A. ÖREROĞLU

METU 2007

IN-VIVO DIRECTED EVOLUTION OF GALACTOSE OXIDASE BY STATIONARY PHASE ADAPTIVE MUTATIONS AND PHYLOGENETIC ANALYSIS OF ERROR-PRONE POLYMERASES

AYLA ÖREROĞLU

NOVEMBER 2007

IN-VIVO DIRECTED EVOLUTION OF GALACTOSE OXIDASE BY STATIONARY PHASE ADAPTIVE MUTATIONS AND PHYLOGENETIC ANALYSIS OF ERROR-PRONE POLYMERASES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYLA ÖREROĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
FOOD ENGINEERING

NOVEMBER 2007

Approval of the thesis:

IN-VIVO DIRECTED EVOLUTION OF GALACTOSE OXIDASE BY STATIONARY PHASE ADAPTIVE MUTATIONS AND PHYLOGENETIC ANALYSIS OF ERROR-PRONE POLYMERASES

submitted by **AYLA ÖREROĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Food Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ______________________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Zümrüt B. Ögel ______________________
Head of Department, **Food Engineering**

Prof. Dr. Zümrüt B. Ögel
Supervisor, **Food Engineering Dept., METU** ______________________

Examining Committee Members:

Prof. Dr. Haluk HAMAMCI ______________________
Food Engineering Dept., METU

Prof. Dr. Zümrüt Begüm ÖGEL ______________________
Food Engineering Dept., METU

Prof. Dr. Ufuk BAKIR ______________________
Chemical Engineering Dept., METU

Prof. Dr. Alev BAYINDIRLI ______________________
Food Engineering Dept., METU

Assoc. Pr. Dr. G. Candan GÜRAKAN ______________________
Food Engineering Dept., METU

Date: _____ _ 28.11.07_______

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ayla Öreroğlu

Signature :

ABSTRACT

IN-VIVO DIRECTED EVOLUTION OF GALACTOSE OXIDASE BY STATIONARY PHASE ADAPTIVE MUTATIONS AND PHYLOGENETIC ANALYSIS OF ERROR-PRONE POLYMERASES

Öreroğlu, Ayla
M.Sc., Department of Food Engineering
Supervisor: Prof. Dr. Zümrüt Begüm Ögel

November 2007, 82 pages

In this study, the novel idea of in-vivo directed evolution was applied in order to achieve variants of the enzyme galactose oxidase with increased activity. This procedure was done under starvation conditions in *Escherichia coli* BL21 Star (DE3). Previous studies have been carried out in order to improve the activity of this enzyme using directed evolution methods. In this study, the same idea was used in-vivo, during stationary phase adaptive mutations inside the host organism, hence called in-vivo directed evolution. This method gave variants with improved enzyme activity as compared with the wild-type enzyme, and some variants showed activities that were even higher than the variants of previous directed evolution studies, hence making this method a promising approach for the random mutagenesis of genes of interest. The above mentioned mutations are carried out by a special group of polymerases, the error-prone polymerases. Phylogenetic analysis of these error-prone polymerases was also carried out in order to investigate the relationship between the number of error-prone polymerases and the level of complexity of organisms, and both the number of error-prone polymerases and the ratio of error-prone polymerases to total DNA polymerases of six organisms were studied. It was found that as the organism gets more complex, the number of error-prone polymerases and their ratio to the total polymerases increase.

Keywords: Directed Evolution, Phylogenetic Analysis, Error-Prone Polymerases, Galactose Oxidase.

ÖZ

GALAKTOZ OKSİDAZ ENZİMİN HÜCRE İÇİ YÖNLENDİRİLMİŞ EVRİMİ VE HATA EĞİLİMLİ POLİMERAZLARININ FİLOGENETİK ANALİZİ

Öreroğlu, Ayla
Yüksek Lisans, Gıda Mühendisliği Bölümü
Tez Danışmanı: Prof. Dr. Zümrüt Begüm Ögel

Kasım 2007, 82 sayfa

Bu çalışmada galaktoz oksidaz enziminin aktivitesini arttırmak amacı ile Hücre-İçi Yönlendirilmiş Evrim denenmiştir. Bu yöntem *Escherichia coli* BL21 Star (DE3) mikroorganizmasında açlık koşulunda gerçekleştirilmiştir. Daha önce yapılan çalışmalarda, bu enzimin aktivitesinin arttırmak amacı ile Yönlendirilmiş Evrim yöntemi kullanılmıştır. Bu çalışmada, aynı düşünce hücre içi olarak uygulanmıştır – durgun faz süresince konakçı organizmada adaptiv mutasyon – ve dolayısıyla Hücre-İçi Yönlendirilmiş Evrim olarak adlandırılmıştır. Bu metod kullanılarak elde edilen varyantların enzim aktiviteleri, hiçbir mutasyona uğramamış enzimin aktivitesinden daha yüksek çıkmıştır. Ayrıca, bazı varyantların aktivitesi, daha önce Yönlendirilmiş Evrim yöntemi ile elde edilen varyantların aktivitesinden de daha yüksek çıkmıştır. Buna gore, önerilen bu yeni yöntemin istenilen gende gelişigüzel mutasyonlar elde etmede ümit verici bir yöntem olduğu sonucuna varılmıştır. Yukarıda belirtilen mutasyonlar, özel bir polimeraz ailesi, hata-eğilimli polimerazlar grubu tarafından yapılmaktadır. Organizmanın kompleksitesi ve hata-eğilimli polimerazlarının sayısı arasındaki ilişkinin anlaşılması amacı ile filogenetik analiz yapılmıştır. Altı tane farklı organizmanın hata-eğilimli polimeraz sayısı ve her bir organizmada bulunan hata-eğilimli polimerazların o organizmadaki toplam DNA polimeraz sayısına oranı incelenmiştir. Yapılan bu inceleme sonucunda, organizmanın kompleksitesi arttıkça, hata-eğilimli polimerazların sayısının ve oranının arttığı bulunmuştur.

Anahtar Kelimer: Yönlendirilmiş Evrim, Filogenetik Analizi, Hata-Eğimli Polimerazlar, Galaktoz Oksidaz.

DEDICATION

Dedicated to My Family…

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Generating Novel Enzymes by Random Mutagenesis

Enzymes are increasingly being sought as alternatives to chemical catalysts, particularly as improved recombinant expression systems make them more cost effective. However, enzymes isolated from natural sources do not always fulfill the unnatural requirements demanded for industrial and biotechnological applications. Protein engineering is a relatively new phenomenon in which science is used to modify enzymes to particular specifications (1). There are two main strategies for protein engineering: *Rational Design* and *Random Mutagenesis.* Rational design is the process in which detailed information about the structure and function of a protein is used to make specific mutations to make the desired changes. However, the need for detailed information about the structure and function of a specific protein makes rational design a difficult method. Random mutagenesis, on the other hand, is the more preferred method since it involves introducing random mutations to the gene of interest and screening for the desired results.

1.1.1 Directed Evolution

1.1.1.1 General Information

Directed evolution is an example of random mutagenesis, and has allowed the generation of enzymes with greatly enhanced characteristics, and in some examples, enzymes with new and completely novel substrate specificities. Directed evolution presents a powerful means for exploring and altering enzyme substrate specificity and catalysis. In contrast to rational design, which generally concentrates on a small number of variants, evolutionary methods for protein engineering rely on the generation of vast molecular diversity by random mutagenesis and recombination, enabling the improvement of structural and functional properties, such as stability and performance under different conditions (e.g. at extreme temperatures and pH, and in organic co-solvents), or changes in their reaction and substrate specificity (1). Directed evolution is comprised of the creation of mutant libraries followed by the identification of desired variants by using a suitable screening or selection tool (2).

Directed evolution implements an iterative Darwinian optimization process, whereby the fittest variants are selected from an ensemble of random mutations. Improved variants are identified by screening or selection for the properties of interest and their encoding genes are then used as parent genes for the following round of evolution. This approach has proven particularly advantageous in cases in which prior knowledge of the protein's structure or mechanism was not available.

The major steps in a typical directed enzyme evolution experiment are as follows (Figure 1.1): the genetic diversity for evolution is created by random mutagenesis using a variety of methods based on error-prone DNA replication and chemical mutagenesis, or by homologous recombination of closely related genes using a process called family shuffling of one or more parent sequences. These altered genes are cloned back into a plasmid for expression in a suitable host organism (bacteria or yeast). Clones expressing improved enzymes are identified in a high-throughput screen, or in some cases, by selection, and the gene(s) encoding those improved enzymes are isolated and recycled to the next round of directed evolution. Screening or selecting for heterologous expression can be done in two ways: screening/selecting for the activity of a reporter protein, or screening/selecting for the protein's own function (3, 4).

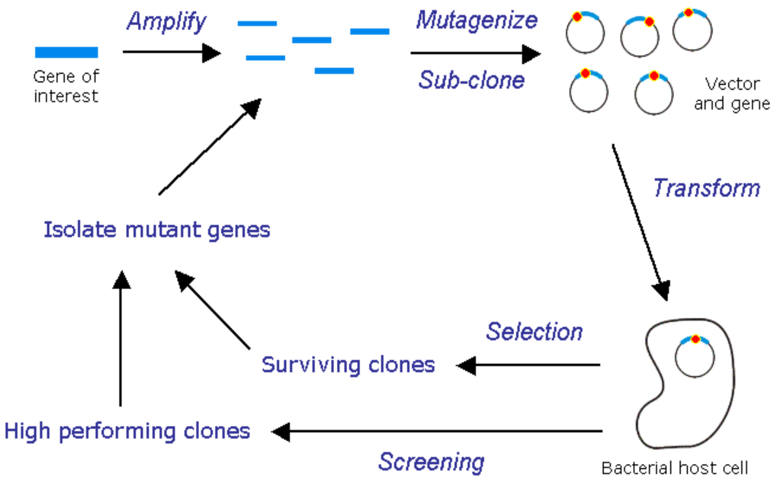


Figure 1.1 Key steps of a typical directed evolution experiment. The gene of interest is amplified and mutagenized to create a pool of mutants. These are then transformed into the host and selection/screening is carried out to identify high performing clones. The process is then repeated (1).

1.1.1.2 Directed Evolution Methods

As mentioned before, genetic diversity can be introduced by point mutagenesis, recombination or a combination of the two methods. Some established strategies for directed evolution are briefly outlined in Table 1.1.

Table 1.1 Examples of established directed evolution methods (2).

Directed evolution method	**Description**
Random point mutagenesis	Genetic diversity introduced by random point mutagenesis using error-prone PCR (EP-PCR).
Homologous recombination-based methods	
DNA shuffling	DNA shuffling methods rapidly recombine mutations that arise from point mutagenesis through a fragmentation and PCR-mediated re-assembly process.
Staggered extension process (StEP)	StEP generates recombination events through PCR amplification of a template with very short extension times, whereby partially elongated oligonucleotides are able to anneal.
Gene site saturation mutagenesis (GSSM)	GSSM uses sets of degenerate oligonucleotides to introduce all 19 amino acid substitutions at every position of the gene.
Non-homologous recombination-based methods	
Incremental truncation for the creation of hybrid enzymes (ITCHY)	ITCHY is a method for generating chimeric gene libraries by non-homologous recombination through a process of nested gene deletions and fusions.
Exon shuffling	*In vitro* exon shuffling is performed using a mixture of oligonucleotides to allow combinations of exons to be spliced together to generate mosaic proteins.
Structure-based combinatorial methods	
Structure-based combinatorial protein engineering (SCOPE)	Multiple PCR primers are designed, based on protein primary, secondary and tertiary structure, and used to create crossover gene libraries from distantly related proteins. SCOPE provides an effective means for the creation of functional gene libraries from gene families that share both low and high sequence homology.
Synthesis of rational design and directed evolution	
Computational design	Incorporation of mutations determined to be functionally relevant based on protein structure, fitness predictions of sequence, and ‘packing’ algorithms.

1.1.1.3 Advantages and Disadvantages of Directed Evolution Methods

Directed evolution has the unique advantage of adapting the biocatalyst to imposed conditions through the design and implementation of genetic selections and screens for function without requiring extensive prior knowledge of gene-to-function relationships. Another important advantage of directed evolution is that unlike rational design, prior knowledge of the protein's structure or mechanism is not needed.

The primary limitation of a majority of the directed evolution methods developed for single genes is that they cannot target the multiple, isolated genetic elements that comprise or control metabolic pathways simultaneously. Another disadvantage of this method is the time consuming and laborious step of creating gene libraries (5).

1.1.2 Error-Prone Polymerases used in Directed Evolution

As mentioned, one method of directed evolution is the use of error prone PCR to introduce mutations to the gene of interest. The enzyme used is Taq polymerase, and this enzyme has no proofreading ability. Taq polymerase is made error prone by changing the concentrations of dNTP's (increasing dTTP and dCTP, and decreasing dATP and dGTP), increasing magnesium concentrations and adding manganese to the PCR reaction mixture. The changes mentioned above cause the enzyme to work in an error-prone manner, thus introducing mutations during the PCR cycles.

Considering the above information, directed evolution aims at introducing random mutations to the gene of interest in order to develop better characteristics. However, mutations do occur naturally in cells in response to stress, such as starvation or damage, by a specific class of polymerases.

1.2 Error-Prone Polymerases and Mutagenesis in Cells

1.2.1 DNA Polymerases

Any living cell is faced with the fundamental task of keeping the genome intact in order to develop in an organized manner, to function in a complex environment, to divide at the right time, and to die when it is appropriate. To achieve this goal, efficient machinery is required to maintain the genetic information encoded in DNA during cell division, DNA repair, DNA recombination, and the bypassing of damage in DNA. DNA polymerases (pols) are the key enzymes required to maintain the integrity of the genome under all these circumstances (6). One class of DNA polymerases has gathered much attention during recent years due to its difference from the rest: error-prone Y-family polymerases. This family will be discussed in detail in the sections that follow.

1.2.1.1 Functions

DNA Polymerases function by adding nucleotide triphosphate (dNTP) residues to the 3'-end of the growing chain of DNA, using a complementary DNA chain as a template. Small RNA molecules are generally used as primers for chain elongation (7). DNA polymerases function in repeated cycles during which they move along the partly single-stranded, partly double-stranded DNA chain. In each catalytic cycle, a single nucleotide, complementary to the nucleotide on the template strand (if no error is introduced), is added to the primer strand of the double-stranded DNA. Thus, the double-stranded DNA chain with N base-pairs changes to one with N + 1 base-pairs (8). The steps of DNA replication are summarized as follows:

Initiation of Replication

- Synthesis of the first new phosphodiester bond.
- Primer synthesis and assembly of replication fork. Includes:
    - Pre-priming - steps that occur prior to primer synthesis
    - Priming - formation of first primer

Elongation

- The process of DNA synthesis.
- Replication forks moving through DNA synthesizing new DNA strands.

Termination

- When replication stops. Usually occurs when replication forks meet (9).

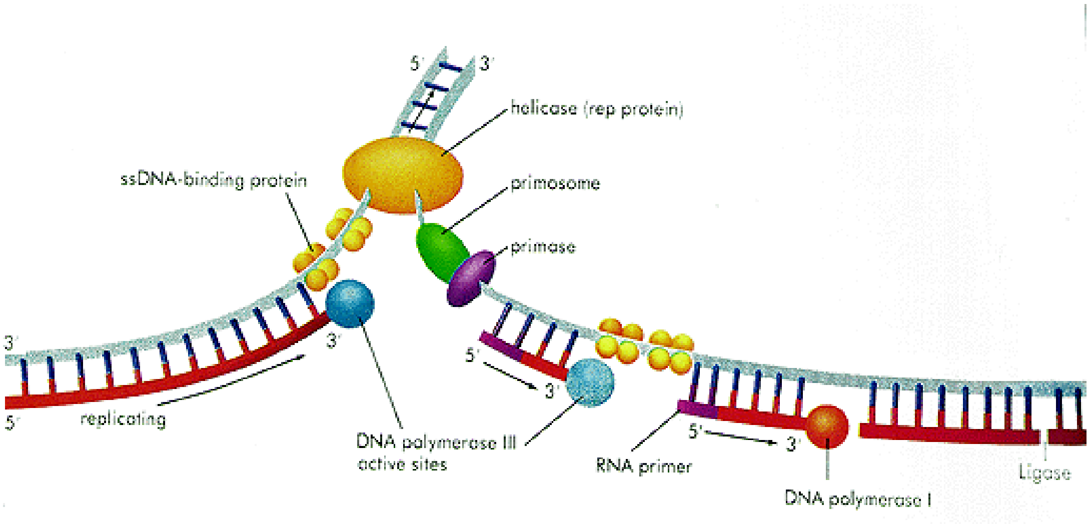


Figure 1.2 DNA Replication. The replication process is initiated by the primer, which is then continued by DNA polymerase, and finally terminated when the forks meet (10).

1.2.1.2 Structures

DNA polymerases have highly-conserved structure, which means that their overall catalytic subunits vary, on a whole, very little from species to species (Figure 1.3).

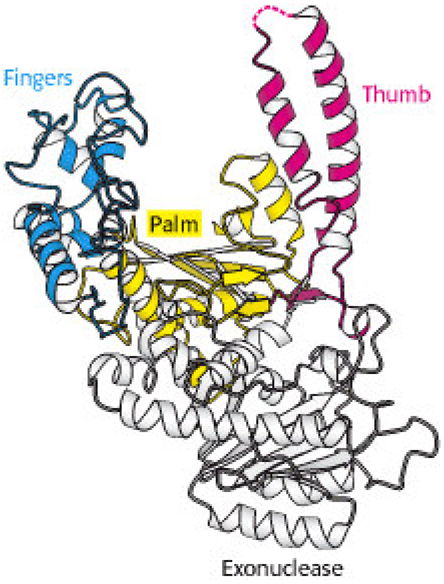


Figure 1.3 DNA Polymerase Structure. The fingers, palm, and thumb subdomains are arranged to form a deep DNA binding cleft, with the catalytic site in the palm, at the base of the cleft. The proofreading exonuclease lies adjacent to the thumb domain and extends south of the polymerase domain in this view. The position of the exonuclease domain varies with respect to the polymerase domain in different enzymes (11).

Among the DNA polymerase families, significant structural and sequence divergence is observed in finger and thumb subdomains (12). The palm domains of all DNA polymerases, with the exception of DNA polymerase β (pol β), have a common structural core, containing the two universally conserved aspartate residues which, together with the dNTP, bind the two divalent metal ions that catalyze the polymerase reaction (13).

Nucleotide incorporation by polymerases occurs through an orchestrated sequence of steps in both family A and B DNA polymerases. Binding of the incoming dNTP is associated with the polymerase adopting a 'closed conformation', characterized by closing of the nucleotide binding site and produced by the rotation of helices found in the fingers subdomain. Incorporation of each nucleotide involves precise positioning of hydrophobic and hydrophilic interactions with the dNTP, two divalent cations, the 3'-primer hydroxyl group, and Watson-Crick base pairing. Together, these events stabilize the 'closed conformation', restrict conformations and structures of the incoming nucleotide, promote the efficiency of correct nucleotide incorporation, and facilitate phosphodiester bond formation between the 3' primer hydroxyl group and the α-phosphate of the incoming dNTP. Incorporation of an incorrect nucleotide at the primer terminus in family A and B polymerases slows extension and allows time for the 3' primer terminus to partition into the 3' exonuclease site. Partitioning of DNA from the polymerase active site into the exonuclease site (separated by 30–40 Å) is accompanied by a rotation of the DNA along the helix axis and melting of several base pairs at the 3' primer terminus axis. The tip of the thumb subdomain rotates with the DNA (14).

