0	0	0	0	0	0	0
SW34	Q486H,E517G,A605K	938703	625650	493365	395906	401123	63888	30127	0	0	0	0
SW36	D575T,N580Q,R584V	40726	0	34646	35920	44580	0	0	0	0	0	0
SW41	L606S,R657D,E739R	46925	23047	0	29686	29736	0	0	0	0	0	0
SW47	T511V,A594C,L606S,A740R	66112	26538	27445	35249	30021	22991	0	0	0	0	0
SW72	T511V,E517G,L606C,F664Y	499069	309148	155466	97099	55983	17934	12553	0	0	0	0
SW76	S510I,T511V,L613I,E739R	216700	172538	119075	121527	52290	26955	0	0	0	0	0

Table 3-5. Mutations present after selection for active polymerases.

Cell Line	Mutations Present	Cell Line	Mutations Present
SW79	-	SW104	P336L,M371T,N580S,L613A
SW80	E517G,A597S,A605G,D622A	SW105	-
SW81	E468G	SW106	E504G,R533I,R584V,F664L,F697S
SW82	L11P	SW107	G195D,L230P,P552Q,D575T,F664L
SW83	-	SW108	L606P,I611E,E739R,R743A
SW84	A594C,F664Y	SW109	-
SW85	Q579A	SW110	-
SW86	-	SW111	K257E,S510I,D575F,L606C
SW87	F44S	SW112	-
SW88	-	SW113	-
SW89	T31M	SW114	M314I
SW90	-	SW115	A201E
SW91	V61L	SW116	N480R,T511A
SW92	-	SW117	-
SW93	V646I	SW118	Q486H,F595V,D622A,F664Y
SW94	D575T,F664H,F721L,E742V,R743A,E817G	SW119	S510I
SW95	I60T	SW120	-
SW96	-	SW121	G327S,H330Y,N480R,R533I
SW97	-	SW122	-
SW98	R743A	SW123	S510I,A605E,E612I
SW99	I543T,A594C,E631G,F664Y,W703R	SW124	-
SW100	Q486H,F595V,D622A	SW125	-
SW101	-	SW126	-
SW102	-	SW127	R533I,K537I,Q579A,E739R
SW103	Q486H,D575T,N580S	SW128	-

* The dashes (-) represent polymerases with no amino acid mutations relative to the co-*Taq* sequence.

Table 3-6. Breakdown of types of mutations present after selection.

Cell Line	Codon-Optimized Taq	No RD Mutations (Random Taq Mutations)	RD Library Variants	RD Variants + Add'l Mutations	RD Variants with Reversions	RD Recombinants with 1 Crossover	RD Recombinants with 2 Crossovers	# of DNA Mutations		
								Silent	Non-silent	Total
SW79	X	-	-	-	-	-	-	0	0	0
SW80	-	-	-	E517G,A597S, A605G,D622A	-	E517G,A597S, A605G,D622A	-	0	8	8
SW81	-	E468G	-	-	-	-	-	0	1	1
SW82	-	L11P	-	-	-	-	-	0	1	1
SW83	X	-	-	-	-	-	-	0	0	0
SW84	-	-	-	-	A594C,F664Y	-	-	0	3	3
SW85	-	-	-	-	Q579A	-	-	1	3	4
SW86	X	-	-	-	-	-	-	0	0	0
SW87	-	F44S	-	-	-	-	-	0	1	1
SW88	X	-	-	-	-	-	-	1	0	1
SW89	-	T31M	-	-	-	-	-	0	1	1
SW90	X	-	-	-	-	-	-	0	0	0
SW91	-	V61L	-	-	-	-	-	0	1	1
SW92	X	-	-	-	-	-	-	0	0	0
SW93	-	V646I	-	-	-	-	-	0	1	1
SW94	-	-	-	D575T,F664H, F721L,E742V, R743A,E817G	-	-	-	0	16	16
SW95	-	I60T	-	-	-	-	-	1	1	2
SW96	X	-	-	-	-	-	-	0	0	0
SW97	X	-	-	-	-	-	-	0	0	0
SW98	-	-	-	-	R743A	-	-	0	3	3
SW99	-	-	-	I543T,A594C, E631G,F664Y, W703R	I543T,A594C, E631G,F664Y, W703R	-	-	0	6	6
SW100	-	-	-	-	Q486H,F595V, D622A	-	-	0	4	4
SW101	X	-	-	-	-	-	-	0	0	0
SW102	X	-	-	-	-	-	-	0	0	0
SW103	-	-	Q486H,D575T, N580S	-	-	-	-	0	7	7
SW104	-	-	-	P336L,M371T, N580S,L613A	P336L,M371T, N580S,L613A	-	-	0	7	7
SW105	X	-	-	-	-	-	-	0	0	0
SW106	-	-	-	E504G,R533I, R584V,F664L, F697S	-	-	-	1	8	9
SW107	-	-	-	G195D,L230P, P552Q,D575T, F664L	G195D,L230P, P552Q,D575T, F664L	G195D,L230P, P552Q,D575T, F664L	-	1	8	9
SW108	-	-	L606P,I611E, E739R,R743A	-	-	-	-	1	10	11
SW109	X	-	-	-	-	-	-	0	0	0
SW110	X	-	-	-	-	-	-	0	0	0
SW111	-	-	-	K257E,S510I, D575F,L606C	K257E,S510I, D575F,L606C	K257E,S510I, D575F,L606C	-	1	9	10
SW112	X	-	-	-	-	-	-	0	0	0
SW113	X	-	-	-	-	-	-	0	0	0
SW114	-	M314I	-	-	-	-	-	0	1	1
SW115	-	A201E	-	-	-	-	-	0	1	1
SW116	-	-	-	-	N480R,T511A	-	-	0	4	4
SW117	X	-	-	-	-	-	-	0	0	0
SW118	-	-	Q486H,F595V, D622A,F664Y	-	-	-	-	0	5	5
SW119	-	-	-	-	S510I	-	-	0	3	3
SW120	X	-	-	-	-	-	-	0	0	0
SW121	-	-	-	-	G327A,H330Y, N480R,R533I	-	-	0	7	7
SW122	-	-	-	-	-	-	-	0	1	1
SW123	-	-	S510I,A605E, E612I	-	-	-	-	0	7	7
SW124	X	-	-	-	-	-	-	0	0	0
SW125	X	-	-	-	-	-	-	0	0	0
SW126	X	-	-	-	-	-	-	0	0	0
SW127	-	-	-	-	-	-	R533I,K537I, Q579A,E739R	0	10	10
SW128	X	-	-	-	-	-	-	0	0	0

* An "X" indicates polymerases with no amino acid mutations relative to the co-*Taq* sequence.

CHAPTER 4
DISTRIBUTION OF THERMOSTABILITY IN POLYMERASE MUTATION SPACE

Introduction

Recent years have seen a dramatic increase in the number of experiments being performed to optimize protein function utilizing directed evolution. With the rise of directed evolution, there is a proportional escalation in the number and type of approaches used to create the libraries for these selections.

Many different theories exist on the best methods to create the best library, one that contains a large number of diverse, yet active clones (Hibbert and Dalby, 2005, Arnold and Georgiou, 2003b). These theories contradict each other on fundamental levels; for example, some say it is best to use random mutagenesis throughout the entire gene (Drummond et al., 2005), and others think that it is better to perform random mutagenesis only within the region containing the active site of the protein (Park et al., 2005, Dalby, 2003). Conversely, some researchers believe that site-saturation mutagenesis at carefully selected sites generates the best results (Parikh and Matsumura, 2005), while a few consider that mutagenesis at specific sites with specific amino acids will allow for the creation of an optimal library (Crameri et al., 1998, Crameri et al., 1996, Castle et al., 2004).

Our laboratory is interested in pursuing the directed evolution of polymerases to incorporate non-standard nucleotides (NSBs), specifically those exhibiting a C-glycosidic linkage (Fig. 2-3), such as 2'-deoxypseudouridine (dψU) and 2'-deoxypseudothymidine (dψT). To determine what type of mutagenic library would best suit our needs, we compared two libraries for their ability to perform at high temperatures, a prerequisite for selection in emulsions under the Ghadessy *et al.* conditions, as well as being desired for a synthetic biology(Ghadessy et al., 2004, Ghadessy et al., 2001).

The first was the rationally designed (RD) polymerase library, designed by Dr. Eric Gaucher using the REAP approach as discussed in the previous chapter, where carefully selected residues were changed into other specific amino acids, and a random library (L4) with mutations spread across the whole polymerase sequence for their ability to function at various temperatures. The second was a randomly generated library (L4) with mutations spread across the entire polymerase sequence. The L4 library was created using error-prone PCR with *Taq* polymerase and manganese chloride serving as the mutators (Arnold and Georgiou, 2003b). The starting gene was derived from the co-*taq* polymerase gene, which is the $His_{(6)}$-*wt taq* polymerase gene whose sequence had been optimized for codon usage in *E. coli* cells (Gustafsson et al., 2004).

The 74 *Taq* polymerase mutants in each library were first tested for their ability to incorporate dNTPs at various temperatures in a PCR reaction to determine the optimal temperature at which individual polymerases performed, judging by the generation of full-length PCR products. In this case, it appeared that random mutagenesis was better able to yield thermostable variants than rational design method, but our RD library was specifically modified for identifying polymerases with altered catalytic activities, and therefore targeted sites where changes would be more likely to decrease thermostability.

Then, variants from the RD library were tested for their ability to incorporate C-glycosides using mixtures of dψUTP and TTP in different ratios, both at 94.0 °C and at their optimal temperature. We identified only one mutant with enhanced abilities over the co-*Taq* polymerase to incorporate dψUTP.

Finally, the ability of dψUTP to epimerize at high temperatures was of concern. Epimerization would lower the concentration of the β-anomer, which is the desired substrate for

the polymerase (Fig. 4-1). Therefore, parallel experiments were performed with dψTTP, which is known not to epimerize (Wellington and Benner, 2006, Cohn, 1960, Chambers et al., 1963). These results suggested that dψU is epimerizing to generate the α-epimer, suggesting that dψU is not as suitable as a C-glycoside substrate in PCR experiments generally, as well as in directed evolution studies to develop thermostable polymerases having new catalytic activities.

Materials and Methods

DNA Sequencing and Analysis

DNA sequencing was carried out by the University of Florida Interdisciplinary Center for Biotechnology Research Sequencing DNA Core Facility using an ABI 3130*xl* Genetic Analyzer (Applied Biosystems, Foster City, California) using primers P-6 through P-9 (Table 3-1). BLAST 2 Sequences software was used for sequence similarity searching (Tatusova and Madden, 1999); Derti's Reverse and/or complement DNA sequences website was used to find the reverse complement of various DNA strands (Derti, 2003); and ExPASy's translate tool was used to translate DNA sequences into their amino acid counterparts (Swiss Institute of Bioinformatics, 1999).

Bacterial Growth Conditions and Strains

The bacterial strains used in this study are listed in Table 3-3 (SW1, SW4 – SW78, and SW134) and those in Table 4-4. The rich media used in these studies was Luria-Bertani (LB) medium (Difco Laboratories, Detroit, Michigan) (Miller, 1972). Ampicillin was provided in liquid or solid medium at a final concentration of 100 μg/mL. Plasmids were transformed into the *E. coli* TG-1 cell line according to manufacturer's protocol (Zymo Research, Orange, California). Cell growth was determined by measuring optical density at 550 nm using a SmartSpec Plus Spectrophotometer (Bio-Rad, Hercules, California). Anhydrotetracycline (2

mg/mL stock in *N,N*-dimethylformamide) was used at a final concentration of 0.2 ng/μL to induce expression.

Synthesis of Triphosphates and Oligonucleotides

Dr. Shuichi Hoshika, from the Foundation for Applied Molecular Evolution (FfAME, Gainesville, Florida), synthesized the pseudothymidine precursor as described in Appendix A. Dr. Daniel Hutter (also of FfAME) synthesized 2'-deoxypseudothymidine-5'-triphosphate (dψTTP) as described in Appendix A. 2'-deoxypseudouridine-5'-triphosphate (dψUTP) was purchased from TriLink BioTechnologies (San Diego, California). Standard deoxynucleotide triphosphates (dNTPs) of 2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxycytidine-5'-triphosphate (dCTP), 2'-deoxyganosine-5'-triphosphate (dGTP), and thymidine-5'-triphosphate (TTP) were purchased from Promega Corporation (Madison, Wisconsin). dψTNTP solutions were comprised of dATP, dCTP, dGTP, and dψTTP, while dψUNTPs were comprised of dATP, dCTP, d GTP, and dψUTP.

Random Mutagenic Library (L4 Library) Creation

DNA 2.0 (Menlo Park, California) synthesized a form of the $His_{(6)}$-*wt taq* polymerase gene (co-*taq*) that was optimized for the codon-usage of *E. coli*, which was then used to construct the pSW2 plasmid (Fig. 3-3). Mutagenic PCR was performed on the co-*taq* gene to generate a library containing three to four amino acid changes per polymerase in a fashion similar to that described by Arnold and Georgiou (Arnold and Georgiou, 2003b). The PCRs contained the following: 1 X Mutagenic Taq Buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl, 15 mM $MgCl_2$), 0.1 ng/μL pSW2, 200 μM dNTPs, 300 nM P-4, 300 nM P-5, 5 U *Taq* polymerase (New England BioLabs, Beverly, Massachusetts), and $MnCl_2$ (115 μM). PCR reaction conditions were as follows: 5 min, 94 °C; (30 s, 94.0 °C; 20 s, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C.

Products were purified with the QIAquick PCR Purification Kit (Qiagen, Valencia, CA), eluted with Qiagen Buffer EB (50 μL), and quantitated at an absorbance of 260 nm using a SmartSpec Plus Spectrophotometer (Bio-Rad).

The mutagenic PCR products were used in an adaptation of the Miyazaki and Takenouchi megaprimer PCR protocol (Miyazaki and Takenouchi, 2002). Samples (10 ng/μL final concentration) were added to a PCR mixture (1X Native *Pfu* Buffer, 100 ng pSW2, 500 μM dNTPs, 6% DMSO). Mixture was heated to 96 °C for 30 s prior to the addition of 0.05 U/μL Native *Pfu* Polymerase (Stratagene, La Jolla, California). Samples were then subjected to PCR [2 min, 96 °C; (30 s, 96.0 °C; 10 min, 68.0 °C)x25 cycles; 30 min, 72.0 °C].

The host strands of DNA (the pSW2 plasmid in the PCR) were digested with 2 U *Dpn*I (New England BioLabs) at 37 °C for 2.5 hrs. Reactions cooled to room temperature, purified using a QIAquick PCR Purification Kit (Qiagen), and eluted with Qiagen Buffer EB (30 μL). Purified products were transformed into the *E. coli* DH5α cell line according to manufacturer's protocol (Invitrogen, Carlsbad, California). Seventy-nine isolated colonies were selected after the transformation (cell lines SW135 through SW211). Each colony was grown in a separate overnight 5 mL LB-Amp culture (250 rpm, 37 °C) in 14 mL 2059 Falcon Tubes (BD Biosciences, San Jose, California). Their plasmids were isolated using the QIAprep Spin Miniprep Kit (Qiagen). Plasmid constructs were verified both by restriction digest analysis, using the enzymes *Bam*HI and *Nco*I according to the manufacturer's protocol (New England BioLabs), and mutations were determined by sequencing. The 74 L4 Library plasmids containing mutations were transformed into the *E. coli* TG-1 cell line (cell lines SW212 through SW285) according to manufacturer's protocol (Zymo Research, Orange, California).

Incorporation of dNTPs by RD and L4 Libraries at Various Temperatures

A single isolated colony from each of the SW5 through SW78 (RD Library) and SW212 through SW285 (L4 Library) cell lines were used to inoculate a 148 individual cultures (5 mL LB-Amp), and were grown for 14.25 hrs at 250 rpm and 37 °C in 14 mL 2059 Falcon Tubes (BD Biosciences). Approximately 2 x 10^8 colony-forming units (cfu), roughly equal to 500 μL of a culture with an OD_{550nm} of 4.0, were used to inoculate individual 100 mL cultures of LB-Amp in 500 mL baffled flasks. These cultures were grown at 37 °C and 250 rpm for 3.75 hrs to an approximate OD_{550nm} of 1.8, and were then induced with anhydrotetracycline. The cells were allowed to grow for 1 hr longer to an approximate OD_{550nm} of 3.0.

Approximately 1 x 10^6 cfu (~2 μL cells) were used as the sole source of polymerase and template in separate PCR reactions containing final concentrations of these constituents: 1X Modified ThermoPol Buffer (2 mM Tris-HCl, pH 9, 10 mM KCl, 1 mM $(NH_4)_2SO_4$, 2.5 mM $MgCl_2$, 0.2% Tween 20), 500 μM dNTPs, 1.4 μM P-4, 1.4 μM P-5, 1.1 ng/μL RNaseA, and 6% DMSO. The PCRs (50 μL) were run under the following conditions: 5 min, X °C; (1 min, X °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C, where X was a denaturing temperature of 75.0 °C, 75.5 °C, 76.6 °C, 78.1 °C, 80.4 °C, 83.1 °C, 86.3 °C, 89.0 °C, 91.1 °C, 92.6 °C, 93.7 °C, or 94.0 °C. Products were analyzed by agarose gel electrophoresis and quantitated using the GeneTools Software, version 3.07 (SynGene, Cambridge, England).

Incorporation of dψUNTPs by RD Library at Optimal Temperatures

A single isolated colony from each of the cell lines (SW4 through SW78) that were active at one of the temperatures tested, were used to inoculate a 33 individual cultures (5 mL LB-Amp). These were grown for 14.25 hrs at 250 rpm and 37 °C in 14 mL 2059 Falcon Tubes (BD Biosciences). Approximately 2 x 10^8 cfu, roughly equal to 500 μL of a culture with an OD_{550nm}

of 4.0, were used to inoculate individual 100 mL cultures of LB-Amp in 500 mL baffled flasks. These cultures were grown at 37 °C and 250 rpm for 3.75 hrs to an approximate OD_{550nm} of 1.8, and were then induced with anhydrotetracycline. The cells were allowed to grow for 1 hr longer to an approximate OD_{550nm} of 3.0.

Approximately 1 x 10^6 cfu (~2 μL cells) were used as the source of polymerase and template in separate PCR reactions containing final concentrations of these constituents: 1X Modified ThermoPol Buffer, 1.4 μM P-4, 1.4 μM P-5, 1.1 ng/μL RNaseA, and 6% DMSO. One of the following sets of nucleotide triphosphates were added to the reactions (final concentrations): 500 μM dNTPs; 500 μM dATP/dGTP/dCTP; 500 μM dATP/dGTP/dCTP + 450 μM TTP + 50 μM dψUTP; 10 μM dATP/dGTP/dCTP + 400 μM TTP + 100 μM dψUTP; 10 μM dATP/dGTP/dCTP + 350 μM TTP + 150 μM dψUTP; 10 μM dATP/dGTP/dCTP + 300 μM TTP + 200 μM dψUTP; 10 μM dATP/dGTP/dCTP + 250 μM TTP + 250 μM dψUTP; 10 μM dATP/dGTP/dCTP + 200 μM TTP + 300 μM dψUTP; 10 μM dATP/dGTP/dCTP + 150 μM TTP + 350 μM dψUTP; 10 μM dATP/dGTP/dCTP + 100 μM TTP + 400 μM dψUTP; 10 μM dATP/dGTP/dCTP + 50 μM TTP + 450 μM dψUTP; 500 μM dψUTPs. The PCRs (50 μL) were run under the following conditions: 5 min, X °C; (1 min, X °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C, where X was each polymerases optimal denaturing temperature of 86.3 °C, 89.0 °C, 91.1 °C, 92.6 °C, 93.7 °C, or 94.0 °C. Products were analyzed by agarose gel electrophoresis and quantitated using the GeneTools Software, version 3.07 (SynGene, Cambridge, England).

Incorporation of dψUTP and dψTTP by co-*Taq* Polymerase at Various Melting Temperatures

The SW4 cell line was used to inoculate an LB-Amp culture (5 mL in a 14 mL 2059 Falcon Tube: BD Biosciences). The culture was grown for 14.25 hrs (250 rpm shaking at 37 °C). Approximately 2 x 10^8 cfu of the resulting cell suspension (ca. 500 μL of a culture with an OD_{550nm} of 4.0) was used to inoculate a secondary culture of LB-Amp (100 mL in a 500 mL baffled flask). The secondary culture was grown (37 °C, 250 rpm) for 3.75 hrs to an approximate OD_{550nm} of 1.8. Expression of the polymerase was then induced with anhydrotetracycline. The cells were allowed to grow for 1 hr longer to an approximate OD_{550nm} of 3.0.