1.2.1.3 Families

Based on primary sequence similarity, DNA polymerases can be categorized into 7 families (Table 1.2).

Different families of pols are involved in different DNA polymerization processes including not only DNA replication but also repair and recombination, a heterogeneity also reflected by varying polypeptide structures and/or subunit compositions. Some pols complement polymerase activity with 3’ → 5’ exonuclease activity (editing activity) and/or 5’ → 3’ ‘‘structure- specific endonuclease’’ activity, often located in separate structural domains on the same polypeptide chain (15).

The A, B, and C families are typified by *E. coli* DNA pol I (DNA pol I), *E. coli* DNA pol II, and *E. coli* DNA pol III α-catalytic subunit, respectively (16).

Family A includes the most abundant DNA polymerases in eubacterial cells, such as *E. coli* DNA polymerase I (Pol I). The most extensively studied polymerases include those in family A (found in prokaryotes, eukaryotes and bacteriophages) and those in family B (found in prokaryotes, eukaryotes, archea and viruses). All of the replicative DNA polymerases from eukaryotic and eubacterial cells belong to families B and C, respectively. Family X consists of eukaryotic DNA polymerase β and terminal transferases (17).

Collectively, polymerases from both families exhibit considerable sequence diversity and function during replication, recombination and repair.

Table 1.2 Classification of DNA Polymerases from the three domains of life according to their families (15).

Family	Prokaryotic	Eukaryotic	Archea
A	Pol I	Pol γ,θ	
B	Pol II	Pol α,δ,ε,ζ*	Pol BI,BII
C	Pol III		
D			Pol D
X		Pol β,λ*,μ*,Tdt*	
Y*	Pol IV,V	Pol η,ι,κ and Rev1	Dbh/Dpo4
RT		Telomerase	

* error-prone polymerases

1.2.1.3.1 A-Family Polymerases

A-family polymerases have a wide distribution in nature. The most extensively studied A-family polymerase, *E. coli* pol I, is a high-fidelity enzyme that assists in maturation of Okazaki fragments during DNA replication and also participates in gap-filling during base excision repair, nucleotide excision repair, and repair of DNA interstrand crosslinks. Recently, the DNA polymerase activity of an A-family enzyme found in mammalian cells, Pol Q (Pol θ), has been identified (18).

A structural model of the process of DNA polymerization has been generated by comparison of various binary (polymerase and DNA) and ternary (polymerase, primer-template DNA duplex, and dNTP at the polymerase active site) complexes. The initial step of DNA synthesis involves binding of a primer-template duplex DNA to the apo-polymerase, which causes the thumb to close down around the DNA. Structural studies of A-family polymerases have identified two distinct conformational states of the fingers subdomain. In this stage, (that is, in the absence of a dNTP), the fingers are in an open conformation. A dNTP then binds to this binary complex (which is now polymerase·DNA·dNTP ternary complex), inducing a conformational change in the fingers, which rotate toward the polymerase active site, moving from the open conformation adopted in the absence of dNTP to a closed, catalytically competent conformation. Similar open and closed conformations have been observed in X-family polymerases, B-family polymerases, and HIV reverse transcriptase

This conformational change forms a binding site that is sterically complementary to a Watson–Crick base pair. This fingers movement is thought to represent the slow, rate-limiting step prior to catalysis which had been detected previously in kinetic studies. Transfer of the dNTP onto the 3′ end of the primer strand ensues, followed by release of pyrophosphate and translocation of the DNA to begin the cycle again (19, 13).

1.2.1.3.2 B-Family Polymerases

Considerably less is known for the family of type B pols, which are replicative enzymes in eukaryotes and most likely also archaea. The structure of gp43 from bacteriophage RB69 provided an excellent first insight into this family.

Six regions of similarity (numbered from I to VI) are found in all or a subset of the B family polymerases. The most conserved region (I) includes a conserved tetrapeptide with two aspartate residues. Its function is not yet known, however, it has been suggested that it may be involved in binding a magnesium ion. (7) The exonuclease and palm domains share the topology and active site of A family enzymes, implying similar metal-assisted mechanisms for polymerase and exonuclease activities. The thumb and finger domains are apparently unrelated to the other polymerase families. The function of the N-terminal domain remains unknown, but may help assemble the multicomponent replication apparatus (15).

1.2.1.3.3 C-Family Polymerases

The C-family polymerase is the DNA polymerase III (pol III) from *E. coli*. This is the replicative enzyme, also called the replicase. It is responsible for the bulk of DNA replication and is a single protein of molecular weight 130 kDa. It is also referred to as polC, dnaE, or the alpha subunit. Though the molecule has DNA polymerase activity by itself, pol III works to replicate DNA in the bacterial cell in conjunction with other proteins. There are two forms of the enzyme. The core enzyme consists of only those subunits that are required for the basic underlying enzymatic activity, which are the three subunits: alpha (α), epsilon (ε) and theta (θ). It is important to note that these subunits are different from the eukaryotic polymerases in Table 1.2. The holoenzyme is the fully functional form of an enzyme, complete with all of its necessary accessory subunits, which consists of the core enzyme, the β sliding clamp and the clamp-loading complex. Totally the enzyme is a complex of 10 polypeptides. These subunits that associate with pol III in the holoenzyme perform several functions. The most interesting is the β subunit, which forms a donut shaped ring around the DNA and helps to anchor the holoenzyme to the DNA during replication. By acting as a sliding clamp, beta helps the holoenzyme to replicate long stretches of DNA without "falling off" the strand. This makes the enzyme highly processive. The holoenzyme remains associated with the fork until replication terminates. DNA polymerase III possesses high fidelity in base selection and only makes one error in approximately 10^5 base pairs. With the intrinsic exonucleolytic proofreading and postreplicative mismatch correction, the overall error rate of DNA replication is as low as approximately 10^{-10}. These high-fidelity DNA polymerases are, however, often sensitive to minor distortions in the template DNA and incoming nucleoside triphosphate (20, 21, 22).

1.2.1.3.4 D-Family Polymerases

Family D polymerases are still not very well characterized. The D family includes the euryarchaeal hetero-dimeric DNA polymerases, and are thought to be replicative polymerases (16).

1.2.1.3.5 X-Family Polymerases

DNA polymerases from the X family belong to a large protein family, including human pol β, pol λ, pol μ and terminal deoxyribonucleotidyl transferase and yeast Pol4. These enzymes have been suggested to play a role in a variety of DNA repair processes in eukaryotes. Human pol β and yeast Pol4 have been mainly involved in base excision repair. Pol4 is the only Pol X family member in yeast and is required for gap filling in some end configurations but not others. Pol λ can physically and functionally interact with the moving platform proliferating cell nuclear antigen in normal and translesion synthesis suggesting a role as a translesion pol. Finally, pol λ, as well as the related *Saccharomyces cerevisiae* Pol4, were proposed to participate in repair of double-stranded DNA breaks (16, 23, 24).

1.2.1.3.6 Y-Family Polymerases

As mentioned in the C-Family Polymerases, *E. coli* DNA polymerase III is a high fidelity polymerase that is severely blocked by DNA lesions (1) (11). Specialized repair polymerases are required to synthesize across DNA lesions in vivo (12). Bypass of DNA lesions, or "translesion synthesis," is performed by the Y family of DNA polymerases (25).

Y-family polymerases are widespread. In *E. coli*, two out of a total of five DNA polymerases belong to the Y-family, PolIV (DinB) and PolV (UmuCD′). Homologs of UmuC and DinB exist in virtually all eubacteria, and DinB homologs are found even in some archaea. The Y-family DNA pols are characterized by their low-fidelity synthesis on undamaged DNA templates and propensity to bypass normally replication-blocking lesions (19). Originally called the UmuC/DinB/Rev1/Rad30 superfamily (after the prototype of each branch of the family), proteins belonging to this superfamily are now identified as the

Y-family of DNA polymerases. Members of the UmuC subfamily are found in bacteria whereas those from the Rad30 branch are found exclusively in eukaryotes. The DinB subfamily is the most diverse and is found in bacteria, archaea and eukaryotes (26). The Y-family polymerases share five highly conserved sequence motifs within the N-terminal 350 residues, but their overall length and the C-terminus vary considerably. These motifs were identified long before the enzymatic activities were characterized, yet they do not show any significant sequence identity to any previously characterized polymerases from the A-, B-, C-, D-, and X-families. Moreover, eukaryotic homologs are often twice as large as archaeal and bacterial counterparts. The C-terminal half often contains sequence motifs for nuclear localization and for interaction with replication processivity factor (β sliding clamp and PCNA) or other polymerases. Y-family polymerases lack a $3' \rightarrow 5'$ exonuclease activity, which is an integral part of all replicative polymerases and performs a proofreading function. Interestingly, each Y-family polymerase differs in substrate specificity, that is, the type of lesions bypassed and mutation spectra generated.

All crystal structures of Y-family polymerases reported to date consist of finger, thumb and palm domains arranged in a classic “right hand-like” configuration (Figure 1.4A). The Y polymerases differ from replicative polymerases in having rather small finger and thumb domains, which leads to an open and solvent accessible active site in the palm domain. Dpo4 interacts mainly with the DNA backbone and the sugar-phosphate moiety of an incoming deoxynucleotide. There is little contact between the Dpo4 active site and the replicating base pair or preceding DNA duplex either in the major or minor groove, where a perfect or mismatched base pair may be distinguished. Watson–Crick base pairs always have a flat and smooth minor groove with a similar pattern of potential hydrogen bonds, but the minor groove of mismatched base pairs is uneven and presents varied patterns of hydrogen-bond donors and acceptors. The lack of an intimate and complementary interface between a polymerase and replicating base pair provides a structural foundation for the high-error-rate and low-fidelity DNA synthesis (Figure 1.4B).

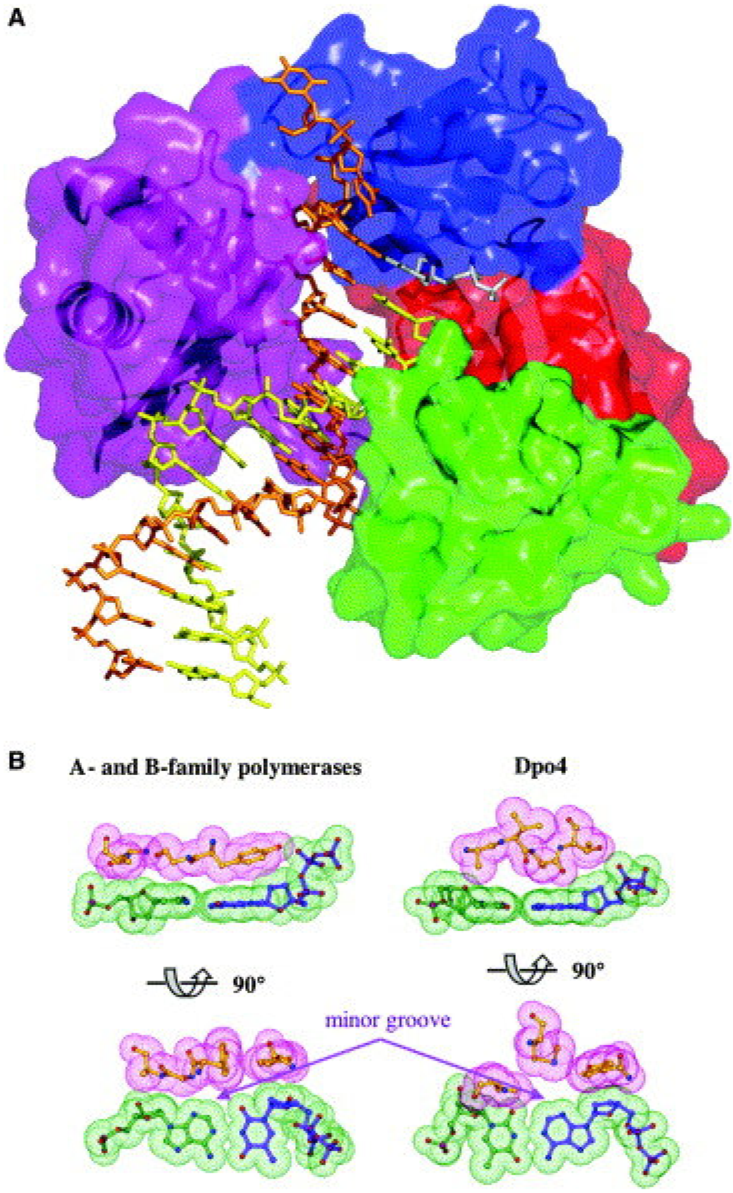


Figure 1.4 The crystal structure of Dpo4 complexed with normal DNA. (A) Dpo4 is shown in molecular surface, and the palm (red), finger (blue), thumb (green) and little finger (purple) domains are color-coded. The template and primer strand are shown as brown and yellow sticks, and the incoming nucleotide is in silver. (B) Comparison of the interface between a replicating base pair (green) and A, B, and Y polymerases (pink). The interface in a high fidelity polymerase (A and B family) is entirely complementary (left), but in Dpo4 it is imperfect and has space to accommodate a wobble base pair (right) (27).

The high fidelity of replicative DNA polymerases also depends on an “induced-fit” conformational change to discriminate against a wrong incoming nucleotide. A correct incoming deoxynucleoside triphosphate makes a Watson–Crick base pair with the templating base and induces structural rearrangement of the finger domain, which secludes the replicating base pair in a closed active site. An incorrect incoming nucleotide or damaged template base hinders this conformational change and reduces the rate of polymerization. In contrast to such an ‘induced-fit” screening of incoming nucleotides, the active site of Dpo4 is preformed regardless of whether an incoming nucleotide is incorrect or the template base is damaged or even absent. Dpo4 is committed and always ready to catalyze the nucleotidyl transfer reaction. A preformed active site is also observed with Dbh, Polι and κ (27).

Expression of Y-family polymerases is limited under normal growth conditions and is induced in the presence of DNA-damaging agents. The main functions of Y polymerases are to rescue-stalled replication forks and enhance cell survival upon DNA damage. The Y-family polymerases replicate DNA in a distributive manner and lack any intrinsic exonucleolytic activity for proofreading. In addition to bypassing damaged template or bulky DNA adducts, these polymerases replicate undamaged DNA with a 10- to 100-fold increased error rate compared to the A-, B-, C- and X-family polymerases. The most extreme example is human polι, which preferentially inserts G opposite T, rather than the canonical Watson-Crick base A, by a factor of 3- to 11-fold, depending upon the template sequence context (20). Hence, members of this family are error-prone polymerases (EP Pols).

1.2.1.3.7 RT-Family Polymerases

Finally, the reverse transcriptase family contains examples from both retroviruses and eukaryotic polymerases. The eukaryotic polymerases are usually restricted to telomerases. These polymerases use a RNA strand template to synthesize the DNA strand.

1.2.1.4 Error-Prone DNA Polymerases

1.2.1.4.1 Eukarya

As listed in Table 1, the eukaryotic EP Y-family polymerases are Pol η, Pol ι Pol κ and Rev1. Pol η and Pol ι belong to then Rad30 subfamily, Pol κ belongs to the DinB subfamily and Rev1 belongs to the Rev1 subfamily of the Y-family polymerases. Other eukaryotic EP polymerases are Polλ, Polμ and Polζ.

Human pol η is the product of the XPV gene, which is mutated in patients with xeroderma pigmentosum variant (XP – V), who have a predisposition for skin cancer. Cells from these patients are defective in the replication of DNA synthesis over damaged DNA. Even though Pol η is a limited fidelity pol, if compared to the more accurate polymerases, it is a translesion pol with a high fidelity in replicating over damaged DNA with certain types of lesions. It is interesting to note that pol η differs in the ability to bypass lesions in single cellular organism yeast and multicellular organisms.

Pol κ is the product of *DINB1* gene, and has a low fidelity of about 1:200 and performs a predominant T→G transversion mutation. It is homologous to bacterial pol IV, and creates mismatches with high frequency on undamaged DNA. As for pol η, pol κ can pass certain lesions in an error- free and others in an error-prone way. Pol κ has another unique property: on the one hand it possesses a very low fidelity, but on the other hand it is moderately processive. This property suggests and important role in spontaneous mutagenesis.

Pol ι is the gene product of *RAD30B*. In sharp contrast to pol η and to a certain extent pol κ, pol ι is a much less accurate translesion pol. It has been found that pol ι can even violate the Watson-Crick base pairing rule, since it preferentially incorporates a G instead of the correct A opposite a template T (28).

1.2.1.4.2 Bacteria

The EP polymerases in bacteria are pol IV and pol V, which belong to DinB and umuC subfamilies, respectively. The genes encoding pol V (umuC and umuD) are among the SOS genes that are induced. Pol V replicates past gaps in the DNA, and synthesis by Pol V is error-prone; there are none of the proofreading functions of Pol III. It is a relatively poor polymerase that synthesizes DNA distributively. Pol V also requires the β subunit and γ complex of Pol III for optimal activity. β functions as the sliding clamp and is required for processivity and γ is the clamp loader (22).

1.2.1.4.3 Archaea

The distribution of EP polymerases in archaea is very limited. A search of the complete genome of the crenarchaeon *Sulfolobus solfataricus* P2 revealed that it possesses a DinB homolog that has been termed DNA polymerase IV (Dpo4). On damaged DNA templates, Dpo4 can facilitate translesion replication of an abasic site, a cis-syn thymine–thymine dimer, as well as acetyl aminofluorene adducted- and cisplatinated-guanine residues (26).

1.2.2 In-vivo Mutations

1.2.2.1 The SOS Response and Starvation

The bacterial SOS response, studied extensively in *Escherichia coli*, is a global response to DNA damage in which the cell cycle is arrested and DNA repair and mutagenesis are induced (6). SOS is the prototypic cell cycle check-point control and DNA repair system. The LexA-RecA-dependent SOS system in E. coli involves more than 50 (and up to 66) genes that participate in DNA repair, recombination, and translesion DNA synthesis and whose expression is induced in response to DNA damage and arrest of DNA synthesis.