The cells themselves (approximately 1 x 10^6 cfu or ~2 μL cells) were used as the sole source of polymerase and template in separate PCRs. These reactions contained a final concentrations of the following: 1X Modified ThermoPol Buffer, 1.4 μM P-4, 1.4 μM P-5, 1.1 ng/μL RNaseA, and 6% DMSO. The final concentration of nucleotide triphosphates added to the reactions can be found in Table 4-6. The PCRs (50 μL) were run under the following conditions: 5 min, X °C; (1 min, X °C; 1 min, 55.0 °C; 3 min, 72.0 °C)x15 cycles; 7 min, 72.0 °C, where X was a denaturing temperature of 86.3 °C, 89.0 °C, 91.1 °C, 92.6 °C, 93.7 °C, or 94.0 °C. Products were analyzed by agarose gel electrophoresis and quantitated using the GeneTools Software, version 3.07 (SynGene, Cambridge, England).

Results

Random Mutagenic Library (L4 Library) Creation

The L4 mutagenic library was created using the co-*taq* gene as the template sequence, and $MnCl_2$ and *Taq* polymerase as the mutagens (Arnold and Georgiou, 2003b). Conditions were manipulated to create a library with approximately three amino acid changes per gene. After

purification of the mutagenic PCR products, they were used in a variation of the Miyazaki and Takenouchi megaprimer PCR protocol (Miyazaki and Takenouchi, 2002), creating the full-length plasmid (pASK-IBA43plus with insert). Purified products were transformed into the *E. coli* DH5α cell line; 79 clones were isolated, sequenced, and compared to the co-*Taq* amino acid sequence (Table 4-2). Of those 79 clones, only five retained the co-*Taq* polymerase sequence; the remaining 74 contained at least one mutation. The plasmids containing the 74 mutant L4 genes were then transformed into the *E. coli* TG-1 expression cell line.

Incorporation of dNTPs by RD and L4 Libraries at Various Temperatures

To determine the optimal temperature for each of the cell lines in the RD and L4 Libraries, each of these mutant *Taq* polymerases were tested for their ability to incorporate dNTPs in PCR reactions at various temperatures ranging from 75.0 °C to 94.0 °C. Reactions contained induced cells (1×10^6 cfu) as the sole source of polymerase and template plasmid, so active polymerases were forced to replicate their own encoding gene (2603 bp).

Figure 4-2[A-C] shows the difference between the PCR products from the co-*Taq* polymerase screen and representatives of the RD Library (SW17) and the L4 Library (SW251). In these, and all of the other reactions screening various temperatures, no full-length product (FLP) was observed at a temperature lower than 86.3 °C. Based on the product band densities, the optimal temperature for each polymerase in these two libraries was determined (Table 4-3 and Table 4-4).

Of the 74 L4 mutants, 39 were active at one of the temperatures tested, as compared to only 33 of the 74 RD mutants that were active. Figure 4-3[A-B] shows the distribution of the active polymerases in each library at the various temperatures. The average temperature for the RD mutants was 87.5 °C and for the L4 mutants it was 89.0 °C. It is interesting to note that

within the RD Library, if the mutant was active with a 94.0 °C temperature, it was active at the other five temperatures as well. This was not the case, however, with the L4 mutants; two of the mutants that were active at 94.0 °C were not active at the lower end of the spectrum. In addition, more mutants were stable at higher temperatures in the L4 variants as opposed to the RD variants.

Incorporation of dψUNTPs by RD Library at Optimal Temperatures

To find a polymerase that can incorporate and extend beyond dψUs with higher efficiency than the co-*Taq* polymerase, each of the active mutant *Taq* polymerases in the RD Library were tested for their ability to incorporate varying concentrations of dψUTP across from template dA in PCR reactions at their optimal temperature. Those polymerases that were not able to incorporate dNTPs at any temperature were not assayed in this experiment. Reactions contained induced cells (1 x 10^6 cfu) as the sole source of polymerase and template plasmid, again forcing active polymerases to replicate their own encoding gene (2603 bp). Control reactions, containing cells (SW4) expressing the co-*Taq* polymerase, were performed with the temperatures of 86.3 °C and 89.0 °C for comparative purposes.

The raw densities of FLP bands, as measured by GeneTools Software (ver. 3.07, SynGene) of the RD mutants at their optimal temperatures were compared to those generated by the co-*Taq* at that same temperature (Table 4-5). None of these polymerases showed an optimal temperature of 94.0 °C; they all had optimal temperatures of 86.3 °C or 89.0 °C. In addition, all but one of the RD mutants failed to incorporate dψUTP as efficiently as co-*Taq* polymerase at their optimal temperature. The remaining mutant (SW29), however, showed an ability to generate FLP in all dNTP/dψUNTP ratios up to 72% more efficiently, on average, than co-*Taq* polymerase at 86.3 °C (Fig. 4-4[A-C]).

Although the RD mutants were unable to perform as well as co-*Taq* polymerase at their optimal temperatures, as compared to the co-*Taq* FLP at that same temperature, they were able to generate, on average, 40% more FLP at their optimal temperature than when they were tested at 94.0 °C (Table 4-5 versus Table 3-4). Figure 4-5[A-C] shows representative (SW8) PCR products from the polymerase screen at 94.0 °C (Fig. 4-5A) and at the SW8 polymerase's optimal temperature of 86.3 °C (Fig. 4-5B). In the first set of reactions, the polymerase produced FLP with concentrations of 350 μM dψUTP; however, in the second set, the polymerase was able to generate FLP with concentrations of 400 μM d ψUTP. This is graphically represented in Figure 4-5C.

Incorporation of dψUTP and dψTTP by co-*Taq* Polymerase at Various Melting Temperatures

We performed a comparative analysis between the ability of co-*Taq* polymerase to incorporate dψUTP or dψTTP in various concentrations at different temperatures. Table 4-6 shows the raw densities, as determined by the GeneTools Software (ver. 3.07, SynGene), of the FLP bands generated by these experiments. Once the dψTTP final concentration reached 300 μM, all of the FLPs generated averaged a 23% higher density than those produced when using dψUTPs at identical concentrations. However, at the temperatures of 91.1 °C and 92.6 °C, all concentrations of dψTTP supported the synthesis of more FLP than dψUTP. In addition, at every temperature tested, FLP was present when the dψTTP concentration was in a 9:1 ratio with TTP, whereas the FLP was only observed at the highest ratio of dψUTP to TTP of 8:2 when the temperature was 86.3 °C or 89.0 °C.

Representative gels of these experiments, at a temperature of 86.3 °C, can be seen in Figure 4-6[A-C]. In the first gel (Fig. 4-6A), we see that FLP was generated up to an 8:2 ratio of

dψUTP to TTP; but was produced at a 9:1 ratio of dψTTP to TTP (Fig. 4-6B). Easily visualized in the chart (Fig. 4-6C), the amount of FLP formed was higher when dψUTP was present, until the final concentration reached 300 μM; then the level of FLP was elevated when in the presence of dψTTP. Graphical representations of the densities of the remaining five temperatures can be found in Figure 4-7[A-E].

Discussion

In previous studies (Chapter 3), we showed that only 24% of the RD Library mutants generated PCR products in the presence of dNTPs when the highest cycle temperature was 94.0 °C (Table 3-4). This was, perhaps, consistent with expectation. A higher percentage of active mutants, if randomly produced, might be expected to yield very few variants (Guo et al., 2004), but this might be balanced through the selection of sites less likely to cause unfolding, according to the REAP hypothesis (Gaucher, 2006). Unexpectedly, we found that at least one of the 35 sites identified by the REAP approach had been previously shown to modify the thermostability of *Taq* polymerase (Ghadessy et al., 2001). It is possible that other mutations affected thermostability as well. Therefore, polymerases in the RD Library were tested for their ability to incorporate dNTPs at temperatures ranging from 75.0 °C to 94.0 °C.

Of the 74 clones tested, none were able to generate FLP below a temperature of 86.3 °C. This is most likely not due to any property of the polymerase, but rather to the inability of the duplex DNA strands to melt at these lower temperatures, considering the temperature of the *taq* gene is ca. 88 °C. Thirty-three of the clones (45%), including the 18 previously shown to have activity (Table 3-4), were now able to incorporate dNTPs at lower temperatures down to 86.3 °C (Table 4-3). This leads us to believe that some of these replacements, located in and around the active site, indeed lowered the thermostability of the co-*Taq* polymerase.

It is also interesting to point out that the SW36 and SW74 cell lines, containing polymerases with the F664H or F664L mutations respectively, refutes Suzuki's theory that this phenylalanine residue can only be mutated to a tyrosine and retain activity (Suzuki et al., 1996). It may also be that the mutation to tyrosine allows the polymerase to retain the thermostability at 94.0 °C; Suzuki *et al.*, however, did not test the thermostability of the mutants in their experiments, and the activity seen with the F664H and F664L mutations in this research was only at lower temperatures.

We wanted to determine if a randomly created mutagenic library would be as likely to create clones that display a decrease in the thermostability of the protein. A random, mutagenic library (L4) was created using $MnCl_2$ and *Taq* polymerase (New England BioLabs) as the mutators (Table 4-2). We identified 74 clones, with unique sequences, that were also tested for their ability to incorporate dNTPs at various temperatures. Thirty-nine (53%) of the clones were able to generate FLP at one of the temperatures tested, and twenty-eight (38%) of these were able to function at 94.0 °C (Table 4-4).

Comparison of the thermostability of polymerases from the two libraries found more thermostability in the L4 random library than in the RD library (Fig. 4-3[A-B]). Thirty-nine of the L4 mutants were active, with an average optimal temperature of 89.0 °C, while only thirty-three of the RD clones retained activity, with an average optimal temperature of 87.5 °C. It remains open whether this difference reflects the slightly greater number of replacements in the RD library members (3.5 residues/protein) over the L4 library members (3 residues/protein).

The ability to function at lower temperatures can be advantageous to the incorporation of non-standard bases. The literature reports examples where NSBs are incorporated more efficiently at lower temperatures (Rappaport, 2004, Horlacher et al., 1995). We, therefore,

decided to test the incorporation of dψUTP at each of the RD mutant's optimal temperatures, comparing the results to those seen at 94.0 °C (Table 3-4).

Table 4-5 shows the ability of the active RD mutants to incorporate various concentrations of dψUTP in a PCR reaction with an optimal temperature of either 86.3 °C or 89.0 °C. Using this assay system, we identified a mutant (SW29) that was able to generate, on average, 72% more product than co-*Taq* polymerase with a temperature of 86.3 °C (Fig. 4-4[A-C]). This mutant, containing the A597S, A740R, and E742V residue changes, produced FLP up to a final concentration of 400 μM dψUTP. The remaining 32 active polymerases that were tested were unable to incorporate the dψUTP as well as co-*Taq* polymerase; they were, however, able to generate, on average, 40% more FLP at their optimal temperature than when they were tested at 94.0 °C (Table 4-5 versus Table 3-4). This lends strength to the theory that NSBs are easier to incorporate at lower temperatures.

Another reason dψUTP is more readily incorporated at lower temperatures could be due to its ability to epimerize (Fig. 4-1) (Wellington and Benner, 2006, Cohn, 1960, Chambers et al., 1963). Epimerization can occur at two stages. First, the dψUTP might epimerize, converting β-dψUTP (the substrate) to α-dψUTP (not a substrate). At worst, the α-dψUTP might be an inhibitor, but this possibility is not considered. The consequence of this conversion is to lower the concentration of β-epimer, as well as its amount. If the concentration of the triphosphate is less than its Michaelis constant, this would slow the rate of primer extension. Here, the concentrations are 500 μM, not dramatically higher than *Taq* polymerase's reported K_ms for dNTPs (16 μM) (Kong et al., 1993), but given the long extension time (3 min), it is doubtful that this is the origin of any effect seen. Alternatively, if the PCR is run to the point of exhaustion of the triphosphates, then a lower yield of PCR product is expected simply because the conversion

of β-dψUTP to α-dψUTP leads to earlier exhaustion. Since the amplicon is about 56% GC, this is considered not to be likely, as the triphosphates are present in equal concentrations, implying that dGTP and/or dCTP would be the first to be exhausted.

The alternative possibility is that the β-dψUTP is epimerizing at high temperatures after it is incorporated into the amplicon. Unlike with epimerization as the triphosphate, epimerized amplicon cannot simply be ignored. Rather, it creates serious problems with read-through by the polymerase; the amplicon may be lost to further PCR even with a single α-dψUTP.

In all experiments, the total concentration (T+dψUTP or T+dψTTP) were kept constant (500 μM). The total primer concentration (1.4 μM each, 7×10^{-11} molecules per 50 μL assay) is approximately five orders of magnitude greater than the number of copies of the plasmids (about 300 copies per cell, and about 10^6 cells per assay). With 2603 nucleotides per amplicon, the triphosphates are nominally consumed after 15.5 PCR cycles; the primer is nominally consumed after 17 PCR cycles. This PCR was carried out for 15 rounds, but it is doubtful that nominal perfection (true doubling each round) was observed here, or anywhere in PCR literature.

The results with dψUTP and dψTTP in carious concentrations and at different temperatures are shown in Table 4-6. While errors in the first column, where the pseudo component concentration is zero, are large, trends across the series are consistent. It appears that dψTTP supports the formation of PCR product at higher concentrations than dψUTP. There is no obvious way to explain this if the limitation on product formation involves the exhaustion of either the triphosphate or the primer. Rather, it is consistent with a slow rate of epimerization of dψUTP once it is incorporated into the amplicon. With dψTTP, eventually at high concentrations, PCR product formation subsides. This is attributed to the accumulative effects of too many unnatural nucleotides present in the amplicon (approaching 25%).

Numerous approaches for library design exist to support directed evolution; this study has compared two of them: A) a rationally designed library based on comparative sequence analysis of the active sites of Family A polymerases, and B) a randomly mutagenized library with no preference to the location of mutations. Our studies have shown that we are more likely to generate active, thermostable mutants with a randomly mutated library than with our rationally designed library. This supports the conclusions drawn by Arnold and colleagues (Drummond et al., 2005), who determined that libraries with high error-rates distributed throughout the entirety of the gene, result in a higher than expected number of active and unique variants.

Our RD Library was designed to identify polymerases with an increased ability to incorporate NSBs, not for thermostability, so the "you get what you select for" theory may be applicable in this situation. We were able to identify one mutant with an increased ability over co-*Taq* polymerase to incorporate dψUTP; this, however, is only one example of a C-glycoside. Since it is possible that dψUTP is epimerizing at the temperatures tested, perhaps the testing of other C-glycosides, such as dψTTP, can assist in the identification of more mutants with the capability of incorporating these, or other NSBs. Additionally, useful information could be gained by testing the ability all of our RD and L4 mutants for their ability to incorporate dψTTP, or other NSBs, not only about incorporation of NSBs, but also in regards to favorable library design.

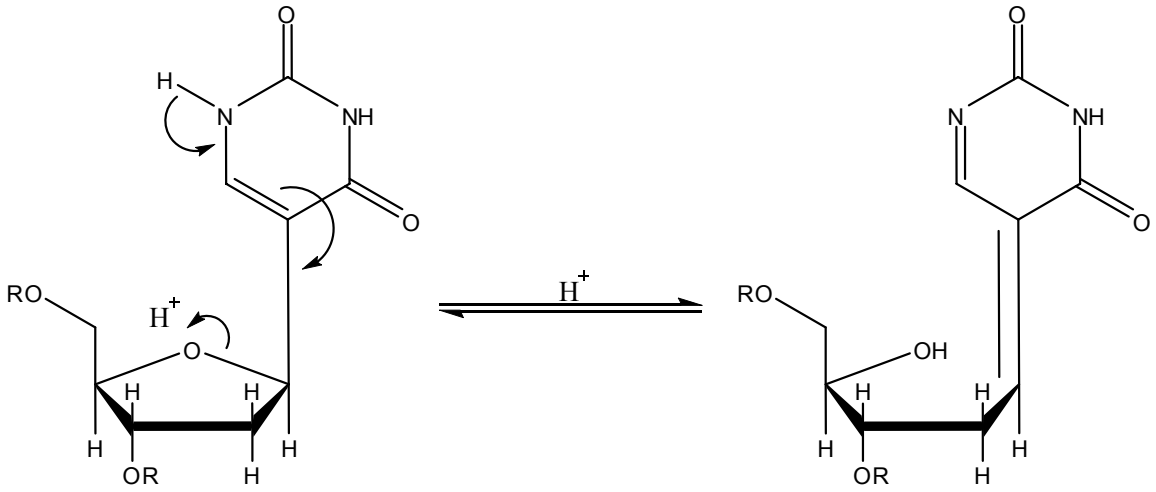


Figure 4-1. Epimerization of 2'-deoxypseudouridine. 2'-deoxypseudouridine can epimerize under acidic, basic, or even neutral conditions over time, in either the nucleoside or the oligonucleotide forms. Polymerases will not incorporate the epimerized form of this nucleotide, therefore use of the non-epimerizing 2'-deoxypseudothymidine is recommended.

Table 4-1. Additional bacterial strains used in this study.