A central part of the SOS response is the de-repression of more than 20 genes under the direct and indirect transcriptional control of the LexA repressor. The LexA regulon includes recombination and repair genes recA, recN, and ruvAB, nucleotide excision repair genes uvrAB and uvrD (that, together with non-induced UvrC and SOS-induced UvrD, take part in nucleotide excision repair), the error-prone DNA polymerase (pol) genes dinB (encoding pol IV) and umuDC (encoding pol V), and DNA polymerase II (which may replicate damaged DNA and induce mutations) in addition to many functions not yet understood. In the absence of a functional SOS response, cells are sensitive to DNA damaging agents. A central role in SOS induction is exerted by RecA, which in the presence of ATP (or dATP) and single-stranded DNA acquires a co-protease activity RecA* and facilitates the proteolytic self-cleavage of the LexA protein (repressor for all the genes of the SOS regulon), thus inducing the LexA regulon. Single-stranded DNA (ssDNA) can be created by processing of DNA damage, stalled replication, and perhaps by other means. The ssDNA acts as a signal that activates an otherwise dormant co-protease activity of RecA. Furthermore, RecA* takes part in trimming UmuD to UmuD', which assists UmuC in pol V polymerase activity and also cleaves the CI repressor of l phage (29). An intriguing feature of the SOS response is inducible mutation. LexA-repressed pol V participates in most UV mutagenesis, by inserting bases across from pyrimidine dimers . Pol IV is required for an indirect mutation phenomenon in which undamaged phage λ DNA is mutated when added to UV-irradiated (SOS-induced) cells. There may be other mutagenic mechanisms induced by the SOS response (30).

Thus, Y-family polymerases are the natural error-prone polymerases that are used by the cell in order to repair, and at the same time, introduce mutations during starvation conditions. The idea that the mutations introduced during these processes are favored towards resulting in genes necessary for the cells survival has gathered much interest in recent years. This phenomena is called “Adaptive Mutation”.

1.2.3 Adaptive Mutations

Adaptive mutation (also called stationary-phase mutation) is a collection of phenomena in which mutations form in stressed or starving, non-growing, or slowly growing cells, and at least some of these mutations allow growth. It is a model for mutational escape of growth-control, such as in oncogenesis, tumor progression, and resistance to chemotherapeutic drugs, and also, like SOS mutagenesis, implies that evolution can be hastened when the need arises. Adaptive mutation has been studied most extensively using an assay for reversion of a lac +1 frameshift allele on an F' sex plasmid in *E. coli* starved on lactose medium. The adaptive mutations are unlike Lac+ mutations in growing cells in that they form during (not before) exposure to selective conditions, and occur via a unique molecular mechanism that requires homologous recombination proteins RecA, RecBC, and RuvABC . The adaptive mutations occur in a hypermutable subpopulation of the starved cells during a transient period of limiting mismatch-repair activity and possess a unique sequence spectrum of −1 deletions in mononucleotide repeats identical to that of mismatch repair defective cells.

As mentioned above the cells undergoing adaptive mutation are transiently differentiated and mutable. The role of the SOS response in adaptive mutation has been examined and it has been found that there is both positive and negative control of adaptive mutation in the Lac system by the LexA repressor. SOS induction of the LexA regulon is required for efficient adaptive mutation. RecF protein is also required for efficient mutation in its SOS-inducing capacity. This implies that the DNA signal provoking SOS during adaptive mutation is not a DNA double-strand break (DSB), and implies that there are ssDNA intermediates in mutation other than at DSBs. PsiB is a protein known to inhibit RecA* activity, and is an SOS-controlled repressor of adaptive mutation.

Studies have suggested that RecA* activity is critical in adaptive mutation, namely when RecA* activity is either too high or too low, mutation is decreased. This indicates a tight control over adaptive mutation by factors modulating the SOS response. The adaptive mutation response appears to occur within a narrow window in the continuum of levels of SOS induction (30).

The study of evolution through adaptive mutagenesis has been based on the ability of strains to acquire mutations that initiate new pathways for catabolism of a carbon source not metabolized by wild-type strains. Commonly observed adaptive mechanisms include constitutive expression of previously inducible enzymes, altered substrate specificities and improved transport of the limiting nutrient.

1.3 Directed Evolution of Galactose Oxidase

1.3.1 General Information about Galactose Oxidase

1.3.1.1 Enzyme Properties

Galactose Oxidase (GAO; E.C. 1.1.3.9), a 68 kDa mononuclear copper-containing enzyme from *Fusarium graminearum*, oxidizes primary alcohols to the corresponding aldehyde with coupled reduction of molecular oxygen to hydrogen peroxide according to the reaction scheme (5):

$$RCH_2OH + O_2 \rightarrow RCHO + H_2O_2$$

This is a two-electron reaction, but with only a single copper at the active site. The second redox active center necessary for the reaction is situated at a tyrosine residue. The crystal structure shows the presence of a novel thioether bond, covalently linking Sγ of Cys-228 and Cε of Tyr-272, with this tyrosine also acting as a ligand to the copper. The side chains of these two residues form an aromatic plane that stacks with the indole ring of Trp-290. Oxidation of Tyr-272 generates a radical, providing the second redox center in the active state of the enzyme (34). The enzyme has a wide range of substrates, but is strictly stereo-specific; whilst D-galactose is a good substrate, galactose oxidase shows no activity with L-galactose or D-glucose. Galactose oxidase is novel in that it utilizes a protein-derived tyrosine free radical to affect catalysis.

1.3.1.2 Uses in Industry

GAO is useful in a wide variety of applications, ranging from analytical and food chemistry to chemoenzymatic synthesis and clinical testing.

Biosensors:

Biological sensors based on GAO have been developed to determine the content of galactose, lactose and other GAO substrates. The biosensors have been applied to detect glucose, galactose and glutamate in human blood serum and fermented solution. (31) Such biosensors have also been used for quality control in dairy industries, online bioprocess monitoring and analysis of blood samples of patients with suspected galactosemia, human nutrition, medicine, and fermentation industry (32).

Industry:

There is interest in the use of GAO for industrial processes such as derivatization of guar gum and related polymers. Guar gum is a galactomannan isolated from Cyamopsis tetragonoloba. It comprises a 1→4-linked ß-D-mannopyranose backbone with 1→6-linked α-D-galactopyranose residues in a ratio of 1 D-galactose to ~1.5–2 D-mannose residues. Guar gum is a complex polymer with molecules that comprise ~10 000 monosaccharide residues. GAO also finds applications in food chemistry. (5) It has also been used in oxidized guar manufacture and to treat the oligosaccharide fraction contained in honey.

Medicine:

Additionally, GAO is also used for the detection of the disaccharide D-galactose-beta-(1,3)--N-acetylgalactosamine (Gal-GalNAc), a tumor marker in colonic cancer and precancer, and provides a cost-effective screening test for patients with neoplasia or at the risk of developing neoplasia.

Others:

The stereospecificity and broad substrate specificity of GAO have been exploited in the chemoenzymatic synthesis of L-sugars from polyols, which are usually difficult to prepare by chemical methods, as well as sugar-containing polyamines and 5-C-(hydroxymethyl)hexoses.

Finally, GAO is used to oxidize the cell surface polysaccharides of membrane-bound glycoproteins containing terminal non-reducing galactose residues: this is an essential step in the successful radiolabeling of these glycoconjugates (33).

1.3.2 Previous Studies on Galactose Oxidase Mutagenesis

Since GAO is a fungal enzyme, its expression in *E. coli* was low, and previous biochemical studies of GAO have been performed on the enzyme obtained from its natural source or from fungal and yeast expression systems not suitable for directed evolution. Expression of GAO (proGON) in *E.coli* has been attempted, but functional enzyme was obtained only as a *lac*Z fusion (34). Proceeding studies carried out resulted in the functional expression of GAO in *E.coli* achieved by directed evolution. These studies have been carried out to improve the activity of GAO towards appropriate substrates, using directed evolution methods based on error-prone PCR (35). The variants exhibited higher activities towards various substrates. Other studies have been performed by using directed evolution methods to increase the activity of GAO (proGOMN) in *E. coli* (36), and the results led to variants with increase in activity.

1.4 Aim of the Study

Galactose oxidase is an enzyme which has many uses in the industry. Previous studies have been carried out and the enhanced activity of galactose oxidase has been achieved using standard directed evolution methods. This study aims at increasing the activity of this enzyme using a novel method – in-vivo directed evolution – by stationary phase adaptive mutations in *E. coli*. The activity of galactose oxidase was indeed increased using this novel method, to such an extent that this approach gave results that were even better than the previous studies on this enzyme.

Considering the error-prone polymerases as the principal factor in introducing such mutations, the phylogenetic analysis of these polymerases were carried out, and the relation between the number of error-prone polymerases and the level of complexity of the organisms were investigated. It was found that as the complexity increases in the organisms (starting from simple archaea to complex eukaryotes) both the number of error-prone polymerases and their ratio to the total polymerases increase, providing an insight into the evolution of organisms.

CHAPTER 2

MATERIALS AND METHODS

2.1 Chemicals

The chemicals used throughout the experiments are listed in APPENDIX A.

2.2 Bacterial strains

The bacterial strains used are outlined in Table 2.1.

Table 2.1 Bacterial strains, genotypes and their specifications (3, 4).

Strain	Genotype	Specifications
***E. coli* XL1 Blue**	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F´ *proAB lacI*^{q}Z∆*M15* Tn*10* (Tet^r)]	Excellent host strain for routine cloning applications using plasmid or lambda vectors
***E. coli* Bl21 Star (DE3)**	*F- ompT hsdS*$_B$ *(r*$_B$ *– m*$_B$ *-) gal dcm rne131 (DE3)*	Suitable for high-level recombinant protein expression λ DE3 lysogen, carries gene for T7 RNA polymerase Requires IPTG to induce expression of the T7 RNA polymerase

Since T7 RNA polymerase synthesizes mRNA faster than E. coli RNA polymerase; transcription from the T7 promoter is not coupled to translation, leaving a pool of unprotected mRNA transcripts in the cell. These unprotected mRNA's are susceptible to enzymatic degradation by endogenous RNases, greatly reducing protein yield. BL21 Star strains contain a mutation in the gene encoding RNaseE (rne131), which is one of the primary enzymes involved with mRNA degradation in E. coli. This mutation significantly improves the stability of mRNA transcripts and thus increases protein production (4).

2.3 Growth media

The growth media used throughout the experiments are outlined below.

- 2TY broth/agar
- M9 minimal medium (1% glucose) broth
- M9 minimal medium (0.005% glucose) broth

The procedure for preparation of growth media, buffers and solutions can be found in APPENDIX B.

2.4 Gene constructs and vector

The galactose oxidase gene constructs proGON and proGOMN were kindly provided by M.J. Phearson from Leeds University, UK.

2.4.1 proGON

Contains the wildtype galactose oxidase with the coding sequence for the 17 amino acid pro-form of the enzyme and contains silent mutations which do not affect the amino acid sequence. It also contains a *Strep*-tag at the C-terminus to allow purification by affinity chromatography and a pre-pro-peptide at the N-terminus.

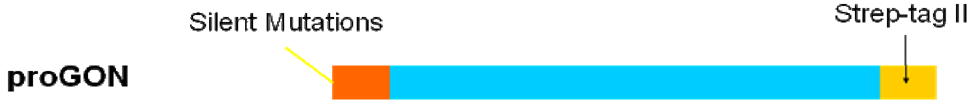


Figure 2.1 Gene construct – proGON. Silent mutations are present at the pro-region which do not affect the amino acid sequence. The *Strep*-tag at the C-terminus allows purification by affinity chromatography.

2.4.2 proGOMN

This construct is the most active variant of directed evolution studies (37), displaying a 30-fold increase in activity resulting from an 18-fold higher expression and 1.7-fold greater catalytic efficiency. This variant, like proGON, contains silent mutations but also has five additional amino acid substitutions (as a result of directed evolution). This construct also contains a *Strep*-tag at the C-terminus and a pre-pro-peptide at the N-terminus.

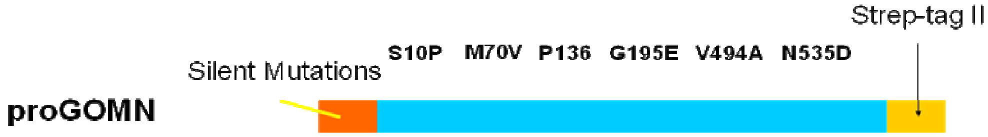
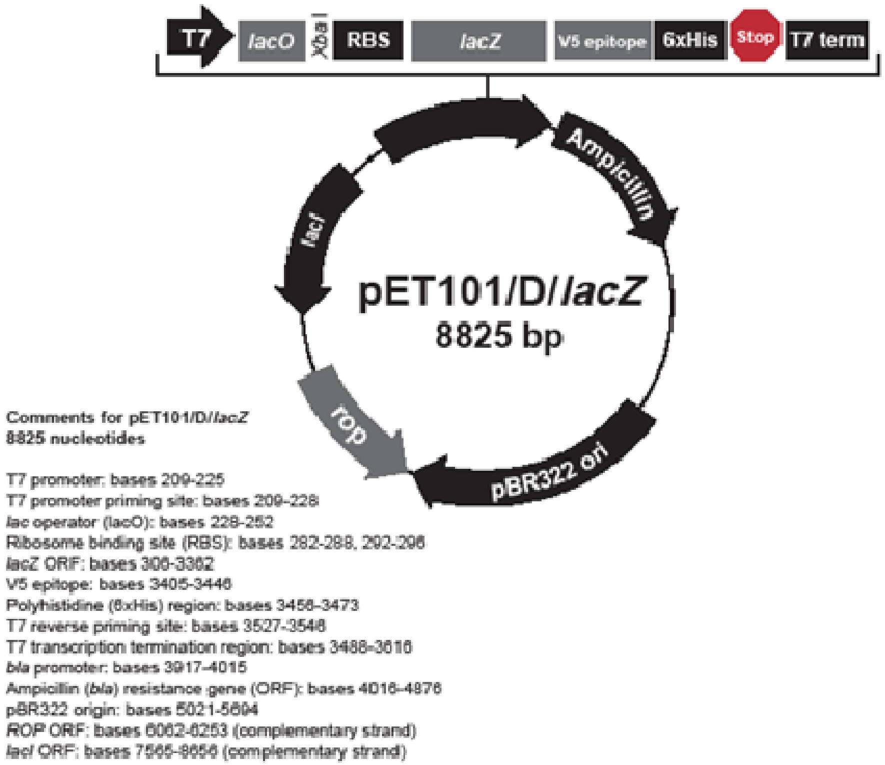


Figure 2.2 Gene construct – proGOMN. Silent mutations are present at the pro-region which do not affect the amino acid sequence. In addition, mutations carried out by directed evolution are present which have resulted in increase in GAO activity. The *Strep*-tag at the C-terminus allows purification by affinity chromatography.

2.4.3 Vector

The vector used was pET101/D/lacZ (Figure 2.3). pET System is one of the most powerful approaches available for producing recombinant proteins. This system uses the bacteriophage T7 promoter to direct the expression of target genes. Since *E. coli* RNA polymerase does not recognize the T7 promoter, there is virtually no transcription of the target gene in the absence of a source of T7 RNA (38). For protein production, a recombinant plasmid is transferred to host E. coli strains containing a chromosomal copy of the gene for T7 RNA polymerase. These hosts are lysogens of bacteriophage DE3, a lambda derivative that has the immunity region of phage 21 and carries a DNA fragment containing the lacI gene, the lacUV5 promoter, and the gene for T7 RNA polymerase (Studier and Moffatt, 1986). This fragment is inserted into the int gene, preventing DE3 from integrating into or excising from the chromosome without a helper phage. Once a DE3 lysogen is formed, the only promoter known to direct transcription of the T7 RNA polymerase gene is the lacUV5 promoter, which is inducible by isopropyl-b-D-thiogalactopyranoside (IPTG). Addition of IPTG to a growing culture of the lysogen induces T7 RNA polymerase, which in turn transcribes the target DNA in the plasmid (39).



Comments for pET101/D/*lacZ*
8825 nucleotides

T7 promoter: bases 209-225
T7 promoter priming site: bases 209-228
lac operator (lacO): bases 228-252
Ribosome binding site (RBS): bases 282-288, 292-298
lacZ ORF: bases 306-3362
V5 epitope: bases 3405-3446
Polyhistidine (6xHis) region: bases 3456-3473
T7 reverse priming site: bases 3527-3548
T7 transcription termination region: bases 3488-3616
bla promoter: bases 3917-4015
Ampicillin (*bla*) resistance gene (ORF): bases 4016-4876
pBR322 origin: bases 5021-5694
ROP ORF: bases 6062-6253 (complementary strand)
lacI ORF: bases 7565-8656 (complementary strand)

Figure 2.3 pET101D vector showing the T7 promoter, lac operator, terminator and other specifications (39).

2.5 Methods

2.5.1 Cultivation of Strains

Glycerol stocks were prepared for *E. coli* XL1 Blue and *E. coli* BL21 star (DE3) strains and also for *E. coli* XL1 Blue + proGON, *E. coli* XL1 Blue + proGOMN, *E. coli* BL21 star (DE3) + proGON and , *E. coli* BL21 star (DE3) + proGOMN. Strains were also kept on 2-TY agar plates containing antibiotics were necessary at 4°C.

2.5.2 Preparation of *E. coli* Competent Cells

1 colony E. coli was taken from the agar plate (with the necessary antibiotic) using a sterile loop, and inoculated in 5 ml 2-TY broth and incubated at 37°C overnight in orbital incubator at 200 rpm. 100 ml 2-TY was inoculated with 1 ml of this overnight culture and incubated at 37°C until the optical density at 550 nm was in the range 0.4 – 0.5. The culture was then dispensed into two 50 ml falcon tubes and chilled on ice for 10 minutes. Then the tubes were centrifuged at 6000 rpm for 5 minutes at 4°C. The supernatant was discarded and the cells were resuspended in a total volume of 25 ml ice-cold solution A each. (See APPENDIX B). The falcons were kept on ice for 15 minutes and centrifuged at 6000 rpm for 5 minutes at 4°C. The supernatant was discarded and the pellets were resuspended in a total volume of 3.5 ml ice-cold solution B each. (See APPENDIX B). 300 µl aliquots were prepared in sterile eppendorfs and stored at -80°C.

2.5.3 Transformation of *E. coli* Competent Cells

A 100 µ aliquot of competent cells was thawed on ice. 1 µl plasmid was mixed with 50µl TE buffer, added to competent cells and thawed on ice for 30 minutes. (For negative control ddH$_2$O was used instead of plasmid). The cells were then kept in water bath at 42°C for 90 seconds and transferred onto ice for 2 minutes. 500 µl preheated 2-TY broth was added and incubated at 37°C for 1 hour in orbital incubator at 200 rpm. 100 µl was pour plated on 2-TY agar containing the necessary antibiotic, and incubated at 37°C overnight. The negative controls, not having any plasmid, would be unable to grow in the antibiotic medium, and this would be a proof that there was no contamination.