Name	Strain	Genotype
SW135	*E. coli* DH5α	SW1/pL4Mut1 (pASK-IBA43+ with co *taq* L4 mut 1, Ap^r)
SW136	*E. coli* DH5α	SW1/pL4Mut2 (pASK-IBA43+ with co *taq* L4 mut 2, Ap^r)
SW137	*E. coli* DH5α	SW1/pL4Mut3 (pASK-IBA43+ with co *taq* L4 mut 3, Ap^r)
SW138	*E. coli* DH5α	SW1/pL4Mut4 (pASK-IBA43+ with co *taq* L4 mut 4, Ap^r)
SW139	*E. coli* DH5α	SW1/pL4Mut5 (pASK-IBA43+ with co *taq* L4 mut 5, Ap^r)
SW140	*E. coli* DH5α	SW1/pL4Mut6 (pASK-IBA43+ with co *taq* L4 mut 6, Ap^r)
SW141	*E. coli* DH5α	SW1/pL4Mut7 (pASK-IBA43+ with co *taq* L4 mut 7, Ap^r)
SW142	*E. coli* DH5α	SW1/pL4Mut8 (pASK-IBA43+ with co *taq* L4 mut 8, Ap^r)
SW143	*E. coli* DH5α	SW1/pL4Mut9 (pASK-IBA43+ with co *taq* L4 mut 9, Ap^r)
SW144	*E. coli* DH5α	SW1/pL4Mut10 (pASK-IBA43+ with co *taq* L4 mut 10, Ap^r)
SW145	*E. coli* DH5α	SW1/pL4Mut11 (pASK-IBA43+ with co *taq* L4 mut 11, Ap^r)
SW146	*E. coli* DH5α	SW1/pL4Mut12 (pASK-IBA43+ with co *taq* L4 mut 12, Ap^r)
SW147	*E. coli* DH5α	SW1/pL4Mut13 (pASK-IBA43+ with co *taq* L4 mut 13, Ap^r)
SW148	*E. coli* DH5α	SW1/pL4Mut14 (pASK-IBA43+ with co *taq* L4 mut 14, Ap^r)
SW149	*E. coli* DH5α	SW1/pL4Mut15 (pASK-IBA43+ with co *taq* L4 mut 15, Ap^r)
SW150	*E. coli* DH5α	SW1/pL4Mut16 (pASK-IBA43+ with co *taq* L4 mut 16, Ap^r)
SW151	*E. coli* DH5α	SW1/pL4Mut17 (pASK-IBA43+ with co *taq* L4 mut 17, Ap^r)
SW152	*E. coli* DH5α	SW1/pL4Mut18 (pASK-IBA43+ with co *taq* L4 mut 18, Ap^r)
SW153	*E. coli* DH5α	SW1/pL4Mut19 (pASK-IBA43+ with co *taq* L4 mut 19, Ap^r)
SW154	*E. coli* DH5α	SW1/pL4Mut20 (pASK-IBA43+ with co *taq* L4 mut 20, Ap^r)
SW155	*E. coli* DH5α	SW1/pL4Mut21 (pASK-IBA43+ with co *taq* L4 mut 21, Ap^r)
SW156	*E. coli* DH5α	SW1/pL4Mut22 (pASK-IBA43+ with co *taq* L4 mut 22, Ap^r)
SW157	*E. coli* DH5α	SW1/pL4Mut23 (pASK-IBA43+ with co *taq* L4 mut 23, Ap^r)
SW158	*E. coli* DH5α	SW1/pL4Mut24 (pASK-IBA43+ with co *taq* L4 mut 24, Ap^r)
SW159	*E. coli* DH5α	SW1/pL4Mut25 (pASK-IBA43+ with co *taq* L4 mut 25, Ap^r)
SW160	*E. coli* DH5α	SW1/pL4Mut26 (pASK-IBA43+ with co *taq* L4 mut 26, Ap^r)
SW161	*E. coli* DH5α	SW1/pL4Mut27 (pASK-IBA43+ with co *taq* L4 mut 27, Ap^r)
SW162	*E. coli* DH5α	SW1/pL4Mut28 (pASK-IBA43+ with co *taq* L4 mut 28, Ap^r)
SW163	*E. coli* DH5α	SW1/pL4Mut29 (pASK-IBA43+ with co *taq* L4 mut 29, Ap^r)
SW164	*E. coli* DH5α	SW1/pL4Mut30 (pASK-IBA43+ with co *taq* L4 mut 30, Ap^r)
SW165	*E. coli* DH5α	SW1/pL4Mut31 (pASK-IBA43+ with co *taq* L4 mut 31, Ap^r)
SW166	*E. coli* DH5α	SW1/pL4Mut32 (pASK-IBA43+ with co *taq* L4 mut 32, Ap^r)
SW167	*E. coli* DH5α	SW1/pL4Mut33 (pASK-IBA43+ with co *taq* L4 mut 33, Ap^r)
SW168	*E. coli* DH5α	SW1/pL4Mut34 (pASK-IBA43+ with co *taq* L4 mut 34, Ap^r)
SW169	*E. coli* DH5α	SW1/pL4Mut35 (pASK-IBA43+ with co *taq* L4 mut 35, Ap^r)
SW170	*E. coli* DH5α	SW1/pL4Mut36 (pASK-IBA43+ with co *taq* L4 mut 36, Ap^r)
SW171	*E. coli* DH5α	SW1/pL4Mut37 (pASK-IBA43+ with co *taq* L4 mut 37, Ap^r)
SW172	*E. coli* DH5α	SW1/pL4Mut38 (pASK-IBA43+ with co *taq* L4 mut 38, Ap^r)
SW173	*E. coli* DH5α	SW1/pL4Mut39 (pASK-IBA43+ with co *taq* L4 mut 39, Ap^r)
SW174	*E. coli* DH5α	SW1/pL4Mut40 (pASK-IBA43+ with co *taq* L4 mut 40, Ap^r)
SW175	*E. coli* DH5α	SW1/pL4Mut41 (pASK-IBA43+ with co *taq* L4 mut 41, Ap^r)
SW176	*E. coli* DH5α	SW1/pL4Mut42 (pASK-IBA43+ with co *taq* L4 mut 42, Ap^r)
SW177	*E. coli* DH5α	SW1/pL4Mut43 (pASK-IBA43+ with co *taq* L4 mut 43, Ap^r)
SW178	*E. coli* DH5α	SW1/pL4Mut44 (pASK-IBA43+ with co *taq* L4 mut 44, Ap^r)
SW179	*E. coli* DH5α	SW1/pL4Mut45 (pASK-IBA43+ with co *taq* L4 mut 45, Ap^r)
SW180	*E. coli* DH5α	SW1/pL4Mut46 (pASK-IBA43+ with co *taq* L4 mut 46, Ap^r)
SW181	*E. coli* DH5α	SW1/pL4Mut47 (pASK-IBA43+ with co *taq* L4 mut 47, Ap^r)
SW182	*E. coli* DH5α	SW1/pL4Mut48 (pASK-IBA43+ with co *taq* L4 mut 48, Ap^r)
SW183	*E. coli* DH5α	SW1/pL4Mut49 (pASK-IBA43+ with co *taq* L4 mut 49, Ap^r)
SW184	*E. coli* DH5α	SW1/pL4Mut50 (pASK-IBA43+ with co *taq* L4 mut 50, Ap^r)
SW185	*E. coli* DH5α	SW1/pL4Mut51 (pASK-IBA43+ with co *taq* L4 mut 51, Ap^r)
SW186	*E. coli* DH5α	SW1/pL4Mut52 (pASK-IBA43+ with co *taq* L4 mut 52, Ap^r)
SW187	*E. coli* DH5α	SW1/pL4Mut53 (pASK-IBA43+ with co *taq* L4 mut 53, Ap^r)
SW188	*E. coli* DH5α	SW1/pL4Mut54 (pASK-IBA43+ with co *taq* L4 mut 54, Ap^r)
SW189	*E. coli* DH5α	SW1/pL4Mut55 (pASK-IBA43+ with co *taq* L4 mut 55, Ap^r)
SW190	*E. coli* DH5α	SW1/pL4Mut56 (pASK-IBA43+ with co *taq* L4 mut 56, Ap^r)
SW191	*E. coli* DH5α	SW1/pL4Mut57 (pASK-IBA43+ with co *taq* L4 mut 57, Ap^r)
SW192	*E. coli* DH5α	SW1/pL4Mut58 (pASK-IBA43+ with co *taq* L4 mut 58, Ap^r)
SW193	*E. coli* DH5α	SW1/pL4Mut59 (pASK-IBA43+ with co *taq* L4 mut 59, Ap^r)
SW194	*E. coli* DH5α	SW1/pL4Mut60 (pASK-IBA43+ with co *taq* L4 mut 60, Ap^r)
SW195	*E. coli* DH5α	SW1/pL4Mut61 (pASK-IBA43+ with co *taq* L4 mut 61, Ap^r)
SW196	*E. coli* DH5α	SW1/pL4Mut62 (pASK-IBA43+ with co *taq* L4 mut 62, Ap^r)
SW197	*E. coli* DH5α	SW1/pL4Mut63 (pASK-IBA43+ with co *taq* L4 mut 63, Ap^r)
SW198	*E. coli* DH5α	SW1/pL4Mut64 (pASK-IBA43+ with co *taq* L4 mut 64, Ap^r)
SW199	*E. coli* DH5α	SW1/pL4Mut65 (pASK-IBA43+ with co *taq* L4 mut 65, Ap^r)
SW200	*E. coli* DH5α	SW1/pL4Mut66 (pASK-IBA43+ with co *taq* L4 mut 66, Ap^r)
SW201	*E. coli* DH5α	SW1/pL4Mut67 (pASK-IBA43+ with co *taq* L4 mut 67, Ap^r)
SW202	*E. coli* DH5α	SW1/pL4Mut68 (pASK-IBA43+ with co *taq* L4 mut 68, Ap^r)
SW203	*E. coli* DH5α	SW1/pL4Mut69 (pASK-IBA43+ with co *taq* L4 mut 69, Ap^r)
SW202	*E. coli* DH5α	SW1/pL4Mut70 (pASK-IBA43+ with co *taq* L4 mut 70, Ap^r)
SW203	*E. coli* DH5α	SW1/pL4Mut71 (pASK-IBA43+ with co *taq* L4 mut 71, Ap^r)
SW204	*E. coli* DH5α	SW1/pL4Mut72 (pASK-IBA43+ with co *taq* L4 mut 72, Ap^r)
SW205	*E. coli* DH5α	SW1/pL4Mut73 (pASK-IBA43+ with co *taq* L4 mut 73, Ap^r)
SW206	*E. coli* DH5α	SW1/pL4Mut74 (pASK-IBA43+ with co *taq* L4 mut 74, Ap^r)
SW207	*E. coli* DH5α	SW1/pL4Mut75 (pASK-IBA43+ with co *taq* L4 mut 75, Ap^r)
SW208	*E. coli* DH5α	SW1/pL4Mut76 (pASK-IBA43+ with co *taq* L4 mut 76, Ap^r)
SW209	*E. coli* DH5α	SW1/pL4Mut77 (pASK-IBA43+ with co *taq* L4 mut 77, Ap^r)
SW210	*E. coli* DH5α	SW1/pL4Mut78 (pASK-IBA43+ with co *taq* L4 mut 78, Ap^r)
SW211	*E. coli* DH5α	SW1/pL4Mut79 (pASK-IBA43+ with co *taq* L4 mut 79, Ap^r)
SW212	*E. coli* TG-1	SW1/pL4Mut1 (pASK-IBA43+ with co *taq* L4 mut 1, Ap^r)
SW213	*E. coli* TG-1	SW1/pL4Mut2 (pASK-IBA43+ with co *taq* L4 mut 2, Ap^r)
SW214	*E. coli* TG-1	SW1/pL4Mut3 (pASK-IBA43+ with co *taq* L4 mut 3, Ap^r)
SW215	*E. coli* TG-1	SW1/pL4Mut4 (pASK-IBA43+ with co *taq* L4 mut 4, Ap^r)
SW216	*E. coli* TG-1	SW1/pL4Mut5 (pASK-IBA43+ with co *taq* L4 mut 5, Ap^r)
SW217	*E. coli* TG-1	SW1/pL4Mut6 (pASK-IBA43+ with co *taq* L4 mut 6, Ap^r)
SW218	*E. coli* TG-1	SW1/pL4Mut7 (pASK-IBA43+ with co *taq* L4 mut 7, Ap^r)
SW219	*E. coli* TG-1	SW1/pL4Mut8 (pASK-IBA43+ with co *taq* L4 mut 8, Ap^r)
SW220	*E. coli* TG-1	SW1/pL4Mut9 (pASK-IBA43+ with co *taq* L4 mut 9, Ap^r)
SW221	*E. coli* TG-1	SW1/pL4Mut10 (pASK-IBA43+ with co *taq* L4 mut 10, Ap^r)
SW222	*E. coli* TG-1	SW1/pL4Mut11 (pASK-IBA43+ with co *taq* L4 mut 11, Ap^r)
SW223	*E. coli* TG-1	SW1/pL4Mut12 (pASK-IBA43+ with co *taq* L4 mut 12, Ap^r)
SW224	*E. coli* TG-1	SW1/pL4Mut13 (pASK-IBA43+ with co *taq* L4 mut 13, Ap^r)
SW225	*E. coli* TG-1	SW1/pL4Mut14 (pASK-IBA43+ with co *taq* L4 mut 14, Ap^r)
SW226	*E. coli* TG-1	SW1/pL4Mut15 (pASK-IBA43+ with co *taq* L4 mut 15, Ap^r)
SW227	*E. coli* TG-1	SW1/pL4Mut16 (pASK-IBA43+ with co *taq* L4 mut 16, Ap^r)
SW228	*E. coli* TG-1	SW1/pL4Mut17 (pASK-IBA43+ with co *taq* L4 mut 17, Ap^r)
SW229	*E. coli* TG-1	SW1/pL4Mut18 (pASK-IBA43+ with co *taq* L4 mut 18, Ap^r)
SW230	*E. coli* TG-1	SW1/pL4Mut20 (pASK-IBA43+ with co *taq* L4 mut 20, Ap^r)
SW231	*E. coli* TG-1	SW1/pL4Mut21 (pASK-IBA43+ with co *taq* L4 mut 21, Ap^r)
SW232	*E. coli* TG-1	SW1/pL4Mut22 (pASK-IBA43+ with co *taq* L4 mut 22, Ap^r)
SW233	*E. coli* TG-1	SW1/pL4Mut23 (pASK-IBA43+ with co *taq* L4 mut 23, Ap^r)
SW234	*E. coli* TG-1	SW1/pL4Mut26 (pASK-IBA43+ with co *taq* L4 mut 26, Ap^r)
SW235	*E. coli* TG-1	SW1/pL4Mut27 (pASK-IBA43+ with co *taq* L4 mut 27, Ap^r)
SW236	*E. coli* TG-1	SW1/pL4Mut28 (pASK-IBA43+ with co *taq* L4 mut 28, Ap^r)
SW237	*E. coli* TG-1	SW1/pL4Mut29 (pASK-IBA43+ with co *taq* L4 mut 29, Ap^r)
SW238	*E. coli* TG-1	SW1/pL4Mut30 (pASK-IBA43+ with co *taq* L4 mut 30, Ap^r)
SW239	*E. coli* TG-1	SW1/pL4Mut31 (pASK-IBA43+ with co *taq* L4 mut 31, Ap^r)
SW240	*E. coli* TG-1	SW1/pL4Mut32 (pASK-IBA43+ with co *taq* L4 mut 32, Ap^r)
SW241	*E. coli* TG-1	SW1/pL4Mut33 (pASK-IBA43+ with co *taq* L4 mut 33, Ap^r)
SW242	*E. coli* TG-1	SW1/pL4Mut34 (pASK-IBA43+ with co *taq* L4 mut 34, Ap^r)
SW243	*E. coli* TG-1	SW1/pL4Mut35 (pASK-IBA43+ with co *taq* L4 mut 35, Ap^r)
SW244	*E. coli* TG-1	SW1/pL4Mut36 (pASK-IBA43+ with co *taq* L4 mut 36, Ap^r)
SW245	*E. coli* TG-1	SW1/pL4Mut37 (pASK-IBA43+ with co *taq* L4 mut 37, Ap^r)
SW246	*E. coli* TG-1	SW1/pL4Mut38 (pASK-IBA43+ with co *taq* L4 mut 38, Ap^r)
SW247	*E. coli* TG-1	SW1/pL4Mut39 (pASK-IBA43+ with co *taq* L4 mut 39, Ap^r)
SW248	*E. coli* TG-1	SW1/pL4Mut40 (pASK-IBA43+ with co *taq* L4 mut 40, Ap^r)
SW249	*E. coli* TG-1	SW1/pL4Mut41 (pASK-IBA43+ with co *taq* L4 mut 41, Ap^r)
SW250	*E. coli* TG-1	SW1/pL4Mut42 (pASK-IBA43+ with co *taq* L4 mut 42, Ap^r)
SW251	*E. coli* TG-1	SW1/pL4Mut43 (pASK-IBA43+ with co *taq* L4 mut 43, Ap^r)
SW252	*E. coli* TG-1	SW1/pL4Mut44 (pASK-IBA43+ with co *taq* L4 mut 44, Ap^r)
SW253	*E. coli* TG-1	SW1/pL4Mut45 (pASK-IBA43+ with co *taq* L4 mut 45, Ap^r)
SW254	*E. coli* TG-1	SW1/pL4Mut46 (pASK-IBA43+ with co *taq* L4 mut 46, Ap^r)
SW255	*E. coli* TG-1	SW1/pL4Mut48 (pASK-IBA43+ with co *taq* L4 mut 48, Ap^r)
SW256	*E. coli* TG-1	SW1/pL4Mut50 (pASK-IBA43+ with co *taq* L4 mut 50, Ap^r)
SW257	*E. coli* TG-1	SW1/pL4Mut51 (pASK-IBA43+ with co *taq* L4 mut 51, Ap^r)
SW258	*E. coli* TG-1	SW1/pL4Mut52 (pASK-IBA43+ with co *taq* L4 mut 52, Ap^r)
SW259	*E. coli* TG-1	SW1/pL4Mut53 (pASK-IBA43+ with co *taq* L4 mut 53, Ap^r)
SW260	*E. coli* TG-1	SW1/pL4Mut54 (pASK-IBA43+ with co *taq* L4 mut 54, Ap^r)
SW261	*E. coli* TG-1	SW1/pL4Mut55 (pASK-IBA43+ with co *taq* L4 mut 55, Ap^r)
SW262	*E. coli* TG-1	SW1/pL4Mut56 (pASK-IBA43+ with co *taq* L4 mut 56, Ap^r)
SW263	*E. coli* TG-1	SW1/pL4Mut57 (pASK-IBA43+ with co *taq* L4 mut 57, Ap^r)
SW264	*E. coli* TG-1	SW1/pL4Mut58 (pASK-IBA43+ with co *taq* L4 mut 58, Ap^r)
SW265	*E. coli* TG-1	SW1/pL4Mut59 (pASK-IBA43+ with co *taq* L4 mut 59, Ap^r)
SW266	*E. coli* TG-1	SW1/pL4Mut60 (pASK-IBA43+ with co *taq* L4 mut 60, Ap^r)
SW267	*E. coli* TG-1	SW1/pL4Mut61 (pASK-IBA43+ with co *taq* L4 mut 61, Ap^r)
SW268	*E. coli* TG-1	SW1/pL4Mut62 (pASK-IBA43+ with co *taq* L4 mut 62, Ap^r)
SW269	*E. coli* TG-1	SW1/pL4Mut63 (pASK-IBA43+ with co *taq* L4 mut 63, Ap^r)
SW270	*E. coli* TG-1	SW1/pL4Mut64 (pASK-IBA43+ with co *taq* L4 mut 64, Ap^r)
SW271	*E. coli* TG-1	SW1/pL4Mut65 (pASK-IBA43+ with co *taq* L4 mut 65, Ap^r)
SW272	*E. coli* TG-1	SW1/pL4Mut66 (pASK-IBA43+ with co *taq* L4 mut 66, Ap^r)
SW273	*E. coli* TG-1	SW1/pL4Mut67 (pASK-IBA43+ with co *taq* L4 mut 67, Ap^r)
SW274	*E. coli* TG-1	SW1/pL4Mut68 (pASK-IBA43+ with co *taq* L4 mut 68, Ap^r)
SW275	*E. coli* TG-1	SW1/pL4Mut69 (pASK-IBA43+ with co *taq* L4 mut 69, Ap^r)
SW276	*E. coli* TG-1	SW1/pL4Mut70 (pASK-IBA43+ with co *taq* L4 mut 70, Ap^r)
SW277	*E. coli* TG-1	SW1/pL4Mut71 (pASK-IBA43+ with co *taq* L4 mut 71, Ap^r)
SW278	*E. coli* TG-1	SW1/pL4Mut72 (pASK-IBA43+ with co *taq* L4 mut 72, Ap^r)
SW279	*E. coli* TG-1	SW1/pL4Mut73 (pASK-IBA43+ with co *taq* L4 mut 73, Ap^r)
SW280	*E. coli* TG-1	SW1/pL4Mut74 (pASK-IBA43+ with co *taq* L4 mut 74, Ap^r)
SW281	*E. coli* TG-1	SW1/pL4Mut75 (pASK-IBA43+ with co *taq* L4 mut 75, Ap^r)
SW282	*E. coli* TG-1	SW1/pL4Mut76 (pASK-IBA43+ with co *taq* L4 mut 76, Ap^r)
SW283	*E. coli* TG-1	SW1/pL4Mut77 (pASK-IBA43+ with co *taq* L4 mut 77, Ap^r)
SW284	*E. coli* TG-1	SW1/pL4Mut78 (pASK-IBA43+ with co *taq* L4 mut 78, Ap^r)
SW285	*E. coli* TG-1	SW1/pL4Mut79 (pASK-IBA43+ with co *taq* L4 mut 79, Ap^r)

Table 4-2. L4 Mutant Library.

Plasmid Name	Mutations Present in L4 *Taq* Library	Plasmid Name	Mutations Present in L4 *Taq* Library
pL4Mut1	L4Q,G16S,R91H,E292G,D575N,S620P	pL4Mut41	E794G,M804V
pL4Mut2	V110A	pL4Mut42	L530P,K539N,L654P
pL4Mut3	G197C,F269S,K790R	pL4Mut43	F44I,E167G
pL4Mut4	L409P,V615I,K828R	pL4Mut44	Y392F,N412D,N562D,E649G
pL4Mut5	L27Q,L30Q,R263S,L273R,L409P	pL4Mut45	P809T,E227K
pL4Mut6	F89S,I160T,P261L,*GAP*	pL4Mut46	*GAP*
pL4Mut7	V38G,K222E,F255I,E407A,E691A	pL4Mut47	NONE
pL4Mut8	P552L,L765P	pL4Mut48	*GAP*
pL4Mut9	N482I	pL4Mut49	NONE
pL4Mut10	A83G,I135N,L285Q,Y336H,*GAP*	pL4Mut50	K216I,A455D,V651E
pL4Mut11	G393S,T444P,M670T,E710G	pL4Mut51	L108P,G197S,L377P,R390C,T503A,GAP
pL4Mut12	L789P	pL4Mut52	K125M
pL4Mut13	E6D,K351R,A797D,G821D	pL4Mut53	E167G,S309P,T719A,L814Q
pL4Mut14	L122Q,D341G,A411V	pL4Mut54	W315C,T506P
pL4Mut15	I596M,M643V	pL4Mut55	*GAP*
pL4Mut16	A115P,L458R,H558P,H617R,L654P,K801E	pL4Mut56	Y158C,S309T,A404T,S540G,M758T
pL4Mut17	Y113H,G276V,L409P	pL4Mut57	V38A,K337E,V796D
pL4Mut18	R693C,E731G,V812A	pL4Mut58	D101G
pL4Mut19	NONE	pL4Mut59	H330P,*GAP*
pL4Mut20	F44L,T183A,K194E,L291P	pL4Mut60	R91H,F561S
pL4Mut21	K337E	pL4Mut61	D493G
pL4Mut22	L362P,M441I,E517V	pL4Mut62	S121T,E599V,Y808H
pL4Mut23	G81D,K203E,V446A,D634G	pL4Mut63	A80V,R220C,G367C,D378N,R389L,S574G,G752D,M776L
pL4Mut24	NONE	pL4Mut64	L4R,I150N,K203R,K337E,Q563Stop,P647L,A774V
pL4Mut25	NONE	pL4Mut65	W425R,E771G,V796D
pL4Mut26	*GAP*	pL4Mut66	L13P,A565V
pL4Mut27	L279P,S287T	pL4Mut67	L93P,E156G,A213T,P299S,N580S,E771G,V796D
pL4Mut28	L12P,V133M,N217D,L266P,E300G,L546P,W703R,L825P	pL4Mut68	E109K,G209C,W240R,L373Q
pL4Mut29	A231V,*GAP*	pL4Mut69	K46E,V118A,L218P,I529T,Q579H
pL4Mut30	K203I,A268G,D544N,R633L,T753A,V763A	pL4Mut70	G184C,L491P,R556H
pL4Mut31	L285P,*GAP*	pL4Mut71	S309Y,D369G,F479S,I581V,A605T
pL4Mut32	R91L,E534G	pL4Mut72	E420D,E678G,K828E
pL4Mut33	L221P,E264V,K528R,GAP	pL4Mut73	K216I,T503A,T511M,Q589R,K759E
pL4Mut34	L777P,Stop830	pL4Mut74	V780A
pL4Mut35	N624Y	pL4Mut75	K759R
pL4Mut36	K337E	pL4Mut76	E109K,G209C,W240R,L373Q
pL4Mut37	A126P,*GAP*	pL4Mut77	F721L
pL4Mut38	D185V,L491P,M643V	pL4Mut78	Y42H,R220C,W425Stop,R590W
pL4Mut39	Y75C,K216E,S377T,A565V,M758T,E770D	pL4Mut79	Y169C,T247A,D248G,E638G
pL4Mut40	*GAP,GAP*		

*All are derivatives of the co-*taq* gene, and all are inserted into the pASK-IBA43plus vector. "NONE" means no mutations were found relative to the co-*Taq* sequence, "*GAP*" denotes the presence of a frameshift mutation within the protein; and "Stop" indicates the presence of a Stop codon in the sequence.