2.5.4 Plasmid Isolation using Alkaline Lysis Procedure

5 ml overnight culture of bacterial cells carrying the plasmid of interest was centrifuged at 6500 rpm for 5 minutes at 4°C. The supernatant was discarded and the cells were resuspended in 200 µl solution I (See APPENDIX B) and incubated at room temperature for 15 minutes. 200 µl solution II was then added and the mixture was inverted 7-8 times and incubated on ice for *precisely* 5 minutes. Solution III was then added and mixed by gently inverting 7-8 times, and incubated on ice for 15 minutes. The mixture was then centrifuged at 13000 rpm for 10 minutes at 4°C. The supernatant was transferred into a new 1.5 ml microcentrifuge tube, and 2 volumes of cold ethanol (96%) was added. The mixture was then kept at -80°C for 1 hour. Next, the mixture was again centrifuged at 13000 rpm for 10 minutes at 4°C, and the supernatant was discarded. The pellets were resuspended in 200 µl NE buffer and incubated on ice for 1 hour. 5 µl plasmid was loaded unto agarose gel in order to detect plasmid DNA at this step. The suspension was centrifuged at 13000 rpm for 15 minutes at 4°C. The supernatant was transferred into a new tube, and 2 volumes of cold ethanol (96%) was added. The mixture was then kept at -20°C for 30 minutes. Next the mixture was centrifuged at 13000 rpm for 10 minutes at 4°C. The supernatant was discarded and the pellets were vacuum-dried for 2 minutes and resuspended in 20 µl ddH$_2$O and stored at -20°C.

2.5.5 Agarose Gel Electrophoresis

Inorder to see the plasmids obtained by plasmid purification, 1% agarose was dissolved in 1 X TAE buffer (See APPENDIX B) and boiled for about 3 minutes in microwave. The solution was cooled to about 50°C, and ethidium bromide was added to a final concentration of 0.5 μ/ml. This is done to stain the plasmids, so that they become visible under UV light. The gel was then poured into a mould and a comb was placed in the mould and left for solidification of the gel as it cools down. Next, the mould was placed into the electrophoresis tank containing 1 X TAE buffer, and the combs were removed. A mixture of 4 μl plasmid 2 μl dye was loaded in each well. In order to have an idea about the size of the plasmids, a mixture of 2 μl λ *Hind* III marker and 2 μl dye was also loaded in a well. The lid of the tank was placed and the power turned on at 100 V for about 1 hour. The plasmids were then observed under the UV light at 320 nm, and photographed.

2.5.6 Determination of Biomass

The determination of biomasswas done by plotting the growth curve of *E. coli* BL21 Star (DE3) as follows:

1 colony fresh *E. coli* BL21 Star (DE3) was inoculated in 5 ml M9 medium containing 1% glucose, and incubated at 37°C for 16 hrs at 200 rpm. 1 ml of this culture was then added to 100 ml M9 medium containing 1% glucose, and incubated at 37°C and 200 rpm, and the optical density was measured at 600 nm at 30 min intervals. A replica was also done and the average values were used to draw the growth curve. The growth curve was prepared by plotting the values of optical density at 600 nm versus time.

2.5.7 Determination of Glucose Concentration

The measurement of glucose concentration was done from the same samples of biomass experiment, except that the biomass experiment was done every 30 minutes while glucose concentration experiment was done every hour, for a total of 24 hours. 1 ml glucose reagent was added to 200 μl cell suspension and waited for 10 minutes, during which the color of the medium changed to pink, and measurement of optical destiny was done at 500 nm. A standard curve was also prepared by measuring optical densities for a series of solutions with known glucose concentrations, and the graph of glucose concentration versus optical density was drawn. The equation of this graph was used to calculate the concentration of glucose in each sample with known optical density. The graph of glucose consumption was finally prepared by plotting the values of glucose concentration versus time.

2.5.8 Starvation Procedure

The starvation methods were classified according to the medium: liquid (broth) and solid (agar). The methods in which liquid media were used are designated as **L1** and **L2**, and those in which Solid media were used are designated as **S1** and **S2**.

2.5.8.1 Liquid Medium

The starvation procedure in broth is as follows:

- 1 colony *E.coli BL21 Star (DE3)* + proGON was inoculated in 5 ml M9 (1% glucose) medium containing 75mg/ml ampicillin (M9 1%, 75amp), and incubated at 37°C at 200rpm for 16 hours overnight (O/N).
- 1 ml was taken and added to 100 ml M9 1%, 75amp, and incubated at 37°C at 200rpm for 12 hours (to reach stationary phase).
- After this step, the procedure divides to two (**L1** and **L2**)
- For **L1,** the cells were recollected by centrifuging at 6000 rpm for 5 minutes
- Recollected cells were resuspended in 100 ml M9 0.05%, 75amp, which was the starvation medium, designated as **L1**
- For **L2**, the cells were not recollected after reaching stationary phase, but kept at 37°C and 200rpm for 3 days in the same medium with the supplement of extra ampicillin per day to ensure plasmid stability. During

this period, the medium would get depleted of nutrients and starvation would be reached during the experiment.

- At the end of each day, 2x5ml samples were taken for plasmid purification, and the original medium was refreshed with ampicillin.
- This procedure was continued until day 3.

2.5.8.2 Solid Medium

The starvation procedure in agar is as follows:

- 1 colony *E.coli BL21 Star (DE3)* + proGON was streaked unto M9 (1% glucose) agar containing 75mg/ml ampicillin (M9 1%, 75amp agar), and incubated at 37°C for 16 hours overnight (O/N).
- For **S1**, 1 colony of this plate was then streaked unto M9 0.05%, 75amp agar, which was the starvation agar medium.
- For **S2**, the same plate was incubated for 3 days (but would get depleted on nutrients during this period).
- This procedure was done for five other plates each (resulting is a total of 6 plates M0 and 6 plates M1), and incubated at 37°C for 3 days.
- At the end of each day, 2 plates were taken for plasmid purification.
- This procedure was continued until day 3.

2.5.9 Enzyme Assay Procedure

In order to induce the production of GAO by the cell, IPTG was used (See 2.4.3 Vector). This was done by adding 0.1M IPTG to the growth medium and incubating overnight at 37°C. After the induction of GAO gene, the production of this enzyme was assayed as follows:

- A sterile microtiter plate was used and 80µl sterile NaPi buffer (100mM, pH 7.0) is poured into each well.
- Each colony was taken with a sterile loop and suspended in one microtiter well, and the wells were numbered.
- The loop was then streaked unto the corresponding numbered region of a fresh 2TY 100amp agar. This was done in order to have a replica of the cells that are being assayed, such that the good ones could be used again for further analysis.
- After taking all the colonies, the microtiter plate was closed with its lid and parafilmed.
- The plate was then subjected to liquid nitrogen for 2 minutes (in order to burst the cells) (40).
- The plate was thawed at room temperature.
- CuSO4 was added to a concentration of 50 µM.

200µl galactose oxidase assay solution (APPENDIX B) was added to each well using a micropipette, and the plate was observed for the appearance of green color.

Galactose oxidase solution contains horse-radish peroxidase enzyme (HRP) and galactose. When this solution is added to the sample, if GAO is present, it acts on galactose and converts it into D-galacto-hexodialdos and hydrogen peroxide. HRP then acts on ABTS in the presence of hydrogen peroxide and results in the production of green color. The intensity of green color reflects the amount of hydrogen peroxide, and consequently, the amount on GAO present.

2.5.10 Phylogenetic Methods

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history (their phylogeny). The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree. Traditionally, phylogenies have been constructed from morphological data but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as molecular phylogeny. The data is most commonly in the form of DNA or protein sequences (41).

Tree building methods can be categorized by two ways: by how they handle *data* or by the *approach* taken when building trees. The classification of tree building methods is depicted in Table 2.2.

2.5.10.1 Data

Tree building methods can be divided based on how the data is treated, this being distance or discrete. *Distance methods* first convert aligned sequences into a pairwise distance matrix, then input that matrix into a tree building method, whereas *discrete methods* consider each nucleotide site (or some function of each site) directly and attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids) (42).

2.5.10.2 Approach

Another way of dividing tree building methods is by the way they construct trees. *Cluster methods* follow a set of steps (an algorithm) and arrive at a tree. *Optimality criteria*, however, chooses from among the set of all possible trees. This method defines an optimality criterion for comparing alternative phylogenies to one another and deciding which one is better (43). This criterion is used to assign to each tree a “score” or rank which is a function of the relationship between tree and data.

Table 2.2 Classification of tree building methods based on the type of data used and the method of calculation (43).

		Type of data	
		distances	***nucleotide sites***
Tree building method	***clustering algorithm***	UPGMA Neighbour joining	
	optimality criterion	Minimum evolution	Maximum parsimony Maximum likelihood

2.5.10.3 Neighbour joining method

The neighbor-joining method is a distance based method for constructing evolutionary trees.

It is a greedy algorithm which attempts to minimize the sum of all branch-lengths on the constructed phylogenetic tree. Conceptually, it starts out with a star-formed tree where each leaf corresponds to a species, and iteratively picks two nodes adjacent to the root and joins them by inserting a new node between the root and the two selected nodes. When joining nodes, the method selects the pair of nodes i, j that minimizes the branch-length sum of the resulting new tree (44). In other words, this method constructs trees by linking the least distant pairs of nodes as defined by a modified matrix. This method is generally considered to be fairly good and is widely used (41).

2.5.10.4 UPGMA method

A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA). The algorithm assumes that the distance data has the so-called molecular clock property (the divergence of sequences occur at the same constant rate at all parts of the tree). This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure (41). This method is simple under certain circumstances, but the assumption of molecular clock hypothesis makes it unsuitable for this study.

2.5.10.5 Maximum Parsimony Method

Maximum parsimony is a character-based method (discrete method) that infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data, or in other words by minimizing the total tree length (45). This is done by giving each a criterion or score that is used to choose between different trees. Parsimony implies that simpler hypotheses are preferable to more complicated ones (46). This method is very time consuming, even on very fast computers.

2.5.10.6 Maximum Likelihood Method

This method uses explicit models of molecular evolution and allows for rigorous statistical inference. However, it is very computer intensive. A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference then consists of finding the tree which assigns the highest probability (likelihood) to the data (41). This method appears to estimate the actual amount of change according to the evolutionary model in place. It works with a prior nucleotide substitution model to compute a likelihood score for each tree given the original data.

2.5.10.7 Minimum Evolution Method

In minimum evolution method the data are transformed to pair-wise distances, and the score is calculated from those. The best tree is the tree with the lowest score (i.e. shortest length) (47).

CHAPTER 3

RESULTS AND DISCUSSION

Bacteria produce error-prone polymerases under stress conditions in order to undergo adaptive mutations. The strategy in this study was to use starvation stress and their corresponding in-vivo events to produce mutations in the gene of interest as an alternative method for directed evolution. Due to its nature, this novel approach was named as “in-vivo directed evolution”. Such a strategy is expected to generate enzymes with improved/novel properties. Furthermore, the host in which mutations take place may generate mutations such that the mutated form of the gene is better expressed in this heterologous host. This is expected since mutations will not be completely random, in the sense that they will be generated by the host itself. If this indeed turns out to be the case, it is likely to be of value in developing efficient expression systems for eukaryotic genes in simple organisms, such as *E. coli*.

3.1 Experimental Strategy

3.1.1 General Strategy

The general strategy is as follows:

Transform plasmid with gene of interest → Keep under starvation → Expected mutations by EP Pols → Recollect plasmids → Re-transformation and screening

In this study galactose oxidase gene was used as a model system. The cloning of this gene had been done previously on a suitable *E. coli* vector (see Section 2.4). In later studies, this gene was subjected to directed evolution methods to express this enzyme in functional form in *E. coli.* The evolved enzymes retained the activity and substrate specificity of the native fungal oxidase, but were more thermostable, and were expressed at a much higher level. (36)

The unmutated pET101D-proGON vector (from now on referred to as proGON) was first transformed into competent *E.coli XL1 Blue* cells for enrichment, followed by plasmid purification. The purified plasmid was then transformed into *E.coli BL21 Star (DE3)* for starvation experiment. *E.coli BL21 Star (DE3)* host was selected since is contains the gene for T7 RNA polymerase in the DE3 lysogen, and the only promoter known to direct transcription of the T7 RNA polymerase gene is the lacUV5 promoter, which is inducible by isopropyl-b-D-thiogalactopyranoside (IPTG). Addition of IPTG to a growing culture of the lysogen induces T7 RNA polymerase, which in turn transcribes the target DNA in the plasmid. (See Section 2.4.3)

Plasmids were purified at days 1, 2 and 3 of starvation, and were then retransformed into *E.coli BL21 Star (DE3)* to induce the expression of GAO using IPTG, and to analyse the activity of the enzyme.

The experimental strategy can be classified based on the media used, that is, broth (liquid) and agar (solid). (See Section 2.5.8)

3.1.2 Experiments in Liquid Medium

Experiments in liquid media are named as procedures **L1** and **L2**. Figure 3.1 depicts the strategy used for **L1**, and Figure 3.2 depicts the strategy used for **L2**.

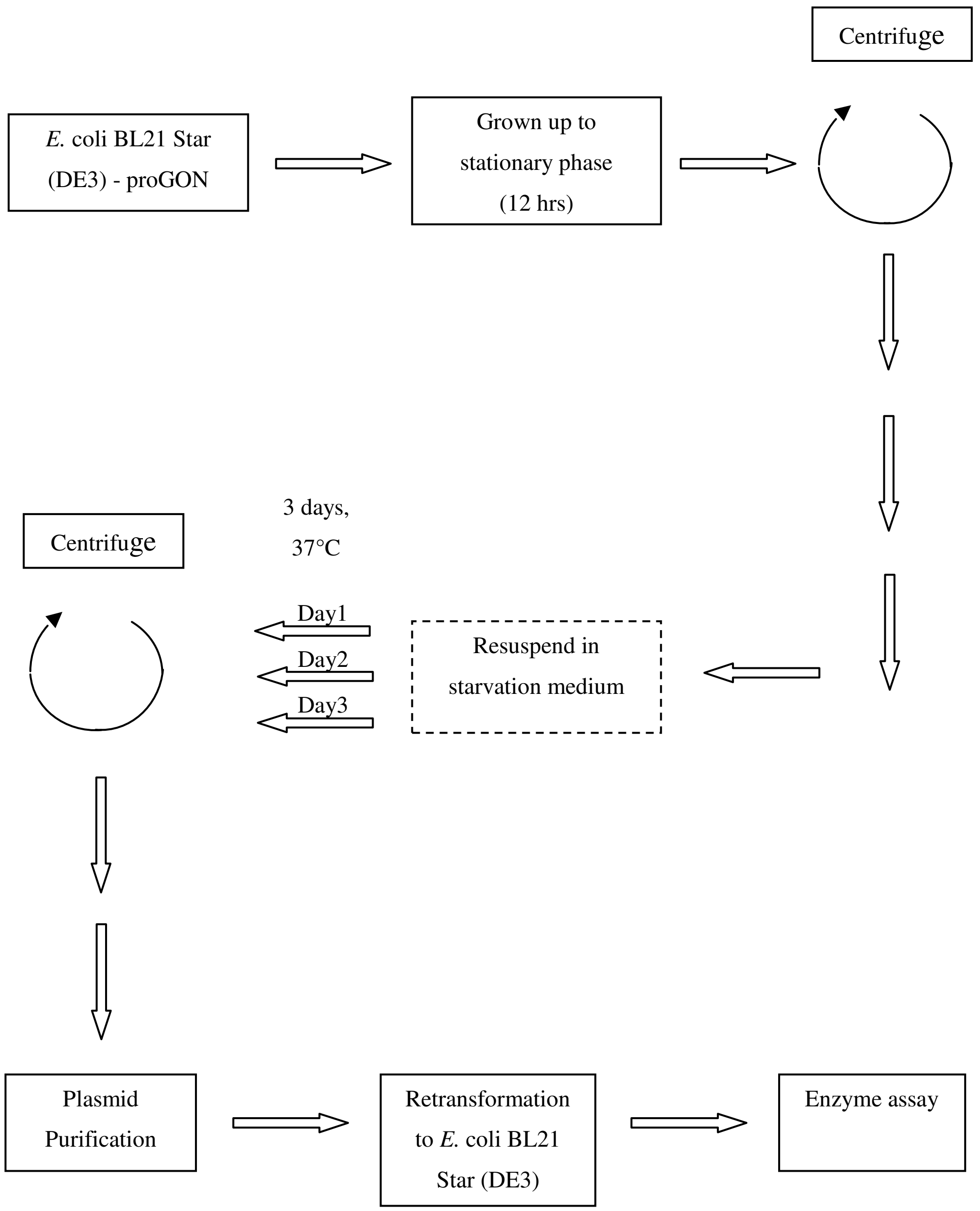

E. coli BL21 Star (DE3) - proGON
Grown up to stationary phase (12 hrs)
Centrifuge
Resuspend in starvation medium
3 days, 37°C
Day1
Day2
Day3
Centrifuge
Plasmid Purification
Retransformation to E. coli BL21 Star (DE3)
Enzyme assay

Figure 3.1 Experimental Strategy for **L1** (See Section 2.5.8.1)

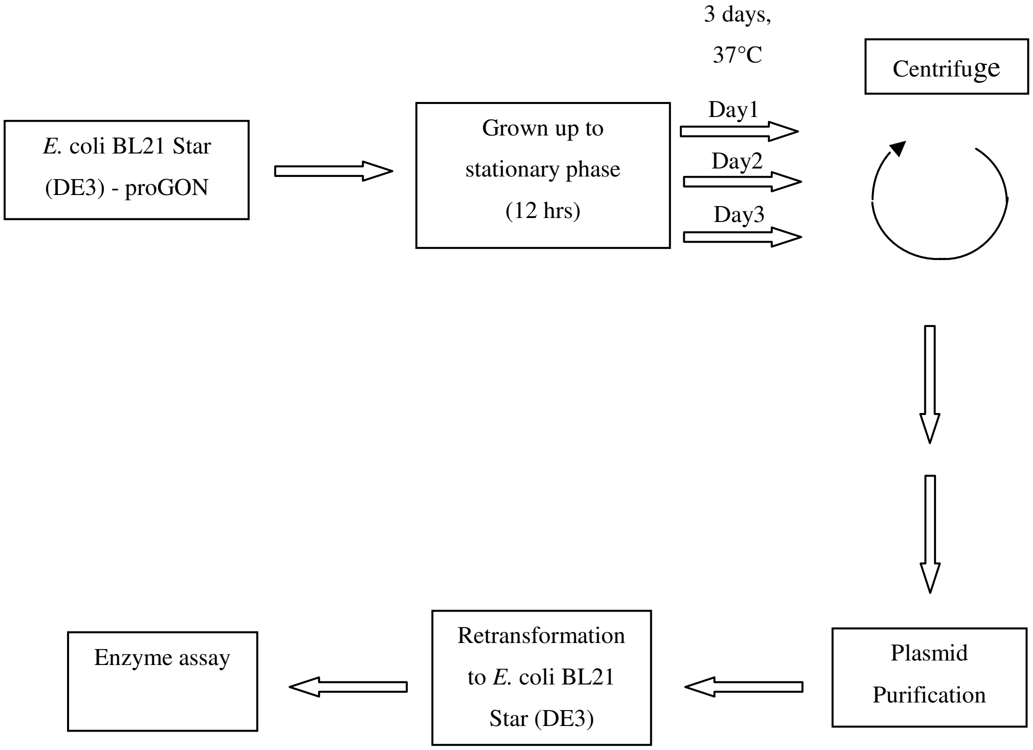

3 days,
37°C
E. coli BL21 Star (DE3) - proGON
Grown up to stationary phase (12 hrs)
Day1
Day2
Day3
Centrifuge
Plasmid Purification
Retransformation to E. coli BL21 Star (DE3)
Enzyme assay

Figure 3.2 Experimental Strategy for **L2** (See Section 2.5.8.1)

3.1.3 Experiments on Solid Medium

Experiments in agar media are named as procedures **S1** and **S2**. Figure 3.3 depicts the strategy used for **S1**, and Figure 3.4 depicts the strategy used for **S2**.