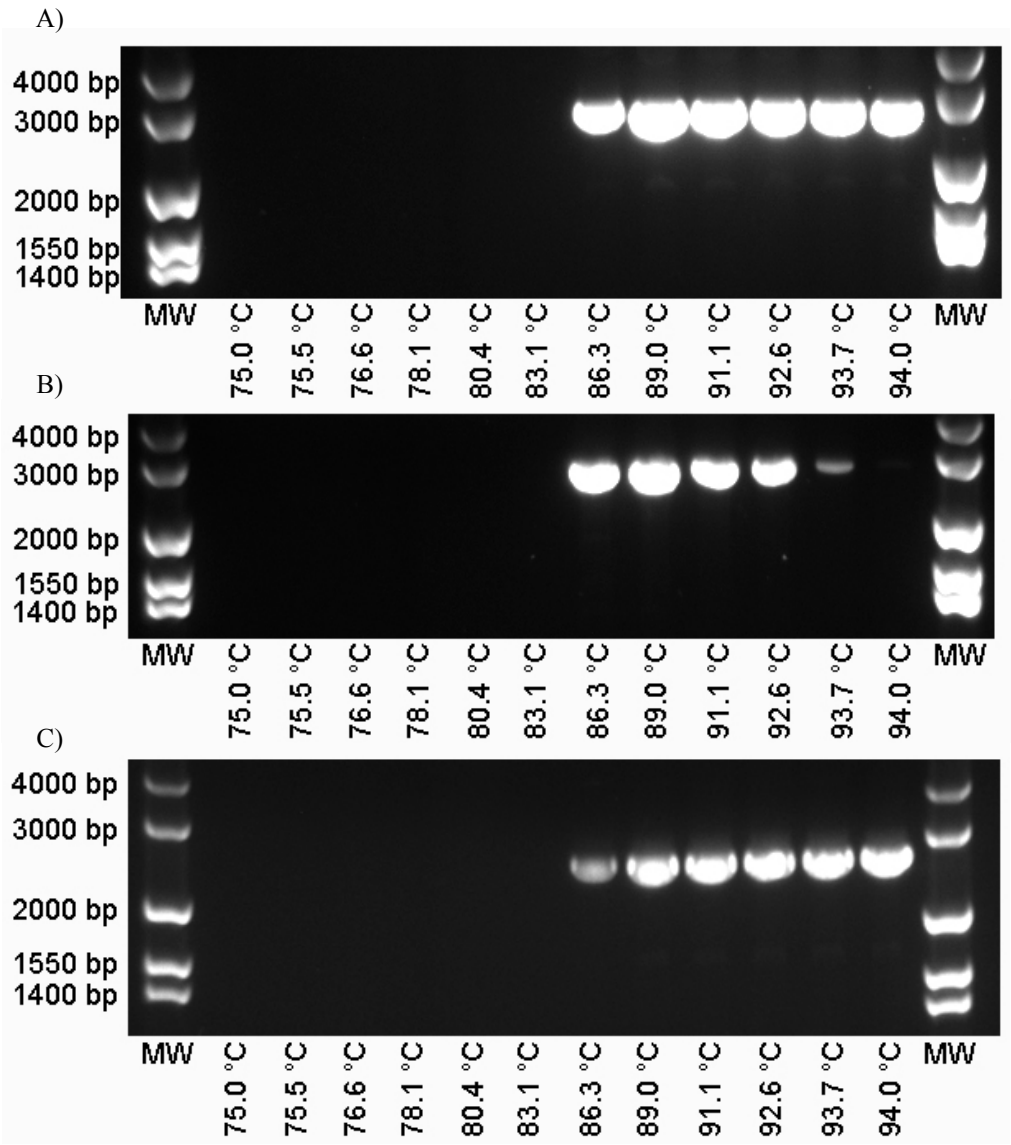


Figure 4-2. Representative images of ethidium-bromide stained agarose gels resolving products arising from PCR amplification using standard dNTPs and three different polymerases. Cells expressing the indicated polymerase provided both the polymerase and the template plasmid for the reaction. Polymerases were therefore forced to replicate their own encoding gene (2603 bp) using primers P-4 and P-5. Optimal temperatures for each polymerase were determined by identifying the FLP band having the highest density. A) The co-*Taq* polymerase having an optimal temperature of 89.0 °C. B) The polymerase expressed in SW17 cells having an optimal temperature of 89.0 °C. C) The polymerase expressed in SW251 cells having an optimal temperature of 94.0 °C.

Table 4-3. Generation of full length PCR products from dNTPs by individual polymerases from the rationally designed (RD) Library at the indicated temperatures.

Cell Line	Substitutions	Optimal Temp	RawDensities(CNT/mm^2)					
			86.3	89.0	91.1	92.6	93.7	94.0
SW4	Codon-Optimized (co) *wt* Taq	89.0	2681942	2925066	2705135	2570101	2474721	2364200
SW5	S573E,Y668F,A740S	86.3	316793	56030	0	0	0	0
SW6	Q486H,K537I,M670G	-	0	0	0	0	0	0
SW7	A605G,L613A,E739P	86.3	2602073	2510025	1400114	0	0	0
SW8	D575F,L606C,A740S	86.3	3512755	3345122	3215369	3155301	3002691	2914907
SW9	T511V,R584V,I611E	-	0	0	0	0	0	0
SW10	N480R,F595V,E742H	86.3	1788759	741832	1037298	1071869	605564	354588
SW11	E517I,V583K,A597S	89.0	675098	995847	868690	798423	724526	781743
SW12	D575F,V583K,M670A	86.3	33298	27057	26676	27017	27803	30258
SW13	E517I,D607W,D622S	-	0	0	0	0	0	0
SW14	A594C,F664Y,A774H	86.3	1687977	1529299	1178437	873870	516390	284889
SW15	F595W,L606P,D622S	86.3	31072	0	0	0	0	0
SW16	S573E,D575F,F595V	89.0	878799	978974	430585	112562	0	0
SW17	S510I,A605K,L606S	89.0	2064807	2250615	1780581	1447476	266259	39332
SW18	S573E,D622L,E742H	-	0	0	0	0	0	0
SW19	N480R,T511V,Y542E	-	0	0	0	0	0	0
SW20	A594C,F664H,M670G	-	0	0	0	0	0	0
SW21	Q486H,D575T,N580S	86.3	2516914	2377642	2415212	2001995	2098686	1867511
SW22	S510I,A605E,E612I	-	0	0	0	0	0	0
SW23	A594C,E612I,M670A	-	0	0	0	0	0	0
SW24	S510I,Q579A,I611Q	-	0	0	0	0	0	0
SW25	A594T,L606C,R657D	89.0	2343694	2414686	2041814	1897633	1489840	1095528
SW26	T541A,A605G,L606S	86.3	1957603	1846837	1080938	96782	0	0
SW27	E517G,K537I,L613A	86.3	1728825	1468864	1031245	801755	537510	465426
SW28	K537I,Q579A,E742V	-	0	0	0	0	0	0
SW29	A597S,A740R,E742V		2669939	2672652	2649972	2308144	2293993	2126919
SW30	N580Q,A605E,L613I	89.0	1055603	1731735	1422463	461004	72848	28402
SW31	N580S,F595V,A605G	86.3	1521229	1380438	1194735	461630	105282	43184
SW32	N580S,D622L,A774H	-	0	0	0	0	0	0
SW33	R533I,R584V,F664L	-	0	0	0	0	0	0
SW34	Q486H,E517G,A605K	89.0	2582963	2673515	2412059	2295835	2319794	2054398
SW35	S573H,F664Y,R743A	-	0	0	0	0	0	0
SW36	D575T,N580Q,R584V	89.0	116303	191723	167107	148572	101952	77632
SW37	T541A,F664L,R743A	-	0	0	0	0	0	0
SW38	T511V,R533I,D622A	-	0	0	0	0	0	0
SW39	A597S,I611E,Y668F	89.0	415640	95546	0	0	0	0
SW40	Y542E,F595W,L606C	-	0	0	0	0	0	0
SW41	L606S,R657D,E739R	89.0	727359	2788978	2565988	2384196	1847172	1518298
SW42	S573H,D575T,L613I	-	0	0	0	0	0	0
SW43	T541A,L606P,L613D	-	0	0	0	0	0	0
SW44	Y542E,V583K,A605E	-	0	0	0	0	0	0
SW45	E517I,F595W,A605E,I611E	-	0	0	0	0	0	0
SW46	T541A,D575F,L613A,D622A	89.0	51480	186191	0	0	0	0
SW47	T511V,A594C,L606S,A740R	89.0	1646260	2387801	2248001	2162871	1823957	1588298
SW48	Q486H,R533I,L606C,L613A	-	0	0	0	0	0	0
SW49	Q486H,F595V,D622A,F664Y	89.0	188922	206661	28711	0	0	0
SW50	E517I,S573H,A605G,E612I	-	0	0	0	0	0	0
SW51	Y542E,R584V,A605K,E612I	-	0	0	0	0	0	0
SW52	D575T,A605E,L606C,D622A	86.3	70860	0	0	0	0	0
SW53	A594T,L613A,F664Y,E742H	86.3	2125296	2020864	1487736	225254	0	0
SW54	D575F,N580Q,W601G,D622S	86.3	1984853	107072	0	0	0	0
SW55	K537I,L606P,A740S,E742H	86.3	2967102	2375974	566103	0	0	0
SW56	A597S,W601G,L606S,F664H	86.3	604223	161850	0	0	0	0
SW57	S510I,E517G,D607W,I611E	-	0	0	0	0	0	0
SW58	S510I,V583K,R584V,L606P	-	0	0	0	0	0	0
SW59	N480R,R533I,A597S,M670G	-	0	0	0	0	0	0
SW60	E612I,D622L,F664L,E739P	-	0	0	0	0	0	0
SW61	I611Q,M670G,E739P,E742H	-	0	0	0	0	0	0
SW62	F595W,F664H,Y668F,E739P	-	0	0	0	0	0	0
SW63	A597S,A605G,D622A,F664L	-	0	0	0	0	0	0
SW64	L606P,I611E,E739R,R743A	-	0	0	0	0	0	0
SW65	D607W,I611Q,R657D,E742V	-	0	0	0	0	0	0
SW66	T541A,I611Q,L613I,D622L	-	0	0	0	0	0	0
SW67	K537I,S573H,N580S,D622S	-	0	0	0	0	0	0
SW68	N480R,S573E,D607W,A740R	-	0	0	0	0	0	0
SW69	D575T,L613D,E739R,A774H	86.3	35415	0	0	0	0	0
SW70	Q579A,R657D,F664Y,A740R	-	0	0	0	0	0	0
SW71	R533I,K537I,A605K,L613I	-	0	0	0	0	0	0
SW72	T511V,E517G,L606C,F664Y	89	607861	899097	748202	686452	668056	716970
SW73	D575T,F664H,E742V,R743A	-	0	0	0	0	0	0
SW74	A594C,I611E,F664L,A740S	86.3	1467239	916262	0	0	0	0
SW75	N580S,L613A,A740S,R743A	-	0	0	0	0	0	0
SW76	S510I,T511V,L613I,E739R	89	1011815	1256969	984215	820803	736008	655700
SW77	V583K,E612I,L613D,Y668F	-	0	0	0	0	0	0
SW78	S573E,R584V,A594C,D622S	-	0	0	0	0	0	0

*The pink rows indicate the sequences of polymerases that generated full-length PCR product at a temperature at 94.0 °C and below. The green rows indicate the sequences of polymerases that generated full-length PCR product at a temperature between 86.3 °C and 93.7 °C, but not at 94.0 °C, suggesting thermal instability. The blue rows indicate the sequences of polymerases that lack evidence of activity at any temperature.

Table 4-4. Generation of full length PCR products from dNTPs by individual polymerases from the randomly generated (L4) Library at the indicated temperatures.

Cell Line	Substitutions	Optimal Temp	Raw Densities(CNT/mm^2)					
			86.3	89.0	91.1	92.6	93.7	94.0
SW4		89.0	2681942	2925066	2705135	2570101	2474721	2364200
SW212	pL4Q,G16S,R91H,E292G,D575N,S620P	-	0	0	0	0	0	0
SW213	V110A	94.0	0	0	629162	696297	776068	789298
SW214	G197C,F269S,K790R	-	0	0	0	0	0	0
SW215	pL409P,V615I,K828R	-	0	0	0	0	0	0
SW216	L27Q,L30Q,R263S,L273R,pL409P	-	0	0	0	0	0	0
SW217	F89S,I160T,P261L,*GAP*	-	0	0	0	0	0	0
SW218	V38G,K222E,F255I,E407A,E691A	-	0	0	0	0	0	0
SW219	P552L,L765P	86.3	75495	0	0	0	0	0
SW220	N482I	91.1	294562	308802	352467	341925	313543	303106
SW221	A83G,I135N,L285Q,Y336H,*GAP*	-	0	0	0	0	0	0
SW222	G393S,T444P,M670T,E710G	-	0	0	0	0	0	0
SW223	L789P	89.0	196706	235080	36241	0	0	0
SW224	E6D,K351R,A797D,G821D	-	0	0	0	0	0	0
SW225	L122Q,D341G,A411V	94.0	276216	318539	401690	406474	409339	439202
SW226	I596M,M643V	89.0	1976511	2266875	2145291	2027642	1963816	1765708
SW227	A115P,pL458R,H558P,H617R,L654P,K801E	-	0	0	0	0	0	0
SW228	Y113H,G276V,pL409P	-	0	0	0	0	0	0
SW229	R693C,E731G,V812A	89.0	2295907	2574544	1929124	287351	0	0
SW230	F44L,T183A,K194E,L291P	91.1	0	152790	181860	171210	136258	109580
SW231	K337E	89.0	155650	2830256	2594346	2465070	2453482	2217280
SW232	L362P,M441I,E517V	-	0	0	0	0	0	0
SW233	G81D,K203E,V446A,D634G	-	0	0	0	0	0	0
SW234	*GAP*	-	0	0	0	0	0	0
SW235	L279P,S287T	92.6	1919302	2036730	2023186	2141827	1991595	1881015
SW236	L12P,V133M,N217D,L266P,E300G,L546P,W703R,L825P	-	0	0	0	0	0	0
SW237	A231V,*GAP*	-	0	0	0	0	0	0
SW238	K203I,A268G,D544N,R633L,T753A,V763A	86.3	125304	45544	0	0	0	0
SW239	L285P,*GAP*	-	0	0	0	0	0	0
SW240	R91L,E534G	91.1	589165	2202362	2313418	2089636	2026448	1996181
SW241	L221P,E264V,K528R,GAP	-	0	0	0	0	0	0
SW242	L777P,Stop830,Stop831	89.0	751487	1105384	832476	504161	188014	97316
SW243	N624Y	89.0	337177	805380	736375	693938	672573	595106
SW244	K337E	89.0	305553	2494404	2346465	2180618	2073938	2098704
SW245	A126P,*GAP*	-	0	0	0	0	0	0
SW246	D185V,pL491P,M643V	-	0	0	0	0	0	0
SW247	Y75C,K216E,S377T,A565V,M758T,E770D	89.0	1586157	2288328	1973994	1367562	456778	125527
SW248	*GAP,GAP*	-	0	0	0	0	0	0
SW249	E794G,M804V	86.3	956992	887580	701751	485211	214733	116760
SW250	L530P,K539N,L654P	86.3	544325	41300	0	0	0	0
SW251	F44I,E167G	94.0	450335	1072266	1113110	1170972	1165528	1283853
SW252	Y392F,N412D,N562D,E649G	89.0	57813	546638	440068	335711	216585	156511
SW253	P809T,E227K	89.0	82899	2184044	2140841	2005506	1872290	1813403
SW254	*GAP*	-	0	0	0	0	0	0
SW255	*GAP*	-	0	0	0	0	0	0
SW256	K216I,A455D,V651E	86.3	385109	294115	251941	130359	47879	0
SW257	L108P,G197S,L377P,R390C,T503A,GAP	-	0	0	0	0	0	0
SW258	K125M	89.0	495313	796179	723534	627670	527588	704237
SW259	E167G,S309P,T719A,L814Q	-	0	0	0	0	0	0
SW260	W315C,T506P	86.3	231973	63284	23784	0	0	0
SW261	*GAP*	-	0	0	0	0	0	0
SW262	Y158C,S309T,A404T,S540G,M758T	86.3	1952541	1879977	1716486	1218814	828628	460152
SW263	V38A,K337E,V796D	86.3	1247145	935335	428651	50958	0	0
SW264	D101G	89.0	2575030	2635450	2621795	2569490	2596392	2453437
SW265	H330P,*GAP*	-	0	0	0	0	0	0
SW266	R91H,F561S	86.3	2131870	1821485	773651	51273	0	0
SW267	D493G	86.3	2861017	2845538	2527108	2319857	2412747	2320855
SW268	S121T,E599V,Y808H	86.3	1142750	1114711	1042342	1019279	945295	915580
SW269	A80V,R220C,G367C,D378N,R389L,S574G,G752D,M776L	-	0	0	0	0	0	0
SW270	pL4R,I150N,K203R,K337E,Q563Stop,P647L,A774V	86.3	41034	32615	21904	0	0	0
SW271	W425R,E771G,V796D	-	0	0	0	0	0	0
SW272	L13P,A565V	-	0	0	0	0	0	0
SW273	L93P,E156G,A213T,P299S,N580S,E771G,V796D	86.3	33845	22121	0	0	0	0
SW274	E109K,G209C,W240R,L373Q	91.1	662117	891592	1018264	961926	605490	705172
SW275	K46E,V118A,L218P,I529T,Q579H	-	0	0	0	0	0	0
SW276	G184C,pL491P,R556H	-	0	0	0	0	0	0
SW277	S309Y,D369G,F479S,I581V,A605T	-	0	0	0	0	0	0
SW278	E420D,E678G,K828E	86.3	1392467	1227461	980922	540006	129694	40950
SW279	K216I,T503A,T511M,Q589R,K759E	-	0	0	0	0	0	0
SW280	V780A	89.0	1463604	1525326	1383685	1305210	1239626	1132704
SW281	K759R	89.0	1713814	1736544	1434711	1493360	1378510	1262243
SW282	E109K,G209C,W240R,L373Q	92.6	163899	229883	278944	333824	180119	193040
SW283	F721L	93.7	1042514	1510617	1497583	1505482	1558579	1448625
SW284	Y42H,R220C,W425Stop,R590W	-	0	0	0	0	0	0
SW285	Y169C,T247A,D248G,E638G	91.1	552304	643967	674392	593949	395651	340181

*The pink rows indicate the sequences of polymerases that generated full-length PCR product at a temperature at 94.0 °C and below. The green rows indicate the sequences of polymerases that generated full-length PCR product at a temperature between 86.3 °C and 93.7 °C, but not at 94.0 °C, suggesting thermal instability. The blue rows indicate the sequences of polymerases that lack evidence of activity at any temperature.

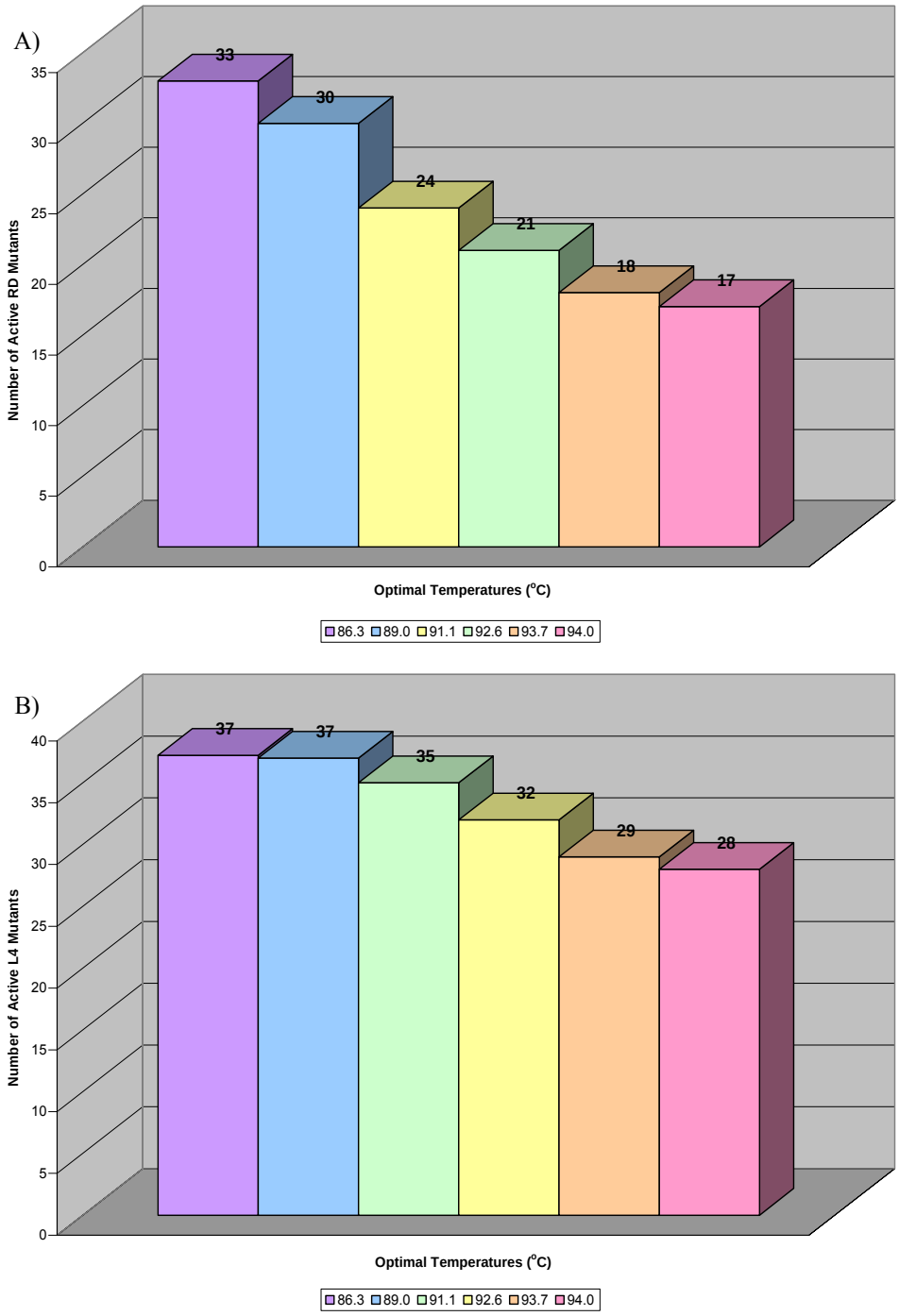


Figure 4-3. Number of active RD and L4 mutants at various temperatures. A) The number of polymerases from the RD Library (a total of 74) that show a FLP band after a PCR run at the indicated temperature. B) The number of polymerases from the L4 Library (a total of 74) that show a FLP band after a PCR run at the indicated temperature.

Table 4-5. Incorporation of dψUTP by RD Library at optimal temperatures.

Cell Line	Substitutions	Optimal Temp (°C)	All dNTPs	Raw Densities (CNT/mm^2) 9 mM dT/ 1 mM dψU	8 mM dT/ 2 mM dψU	7 mM dT/ 3 mM dψU	6 mM dT/ 4 mM dψU	5 mM dT/ 5 mM dψU	4 mM dT/ 6 mM dψU	3 mM dT/ 7 mM dψU	2 mM dT/ 8 mM dψU	1 mM dT/ 9 mM dψU	All dψUNTPs
SW4	Codon-Optimized (co) *wt* Taq	86.3	2312987	2245790	2118682	1882497	1986607	1467054	904781	278005	37474	0	0
SW4	Codon-Optimized (co) *wt* Taq	89.0	2897387	2755827	2618308	2278423	1956256	1229973	719459	165756	32937	0	0
SW5	S573E,Y668F,A740S	86.3	477902	554250	131246	175725	136720	25594	24542	0	0	0	0
SW7	A605G,L613A,E739P	86.3	1602269	1646264	1401268	1060751	775918	415561	175702	56700	32003	0	0
SW8	D575F,L606C,A740S	86.3	1651718	1160344	1175086	1256105	1490172	935859	595203	157066	49287	0	0
SW10	N480R,F595V,E742H	86.3	2447053	1883909	0	1937412	2114111	1331645	924521	223821	47201	0	0
SW11	E517I,V583K,A597S	89.0	1522581	908822	895675	804252	1019994	732645	390750	119777	20402	0	0
SW12	D575F,V583K,M670A	86.3	24450	0	0	0	0	0	0	0	0	0	0
SW14	A594C,F664Y,A774H	86.3	2042519	1602378	1425098	1258617	1222901	625304	221786	59112	33367	0	0
SW15	F595W,L606P,D622S	86.3	0	0	0	0	0	0	0	0	0	0	0
SW16	S573E,D575F,F595V	89.0	499101	301111	229593	148461	47253	55941	0	0	0	0	0
SW17	S510I,A605K,L606S	89.0	2533093	2081443	2281308	1758469	1347369	451647	196131	45945	0	0	0
SW21	Q486H,D575T,N580S	86.3	1962394	1622002	1646617	1759955	1806139	1022359	546130	161926	34784	0	0
SW25	A594T,L606C,R657D	89.0	844336	562078	40359	0	220913	125825	54514	0	0	0	0
SW26	T541A,A605G,L606S	86.3	2092635	2207538	2114500	1849280	1653179	1163636	571355	166205	41491	0	0
SW27	E517G,K537I,L613A	86.3	1096591	553051	547387	434227	543088	235326	94673	49521	0	0	0
SW29	A597S,A740R,E742V	86.3	2793753	2548317	2716809	2354623	2534067	1976105	1338144	471115	79122	0	0
SW30	N580Q,A605E,L613I	89.0	722711	454155	430993	346276	219960	136838	71434	35934	0	0	0
SW31	N580S,F595V,A605G	86.3	1423097	1483595	1310419	1238333	1629658	1226728	713571	193336	58457	0	0
SW34	Q486H,E517G,A605K	89.0	2258399	2192163	1695177	1831735	1732804	1128460	522743	149657	39295	0	0
SW36	D575T,N580Q,R584V	89.0	155345	49500	46602	52393	54069	29149	14770	0	0	0	0
SW39	A597S,I611E,Y668F	86.3	756446	575304	348163	210839	139377	47287	28160	0	0	0	0
SW41	L606S,R657D,E739R	89.0	423788	423788	423788	423788	423788	423788	423788	423788	0	0	0
SW46	T541A,D575F,L613A,D622A	89.0	54135	38017	26102	43280	36751	29931	0	0	0	0	0
SW47	T511V,A594C,L606S,A740R	89.0	543410	447310	396338	382912	321533	139661	55051	20913	0	0	0
SW49	Q486H,F595V,D622A,F664Y	89.0	453564	321305	309288	205921	99846	38405	38184	0	0	0	0
SW52	D575T,A605E,L606C,D622A	86.3	338212	332286	295345	255735	205285	115309	44621	18348	0	0	0
SW53	A594T,L613A,F664Y,E742H	86.3	1782420	1289452	1101912	571901	372153	113447	36870	0	0	0	0
SW54	D575F,N580Q,W601G,D622S	86.3	1583012	1815276	1940698	1651094	1410232	846848	395770	80472	36439	0	0
SW55	K537I,L606P,A740S,E742H	86.3	1754598	1788506	1632844	1278879	1102959	424994	166452	45263	21652	0	0
SW56	A597S,W601G,L606S,F664H	86.3	151962	0	0	0	0	0	0	0	0	0	0
SW69	D575T,L613D,E739R,A774H	86.3	37341	0	0	0	0	0	0	0	0	0	0
SW72	T511V,E517G,L606C,F664Y	89.0	244721	186500	161051	135930	96028	47852	24177	12878	0	0	0
SW74	A594C,I611E,F664L,A740S	86.3	1306976	1075033	738781	473835	273745	104717	44540	29520	0	0	0
SW76	S510I,T511V,L613I,E739R	89.0	437213	262515	229710	99167	168908	101917	37432	22388	0	0	0

*The pink rows indicate the polymerases that showed activity with a temperature of 94.0 °C and lower; these data can be compared to those in Table 3-4. The green rows indicate the polymerases that showed activity with a temperature between 86.3 °C and 93.7 °C; these had no activity at 94.0 °C, suggesting thermal instability.

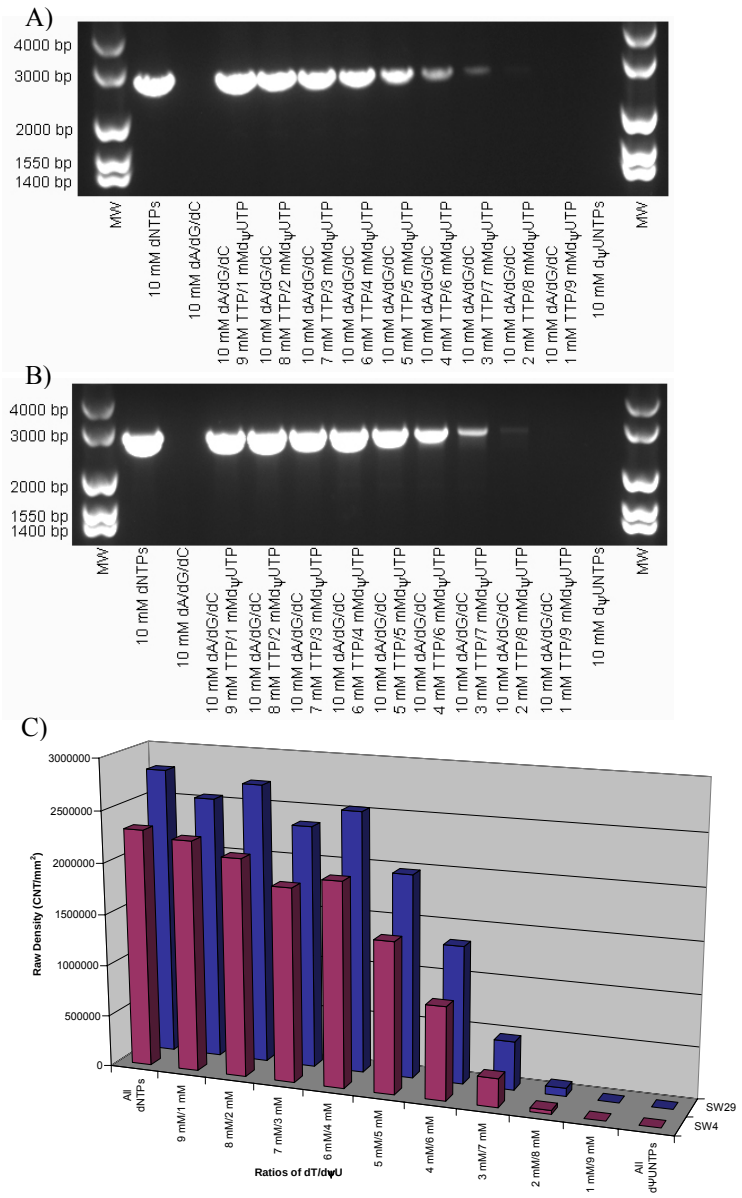


Figure 4-4. Generation of full length PCR product at 86.3 °C using dψUTP by the co-*Taq* polymerase and the RD polymerase in the SW29 cell line. Concentrations of dNTPs/dψUNTPs listed are the starting concentrations (see Materials and Methods for listing of final concentrations). A) Incorporation of various dNTP/dψUNTP ratios by co-*Taq* polymerase. FLP is not generated beyond the ratio of 2 mM TTP/8 mM dψUTP. B) Incorporation of various dNTP/dψUNTP ratios by SW29 cells. FLP is not generated beyond the ratio of 2 mM TTP/8 mM dψUTP. C) A graphical comparison of the band densities in each of these gels. The red columns correlate to the bands in gel A; the blue columns represent those in gel B. Densities can also be found in Table 4-5.

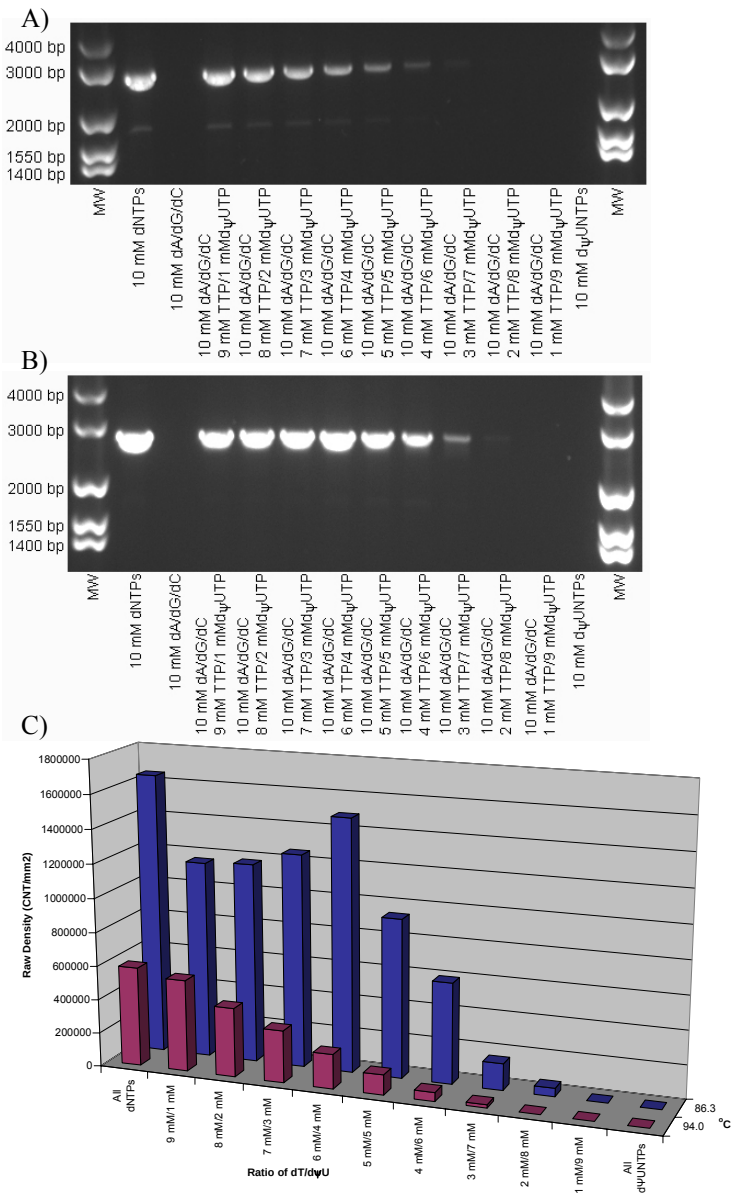


Figure 4-5. Generation of full length PCR product at 94.0 °C and 86.3 °C using dψUTP by the RD polymerase in the SW8 cell line. Concentrations of dNTPs/dψUNTPs listed are the starting concentrations (see Materials and Methods for listing of final concentrations). A) Incorporation of various dNTP/dψUNTP ratios by SW8 cells at 94.0 °C. FLP is not generated beyond the ratio of 3 mM TTP/7 mM dψUTP. B) Incorporation of various dNTP/dψUNTP ratios by SW8 cells at their optimal temperature of 86.3 °C. FLP is not generated beyond the ratio of 2 mM TTP/8 mM dψUTP. C) A graphical comparison of the band densities in each of these gels. The red columns correlate to the bands in gel A; the blue columns represent those in gel B. Densities can also be found in Table 3-4 and Table 4-3.

Table 4-6. Incorporation of dψUTP and dψTTP by co-*Taq* Polymerase at various temperatures.

Cell Line	Melting Temp (°C)	Raw Densities (CNT/mm^2)										
		500 mM dNTPs	**450 mM dT/ 50 mM dψU**	**400 mM dT/ 100 mM dψU**	**350 mM dT/ 150 mM dψU**	**300 mM dT/ 200 mM dψU**	**250 mM dT/ 250 mM dψU**	**200 mM dT/ 300 mM dψU**	**150 mM dT/ 350 mM dψU**	**100 mM dT/ 400 mM dψU**	**50 mM dT/ 450 mM dψU**	**500 mM dψUNTPs**
SW4	86.3	2312987	2245790	2118682	1882497	1986607	1467054	904781	278005	37474	0	0
SW4	89.0	2897387	2755827	2618308	2278423	1956256	1229973	719459	165756	32937	0	0
SW4	91.1	637750	387243	471969	492414	249895	124369	55590	27124	0	0	0
SW4	92.6	455238	323427	258032	222532	153441	62379	37008	24186	0	0	0
SW4	93.7	643793	350500	304969	131233	172586	70910	36048	25309	0	0	0
SW4	94.0	2244256	2005371	1995649	1535822	1255379	589637	188752	64360	0	0	0
Cell Line	**Melting Temp (°C)**	**500 mM dNTPs**	**450 mM dT/ 50 mM dψT**	**400 mM dT/ 100 mM dψT**	**350 mM dT/ 150 mM dψT**	**300 mM dT/ 200 mM dψT**	**250 mM dT/ 250 mM dψT**	**200 mM dT/ 300 mM dψT**	**150 mM dT/ 350 mM dψT**	**100 mM dT/ 400 mM dψT**	**50 mM dT/ 450 mM dψT**	**500 mM dψTNTPs**
SW4	86.3	1315056	1705027	1783891	1837846	1664575	1361746	1143128	1160086	793612	159248	0
SW4	89.0	1383077	1499021	1682736	1506228	1719608	1656790	1374745	915308	543815	72910	0
SW4	91.1	805345	860275	820663	804330	826318	729993	593906	425430	278006	58899	0
SW4	92.6	714674	1074727	1132278	782801	1109007	713425	545528	463738	200326	37858	0
SW4	93.7	549273	525350	490532	450728	486603	409646	290744	211733	121997	36127	0
SW4	94.0	364939	431868	517729	363169	446363	332797	302432	239203	112759	34763	0

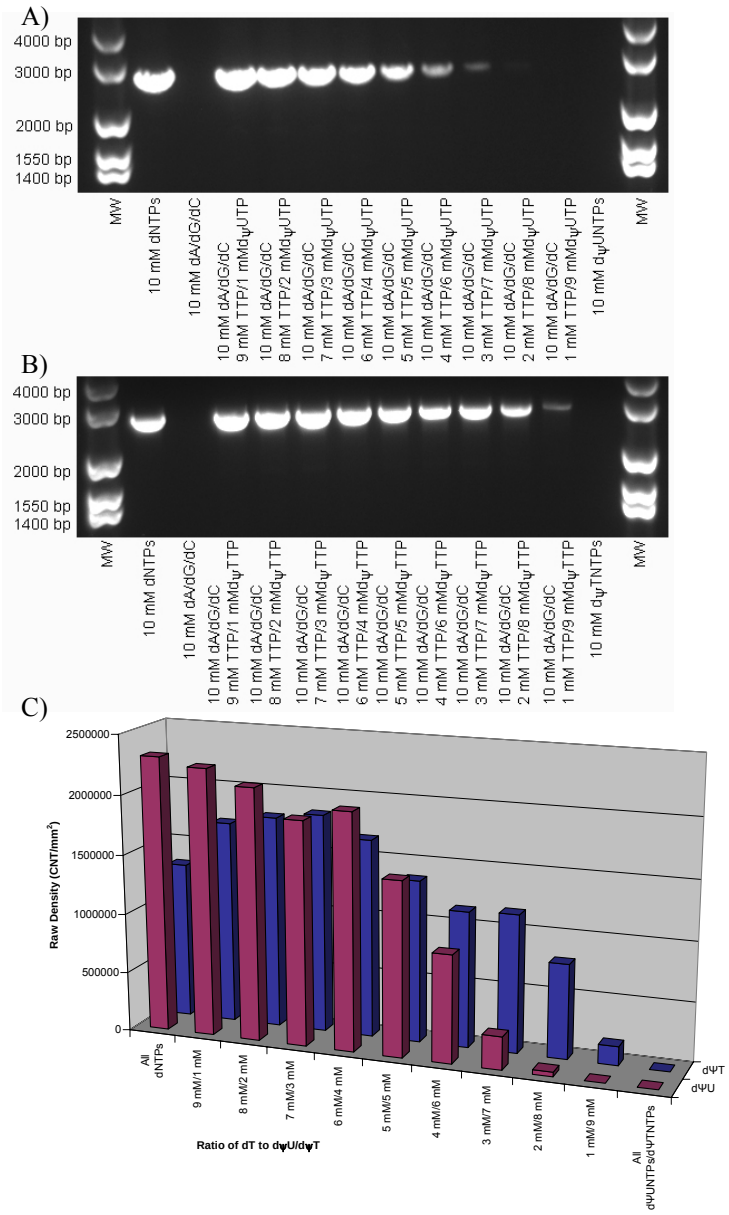


Figure 4-6. Generation of full length PCR product at 86.3 °C by co-*Taq* polymerase using various TTP:dψUTP and TTP:dψTTP ratios. Concentrations of dNTPs/dψUNTPs/dψTNTPs listed are the starting concentrations (see Materials and Methods for listing of final concentrations). A) Incorporation of various dNTP/dψUNTP ratios by co-*Taq* polymerase at 86.3 °C. FLP is not generated beyond the ratio of 2 mM TTP/8 mM dψUTP. B) Incorporation of various dNTP/dψTNTP ratios by co-*Taq* polymerase at 86.3 °C. FLP is not generated beyond the ratio of 1 mM TTP/9 mM dψTTP. C) A graphical comparison of the band densities in each of these gels. The red columns correlate to the bands in gel A; the blue columns represent those in gel B. Densities can also be found in Table 4-6.