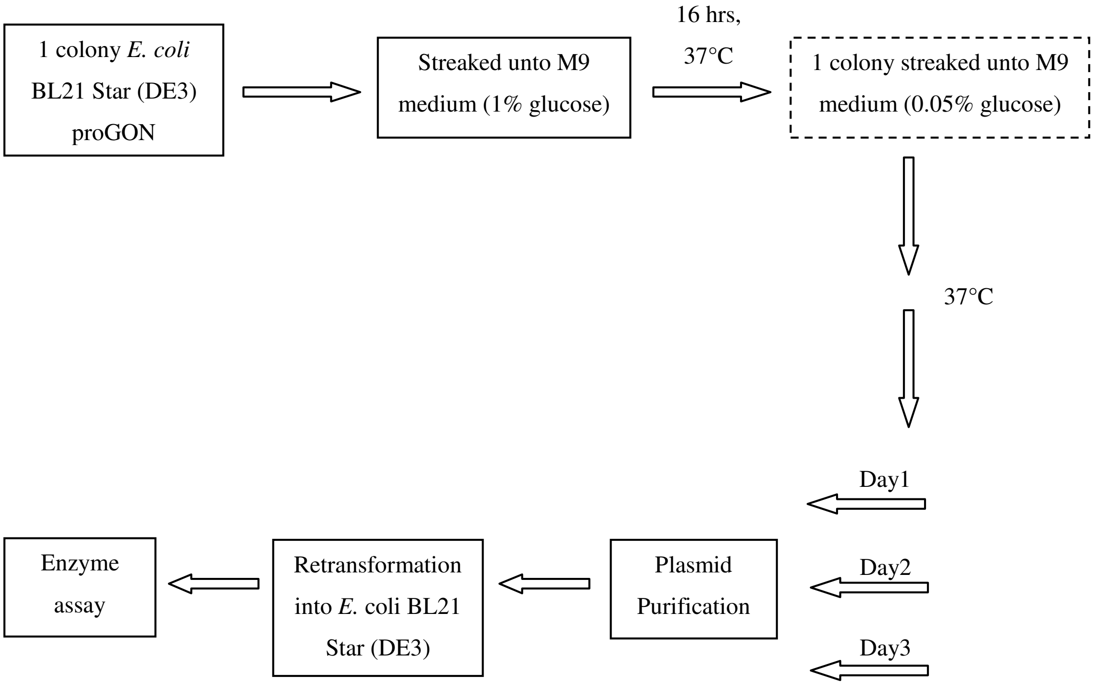

1 colony *E. coli* BL21 Star (DE3) proGON
Streaked unto M9 medium (1% glucose)
16 hrs, 37°C
1 colony streaked unto M9 medium (0.05% glucose)
37°C
Day1
Day2
Day3
Plasmid Purification
Retransformation into *E.* coli BL21 Star (DE3)
Enzyme assay

Figure 3.3 Experimental Strategy for **S1** (See Section 2.5.8.2)

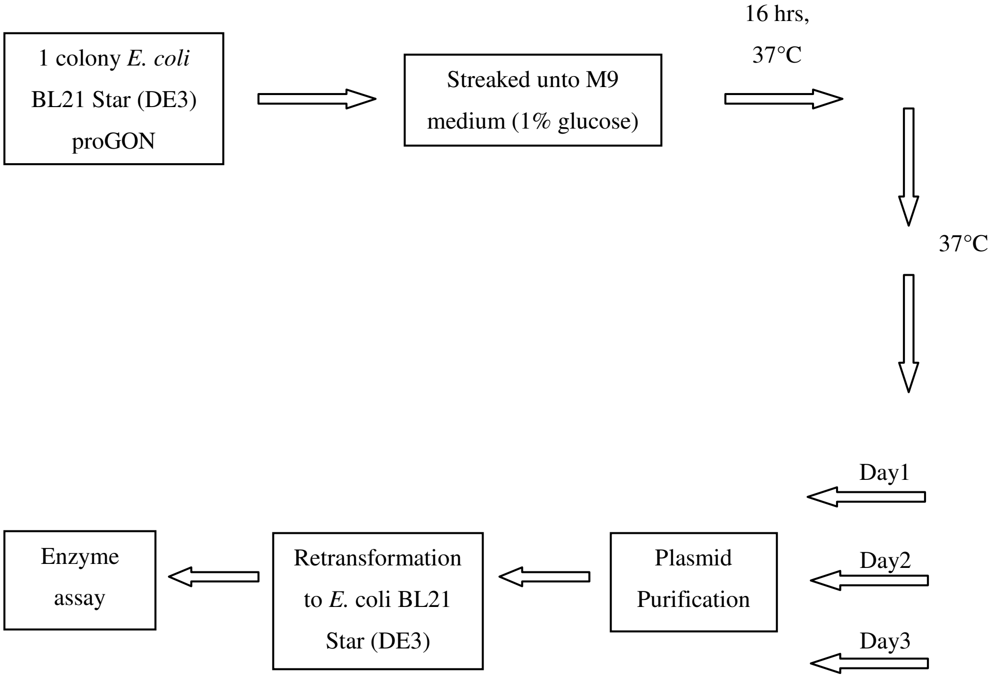

1 colony *E. coli* BL21 Star (DE3) proGON
Streaked unto M9 medium (1% glucose)
16 hrs, 37°C
37°C
Day1
Day2
Day3
Plasmid Purification
Retransformation to *E.* coli BL21 Star (DE3)
Enzyme assay

Figure 3.4 Experimental Strategy for **S2** (See Section 2.5.8.2)

3.2 Determination of the Time Course for Cells to Reach Stationary Phase

As mentioned in Section 1.2.2, mutations occur during the stationary phase. Consequently, the experiments must be performed with stationary phase cells. A growth curve of *E. coli* BL21 Star (DE3) in M9 medium (1% glucose) was prepared in order to find the onset of the stationary phase. (See Section 2.5.6)

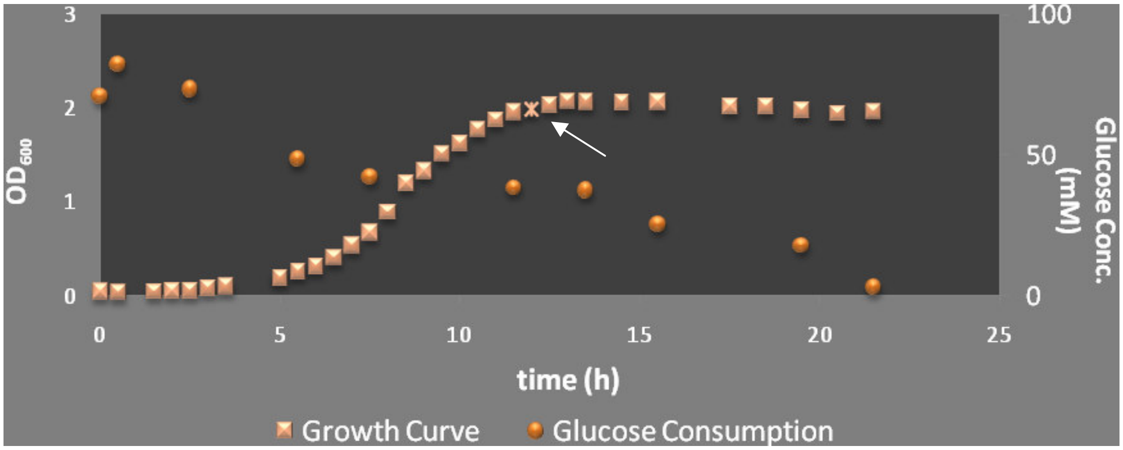


Figure 3.5 Growth of *E. coli* BL21 Star (DE3) cells showing the onset of stationary phase. The amount of glucose consumption is also depicted. (See Sections 2.5.6 and 2.5.7)

As depicted in the graph, the onset of stationary phase was at about 12 hours (shown by an arrow), which is used for the starvation procedure. Glucose consumption curve was also drawn in order to have an idea about the amount of carbon supply remaining during the 24 hour interval.

3.3 Preparation of Plasmids for In-vivo Directed Evolution

After the preparation of *E.coli XL1 Blue* competent cells (See Section 2.5.2), the pET101D-proGON and pET101D-proGOMN plasmids (proGON and proGOMN) were transformed into competent *E.coli XL1 Blue* cells. The plasmids were then purified after growth (Figure 3.6).

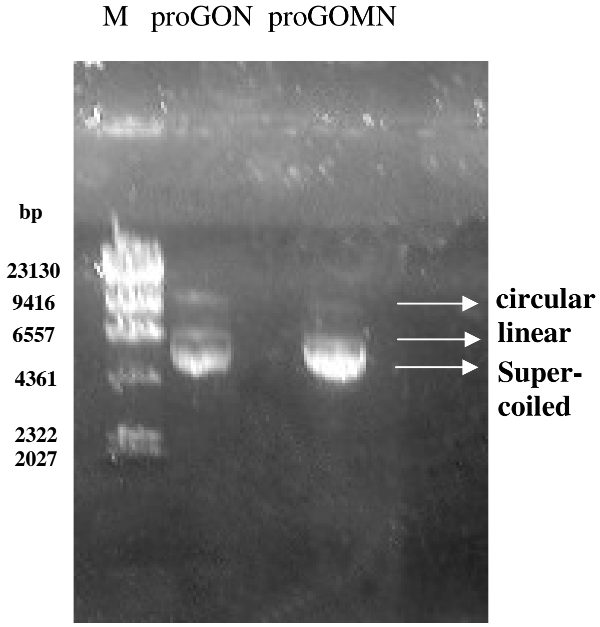


Figure 3.6 Purification of proGON and proGOMN from *E.coli* XL1 Blue cells. The supercoiled form of the plasmids are predominant. The plasmid size is 7801 bp, which corresponds to the linear form of the plasmids.

As shown in Figure 3.6, both proGON and proGOMN plasmids gave three bands in gel electrophoresis. As can be seen from the figure, the third bands of both proGON and proGOMN are brighter than the other two bands, meaning that the super-coiled form of the plasmids is more abundant than the linear and circular forms. The actual size of the plasmids corresponds to the linear form, and is 7801 bp.

3.4 *E. coli* Starvation and Analysis of plasmids from starved cells

3.4.1 Day 1

At the end of Day1, plasmids were purified and analysed by gel electrophoresis. (See Sections 2.5.4 and 2.5.5).

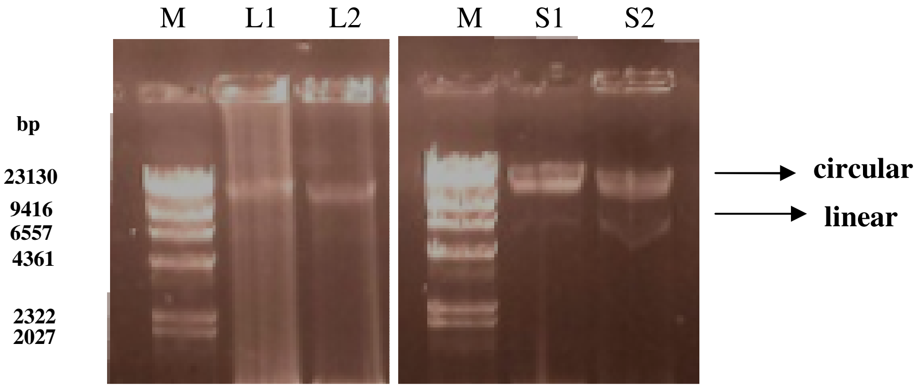


Figure 3.7 Plasmid purification - Day 1. Some changes have occurred after the first day of starvation. Smears have appeared on the gel for samples **L1** and **L2.** The open circular forms of the plasmids seem to become more predominant after the first day of starvation.

As shown in Figure 3.7, the plasmids isolated after the first day of starvation seem to look slightly different from those isolated from cells grown under normal conditions (Figure 3.6). The first band, that is the circular form of the plasmids, seems to be predominant. The second band (linear form) seen in Figure 3.6 is slightly observable, and the third band is not present. The size of the linear form is again about 7800 bp. Moreover, the bands of **S1** and **S2** are brighter and did not generate smears as compared to **L1** and **L2**.

3.4.2 Day 2

At the end of Day2, plasmids were purified and analysed by gel electrophoresis as follows (Figure 3.8).

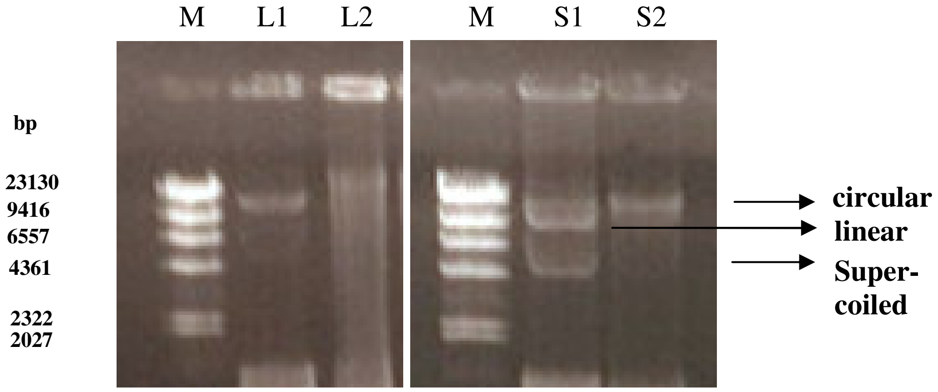


Figure 3.8 Plasmid purification - Day 2. The same trend continues for day 2; smears are still observable and the circular form is predominant.

As can be seen from the figure, the results of solid medium seem to be better than the liquid medium, similar to Day1. This can be due to the fact that cell lysis is more probable in liquid medium as compared to solid medium (due to osmotic pressure that is exerted on the cells in liquid medium). Cell lysis can be the reason for the smears observed in plasmids from liquid medium **L2**. Moreover, the linear form of the plasmid is observable in **S1**, The expected size of plasmids are 7801 bp, and the actual size calculated from Figure 3.8 is the same.

3.4.3 Day 3

The results of plasmid purification of Day3 are as follows (Figure 3.9).

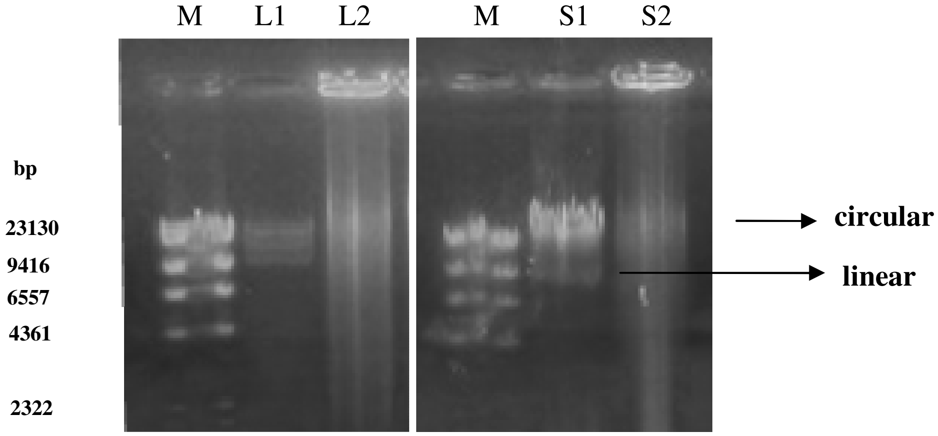


Figure 3.9 Plasmid purification - Day 3. The presence of smears is highest in day 3, indicating plasmid degradation to some extent.

The figure above shows more smears than day 1 and day2, this time in both **L2** and **S2**. The increase in the appearance of smears in the third day further supports the explanation of smears as the result of cell lysis, since cell lysis increases as time passes. This is followed by degradation of plasmids, giving result in smears.

Moreover, 3 bands close to each other appear for **L1**, and it seems like the cells of **L1** at Day 3 have undergone plasmid degradation because of cell lysis.

3.5 Re-transformation of plasmids and expression of GAO mutants

After collecting the plasmids of starvation experiment pertaining to days 1, 2 and 3, these plasmids were transformed into *E.coli BL21 Star (DE3)* and IPTG was used to induce the expression of the galactose oxidase gene (See Sections 2.2 and 2.4).

In order to detect the GAO activity of mutant cells, a new method was developed. The idea was to analyse single colony forming units (cfus) separately for enhanced GAO activity. This would not be possible if the cells were grown in broth medium, since then cfus could not be separated. As a result, mutant cells were induced and grown on agar plates, and the GAO assay was performed on microtiter plates as described in Section 2.5.9.

GAO assay was used to measure the amount of enzyme activity in mutant cells. However, the activity of these cells must be compared with a control in order to have an idea about the extent of change in activity compared with the wildtype gene. For this aim, the wildtype GAO gene (*E. coli* BL21 Star (DE3) + proGON) and the GAO gene with mutations to enhance its enzymatic activity (*E. coli* BL21 Star (DE3) + proGOMN) were both used as controls.

The results of the enzymatic assays performed for controls are shown below (Figure 3.10).

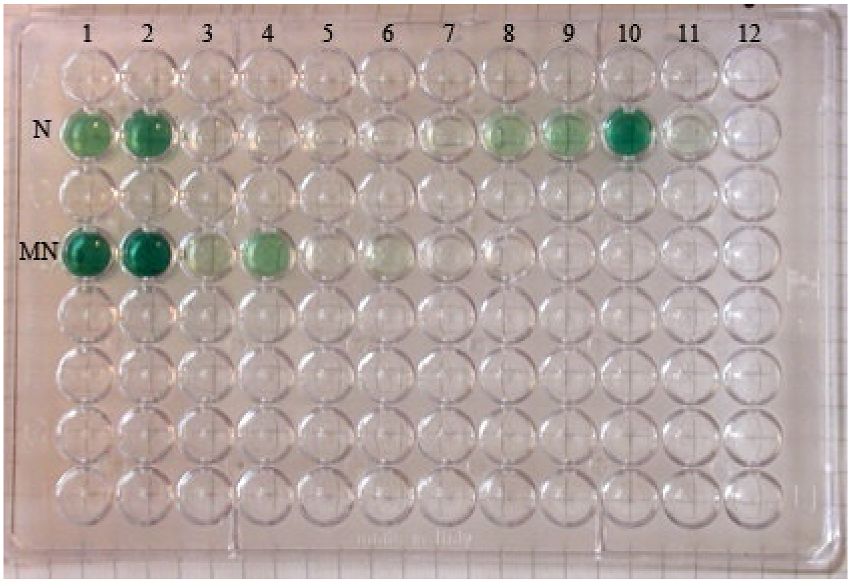


Figure 3.10 Enzyme Assay – Controls. The microtiter plate is numbered from 1 to 12 indicating the number of colonies used for the assay. Each well contains a single colony, which is also replicated on numbered agar plate.

The microtiter plates were numbered from 1 to 12, and labelled as **N** and **MN**, corresponding to wildtype GAO cells (*E. coli* BL21 Star (DE3) + proGON) and directed evolution mutated GAO cells (*E. coli* BL21 Star (DE3) + proGOMN), respectivley.

The results of the enzymatic assays performed for variants are shown in Figure 3.11. The variants are numbered according to the day of the experiment (Day1, Day2 or Day3) and the four types of experiments performed (**L1**, **L2**, **S1** and **S2**). (See Section 2.5.8). Hence a variant obtained from Day2 of **L2**, for instance, is labelled as "**Day2-L2**".

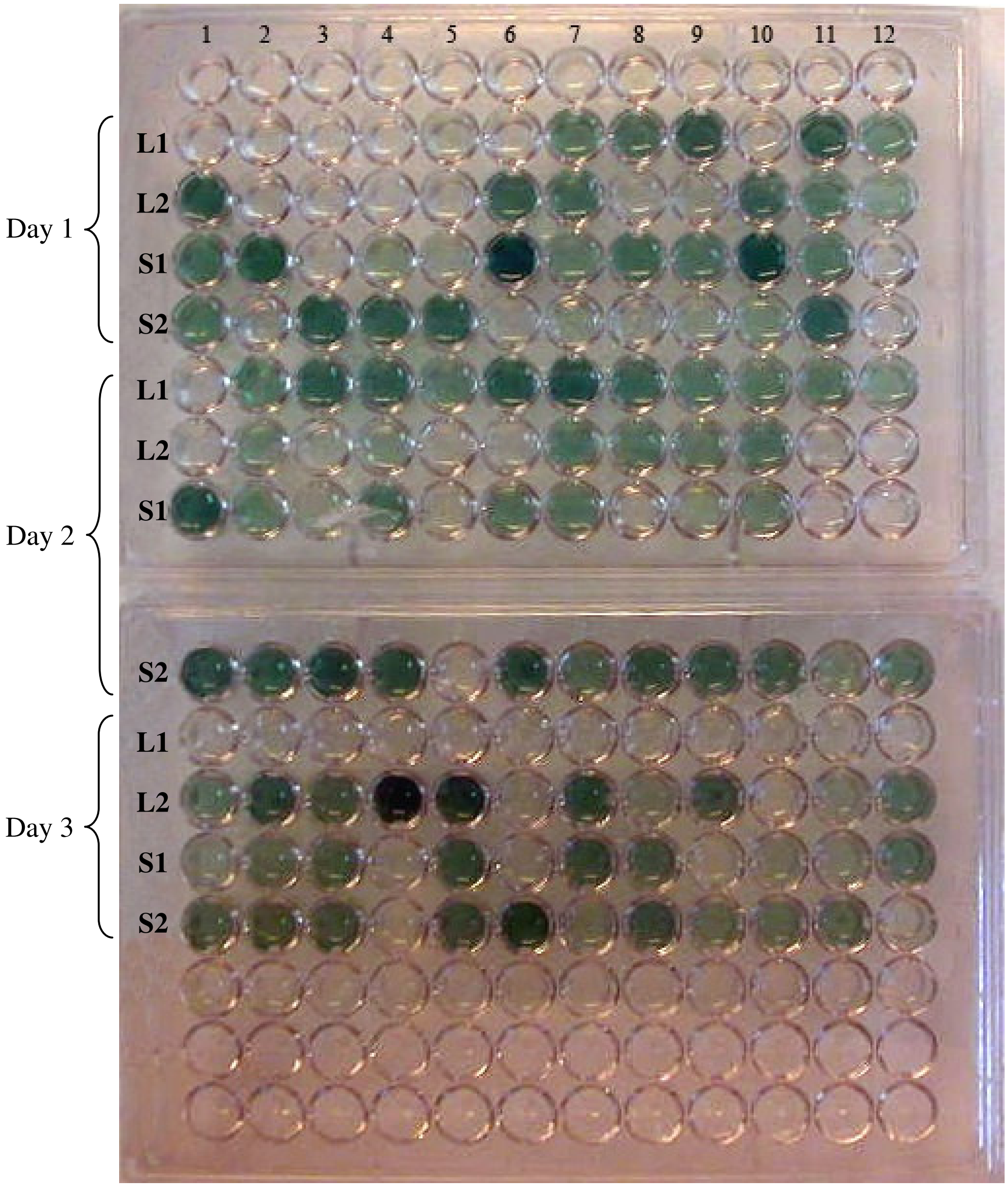


Figure 3.11 Enzyme Assay – Variants. The microtiter plate is numbered from 1 to 12 indicating the number of colonies used for the assay. Each well contains a single colony.

3.6 Quantitative Analysis of the level of GAO expression

In order to perform a quantitative analysis of the results, the RGB color system is used to numerate each color. The RGB color model is an additive model in which red, green, and blue are combined in various ways to reproduce other colors. The name of the model and the abbreviation ‘RGB’ come from the three primary colors, red, green, and blue. (48) Using the RGB color mode of Adobe Photoshop CS2, quantitative data are obtained, where the values *decrease* as the darkness *increase*.

3.6.1 Quantitative analysis of controls

As mentioned in Section 3.5, the activity of the variants must be compared with a control in order to have an idea about the extent of change in activity compared with the wildtype gene. The quantitative results of the controls are shown in Figure 3.12.

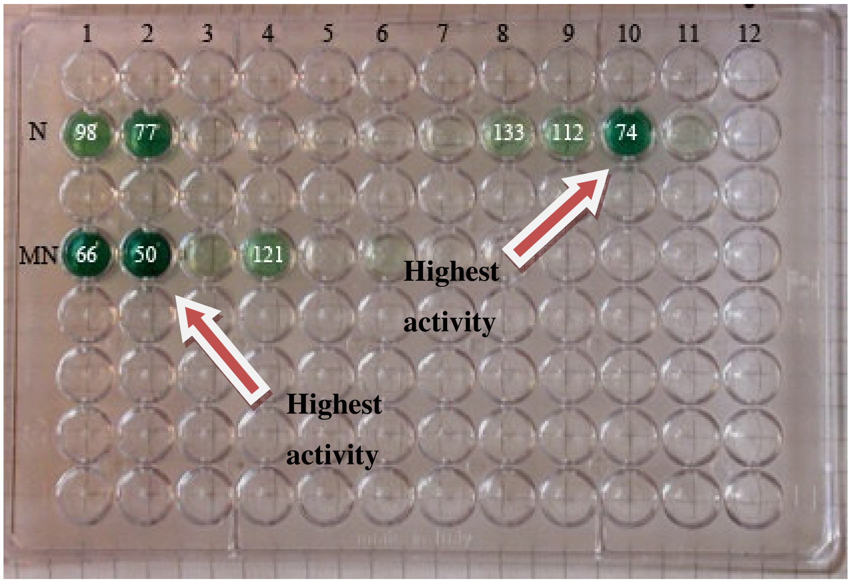


Figure 3.12 Quantitative analysis of results – Controls. The highest activities observed for the controls **N** and **MN** are indicated by the arrows.

As can be observed from the results, proGOMN isolated from single cells has shown a higher activity as compared with proGON as expected (since the activity of proGOMN has been enhanced using directed evolution). These results are consistent with the results of previous work done to enhance the acitivity of GAO using directed evolution methods (33, 37, 49). The cells showing the highest activity are **N10** for **N**, and **MN2** for **MN**.

3.6.2 Quantitative analysis of variants

The quantitative results of the variants are shown in Figure 3.13.

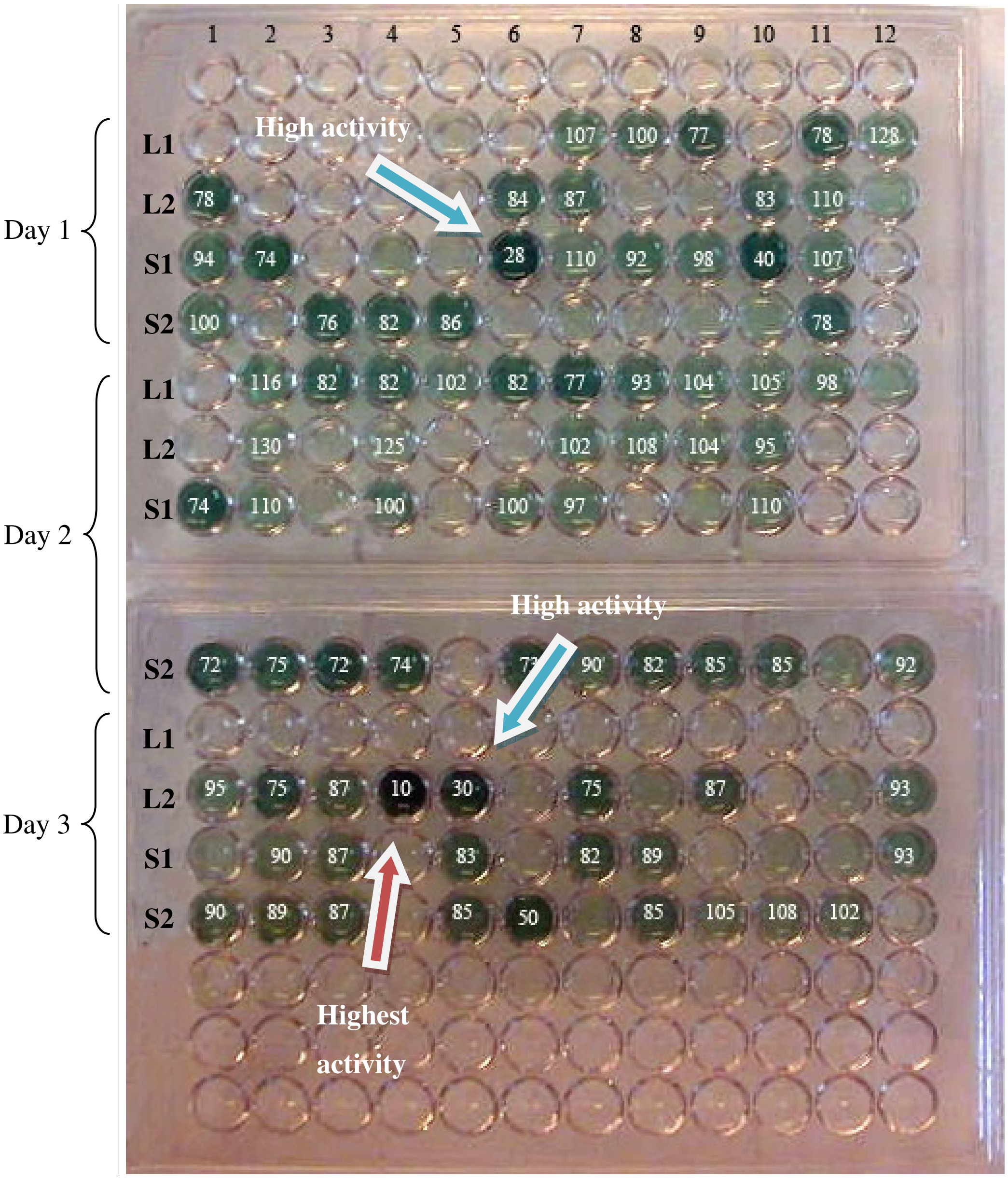


Figure 3.13 Quantitative analysis of results – Variants. High activities observed for the variants are indicated by blue arrows, and the highest activity (best variant) is indicated by the blue arrow.

As can be seen from the results, **L1** variants show activity at Day1 and Day2, but show no activity at Day3. This can be due to the fact that the cells were subjected to a starvation medium from the beginning, making it hard for them to survive at the end of Day3 due to the depletion of nutrients.

On the other hand, **L2** samples not only show activity at the end of Day3, but also one of the variants, **Day3 L2_**(labelled with a red arrow in Figure 3.13 with a value of 10) has the highest activity (even higher than the control proGOMN). This is very important since it shows that during starvation the cells have carried out mutations which has also affected the activity of the gene of interest, to the point that this activity is even higher than that achieved using normal in-vitro directed evolution methods. While the highest score for the controls were 74 for **N** and 50 for **MN** (Figure 3.12), the variant **Day3 L2** has a score of 10, which is about *7 times* that of **N** and *5 times* that of **MN**.

Good results are also obtained for **S1** samples (**Day1 S1**). Again here the activity of this variant is about *3 times* that of **N** and *2 times* that of **MN**.

Considering that the best variant of previous directed evolution studies (**MN**) had a 30-fold increase in activity, the ratios between the color codes of the best results can be calculated to give an approximate estimate of the increase in activity of the best vairant. Thus when a change in score from 74 (in **N**) to 50 (in **MN**) corresponds to about 30-fold increase in activity, a change from 74 (in **N**) to 10 (in best variant of variant **Day3 L2**) will correspond to approximately *150-fold increase* in enzyme activity.

It is interesting that some variants of **S1** gave good results, some of which were better than **MN**, while variants of **S2** gave activities close to **N**. Considering that **S1** was starvation medium with very low carbon source, it seems that the stress arising from starvation has resulted in more mutations in these variants.

The results above show that the induction of error prone polymerases during the SOS response and adaptive mutations in bacteria provides the means for mutation, and literally “hasten evolution”, thus increasing the activity of the desired gene during this process. This is a similar approach to directed evolution.

3.7 Phylogenetic Tree of Error-Prone Polymerases

It was discussed in Section 1.2.3 that adaptive mutations occurring during stressful conditions result in mutations that allow growth of the organism. This phenomena is a form of “hastened evolution”, since the organism carries out changes in order to cope with its environment. These mutations are carried out by EP polymerases, which have been found in organisms from all three domains of life. However, the point that remains unknown is whether there is a relation between the level of complexity of an organism and the EP polymerases it contains. The aim of this study was to generate the phylogenetic tree of EP polymerases of some organisms from all three domains of life, and to see is there is any link between EP polymerases and the level of complexity of the organism.

3.7.1 Tree-building method justification

Each tree building method has its advantages and disadvantages, and in order to choose the right tree, the tree building methods were examined carefully, especially when considering the assumptions of the methods.

Neighbor joining method has the advantage of being fast, and thus is suited for large datasets and for bootstrap analysis. Moreover, it is an efficient method and permits lineages with largely different branch lengths.

UPGMA is a simple method, but is very sensitive to unequal evolutionary rates. In other words, this method assumes the molecular clock assumption, and is thus not suitable for bacteria (this assumption is not valid for bacteria).

Maximum parsimony has the disadvantage of slow computation and poor performance when there is substantial among-site rate heterogeneity. It has a lower efficiency compared with Neighbor joining method.

Considering the discussion above, the neighbor joining method was found to be the most suitable method for the construction of phylogenetic trees of EP polymerases.

3.7.2 Choice of error-prone polymerase and species

Y-family error prone polymerases are present in organisms from all three domains of life: archaea, bacteria and eukarya. As mentioned previously in Section 1.2.1.3.6, Y-family polymerases are divided into four sub-groups: DinB, UmuC, Rad30 and Rev1. It has been found that the UmuC subfamily are found in bacteria whereas those from the Rad30 and Rev1 branch are found exclusively in eukaryotes. DinB subfamily, on the other hand, is the most diverse and is found in archaea, bacteria and eukaryotes. Quite remarkably, these DinB orthologs are often well conserved (26).

3.7.3 Phylogenetic profiling of error-prone polymerases

In order to gain a better understanding about the relationship between error prone polymerases, 2 species from each domain were chosen and the phylogenetic tree of all error-prone polymerases from these species was drawn. The species chosen from archeal domain were *Sulfolobus solfataricus* and *Acidianus infernus*, those chosen from bacterial domain were *Eschericia coli* and *Salmonella typhimurium*, and those chosen from eukaryotic domain were *Homo sapiens* and *Saccharomyces cerevisiae*. (See APPENDIX C). The reason for chosing the above mentioned species is that most studies have been carried out on these species. Table 3.1 shows the species chosen. For each species in the table, all of the error-prone polymerases belonging to each species are included, with information about the sub-family that each polymerase belongs to.

Table 3.1 List of species (and their error-polymerase genes) chosen for phylogenetic analysis. 2 organisms were chosen from each domain of life. DinB, muC, Rad30 and Rev1 are the sub-families of the Y family polymerases. The last row shows error-prone polymerases that do not belong to the Y family.

	Archaea		**Bacteria**		**Eukarya**	
	S. solfataricus	*A.infernus*	*E. coli*	*S.typhimurium*	*S.cerevisiae*	*H.sapiens*
DinB	Dpo4	Dpo4	PolIV	PolIV	-	Polκ
UmuC	-	-	PolV	PolV	-	-
Rad30	-	-	-	-	Polη	Polη Polι
Rev1	-	-	-	-	Rev1	Rev1
Other families*	-	-	-	-	Polζ	Polλ Polμ Polζ

* Error-prone polymerases that do not belong to the Y-family polymerases

Table 3.1 shows the six organisms chosen for phylogenetic analysis. The polymerases are classified according to the sub-family they belong to. All the error-prone polymerases found to date have been considered for each of the organisms above. In *H. sapiens*, for instance, search was done to find all of the error-prone polymerases, some of which did not even belong to the Y-family, but were still included in this study. One point that is immediately observable from the table is the large number of EP polymerases that are present in *H. sapiens*, in contrast to only one EP polymerase in *S. solfataricus.* This will be discussed in the sections that follow.

In order to carry out phylogenetic analysis of the above mentioned organisms, the NCBI database (50) was searched for EP polymerases in these organisms. The results (See APPENDIX C) for EP polymerase sequences of the species) were used to draw phylogenetic trees using MEGA 4.0 program (51).