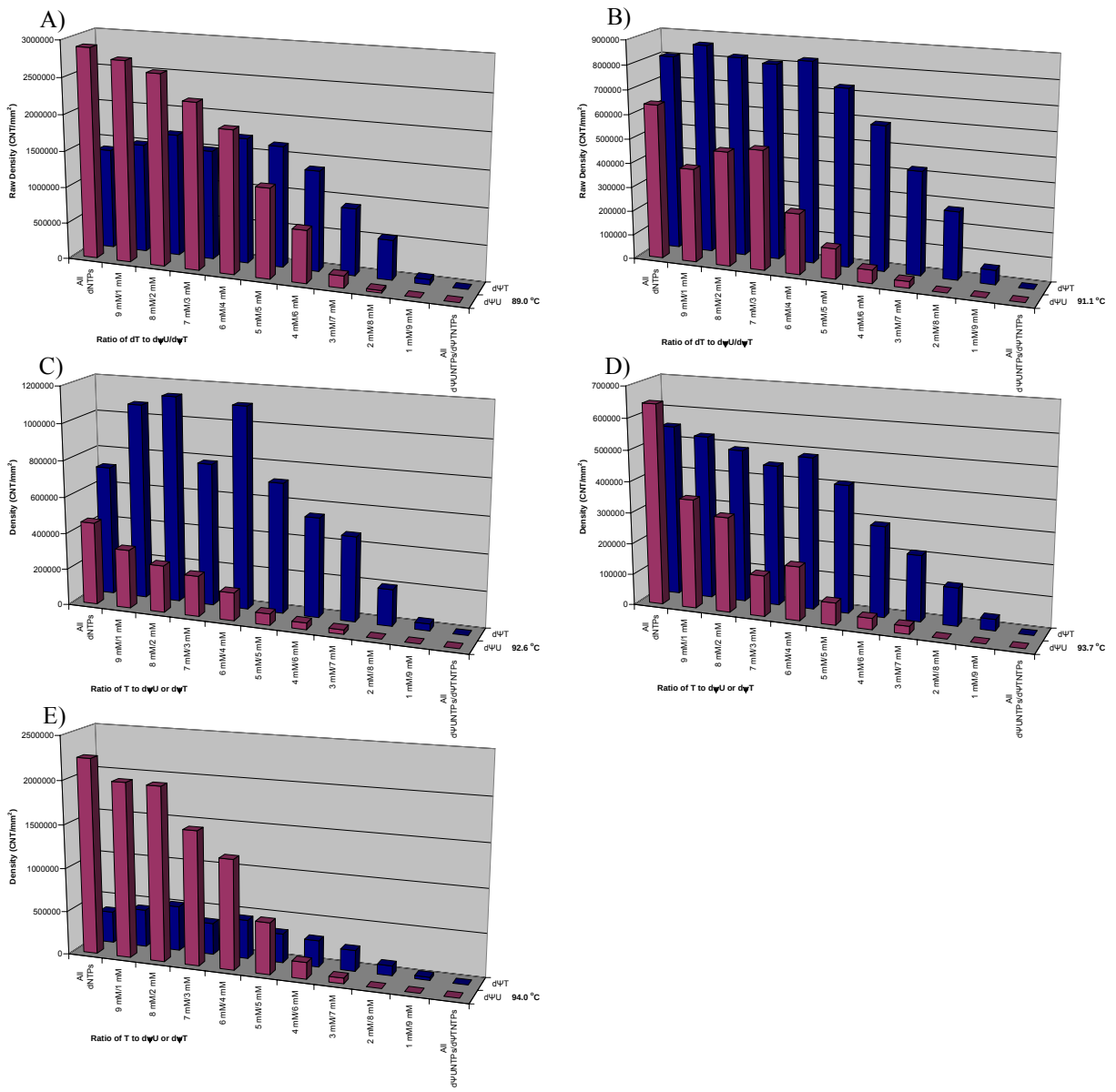


Figure 4-7. Graphical comparisons of the band densities listed in Table 4-6. All of the reactions were identical except for the temperature of the PCR and ratios of dT:dψU or dT:dψT. The red columns correlate to the densities of full-length PCR product bands present on gels containing the dψU studies. The blue columns correlate to the densities of the full-length PCR product bands present on gels containing the dψT studies. Data for the 86.3 °C PCR study is shown in Figure 4-6C. A) A graphical comparison of the band densities at 89.0 °C. B) A graphical comparison of the band densities at 91.1 °C. C) A graphical comparison of the band densities at 92.6 °C. D) A graphical comparison of the band densities at 93.7 °C. E) A graphical comparison of the band densities at 94.0 °C.

CHAPTER 5
CONCLUSIONS

The notion of creating an artificially expanded genetic information system (AEGIS) by adding extra "letters" to the DNA alphabet has sparked interest in determining what features of these non-standard nucleotides (NSBs) could be a hindrance for incorporation by polymerases. Studies have been performed on the incorporation of a variety of different NSBs, such as those lacking minor-groove electrons (Hendrickson et al., 2004), those with a C-glycosidic linkage (Lutz et al., 1999), and those that do not allow for the creation of hydrogen bonds between the nucleobases (Delaney et al., 2003). However, prior to this study, only a limited amount of research has studied the incorporation of multiple, sequential NSBs. This dissertation focused on the stability of the duplex DNA containing multiple C-glycosides, the ability of polymerases to incorporate multiple, sequential nucleotides containing a C-glycosidic linkage, and the directed evolution of polymerases to incorporate these nucleotides more efficiently and faithfully.

DNA Helical Structure in the Presence of C-Glycosides

Using circular dichroism (CD), previous studies of the helical structure of duplex DNA have shown that poly(U)●poly(A) helices favor the A-helical conformation while poly(dT)●poly(dA) helices display B-DNA structure (Ivanov et al., 1973, Saenger, 1984, Chandrasekaran and Radha, 1992). It was necessary to determine if the presence of multiple C-glycosides in double-stranded DNA (dsDNA) would alter the conformation of the helix to a point where there is a phase transition from B-DNA to A-DNA, possibly making it difficult for polymerases evolved to handle B-DNA helices unable to replicate the DNA containing multiple C-glycosidic nucleotides.

CD measurements were used to test dsDNA containing from one to twelve consecutive dψU●dA or dT●dA base pairs. Results indicated that at 25 °C, the addition of more ψU did not

generate a trend in the CD spectra that might indicate a change from a B-helical conformation to an A-helical conformation. Relatively little difference was observed between the CD spectra of duplexes containing increasing numbers of 2'-deoxypseudouridine (dψU) nucleotides and those containing dT. These data suggest that gross conformational change should not present a problem for a polymerase to incorporate and replicate DNA containing these C-glycosides.

Polymerase Screen for the Incorporation of C-glycosides

Previous studies on the incorporation of C-glycosides required only that polymerases incorporate up to three consecutive 2'-deoxypseudothymdidine (dψT) residues into a growing DNA strand (Lutz et al., 1999). To create an artificially expanded alphabet that freely incorporates C-glycosides, including the AEGIS alphabet that has three species with a C-glycosidic linkage (Fig. 1-4), polymerases would be required to incorporate more than three of these NSBs consecutively, efficiently, and faithfully.

In the first part of this study, primer-extension assays were used to screen a number of Family A and Family B polymerases for their ability to incorporate and extend beyond four of the two representative C-glycosides, 2'-deoxypseudouridine (dψU) and 2'-deoxypseudothymidine (dψT). Studies described here showed that although the Klenow (exo-), *Bst* Large Fragment, and Therminator™ polymerases performed exceptionally well in their ability to incorporate and fully extend beyond four consecutive dψT and dψU nucleotides. Klenow (exo-) and *Bst* are not thermostable, and thus cannot support PCR. Further, according to its manufacturer, Therminator is not recommended for any applications except DNA sequencing and primer-extension reactions. This means that none of these three polymerases were likely candidates for future studies. *Taq* polymerase, however, which was also able to incorporate and

extend beyond the four NSBs, albeit with less efficiency, is able to support high temperature PCR. *Taq* was therefore selected as a candidate for further study.

It is also interesting to note that based on full-length product (FLP) band densities, it appears that the incorporation of dψTTP by polymerases was more efficient than the incorporation of dψUTP.

Taq Polymerase Primer-Extension Assays

If *Taq* was used for the starting point to obtain polymerases that accept C-glycosides, it must replicate its own encoding polymerase gene, forcing it to incorporate four consecutive dψT or dψU across from template dA, as this is the longest run of consecutive dA's in the *taq* polymerase gene. Since we have already shown that *Taq* can incorporate and extend beyond four consecutive C-glycosides, as seen in Chapter 2 of this dissertation, we next needed to demonstrate its ability to incorporate and extend beyond up to twelve consecutive dψT-dA or dψU-dA base pairs.

Results showed that the production of FLP was terminated if it required the incorporation of more than five consecutive C-glycosides by *Taq* polymerase. The FLP band densities from this data showed that the dψTTP was incorporated more efficiently than the dψUTP. If this polymerase is to be used as a potential candidate for synthetic biology containing C-glycosides, it must first be modified by directed evolution experiments to allow it to incorporate more consecutive C-glycosides.

Growth and Purification of *Taq* Polymerase

A tightly regulated plasmid containing an N-terminal hexahistidine tagged *wt taq* gene ($His_{(6)}$-*wt Taq*) was constructed and transformed into the *E. coli* TG-1 expression strain (Skerra, 1994). Growth and expression conditions were then optimized prior to using the cells in

selection experiments. Previous studies showed that the expression of polymerases *in vivo* is toxic to the cells (Moreno et al., 2005, Andraos et al., 2004); this was also observed in these studies. Once the most favorable set of expression conditions was ascertained (a 1 hr expression following a late log phase induction), the $His_{(6)}$-*wt Taq* polymerase was purified via nickel chromatography and its activity was tested. Almost identical amounts of FLP were found to be generated in a PCR reaction when identical concentrations (ng/μL) of the purified $His_{(6)}$-*wt Taq* polymerase and *Taq* polymerase purchased from New England BioLabs were used; this signifies that the purified protein isolated was indeed an active polymerase.

It was noted that a low level of $His_{(6)}$-*wt Taq* polymerase was being produced after only 1 hr of induction, most likely best explained by polymerase toxicity. To rectify this situation, the gene encoding $His_{(6)}$-*wt Taq* polymerase was optimized for codon-usage in *E. coli* (co-*taq* gene) by our collaborators, DNA 2.0 Inc (Gustafsson et al., 2004). The codon-optimization does not affect the toxicity of the protein, but it does allow the *E. coli* cells to produce a greater amount of protein in the same amount of time. Once optimized, after one hour of induction, at least three times as much polymerase was produced, as evidenced by the density of the bands on a Coomassie blue stained SDS-PAGE (7.5%) gel (Fig. 3-6C).

The co-*taq* gene was cloned into the tightly regulated plasmid with a histidine tag, transformed into *E. coli* cells, and its growth and expression conditions were compared to those of the $His_{(6)}$-*wt Taq* polymerase. The data revealed that under identical expression conditions, more polymerase was produced by cells expressing the co-*Taq* polymerase than those expressing the $His_{(6)}$-*wt Taq* polymerase. To maximize the formation of product in the directed evolution reactions, the cells containing co-*Taq* polymerase were used in all further experiments. This is the first example of a polymerase that has been optimized for the codon usage of the expression

cell strain. The success in the overproduction of large quantities of active co-*Taq* polymerase in *E. coli*, relative to the overproduction of $His_{(6)}$-*wt Taq* in cells, could be useful for other applications such as structural studies and commercial production, which require large amounts of protein.

Creation of co-*Taq* Polymerase Mutant Libraries

Literature presents many different theories regarding the best methods to create a library most useful for directed evolution experiments. Such a library contains a large number of diverse, yet active clones (Hibbert and Dalby, 2005, Arnold and Georgiou, 2003b, Drummond et al., 2005, Park et al., 2005, Dalby, 2003, Parikh and Matsumura, 2005, Crameri et al., 1998, Crameri et al., 1996, Castle et al., 2004). For this dissertation, two of these methods were selected for comparative analysis. The first was a "rationally designed" (RD) library, generated by Dr. Eric Gaucher (FfAME), through the selection of specific replacement amino acids based on a combination of evolutionary analysis and previous functional studies. In addition, a random library (termed L4) was generated with mutations randomly spread across the whole polymerase sequence.

Creation of the Rationally Designed Mutagenic Library (RD Library)

The reconstructing evolutionary adaptive paths (REAP) approach, was used to create the RD Library, allowing for modification at residues where Type II functional divergence occurred within a family of polymerases. In this approach, sites were identified that, in the historical evolution of the polymerase, had a split "conserved but different" pattern of evolutionary variation, and had previously been suggested to lead to a change in the function or behavior of the polymerase. Using this technique in combination with sequences discussed in a recent review on the evolution of novel polymerase activities (Henry and Romesberg, 2005), a total of 57 amino acid changes at 35 sites in the *Taq* polymerase sequence were chosen. The 57

replacement amino acid residues were selected from those found at those sites within the Family A viral polymerase sequences, as literature has revealed that viral polymerases are more able to incorporate NSBs than other polymerases (Sismour et al., 2004, Leal et al., 2006, Horlacher et al., 1995). The FfAME collaborators at DNA 2.0 then created and synthesized the RD library containing 74 different mutant sequences; the 57 amino acid changes we dictated were used in various combinations to yield three or four amino acid mutations per sequence. This approach to creating mutagenic libraries restricts the diversity based on evolutionary data, but in doing so, was predicted to create a large number of active clones.

Creation of the Random Mutagenic Library (L4 Library)

The L4 random mutagenic library was created from the co-*taq* gene using error-prone PCR with the mutagens $MnCl_2$ and *Taq* polymerase and primers flanking either end of the gene (Arnold and Georgiou, 2003b). This allowed mutations to be located anywhere in the sequence of the gene. Rather than risk losing large quantities of the mutagenic PCR product during digestions, ligations, and purification, a variation of the megaprimer PCR protocol was used to create the full length plasmids with inserts (Miyazaki and Takenouchi, 2002). This procedure was found to be extremely useful when creating libraries, as it generates crossover mutations and reversions, introducing more diversity. The 74 unique clones generated by these techniques contained approximately three amino acid changes per sequence resulting from an average of 4.3 base mutations per gene. These combined procedures are recommended for creating future mutagenic libraries because of their simplicity, their cost, and the ease with which they can be modified to increase or decrease the number of mutations per gene.

Preliminary Studies of the Incorporation of dψUTP by the RD Library

Initially, members of the RD library were individually tested for their ability to incorporate increasing concentrations of dψUTP into PCR products. It was discovered that only 18 of the 74

mutants tested were able to form FLP, even in the presence of only standard dNTPs. At this point, the design of the RD library was questioned, and it was noticed that at least one of the sites we had mutated had been previously shown to be involved in the thermostability of the *Taq* polymerase (Ghadessy et al., 2001).

In addition, we designed our 57 replacement amino acid residues based on the sequences of Family A viral polymerases, which are only thermostable up to 37 °C. Therefore, it is very likely that the mutations introduced in this approach caused a decrease in the thermostability of the RD polymerase variants . The next step was to test the ability of the polymerase variants to function at a variety of temperatures.

Incorporation of dNTPs by RD and L4 Libraries at Various Temperatures

The mutants from each library were individually tested for their ability to form FLP at various temperatures in PCRs containing only standard dNTPs. We found that by lowering the temperature from 94.0 °C to 86.3 °C, the number of mutant generating PCR products increased. In the RD Library, the number of active mutant polymerases increased from 18 to 33; in the L4 Library, 39 mutants were active when the temperature was lowered, compared with only 27 active at a temperature of 94.0 °C. These results suggest that it more likely to generate active, thermostable mutants with a randomly mutated library than with a rationally designed library. This supports the conclusions drawn by Arnold and colleagues (Drummond et al., 2005), who determined that libraries with mutations distributed throughout the entirety of the gene are more likely to result in active and unique variants than if the mutations were limited to the active site.

Based on the generalization that approximately one-third of all random amino acid changes will result in the inactivation of a protein (Guo et al., 2004), and the design we employed to create the RD Library, it was perhaps reasonable to expect that more active mutants would be present in the RD Library than the randomly created L4 Library, when the same number of

clones were tested. Since this was not the case, the design of the RD Library must be examined. It was also reasonable to conclude that by focusing mutations in and around the active site, the risk of knocking out activity was increased, even though the sites chosen were known to be variable, and the residues chosen were known to function in the evolutionary history of the polymerase. This library, however, was designed with the incorporation of NSBs in mind, so the possibility was investigated that the RD variants will have an increased ability to incorporate the C-glycosides when compared to co-*Taq* polymerase.

Incorporation of dψUTP by the RD Library at Optimal Temperatures

After the identification of the optimal temperature for each of the 33 active RD polymerases, we challenged the polymerases to incorporate increasing concentrations of dψUTP in PCR reactions at their optimal temperature. It was discovered that one RD mutant polymerase (pSW27: A597S, A740R, E742V) was able to incorporate dψUTP more efficiently than the co-*Taq* polymerase. The A597S mutation has previously been shown to assist in the incorporation of rNTPs (Xia et al., 2002), while the E742 mutation contributed to the incorporation of various NSBs (Ghadessy et al., 2004). Since dψUTP is closely related to rUTP and is an NSB, it is arguable that these changes contributed greatly to the activity of this mutant. The remaining 32 polymerases tested were unable to incorporate the dψUTP as well as co-*Taq* polymerase; they were, however, able to generate more FLP at their optimal temperature than when they were tested previously at 94.0 °C.

It is possible, that the increased ability to incorporate C-glycosides at lower temperatures can be attributed to the fact that NSBs are sometimes incorporated more efficiently at lower temperatures (Rappaport, 2004, Horlacher et al., 1995). Alternatively, it was considered that the dψUTP is epimerizing at the higher temperatures, thereby making it difficult for the polymerase

to incorporate the base into a growing DNA strand (Wellington and Benner, 2006, Cohn, 1960, Chambers et al., 1963). Therefore, a test was designed to establish which of these theories was actually occurring.

Incorporation of dψUTP and dψTTP by co-*Taq* Polymerase at Various Temperatures

The presence of the methyl group on dψTTP inhibits the epimerization of the C-glycoside (Wellington and Benner, 2006), so it was possible to perform a comparative analysis between co-*Taq* polymerases' ability to cope with dψUTP and dψTTP in various concentrations and at different temperatures. The co-*Taq* polymerase was found able to incorporate final concentrations of dψTTP greater than those with dψUTP at all temperatures tested. This leads to the conclusion that the epimerization of the nucleotide is hindering the incorporation of dψUTP, consequently this should not be used as a model C-glycoside in future studies.

Selection of Thermostable RD Mutants Using Water-In-Oil Emulsions

Selections require that some members of the library perform differently than the original protein of interest (Arnold and Georgiou, 2003a, Lutz and Patrick, 2004). One of the goals of this research is to evolve polymerases to incorporate various C-glycoside triphosphates efficiently and faithfully, thus it makes sense to perform an initial selection to identify mutants able into incorporate C-glycosides . After noting that the dψUTP was most likely epimerizing under our reaction conditions, and considering our available quantities of dψTTP were limited, we decided to select for the eighteen mutant polymerases in the RD Library that exhibited activity with dNTPs at a temperature of 94.0 °C from the pool of the 74 RD mutants. In doing so, we were able to demonstrate our laboratory's ability to perform *in vitro* selections.