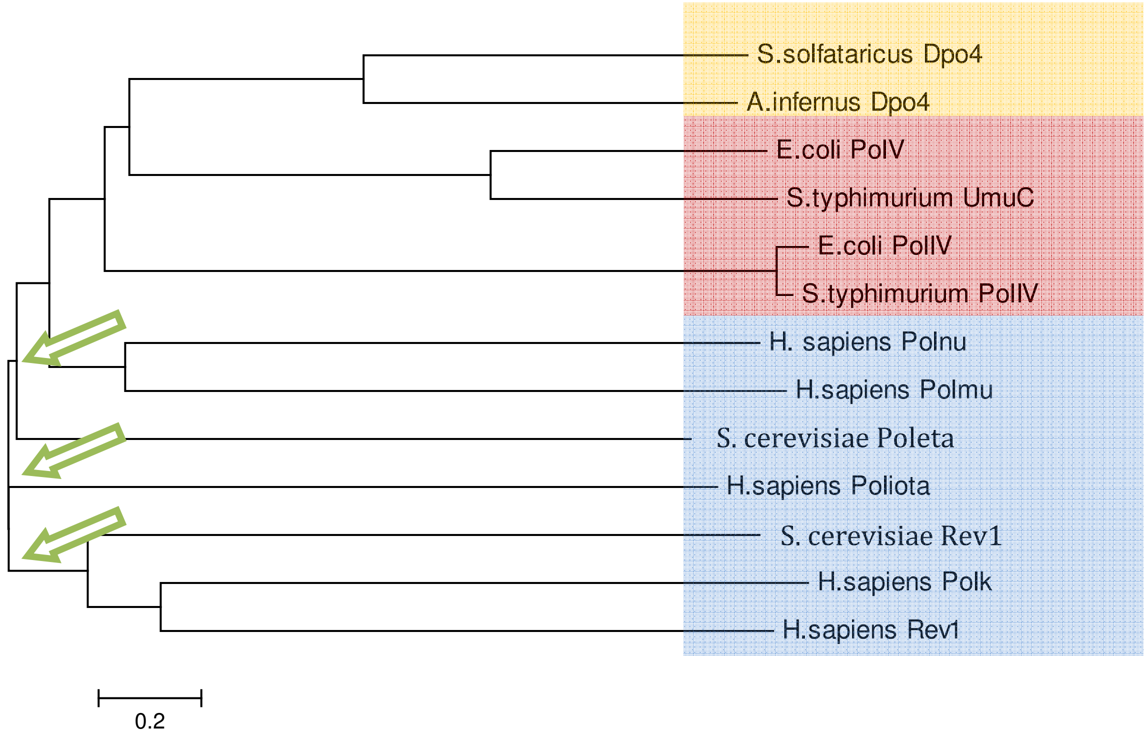


Figure 3.14 Phylogenetic analysis of error-prone polymerases. The tree gave 3 branches, and the species are color coded for ease of analysis. *Yellow:* Archaea, *Pink:* Bacteria, *Blue:* Eukarya.

The phylogenetic tree depicts the error-prone polymerase of the six species according to neighbor joining method. The three domains are colored for convenience. Species belonging to the archaea domain are highlighted in yellow, those of bacteria domain are highlighted in red, and those of eukarya domain are highlighted in blue. Considering Figure 3.14, the phylogenetic grouping of species based on archaea, bacteria and eukarya domains are easily conceivable. This result is as expected, since the resemblance of the archaea *S. solfataricus*, for instance, is much higher to *A. infernus* than to *S. cerevisiae*, since the first also belongs to the archaea domain whereas the second belongs to eukarya domain. The same thing applies for their EP polymerases.

Another important point is the Polι of *H. sapiens*. Previous studies have shown that there is no homologous counterpart of this gene in yeasts, bacteria and archaea. The phylogenetic tree in Figure 3.14 is divided into three branches (shown by the green arrows). The first branch includes the archaea and bacteria domain, in addition to some of eukarya domain. The third branch includes the eukarya domain. The *second* branch, however, contains the EP polymerase of only one species, which is *H. sapiens* Polι. This result is consistent with the studies performed on *H. sapiens* Polι, since as mentioned, there is no homologous counterpart of this gene in archaea, yeast, bacteria and even the other polymerases of *H. sapiens*.

It is interesting to note that there seems to be a trend in the number of EP polymerases in light of the organisms' complexity. Table 3.1 depicts the EP polymerases of species from all three domains of life, archaea, bacteria and eukarya.

It can be easily seen that while archaea species have only one EP polymerase, more complex species belonging to the eukarya domain have many EP polymerases, each having specific functions. (See Section 1.2.1.4).

Humans, for instance, have 14 DNA polymerases, at least 7 of which are error prone. *E. coli*, on the other hand, has a total of 5 DNA polymerases, with only 2 error prone polymerases. At the extreme there are archaea such as *S. solfataricus*, which the only Y-family EP polymerase found is DPo4. Table 3.2 shows the ratio of EP polymerases to normal polymerases:

Table 3.2 Analysis of EP polymerases in three organisms representing the three domains of life, and calculation of the ratio of EP polymerases to total polymerases.

	Archaea ***S. solfataricus***	**Bacteria** ***E. coli***	**Eukarya** ***H.sapiens***
Total polymerases (TP)	4	5	14
EP polymerases (EP)	1	2	7
Ratio (EP / TP)	**0.25**	**0.4**	**0.5**

Thus, as complexity increases in an organism, both the number and ratio of error prone polymerases also increase. This can be explained in light of the fact that complex organisms, like human beings, have cells that are much more differentiated than the simpler organisms, such as bacteria. The more complex an organism is, the more chemical reactions and functions are carried out. In other words, complex organisms have much more complex systems which function together in a complicated matrix, some being induced by others. Hence, such an organism will have a harder time surviving the conditions, and will thus need more ways of surviving them, which makes it reasonable to have more mechanisms of DNA repair and mutagenesis. Considering humans as examples, there are at least four Y-family polymerases, Pol η, ι, κ and Rev1. The most extreme example is human pol ι, which preferentially inserts G opposite T, rather than the canonical Watson-Crick base A, by a factor of 3- to 11-fold, depending upon the template sequence context (20). As for pol κ, it can pass certain lesions in an error- free and others in an error-prone way. Pol κ has another unique property: on the one hand it possesses a very low fidelity, but on the other hand it is moderately processive. This property suggests an important role in spontaneous mutagenesis (6).

All the above comparison shows that as organisms get more complex, starting from the simple archaea to complex human beings, the need for different mechanisms of adaptive mutations increase, hence number of error prone polymerases and their specificity increase, allowing them to survive various stressful conditions.

CHAPTER 4

CONCLUSIONS

In this study, the novel idea of "*in-vivo* directed evolution" was applied to obtain mutant forms with improved galactose oxidase activity. The results of enzyme assays performed on variants of the starvation experiment show that in-vivo directed evolution occurs in an efficient way by the stationary phase adaptive mutation, giving results that are even better that those achieved using previous in-vitro directed evoution methods. The best variant from this novel method showed about *150-fold* increase in activity with respect to the wiltype enzyme, as compared to the 30-fold increase that was achieved using previous directed evolution studies. Moreover, this method has the advantage of being much easier than normal directed evolution methods. Thus it can be concluded that the novel in-vivo directed evolution approach tested in this study is a promising method for the random mutagenesis of genes of interest. Hence stationary phase adaptive mutations may indeed be of practical use for the purpose of directed evolution studies.

The world consists of many organisms; all trying to survive and reproduce under the sometimes harsh conditions that nature poses to them, and adaptive mutations play an important role in cell survival during stressful conditions by inducing a number of error prone Y family polymerases that introduce mutations. The study carried out has shown that there is indeed a link between the EP polymerases and the level of complexity of an organism. Thus, as organisms get more complex, the number and ratio of EP polymerases increase. An organism with a higher number of EP polymerases has a better chance of mutating in the desired way for survival during harsh conditions, and thus, achieves the chance of being the ancestor of many organisms arising from it.

REFERENCES

1. Directed enzyme evolution . *Macquarie University Biotechnology Research Institute.* [Online] 2006. http://biotechnology.mq.edu.au/directed_enzyme.htm.

2. **R. Chatterjeea, L.Yuan.** Directed evolution of metabolic pathways. *Science Direct, Trends in Biotechnology.* 2006, Vol. 24, 1, pp 28-38.

3. Host Strain Genotypes. *University of Bern - Department of Clinical Pharmacology.* [Online] 2007. http://www.ikp.unibe.ch/molbio/ecolistrains.pdf.

4. Competent Cells for Bacterial Expression. *Invitrogen.* [Online] 2007 http://www.invitrogen.com/content/sfs/productnotes/F_Competent%20cells%20expression-041018-MKT-TL-HL0506021.pdf.

5. **D. Wilkinson, N. Akumanyi, R. Hurtado-Guerrero, H. Dawkes, P.F. Knowles, S.E.V. Phillips and M.J. McPherson.** Structural and kinetic studies of a series of mutants of gakactose oxidase identified by directed evolution. *Protein Engineering Design and Selection.* 2004, Vol. 17, 2, pp 141-148.

6. **U. Hübscher, G. Maga, S. Spadari.** Eukaryotic DNA polymerases. *Annu. Rev. Biochem.* 2002, Vol. 71, pp 133-163.

7. EMBL-EBI. *InterPro: IPR006134 DNA polymerase, B region.* [Online] 2007 http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR006134.

8. **I. Andricioaei, A. Goel, D. Herschbach and Martin Karplus.** Dependence of DNA Polymerase Replication Rate on External Forces: A Model Based on Molecular Dynamics Simulations. *Biophysical Journal.* 2004, Vol. 87, pp. 1478-1497.

9. **M. S. Wold.** DNA Replication: Proteins, Genetics, Biochemistry and Replication Mechanisms. *Principles in Mol. and Cell Biol.* [Online] 2006. http://www.medicine.uiowa.edu/biosciences/curriculum/PMCBLectures06/WoldLecture9.pdf.

10. M. Alda and G. Gonzales. *DNA REPLICATION AND GENOME STRUCTURE.* [Online] 2007 http://www.biosci.ohio-state.edu/~mgonzalez/Micro521/04.html.

11. **J. M. Berg, J. L. Tymoczko and L. Stryer.** *Biochemistry Fifth Edition.* s.l. : W. H. Freeman and Co., 2002.

12. **S. J. Johnson, J. S. Taylor and L. S. Beese.** Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proc Natl Acad Sci U S A.* 2003, Vol. 100, 7 pp. 3895–3900.

13. **M. C. Franklin, J. Wang and T. A. Steitz.** Structure of the Replicating Complex of a Pol α Family DNA Polymerase . *Cell.* 2001, Vol. 105, 5, pp 657-667 .

14. **Loeb, P. H. Patel and L. A.** Getting a grip on how DNA polymerases function. *Nature Structural Biology.* 2001, Vol. 8, pp. 656-659.

15. **K. P. Hopfner, A. Eichinger, R. A. Engh, F. Laue, W. Ankenbauer, R. Huber AND B. Angerer.** Crystal structure of a thermostable type B DNA polymerase from Thermococcus gorgonarius. *Proc. Natl. Acad. Sci. USA, Biochemistry.* 1999, Vol. 96, pp. 3600–3605.

16. **C. Savino, L. Federici, K. A. Johnson, B. Vallone, V. Nastopoulos, M. Rossi, F. M. Pisani and D. Tsernoglou.** The Crystal Structure of DNA Polymerase B1 from the Archaeon Sulfolobus solfataricus . *Structure.* 2004, Vol. 12, 11, pp 2001-2008.

17. **Y. Ishino, K. Komori, I. K. O. Cann and Y. Koga.** J Bacteriol. *A Novel DNA Polymerase Family Found in Archaea.* 180, 1998, Vol. 8, pp. 2232–2236.

18. **M. Seki, C. Masutani, L. W. Yang, A. Schuffert, S. Iwai, I. Bahar and R. D. Wood.** High-efficiency bypass of DNA damage by human DNA polymerase Q. *The EMBO Journal* . 2004, Vol. 23, pp. 4484–4494.

19. **S. J. Johnson, J. S. Taylor and L. S. Beese.** Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *PNAS Biochemistry.* 2003, Vol. 100, pp. 3895-3900.

20. **H. Ling, F. Boudsocq, R. Woodgate and W. Yang.** Crystal Structure of a Y-Family DNA Polymerase in Action: A Mechanism for Error-Prone and Lesion-Bypass Replication . *Cell.* 2001, Vol. 107, 1, pp 91-102.

21. E. coli DNA Replication. *Oregon state University.* [Online] 2007. http://oregonstate.edu/instruction/bb492/lectures/DNAII.html.

22. Bacterial DNA polymerases. [Online] 2007. http://www.mun.ca/biochem/courses/3107/Topics/DNA_polymerases.html.

23. **J. M. Daley, R. L. Vander Laan, A. Suresh and T. E. Wilson.** DNA Joint Dependence of Pol X Family Polymerase Action in Nonhomologous End Joining. *J. Biol. Chem.* 2005, Vol. 280, 32, pp 29030-29037.

24. **F. Lecointe, I. V. Shevelev, A. Bailone, S. Sommer and U. Hübscher.** Involvement of an X family DNA polymerase in double-stranded break repair in the radioresistant organism Deinococcus radiodurans. *Molecular Microbiology.* 2004, Vol. 53, 6, pp 1721-1730.

25. **I. Bruck, M. F. Goodman and M. O'Donnell.** The Essential C Family DnaE Polymerase Is Error-prone and Efficient at Lesion Bypass. *J. Biol. Chem.* 2003, Vol. 278, 45, pp 44361-44368.

26. **F. Boudsocq, S.i Iwai, F. Hanaoka and R. Woodgate.** Sulfolobus solfataricus P2 DNA polymerase IV (Dpo4): an archaeal DinB-like DNA polymerase with lesion-bypass properties akin to eukaryotic poln. *Nucleic Acids Research.* 2001, Vol. 29, 22, pp 4607-4616.

27. **Yang, W.** Portraits of a Y-family DNA polymerase . *Federation of European Biochemical Societies.* 2005, Vol. 579, 4, pp 868-872.

28. **U. Hübscher, G. Maga and S. Spadari.** Eukaryotic DNA Polymerases. *Annual Review of Biochemistry.* 2002, Vol. 71, pp 133-163.

29. **C. Janion, A. Sikora, A. Nowosielska and E. Grzesiuk.** Induction of the SOS Response in Starved Escherichia coli. *Environmental and Molecular Mutagenesis.* 2002, Vol. 40, pp 129–133.

30. **G. J. McKenzie, R. S. Harris, P. L. Lee and S. M. Rosenberg.** The SOS response regulates adaptive mutation . *PNAS.* 2000, Vol. 97, 12, pp 6646-6651.

31. **Y. Wang, J. Zhu, R. Zhu, Z. Zhu, Z. Lai, Z. Chen.** Chitosan/Prussian blue-based biosensors. *Meas. Sci. Technol.* 14, 2003, pp 831-836.

32. **J. Tkac, P. Gemeiner and E. Šturdik.** Rapid and sensitive galactose oxidase-peroxidase biosensor for galactose detection with prolonged stability. *Biotechnology Techniques.* 13, 1999, pp 931–936.

33. **F. H. Arnold, I. P. Petrounia, L. Sun.** Directed evolution of galactose oxidase enzymes. *Patent Storm.* [Online] 2006. http://www.patentstorm.us/patents/7115403-desc

34. **L. Sun, I. P. Petrounia, M. Yagasaki, G. Bandara and F. H. Arnold.** Expression and stabilization of galactose oxidase in E. coli by direced evolution. *Protein engineering.* 2001, Vol. 14, 9, pp 699-704.

35. **Delagrave, S.** Directed Evolution of Galactose Oxidase. *Protein engineering.* 2001, Vol. 14, pp 261-267.

36. **L. Sun, I. P. Petrounia, M. Yagasaki, G. Bandara and F. H. Arnold.** Expression and stabilization of galactose oxidase in Escherichia coli by directed evolution. *Protein Engineering.* 2001, Vol. 14, 9, pp 699-704.

37. *Structural and kinetic studies of a series of mutants of galactose oxidase identified by directed evolution.* **D. Wilkinson, N. Akumanyi, R. Hurtado-Guerrero, H. Dawks, P. F. Knowles, S. E. V. Phillips and M. J. McPherson.** 2, pp 141-148, 2004, Vol. 17.

38. *The pET System: Your Choice for Expression.* **R. Mierendorf, K. Yeager and R. Novy.** 1, May 1994, Vol. 1.

39. pET System Manual. [Online] 2006 http://www.fhcrc.org/science/labs/hahn/methods/biochem_meth/pet.pdf.

40. **Pfister, R. A. Smucker and R. M.** Liquid Nitrogen Cryo-Impacting: a New Concept for Cell Disruption. *Applied Microbiology.* 1975, Vol. 30, 3, pp. 445-449.

41. Phylogenetics. *Bioinformatics Solutions.* [Online] 2007 http://www.clcbio.com/index.php?id=46.

42. **Page, R.** Introduction to Tree Building. [book auth.] R. Page and E. Holmes. *Molecular Evolution: A Phylogenetic Approach.* s.l. : Blackwell Science, 1998.

43. **Strickler, P.** Gerstein group - Yale Bioinformatics. *A Brief Review of the Common Tree-building Methods Used in Phylogenetic Inference.* [Online] http://bioinfo.mbb.yale.edu/mbb452a/projects/Patricia-M-Strickler.html.

44. **T. Mailund, G. S. Brodal and R. Fagerberg.** Recrafting the neighbor-joining method. *BMC Bioinformatics.* 2006, Vol. 7, 29.

45. Maximum Parsimony analysis. *Christian de Duve Institute of Cellular Pathology.* [Online] 2006. http://www.icp.ucl.ac.be/~opperd/private/parsimony.html.

46. **Brian Golding.** Parsimony Methods. *Computational Biology.* [Online] http://helix.mcmaster.ca/721/outline2/node50.html.

47. **Bioinformatics.** Phylogeny - Tree Selection Cireteria. [Online] http://artedi.ebc.uu.se/course/BioInfo-10p-2001/Phylogeny/Phylogeny-Criteria/Phylogeny-Criteria.html.

48. Understanding color. *RGB World.* [Online] 2007 http://www.rgbworld.com/color.html.

49. **S. J. Firbank, M. S. Rogers, C. M. Wilmot, D. M. Dooley, M. A. Halcrow, P. F. Knowles, M. J. McPherson, and S. E. V. Phillips.** Crystal structure of the precursor of galactose oxidase: An unusual self-processing enzyme. *Proc Natl Acad Sci U S A* . 2001, Vol. 98, 23, pp 12932–12937.

50. *National Center for Biotechnology Information, NCBI.* [Online] 2007 http://www.ncbi.nlm.nih.gov/.

51. **K. Tamura, J. Dudley, N. Mei and S. Kumar.** Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution.* 2007, Vol. 24, pp. 1596-1599.

52. **M. Vaidyanathan, M. Palaniandavar.** Models for the active site in galactose oxidase: Structure, spectra and redox of copper(II) complexes of certain phenolate ligands. *Proc. Indian Acad. Sci. (Chem. Sci.).* 2000, Vol. 112, 3, pp. 223–238.

53. InterPro: IPR011043 Galactose oxidase/Kelch, beta-propeller. *EMBL- EBI.* [Online] 2006 http://www.ebi.ac.uk/interpro/IEntry?ac=IPR011043.

54. **S. Firbank, P. Knowles, M. McPherson and S. Phillips.** Structural studies of processing in galactose oxidase. [Online] http://www.astbury.leeds.ac.uk/Report/2000/Phillips.6.html.

55. J Bacteriol. 1998 April; . American Society for Microbiology. 180, Vol. (8): , pp. 2232–2236.

56. **E. C. Friedberg, G. C. Walker and W. Siede.** DNA Repair and Mutagenesis. *Trends in Biochemical Sciences.* 1995, Vol. 20, 10.

57. **C. Roodveldt, A. Aharoni and D. S. Tawfik.** Directed evolution of proteins for heterologous expression and stability. *Current Opinion in Structural Biology.* 2005, Vol. 15, pp 50-56.