A variation of the compartmentalized self-replication (CSR) method was used to create water-in-oil emulsions containing all 74 mutants, as a way to link genotype to phenotype (Fig. 1-

13) (Miller et al., 2006, Tawfik and Griffiths, 1998, Ghadessy et al., 2001, Ghadessy et al., 2004). Products from the selection were recloned into the expression vector using the megaprimer PCR method previously discussed (Miyazaki and Takenouchi, 2002). Unfortunately, this protocol caused numerous crossovers, reversions, and additions, so we were not able to determine the true sequences of the all polymerases we isolated using the CSR. However, the megaprimer PCR reveals itself as an effective method for library rediversification between rounds of selection.

We were able to identify one mutant from the 50 clones we sequenced that coded for one of the eighteen variants previously shown to have activity under these reaction conditions. This demonstrates an ability to perform successful *in vitro* selections in our laboratory. If this selection was to be repeated, and products were cloned using the standard digestion, ligation, and purification techniques, it would most likely yield some to all of the eighteen sequences of active polymerases.

Future Experimentation

The results presented in this dissertation open the door for many future experiments. Further structural studies of the DNA containing multiple sequential C-glycosides can expound upon the knowledge gathered here. Given that we now know 2'-deoxypseudouridine is not a good representative of a C-glycoside, we can create duplex DNA containing 2'-deoxypseudothymidine and perform similar circular dichroism studies to distinguish what helical form the DNA assumes. In addition, thermal duplex denaturation studies can be performed with these duplex structures containing dψT to ascertain the stability of the duplex DNA formed when multiple, sequential C-glycosides are present (Geyer et al., 2003).

Additional study of rationally designed library creation is now possible. Since we now know that the use of viral residues to replace those of *Taq* polymerase at some sites may cause a decrease in protein thermostability, in the design of future libraries, we can avoid making mutations at these sites, and possibly increase the percentage of active mutants in the library. Furthermore, in future libraries, we can take into consideration the mutation of sites throughout the polymerase sequence that display Type II functional divergence, not just those in and around the active site.

A distinct decrease in the amount of full-length PCR product was observed when the PCR based assays were performed using increasing concentrations of the C-glycoside and decreasing concentrations of dT. A control reaction using only decreasing levels of dT and no thymidine analogue should be performed to determine how much of the full-length PCR product generated in future reactions actually contains the C-glycosides versus how much is produced only using dT. Another reason we may be seeing this decrease in FLP formation with increasing C-glycoside concentration could be due to the ethidium bromide dye being used to aid in the visualization of the DNA. It is, perhaps, plausible that the ethidium bromide cannot intercalate as efficiently when multiple C-glycosides are present. Therefore, a comparative analysis between the amounts of FLP formed when using ethidium bromide versus another fluorescent dye, such as the SYBR Safe™ DNA Gel Stain, could be performed.

It would also be interesting to test the 74 L4 mutants for their ability to incorporate C-glycosides more efficiently than the co-*Taq* polymerase. This information will also help determine if the mutation of residues not in the active site is beneficial to the incorporation of NSBs. Moreover, if the 74 RD mutants were retested for their ability to incorporate increasing

levels of dψTTP as opposed to dψUTP, we may find more than one polymerase that incorporates the NSBs more efficiently than co-*Taq*.

Once an acceptable library is created, *in vitro* evolution experiments can be performed to identify polymerases able to incorporate high levels of dψTTP, rather than testing each mutant individually. These selections will begin with a moderate ratio of dψT to dT, and increase with each round of selection. Between rounds, we now know that our libraries can be rediversified using the megaprimer PCR protocol, thereby reducing the risk of large quantities of product being lost to the purification steps required for traditional recloning steps. After demonstrating our ability to perform a selection for polymerases that can incorporate C-glycosides, we can begin to apply these techniques to develop polymerases that can incorporate more NSBs.

Directed evolution is already being used in industry to improving the quality of and developing new industrial enzymes and therapeutic treatments (Chirumamilla et al., 2001, Douthwaite and Jermutus, 2006). If the conjunction of the rationally designed library with the modified CSR technique proves to be successful in the isolation of large numbers of active clones, there could be a commercial impact for this system. Right now it takes three to four rounds of selection to isolate clones with a desired trait, and each round takes at least one week; with our system, it is feasible that only one to two rounds of selection would be needed, thereby cutting the time in half. In addition, the use of synthetic gene libraries reduces the amount of time spent creating libraries *de novo*. With an improved ability to produce more clones with desired activity from smaller starting libraries, imagine how many products could be quickly isolated using these techniques.

APPENDIX A
SYNTHESIS OF PSEUDOTHYMIDINE AND PSEUDOTHYMIDINE-CONTAINING OLIGONUCLEOTIDES

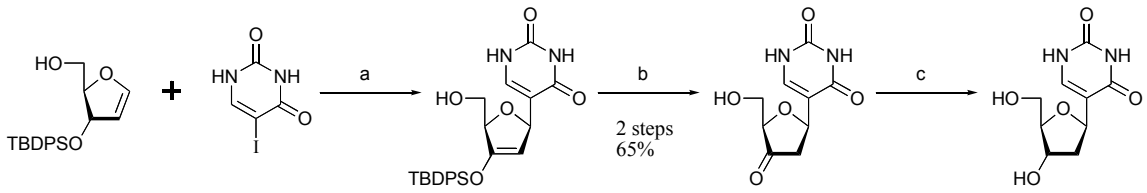
Synthesis of the 2'-deoxypseudothymidine (dψT) precursor was performed by Dr. Shuichi Hoshika according to the procedures previously set forth, with some modifications (actual scheme shown in Figure A-1) (Bhattacharya et al., 1995, Lutz et al., 1999, Zhang and Daves, 1992).

The synthesis of 2'-deoxypseudothymidine-5'-triphosphate (dψTTP) was performed by Dr. Daniel Hutter according to the standard Ludwig-Eckstein procedure for triphosphate synthesis (Ludwig and Eckstein, 1989). It was purified by HPLC on a Waters Delta 600 with Waters 2487 Dual wavelength absorbance detector, controlled by Waters Millennium software. Initial purification was on ion-exchange column [GE Healthcare HiPrep 16/10 DEAE FF column, eluent A = 10 mM NH_4CO_3, eluent B = 1 M NH_4CO_3, gradient from 0 to 80% B in 40 min, flow rate = 3 mL/min, R_t = 22 min] followed by reverse phase HPLC [Waters NovaPak HR C18 column, 19x300 mm, eluent A = 25 mM triethylammonium acetate (TEAA) pH 7, eluent B = 10% CH_3CN in 25 mM TEAA pH 7, gradient from 0 to 80% B in 32 min, flow rate = 5 mL/min, R_t = 16 min]. After lyophilization, it was twice re-dissolved in water and lyophilized again to remove excess TEAA. Analytical HPLC was performed to verify the purification [Waters Alliance 2695 with Waters 2996 PDA detector, controlled by Waters Millennium software; Dionex DNAPac PA-100 column, 4x250 mm, eluent A = 10 mM NH_4CO_3, eluent B = 500 mM NH_4CO_3, gradient from 0 to 40% B in 20 min, flow rate = 0.5 mL/min: R_t = 17 min].

NMR (Varian Mercury 300 MHz spectrometer): ^{1}H-NMR (D_2O, 300 MHz): ∂ (ppm, rel to HDO = 4.65) = 1.97 (ddd, *J* = 5.9, 9.9, 13.3 Hz, 1H); 2.13 (ddd, *J* = 2.6, 5.9, 13.3 Hz, 1H); 3.27 (s, 3H); 3.94-4.00 (m, 3H); 4.39-4.41 (m, 1H); 4.97 (dd, *J* = 5.9, 9.6 Hz, 1H); 7.65 (d, *J* = 0.8 Hz,

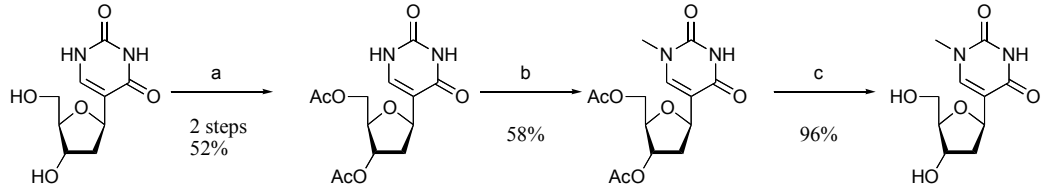
1H). ^{31}P-NMR (D_2O, 121 MHz): ∂ (ppm, rel to external standard H_3PO_4 = 0) = -10.7 (d, *J* = 20 Hz, 1P); -11.2 (d, *J* = 20 Hz, 1P); -23.3 (*t*, *J* = 20 Hz, 1P).

Synthesis of pseudothymidine

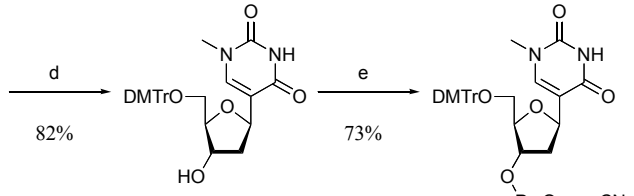


8 **9** **10** **11** **12**

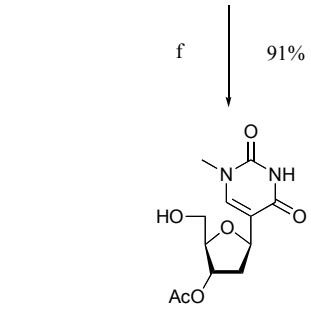
Scheme 2. a) $Pd(OAc)_2$, Ph_3As, Bu_3N, MeCN b) TBAF, AcOH, THF c) $NaBH(OAc)_3$, AcOH, MeCN



12 **13** **14** **15**



16 **18**



17

Scheme 3. a) Ac_2O, DMAP, DMF, –30ºC b) MeI, *N*,*O*-bis(trimethylsilyl)acetamide, CH_2Cl_2 ,reflux or MeI, $(i\text{-Pr})_2NEt$, DMF c) K_2CO_3, MeOH d) DMTrCl, pyridine e) $ClPN(i\text{-Pr})_2OCH_2CH_2CN$, $(i\text{-Pr})_2NEt$, CH_2Cl_2 f) aq. AcOH, THF

Figure A-1. Synthesis of pseudothymidine precursor. This scheme was designed by Dr. Shuichi Hoshika following protocol set forth previously (Bhattacharya et al., 1995, Lutz et al., 1999, Zhang and Daves, 1992).

APPENDIX B
PHYLOGENETIC TREES OF FAMILY A POLYMERASES

The following are insets of the phylogenetic tree seen in Figure 3-1 and a seed alignment of twelve of the 719 Family A polymerases identified in this tree. These trees were generated using Pfam (Bateman, 2006, Finn et al., 2006), and analyzed for sites that underwent Type II functional divergence. In this approach, Dr. Eric Gaucher identified sites that had a split "conserved but different" pattern of historical evolutionary variation, and had been previously suggested to lead to a change in the function or behavior of the polymerase (Henry and Romesberg, 2005). Using Pfam, 57 amino acid changes across 35 sites were identified within the 719 members of Family A polymerases that were available to us (Bateman, 2006, Finn et al., 2006).

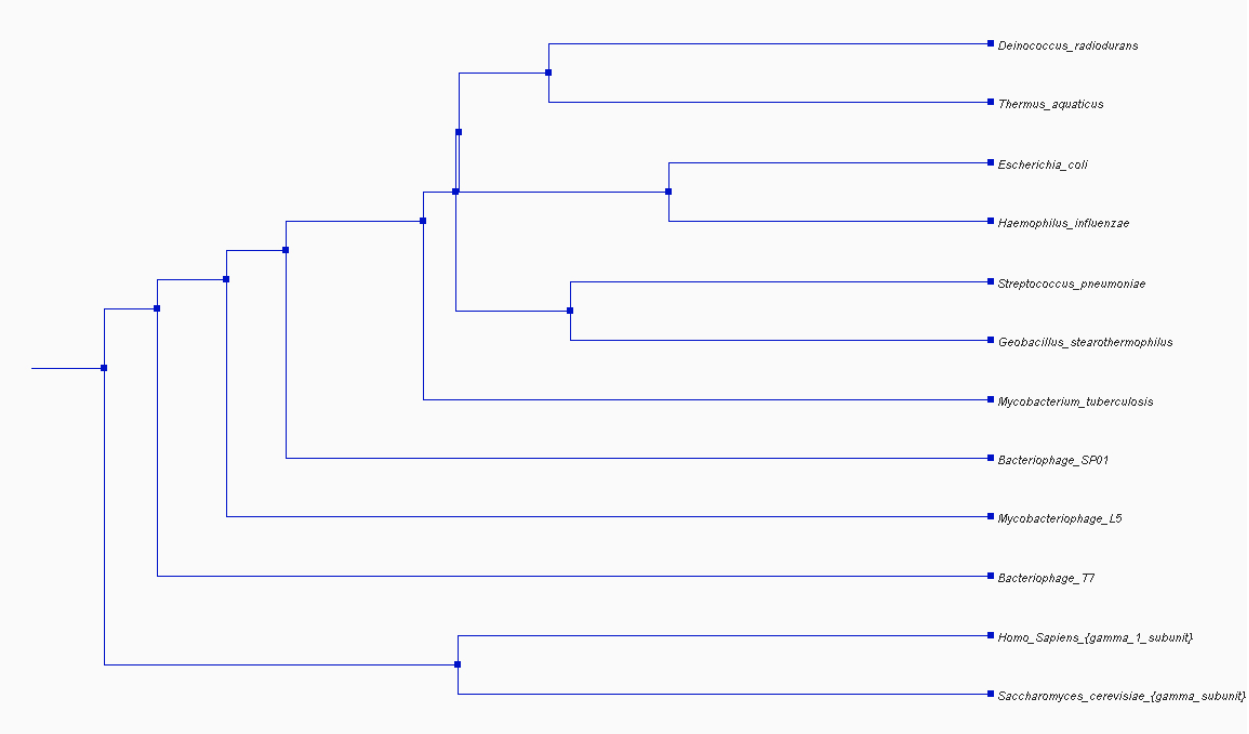


Figure B-1. A seed alignment of the Family A polymerases. This tree was generated using Pfam (Bateman, 2006, Finn et al., 2006), and displays twelve representatives of the major genera found in the 719 Family A polymerase sequences.

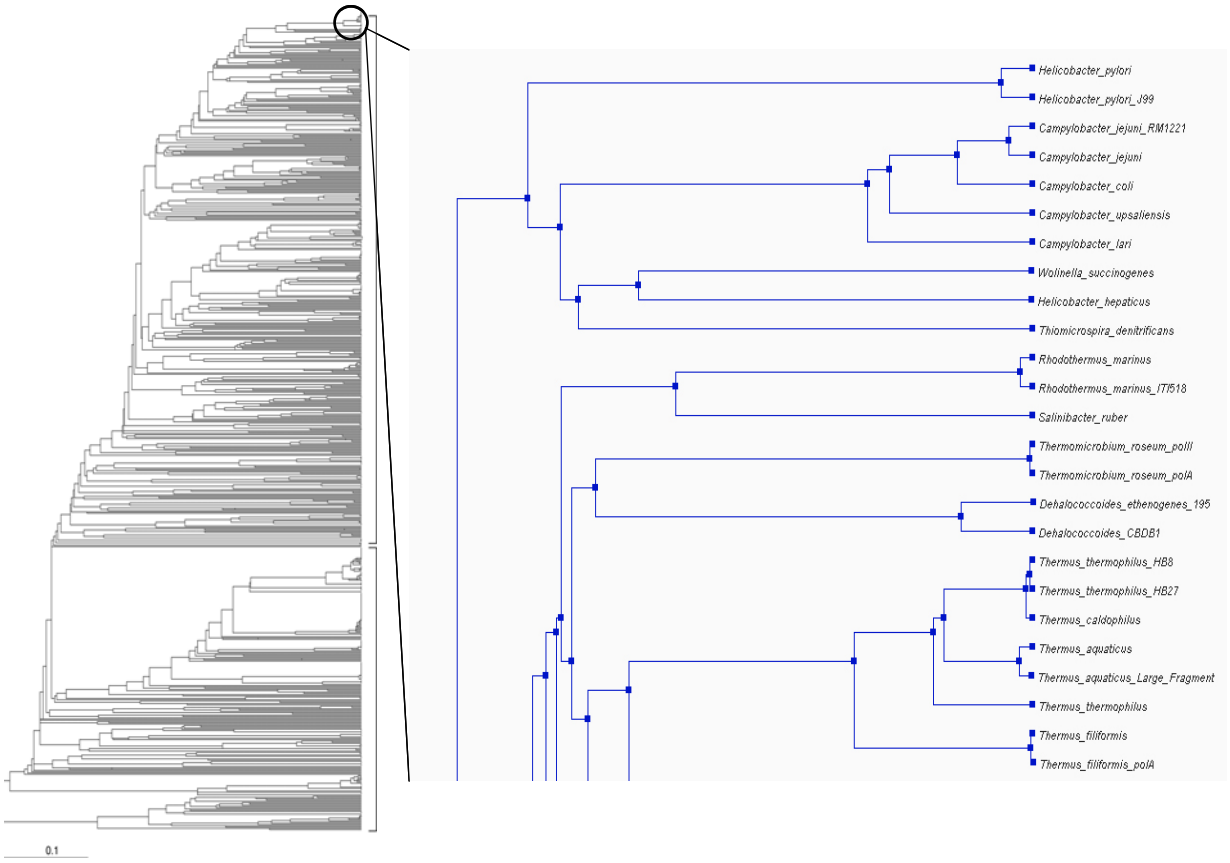


Figure B-2. Inset of the phylogenetic tree of Family A polymerases (from Fig. 3-1) showing the location of *Taq* polymerase. This tree was generated using Pfam (Bateman, 2006, Finn et al., 2006).

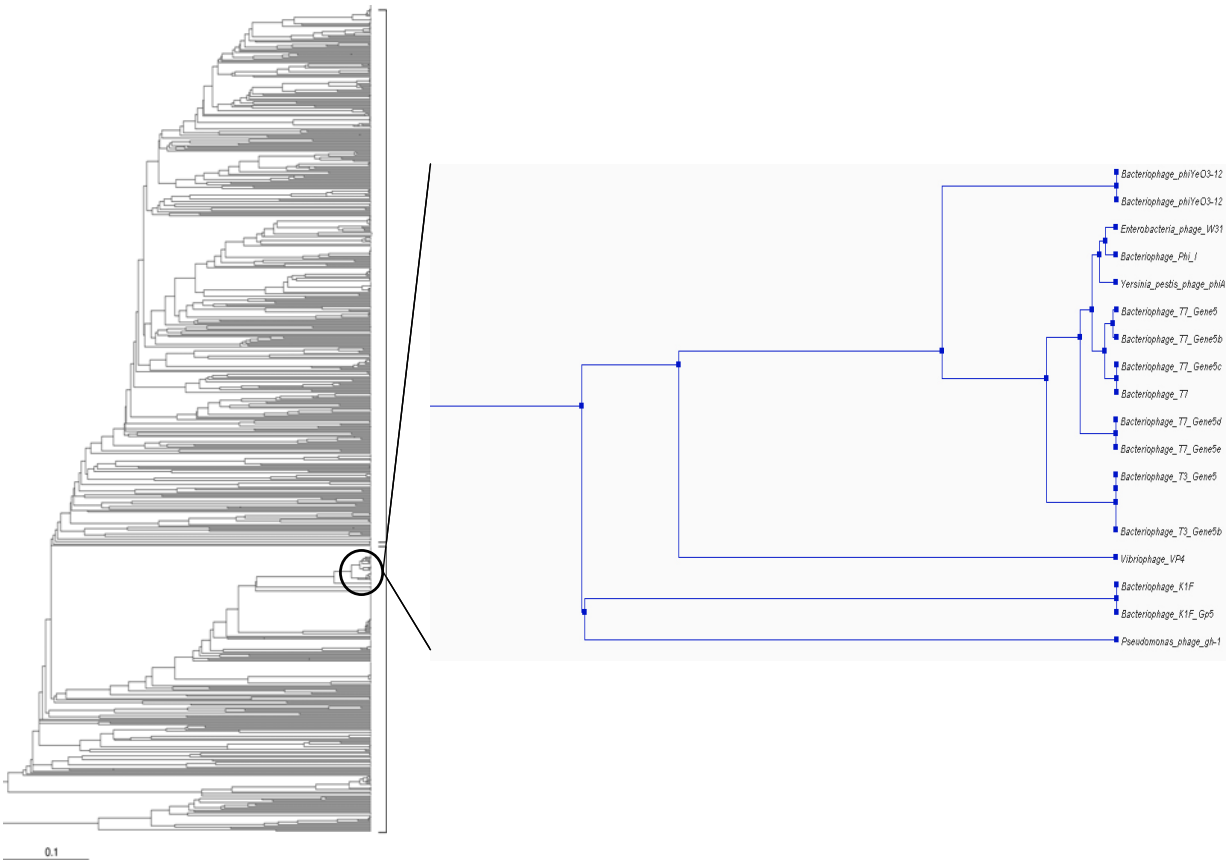


Figure B-3. Inset of the phylogenetic tree of Family A polymerases (from Fig. 3-1) showing the location of some viral polymerases. This tree was generated using Pfam (Bateman, 2006, Finn et al., 2006).

APPENDIX C
GENETIC CODE AND AMINO ACID ABBREVIATIONS

Table C-1. The Genetic Code.

First Letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		Third Letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table C-2. Amino acid abbreviations.

Name	3-Letter Code	1-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

LIST OF REFERENCES

Allemann, R. K., Presnell, S. R. and Benner, S. A. (1991) *Protein Engineering,* **4,** 831-835.

Andraos, N., Tabor, S. and Richardson, C. C. (2004) *Journal of Biological Chemistry,* **279,** 50609-50618.

Argoudelis, A. D. and Mizsak, S. A. (1976) *Journal of Antibiotics,* **29,** 818-823.

Arnez, J. G. and Steitz, T. A. (1994) *Biochemistry,* **33,** 7560-7567.

Arnold, F. H. and Georgiou, G. (Eds.) (2003a) *Directed Enzyme Evolution: Screening and Selection Methods,* Humana Press, Totowa, N.J.

Arnold, F. H. and Georgiou, G. (Eds.) (2003b) *Directed Evolution Library Creation: Methods and Protocols,* Humana Press, Totowa, N.J.

Bain, J. D., Switzer, C., Chamberlin, A. R. and Benner, S. A. (1992) *Nature,* **356,** 537-539.

Bateman, A. (2006), Vol. 2006, *Pfam,* Sanger Institute, http://www.sanger.ac.uk/Software/Pfam/.

Beard, W. A., Shock, D. D., Vande Berg, B. J. and Wilson, S. H. (2002) *Journal of Biological Chemistry,* **277,** 47393-47398.

Beese, L. S., Derbyshire, V. and Steitz, T. A. (1993a) *Science,* **260,** 352-355.

Beese, L. S., Friedman, J. M. and Steitz, T. A. (1993b) *Biochemistry,* **32,** 14095-14101.

Benner, S. A. (2004) *Accounts of Chemical Research,* **37,** 784-797.

Bhattacharya, B. K., Devivar, R. V. and Revankar, G. R. (1995) *Nucleosides & Nucleotides,* **14,** 1269-1287.

Brakmann, S. (2005) *Cellular and Molecular Life Sciences,* **62,** 2634-2646.

Brock, T. D. and Freeze, H. (1969) *Journal of Bacteriology,* **98,** 289-297.

Castle, L. A., Siehl, D. L., Gorton, R., Patten, P. A., Chen, Y. H., Bertain, S., Cho, H. J., Duck, N., Wong, J., Liu, D. L. and Lassner, M. W. (2004) *Science,* **304,** 1151-1154.

Chambers, R. W., Kurkov, V. and Shapiro, R. (1963) *Biochemistry,* **2,** 1192-1203.

Chandrasekaran, R. and Radha, A. (1992) *Journal of Biomolecular Structure & Dynamics,* **10,** 153-168.

Charette, M. and Gray, M. W. (2000) *International Union of Biochemistry and Molecular Biology Life,* **49,** 341-351.

Chien, A., Edgar, D. B. and Trela, J. M. (1976) *Journal of Bacteriology,* **127,** 1550-1557.

Chirumamilla, R. R., Muralidhar, R., Marchant, R. and Nigam, P. (2001) *Molecular and Cellular Biochemistry,* **224,** 159-168.

Cline, J., Braman, J. C. and Hogrefe, H. H. (1996) *Nucleic Acids Research,* **24,** 3546-3551.

Cohn, W. E. (1960) *Journal of Biological Chemistry,* **235,** 1488-1498.

Collins, M. L., Irvine, B., Tyner, D., Fine, E., Zayati, C., Chang, C. A., Horn, T., Ahle, D., Detmer, J., Shen, L. P., Kolberg, J., Bushnell, S., Urdea, M. S. and Ho, D. D. (1997) *Nucleic Acids Research,* **25,** 2979-2984.

Crameri, A., Raillard, S. A., Bermudez, E. and Stemmer, W. P. C. (1998) *Nature,* **391,** 288-291.

Crameri, A., Whitehorn, E. A., Tate, E. and Stemmer, W. P. C. (1996) *Nature Biotechnology,* **14,** 315-319.

Crick, F. (1970) *Nature,* **227,** 561-563.

Dalby, P. A. (2003) *Current Opinion in Structural Biology,* **13,** 500-505.

Davis, D. R. (1995) *Nucleic Acids Research,* **23,** 5020-5026.

Delaney, J. C., Henderson, P. T., Helquist, S. A., Morales, J. C., Essigmann, J. M. and Kool, E. T. (2003) *Proceedings of the National Academy of Sciences of the United States of America,* **100,** 4469-4473.

DeLano, W. L. (2002) *PyMOL,* DeLano Scientific, http://www.pymol.org.

Derti, A. (2003), Vol. 2006, *Reverse and/or complement DNA sequences,* Harvard Medical School, http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/revcomp.html.

Douthwaite, J. and Jermutus, L. (2006) *Current Opinion in Drug Discovery & Development,* **9,** 269-275.

Drummond, D. A., Iverson, B. L., Georgiou, G. and Arnold, F. H. (2005) *Journal of Molecular Biology,* **350,** 806-816.

Egli, M. (2004) *Current Opinion in Chemical Biology,* **8,** 580-591.

Emilsson, G. M. and Breaker, R. R. (2002) *Cellular and Molecular Life Sciences,* **59,** 596-607.

Eom, S. H., Wang, J. M. and Steitz, T. A. (1996) *Nature,* **382,** 278-281.

Fa, M., Radeghieri, A., Henry, A. A. and Romesberg, F. E. (2004) *Journal of the American Chemical Society,* **126,** 1748-1754.

Fairbanks, G., Steck, T. L. and Wallach, D. F. H. (1971) *Biochemistry,* **10,** 2606-2617.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A. (2006) *Nucleic Acids Research,* **34,** D247-D251.

Forterre, P. (2006) *Virus Research,* **117,** 5-16.

Garrett, R. H. and Grisham, C. M. (1999) *Biochemistry,* Harcourt Brace College Publishers, Fort Worth, TX.

Gaucher, E. A. (2006) In *National Institute of Health STTR Phase 1 Grant Number 1 R41 GM074433-01,*Foundation for Applied Molecular Evolution, Gainesville, FL.

Geyer, C. R., Battersby, T. R. and Benner, S. A. (2003) *Structure,* **11,** 1485-1498.

Ghadessy, F. J., Ong, J. L. and Holliger, P. (2001) *Proceedings of the National Academy of Sciences of the United States of America,* **98,** 4552-4557.

Ghadessy, F. J., Ramsay, N., Boudsocq, F., Loakes, D., Brown, A., Iwai, S., Vaisman, A., Woodgate, R. and Holliger, P. (2004) *Nature Biotechnology,* **22,** 755-759.

Ghosh, A. and Bansal, M. (2003) *Acta Crystallographica Section D-Biological Crystallography,* **59,** 620-626.

Goldman, M. and Marcy, D. (2001) In *HIV-1 Reverse Transcriptase Tutorial*, pp. 1-3.

Griffiths, A. D. and Tawfik, D. S. (2006) *Trends in Biotechnology,* **24,** 395-402.

Grosjean, H., Constantinesco, F., Foiret, D. and Benachenhou, N. (1995) *Nucleic Acids Research,* **23,** 4312-4319.

Gu, X. (1999) *Molecular Biology and Evolution,* **16,** 1664-1674.

Gu, X. (2002), Vol. 2006, *DIVERGE 2.0,* Iowa State University, http://xgu.zool.iastate.edu/software.html.

Guo, H. H., Choe, J. and Loeb, L. A. (2004) *Proceedings of the National Academy of Sciences of the United States of America,* **101,** 9205-9210.

Gustafsson, C., Govindarajan, S. and Minshull, J. (2004) *Trends in Biotechnology,* **22,** 346-353.

Hendrickson, C. L., Devine, K. G. and Benner, S. A. (2004) *Nucleic Acids Research,* **32,** 2241-2250.

Henry, A. A., Olsen, A. G., Matsuda, S., Yu, C. Z., Geierstanger, B. H. and Romesberg, F. E. (2004) *Journal of the American Chemical Society,* **126,** 6923-6931.

Henry, A. A. and Romesberg, F. E. (2005) *Current Opinion in Biotechnology,* **16,** 370-377.

Hibbert, E. G. and Dalby, P. A. (2005) *Microbial Cell Factories,* **4**.

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y. and Yokoyama, S. (2006) *Nature Methods,* **3,** 729-735.

Hirao, I., Ohtsuki, T., Fujiwara, T., Mitsui, T., Yokogawa, T., Okuni, T., Nakayama, H., Takio, K., Yabuki, T., Kigawa, T., Kodama, K., Nishikawa, K. and Yokoyama, S. (2002) *Nature Biotechnology,* **20,** 177-182.

Hogrefe, H. H., Cline, J., Lovejoy, A. E. and Nielson, K. B. (2001) In *Hyperthermophilic Enzymes, Pt C*, Vol. 334, pp. 91-116.

Horlacher, J., Hottiger, M., Podust, V. N., Hubscher, U. and Benner, S. A. (1995) *Proceedings of the National Academy of Sciences of the United States of America,* **92,** 6329-33.

Huisse, F. (2004) *Journal of Clinical Virology,* **30,** S26-S28.

Ivanov, V. I., Minchenk, L. E., Schyolki, A. K. and Poletaye, A. I. (1973) *Biopolymers,* **12,** 89-110.

Johnson, S. C., Marshall, D. J., Harms, G., Miller, C. M., Sherrill, C. B., Beaty, E. L., Lederer, S. A., Roesch, E. B., Madsen, G., Hoffman, G. L., Laessig, R. H., Kopish, G. J., Baker, M. W., Benner, S. A., Farrell, P. M. and Prudent, J. R. (2004) *Clinical Chemistry,* **50,** 2019-2027.

Joyce, C. M. and Benkovic, S. J. (2004) *Biochemistry,* **43,** 14317-24.

Kelman, Z., Hurwitz, J. and O'Donnell, M. (1998) *Structure,* **6,** 121-125.

Kim, Y., Eom, S. H., Wang, J. M., Lee, D. S., Suh, S. W. and Steitz, T. A. (1995) *Nature,* **376,** 612-616.

Kong, H. M., Kucera, R. B. and Jack, W. E. (1993) *Journal of Biological Chemistry,* **268,** 1965-1975.

Kornberg, A., Lehman, I. R., Bessman, M. J. and Simms, E. S. (1956) *Biochimica Et Biophysica Acta,* **21,** 197-198.

Kunkel, T. A. and Bebenek, R. (2000) *Annual Review of Biochemistry,* **69,** 497-529.

Laemmli, U. K. (1970) *Nature,* **227,** 680-685.

Lane, B. G., Ofengand, J. and Gray, M. W. (1995) *Biochimie,* **77,** 7-15.

Leal, N. A., Sukeda, M. and Benner, S. A. (2006) *Nucleic Acids Research,* **34,** 4702-4710.

Lewin, B. (1997) *Genes VI,* Oxford University Press, New York.

Li, J. S., Fan, Y. H., Zhang, Y., Marky, L. A. and Gold, B. (2003) *Journal of the American Chemical Society,* **125,** 2084-2093.

Li, J. S., Shikiya, R., Marky, L. A. and Gold, B. (2004) *Biochemistry,* **43,** 1440-1448.

Li, Y., Kong, Y., Korolev, S. and Waksman, G. (1998a) *Protein Science,* **7,** 1116-1123.

Li, Y., Korolev, S. and Waksman, G. (1998b) *European Molecular Biology Organization Journal,* **17,** 7514-7525.

Limbach, P. A., Crain, P. F. and McCloskey, J. A. (1994) *Nucleic Acids Research,* **22,** 2183-2196.

Lin-Goerke, J. L., Robbins, D. J. and Burczak, J. D. (1997) *Biotechniques,* **23,** 409-12.

Ludwig, J. and Eckstein, F. (1989) *Journal of Organic Chemistry,* **54,** 631-635.

Lutz, M. J., Horlacher, J. and Benner, S. A. (1998) *Bioorganic & Medicinal Chemistry Letters,* **8,** 1149-1152.

Lutz, S., Burgstaller, P. and Benner, S. A. (1999) *Nucleic Acids Research,* **27,** 2792-8.

Lutz, S. and Patrick, W. M. (2004) *Current Opinion in Biotechnology,* **15,** 291-297.

Michelet, W. and Genet, J. P. (2005) *Current Organic Chemistry,* **9,** 405-418.

Miller, J. H. (1972) *Experiments in molecular genetics,* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Miller, O. J., Bernath, K., Agresti, J. J., Amitai, G., Kelly, B. T., Mastrobattista, E., Taly, V., Magdassi, S., Tawfik, D. S. and Griffiths, A. D. (2006) *Nature Methods,* **3,** 561-570.

Miyazaki, K. and Takenouchi, M. (2002) *Biotechniques,* **33,** 1033-1038.

Morales, J. C. and Kool, E. T. (2000) *Journal of the American Chemical Society,* **122,** 1001-1007.

Moreno, R., Haro, A., Castellanos, A. and Berenguer, J. (2005) *Applied and Environmental Microbiology,* **71,** 591-593.

Muller, U. F. (2006) *Cellular and Molecular Life Sciences,* **63,** 1278-1293.

Najmudin, S., Cote, M. L., Sun, D. M., Yohannan, S., Montano, S. P., Gu, J. and Georgiadis, M. M. (2000) *Journal of Molecular Biology,* **296,** 613-632.

Neumann, J. M., Bernassau, J. M., Gueron, M. and Trandinh, S. (1980) *European Journal of Biochemistry,* **108,** 457-463.

Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. and Steitz, T. A. (1985) *Nature,* **313,** 762-766.

Ong, J. L., Loakes, D., Jaroslawski, S., Too, K. and Holliger, P. (2006) *Journal of Molecular Biology,* **361,** 537-550.

Parikh, M. R. and Matsumura, I. (2005) *Journal of Molecular Biology,* **352,** 621-628.

Park, S., Morley, K. L., Horsman, G. P., Holmquist, M., Hult, K. and Kazlauskas, R. J. (2005) *Chemistry & Biology,* **12,** 45-54.

Patel, P. H. and Loeb, L. A. (2001) *Nature Structural Biology,* **8,** 656-659.

Paul, N. and Joyce, G. F. (2004) *Current Opinion in Chemical Biology,* **8,** 634-639.

Pavlov, A. R., Pavlova, N. V., Kozyavkin, S. A. and Slesarev, A. I. (2004) *Trends in Biotechnology,* **22,** 253-260.

Perler, F. B., Kumar, S. and Kong, H. M. (1996) In *Advances in Protein Chemistry*, Vol. 48, pp. 377-435.

Piccirilli, J. A., Krauch, T., Moroney, S. E. and Benner, S. A. (1990) *Nature,* **343,** 33-37.

Piccirilli, J. A., Moroney, S. E. and Benner, S. A. (1991) *Biochemistry,* **30,** 10350-6.

Presnell, S. R. and Benner, S. A. (1988) *Nucleic Acids Research,* **16,** 1693-702.

Rappaport, H. P. (2004) *Biochemical Journal,* **381,** 709-717.

Raychaudhuri, S., Conrad, J., Hall, B. G. and Ofengand, J. (1998) *RNA - A Publication of the RNA Society,* **4,** 1407-1417.

Rich, A. and Zhang, S. G. (2003) *Nature Reviews Genetics,* **4,** 566-572.

Rothwell, P. J. and Waksman, G. (2005) In *Fibrous Proteins: Muscle and Molecular Motors*, Vol. 71, pp. 401-440.

Roychowdhury, A., Illangkoon, H., Hendrickson, C. L. and Benner, S. A. (2004) *Organic Letters,* **6,** 489-492.

Saenger, W. (1984) *Principles of Nucleic Acid Structure,* Springer-Verlag, New York.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. (1988) *Science,* **239,** 487-491.

Sambrook, J., Fritsch, E. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sismour, A. M. and Benner, S. A. (2005) *Nucleic Acids Research,* **33,** 5640-5646.

Sismour, A. M., Lutz, S., Park, J. H., Lutz, M. J., Boyer, P. L., Hughes, S. H. and Benner, S. A. (2004) *Nucleic Acids Research,* **32,** 728-735.

Skerra, A. (1994) *Gene,* **151,** 131-135.

Steitz, T. A. (1999) *Journal of Biological Chemistry,* **274,** 17395-17398.

Suzuki, M., Baskin, D., Hood, L. and Loeb, L. A. (1996) *Proceedings of the National Academy of Sciences of the United States of America,* **93,** 9670-9675.

Swiss Institute of Bioinformatics. (1999) Vol. 2006, *Translate,* ExPASy, http://www.expasy.ch/tools/dna.html.

Switzer, C., Moroney, S. E. and Benner, S. A. (1989) *Journal of the American Chemical Society,* **111,** 8322-8323.

Switzer, C. Y., Moroney, S. E. and Benner, S. A. (1993) *Biochemistry,* **32,** 10489-96.

Tatusova, T. A. and Madden, T. L. (1999) *Fems Microbiology Letters,* **177,** 187-188.

Tawfik, D. S. and Griffiths, A. D. (1998) *Nature Biotechnology,* **16,** 652-656.

Tindall, K. R. and Kunkel, T. A. (1988) *Biochemistry,* **27,** 6008-6013.

Vartanian, J. P., Henry, M. and WainHobson, S. (1996) *Nucleic Acids Research,* **24,** 2627-2631.

Watson, J. D. and Crick, F. H. C. (1953a) *Nature,* **171,** 964-967.

Watson, J. D. and Crick, F. H. C. (1953b) *Nature,* **171,** 737-738.

Wellington, K. W. and Benner, S. A. (2006) *Nucleosides, Nucleotides, and Nucleic Acids,* **25,** 1309-1333.

Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S. and Griffiths, A. D. (2006) *Nature Methods,* **3,** 545-550.

Xia, G., Chen, L. J., Sera, T., Fa, M., Schultz, P. G. and Romesberg, F. E. (2002) *Proceedings of the National Academy of Sciences of the United States of America,* **99,** 6597-6602.

Yan, X. H. and Xu, Z. R. (2006) *Drug Discovery Today,* **11,** 911-916.

Zhang, H. C. and Daves, G. D. (1992) *Journal of Organic Chemistry,* **57,** 4690-4696.

Zhao, H. M., Giver, L., Shao, Z. X., Affholter, J. A. and Arnold, F. H. (1998) *Nature Biotechnology,* **16,** 258-261.

Zhou, B. L., Pata, J. D. and Steitz, T. A. (2001) *Molecular Cell,* **8,** 427-437.

Zhou, J., Yang, M. M., Akdag, A. and Schneller, S. W. (2006) *Tetrahedron,* **62,** 7009-7013.

BIOGRAPHICAL SKETCH

Stephanie Ann Havemann was born in Akron, Ohio and raised in Beaufort, South Carolina. She attended Beaufort Academy for primary school, where she began competing in cheerleading, softball, and golf. She participated in these sports throughout her high school career at Beaufort High School, where she graduated in the top 10 of her senior class. She also served on the Science Academic Challenge Team for 3 years, and led her team to one silver and two gold medals.

She attended Mercer University in Macon, Georgia for her undergraduate career, obtaining a Bachelor of Science in Biology and another Bachelor of Science in Environmental Science in 2000. While there, she conducted a year of undergraduate research under Dr. Alan Smith characterizing the lipid transport proteins and pro-phenol oxidase of insects. Another semester of undergraduate research was performed, under the supervision of Dr. David Crowely, in attempts to identify an excision repair gene of the archaea, *Haloferax volcanii*, that was homologous to that of the *E. coli uvr*A gene. She was also the first non-engineering major at Mercer ever to participate in an engineering senior design project. Her three-person team designed and performed the initial construction of the Water Resource Monitor for the City of Macon, allowing the city to monitor the depth, temperature, and pH of the Ocmulgee River.

Her graduate career began in 2000, in the laboratory of Dr. Madeline Rasche at the University of Florida's Department of Microbiology & Cell Science. There, she devised and implemented an assay to detect the levels of methanopterin produced in various methanogenic and methylotrophic cells. She joined Dr. Steven Benner's laboratories in 2002 in the University of Florida's Department of Chemistry where she studied the incorporation of non-standard bases into DNA. Her research focused on the directed evolution of polymerases to incorporate non-

standard bases, exhibiting a C-glycosidic linkage, with efficiency and fidelity. She plans to continue her academic study as a post-doctoral research fellow.