58. **Weissman, K.** Rational or Random? *RSC.* 2004.

59. Directed Enzyme Evolution. *Maqcuarie University Biotechology Research Institute.* [Online] 2006. http://biotechnology.mq.edu.au/directed_enzyme.htm.

60. **F. H. Arnold, G. Georgiou.** Methods in Molecular Biology: Directed Enzyme Evolution: Screening and Selection Methods. *Biochemistry.* 2004, Vol. 5, 3.

61. **Arnold, F. H.** Directed Enzyme Evolution. [Online] http://www.che.caltech.edu/groups/fha/Enzyme/directed.html.

62. **T. Mailund, G. S. Brodal and R. Fagerberg.** Recrafting the neighbor-joining method. *BMC Bioinformatics.* 2006, Vol. 7, 29.

63. Lab color space. *Wikipedia.* [Online] 2007 http://en.wikipedia.org/wiki/Lab_color_space.

64. Color Models: HSB, RGB, CYMK and LAB. *Wowarea.* [Online] 2007 http://www.wowarea.com/english/help/color.htm.

65. Lab Color Space Model. *MacDermid ColorSpan.* [Online] 2007 http://www.colorspan.com/support/tutorials/cmpl/lab.asp.

66. EBI Tools: Clustal W. *Europian Bioinformatics Institute.* [Online] 2007 http://www.ebi.ac.uk/Tools/clustalw/index.html.

APPENDIX A

CHEMICALS, ENZYMES AND THEIR SUPPLIERS

Agar	Merck
Agarose	Sigma
ABTS	Applichem
$CaCl_2$	Sigma
$CuSO_4$	Merck
D-Galactose	Sigma
EDTA	Merck
Ethidium Bromide	Merck
Glacial Acetic Acid	Merck
Glucose	Sigma
Glycerol	Merck
HCl	Merck
Horseradish Peroxidase	Sigma
IPTG	Fermentas
KCl	Merck
λ DNA/*Hind* III DNA Marker	Fermentas
NaCl	Merck
$Na_2HPO_4.7H_2O$	Merck
NaH_2PO_4	Merck
NaOH	Merck
SDS	Merck
Sodium Acetate	Merck
Tris Base	Merck
Yeast Extract	Merck

APPENDIX B

PREPARATION PROCEDURES

2-TY Broth (1 L)

Tryptone	16 g
Yeast Extract	10 g
NaCl	5 g

The above chemicals are dissolved in 900 ml dH_2O, the pH is adjusted to 7.0 using 5 N NaOH (approximately 2 ml), and dH_2O is added to a final volume of 1 L. The solution is autoclaved and stored at 0°C.

2-TY Agar (1 L)

Tryptone	10 g
Yeast Extract	5 g
NaCl	10 g

The above chemicals are dissolved in 900 ml dH_2O, the pH is adjusted to 7.0 using 5 N NaOH (approximately 2 ml). 15 g agar is added and dissolved, and dH_2O is added to a final volume of 1 L. The solution is autoclaved and stored at 0°C.

10 X M9 Salts (100 ml)

$Na_2HPO_4.7H_2O$	10.95 g
KH_2PO_4	3 g
NaCl	0.5 g
NH_4Cl	1 g

The above chemicals are dissolved up to a total volume of 100 ml using dH_2O, and autoclaved.

Agarose (1%)

1 g agarose is dissolved in 100 ml 1xTAE buffer.

Ampicillin (100 mg/ml)

1 g ampicillin is dissolved in 5 ml ddH_2O as stock solution. Aliquotes of 100 mg/ml concentration are prepared by adding 250 μl of this stock solution to 750 μl ddH_2O and stored at -20°C.

Calcium Chloride (1 M)

54 g $CaCl_2.6H_2O$ is dissolved in 120 ml dH_2O, and brought to a final volume of 200 ml. The solution is filter sterilized, and stored in 1 ml aliquots at -20°C.

Copper Sulfate

15.95 g is dissolved in 100 ml H_2O.

EDTA (0.5 M, pH 8.0)

18.61 g ehtylenedinitrilotetraacetic acid disodium salt dehydrate is added to 80 ml distilled H_2O. It is stirred vigorously on a magnetic stirrer and the pH is adjusted to 8.0 by NaOH pellets. The solution is dispensed into aliquots and sterilized by autoclaving.

Ethidium Bromide solution (10 mg/ml)

0.2 g EtBr is dissolved into 20 ml H_2O by carefully stirring for several hours. The solution is stored by wrapping in aluminum foil or in dark bottle at room temperature.

Galactose Oxidase Assay Solution

D-galactose	5.4 g
ABTS	22 mg
HrP (90 U/mg)	8.25 mg
100 mM NaPi (pH = 7)	50 ml

Reagents are mixed and the solution is stored in dark bottles at 0°C.

Glucose (20%)

20 g glucose is dissolved in 70 ml dH_2O, and brought to a final volume of 100 ml. The solution is filter sterilized and stored at 4°C.

IPTG

2.4 g IPTG is dissolved in 10 ml H_2O, filter sterilized, dispensed into aliquots and stored at -20°C.

M9 Minimal Medium Broth (1 L)

	1% Glucose	**0.05% Glucose**
dH_2O	848 ml	895.6 ml
10XM9 Salt	100 ml	100 ml
20% Glucose	50 ml	2.5 ml
1 M $MgSO_4$	1 ml	1 ml
Vit. B1	1 ml	1 ml

The above solutions are autoclaved separately and mixed when the temperature of the solutions reach below 50°C.

M9 Minimal Medium Agar (1 L)

	1% Glucose	**0.05% Glucose**
10XM9 Salt	100 ml	100 ml
20% Glucose	50 ml	2.5 ml
1 M $MgSO_4$	1 ml	1 ml
Vit. B1	1 ml	1 ml

The above solutions are autoclaved separately.

	1% Glucose	**0.05% Glucose**
dH_2O	833 ml	880.6 ml
Agar	15 g	15 g

The agar is dissolved in dH_2O as outlined above, and autoclaved.
All the solutions are mixed when the temperature of the solutions reach below 50°C.

Magnesium Chloride

19 g $MgCl_2$ is dissolved in 90 ml dH_2O, and the volume is made up to 100 ml with dH_2O, and sterilized by autoclaving.

Magnesium Sulfate (1 M)

12.32 g $MgSO_4.7H_2O$ is dissolved in 30 ml dH_2O, and brought to a final volume of 50 ml. The solution is then autoclaved.

Sodium Chloride (5 M)

29.22 g NaCl is dissolved in 70 ml dH_2O, and adjusted to a final volume of 100 ml. The solution is autoclaved and stored at room temperature.

Sodium Hydroxide (10 N)

20 g NaOH pellets are dissolved in 50 ml dH_2O and stored at room temperature.

NE Buffer

0.3 M NaAC (pH = 7.0)

SDS (10%)

10 g SDS is dissolved in 100 ml dH_2O carefully wearing a mask, autoclaved and stored at room temperature.

Sodium Acetate (3M, pH 5.2 and 7.0)

40.81 g $NaAc.3H_2O$ is dissolved in 80 ml H_2O. The pH is adjusted to 5.2 using glacial acetic acid or to 7.0 using dilute acetic acid. The volume is adjusted to 100 ml with dH_2O, autoclaved and stored at room temperature.

Solution I (Alkaline Lysis)

50 mM Glucose

25 mM Tris.HCL (pH 8.0)

10 mM EDTA

Solution II (Alkaline Lysis)

0.2 N NaOH

1% SDS

Solution III (Alkaline Lysis)

3.0 M NaAC (pH 4.8).

Solution A (Competent Cell Preparation)

$CaCl_2$ (1 M)	2.5 ml
Tris-HCL (1 M, pH 8.0)	500 µl
dH_2O	47 ml

Solution B (Competent Cell Preparation)

$CaCl_2$ (1 M)	500 µl
Tris-HCL (1 M, pH 8.0)	100 µl
Glycerol	2 ml
dH_2O	7.4 ml

TAE Buffer (50X, per liter)

242 g Tris base is dissolved in 600 ml dH_2O. The pH is adjusted to 8.0 using approximately 57 ml glacial acetic acid. 100 ml EDTA (0.5 M, pH 8.0) is added and the volume is adjusted to 1 L with dH_2O.

TE Buffer

	pH 7.4	**pH 7.6**	**pH 8.0**
10 mM Tris-HCL	pH 7.4	pH 7.6	pH 8.0
1 mM EDTA	pH 8.0	pH 8.0	pH 8.0

Tetracycline (10 mg/ml)

100 mg tetracycline is dissolved in 10 ml 50% ethanol, aliquoted and stored at -20°C.

Tris-HCL Buffer (1 M, pH 8.0, 1 L)

121.1 g Tris base is dissolved in 800 ml dH_2O. The pH is adjusted to 8.0 using concentrated HCL. dH_2O is added to a final volume of 1 L, and the solution is autoclaved and stored at room temperature.

Vitamin B1

100 mg thiamin hydrochloride is dissolved in 40 ml dH_2O and H_2O is added to a final volume of 50 ml. The solution is filter sterilized using 0.2 µm filter and stored at 4°C.

APPENDIX C

LIST OF SPECIES USED FOR TREE CONSTRUCTION

Sulfolobus solfataricus P2

Archaea; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae; Sulfolobus.

```
>gi|13815749|gb|AAK42588.1| DNA polymerase IV (family Y) (dpo4)
[Sulfolobus solfataricus P2]
MIVLFVDFDYFYAQVEEVLNPSLKGKPVVVCVFSGRFEDSGAVATANYEARKFGVKAGIPIVEAKKI
LPNAVYLPMRKEVYQQVSSRIMNLLREYSEKIEIASIDEAYLDISDKVRDYREAYNLGLEIKNKILE
KEKITVTVGISKNKVFAKIAADMAKPNGIKVIDDEEVKRLIRELDIADVPGIGNITAEKLKKLGINK
LVDTLSIEFDKLKGMIGEAKAKYLISLARDEYNEPIRTRVRKSIGRIVTMKRNSRNLEEIKPYLFRA
IEESYYKLDKRIPKAIHVVAVTEDLDIVSRGRTFPHGISKETAYSESVKLLQKILEEDERKIRRIGV
RFSKFIEAIGLDKFFDT
```

Acidianus infernus

Archaea; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae; Acidianus.

```
>gi|74418666|gb|ABA03146.1| Dpo4 [Acidianus infernus]
MIVLFVDFDYFFAQVEEVLNPELKGKPVAVCVFSGRFKDSGAIATANYEARKLGIKSGMPIPKAKEI
APNAIYLPIRKDLYKQVSDRIMYGILSKYSSKIEIASIDEAYLDITDRVKDYYEAYQLGKKIKDEIY
QKEKITVTIGIAPNKVFAKIIAEMNKPNGLGILKPEEVEGFIRSLPIEEVPGVGDSIYSKLKEMEIK
YLYDVLKVDFEKLKKEIGKSKASYLYSLANNTYAEPVKEKVRKHIGRYVTMKKNSRDIKEILPYLKR
AIDEAYSKTNGGIPKTLAVVAIMEDLDIVSREKTFNFGISKDRAYLEAEKLLEEIIKSDKRRLRRVG
VRLGKIYKSTTLDNFFNNV
```

Escherichia coli K12

Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.

```
>gi|2501652|sp|Q47155.1|DPO4_ECOLI DNA polymerase IV (Pol IV)
MRKIIHVDMDCFFAAVEMRDNPALRDIPIAIGGSRERRGVISTANYPARKFGVRSAMPTGMALKLCP
HLTLLPGRFDAYKEASNHIREIFSRYTSRIEPLSLDEAYLDVTDSVHCHGSATLIAQEIRQTIFNEL
QLTASAGVAPVKFLAKIASDMNKPNGQFVITPAEVPAFLQTLPLAKIPGVGKVSAAKLEAMGLRTCG
DVQKCDLVMLLKRFGKFGRILWERSQGIDERDVNSERLRKSVGVERTMAEDIHHWSECEAIIERLYP
ELERRLAKVKPDLLIARQGVKLKFDDFQQTTQEHVWPRLNKADLIATARKTWDERRGGRGVRLVGLH
VTLLDPQMERQLVLGL
```

Escherichia coli K12
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Escherichia.
>gi|16129147|ref|NP_415702.1| DNA polymerase V, subunit C
[Escherichia coli K12]
MFALCDVNAFYASCETVFRPDLWGKPVVVLSNNDGCVIARNAEAKALGVKMGDPWFKQKDLFRRCGV
VCFSSNYELYADMSNRVMSTLEELSPRVEIYSIDEAFCDLTGVRNCRDLTDFGREIRATVLQRTHLT
VGVGIAQTKTLAKLANHAAKKWQRQTGGVVDLSNLERQRKLMSALPVDDVWGIGRRISKKLDAMGIK
TVLDLADTDIRFIRKHFNVVLERTVRELRGEPCLQLEEFAPTKQEIICSRSFGERITDYPSMRQAIC
SYAARAAEKLRSEHQYCRFISTFIKTSPFALNEPYYGNSASVKLLTPTQDSRDIINAATRSLDAIWQ
AGHRYQKAGVMLGDFFSQGVAQLNLFDDNAPRPGSEQLMTVMDTLNAKEGRGTLYFAGQGIQQQWQM
KRAMLSPRYTTRSSDLLRVK

Salmonella typhimurium
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Salmonella.
>gi|54040957|sp|P63989.1|DPO4_SALTY DNA polymerase IV (Pol IV)
MRKIIHVDMDCFFAAVEMRDNPALRDIPIAIGGSRERRGVISTANYPARQFGVRSAMPTAMALKLCP
HLTLLPGRFDAYKEASRHVRDIFSRYTSLIEPLSLDEAWLDVTDSPHCYGSATLIAREIRQTIFNEL
QLTASAGVAPVKFLAKIASDLNKPNGQYVITPADVPDFLKTLPLAKIPGVGKVSAAKLENMGLRTCG
DIQQCDLAMLLKRFGKFGRVLWERSQGIDERDVNSERLRKSVGVERTLAEDIHEWSDCEAIIEHLYP
ELERRLAIVKPDLLIARQGVKLKFNDFQQTTQEHVWPQLNKEDLITTARKTWDERRGERGVRLVGLH
VTLLDPQLERQLVLGL

Salmonella typhimurium
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
Enterobacteriaceae; Salmonella.
>gi|217089|dbj|BAA14226.1| UmuC [Salmonella typhimurium]
MFALADVNSFYASCEKVFRPDLRDRSVVVLSNNDGCVIPRSAEAKKLGIKMGVPWFQLRSAKFPEPV
IAFSSNYALYASMSNRVMVHLEELAPRVEQYSIDEMFLDIRGIDSCIDFEDFGRQLREHVRSGTGLT
IGVGMGPTKTLAKSAQWASKEWSQFGGVLALTLHNQKRTEKLLSLQPVEEIWGVGRRISKKLNTMGI
TTALQLARANPTFIRKNFNVVLERTVRELNGESCISLEEAPPPKQQIVCSRSFGERVTTYEAMRQAV
CQHAERAAEKLRGERQFCRHIAVFVKTSPFAVTEPYYGNLASEKLLIPTQDTRDIIAAAVRALDRIW
VDGHRYAKAGCMLNDFTPTGVSQLNLFDEVQPRERSEQLMQVLDGINHPGKGKIWFAGRGIAPEWQM
KRELLSPAYTTRWADIPAAKLT

Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
>gi|30048187|gb|AAH50718.1| POLK protein [Homo sapiens]
MDSTKEKCDSYKDDLLLRMGLNDNKAGMEGLDKEKINKIIMEATKGSRFYGNELKKEKQVNQRIENM
MQQKAQITSQQLRKAQLQVDRFAMELEQSRNLSNTIVHIDMDAFYAAVEMRDNPELKDKPIAVGSMS
MLSTSNYHARRFGVRAAMPGFIAKRLCPQLIIVPPNFDKYRAVSKEVKEILADYDPNFMAMSLDEAY
LNITKHLEERQNWPEDKRRYFIKMGSSVENDNPGKEVNKLSEHERSISPLLFEESPSDVQPPGDPFQ
VNFEEQNNPQILQNSVVFGTSAQEVVKEIRFRIEQKTTLTASAGIAPNTMLAKVCSDKNKPNGQYQI
LPNRQAVMDFIKDLPIRKVSGIGKVTEKMLKALGIITCTELYQQRALLSLLFSETSWHYFLHISLGL

GSTHLTRDGERKSMSVERTFSEINKAEEQYSLCQELCSELAQDLQKERLKVLYFDMVSLVFKFFNSK
MLP

Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
>gi|112420958|ref|NP_861524.2| polymerase (DNA directed) nu [Homo
sapiens]
MENYEALVGFDLCNTPLSSVAQKIMSAMHSGDLVDSKTWGKSTETMEVINKSSVKYSVQLEDRKTQS
PEKKDLKSLRSQTSRGSAKLSPQSFSVRLTDQLSADQKQKSISSLTLSSCLIPQYNQEASVLQKKGH
KRKHFLMENINNENKGSINLKRKHITYNNLSEKTSKQMALEEDTDDAEGYLNSGNSGALKKHFCDIR
HLDDWAKSQLIEMLKQAAALVITVMYTDGSTQLGADQTPVSSVRGIVVLVKRQAEGGHGCPDAPACG
PVLEGFVSDDPCIYIQIEHSAIWDQEQEAHQQFARNVLFQTMKCKCPVICFNAKDFVRIVLQFFGND
GSWKHVADFIGLDPRIAAWLIDPSDATPSFEDLVEKYCEKSITVKVNSTYGNSSRNIVNQNVRENLK
TLYRLTMDLCSKLKDYGLWQLFRTLELPLIPILAVMESHAIQVNKEEMEKTSALLGARLKELEQEAH
FVAGERFLITSNNQLREILFGKLKLHLLSQRNSLPRTGLQKYPSTSEAVLNALRDLHPLPKIILEYR
QVHKIKSTFVDGLLACMKKGSISSTWNQTGTVTGRLSAKHPNIQGISKHPIQITTPKNFKGKEDKIL
TISPRAMFVSSKGHTFLAADFSQIELRILTHLSGDPELLKLFQESERDDVFSTLTSQWKDVPVEQVT
HADREQTKKVVYAVVYGAGKERLAACLGVPIQEAAQFLESFLQKYKKIKDFARAAIAQCHQTGCVVS
IMGRRRPLPRIHAHDQQLRAQAERQAVNFVVQGSAADLCKLAMIHVFTAVAASHTLTARLVAQIHDE
LLFEVEDPQIPECAALVRRTMESLEQVQALELQLQVPLKVSLSAGRSWGHLVPLQEAWGPPPGPCRT
ESPSNSLAAPGSPASTQPPPLHFSPSFCL

Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
>gi|154350220|ref|NP_009126.2| polymerase (DNA directed) iota [Homo
sapiens]MEKLGVEPEEEGGGDDDEEDAEAWAMELADVGAAASSQGVHDQVLPTPNASSRVIVHVD
LDCFYAQVEMISNPELKDKPLGVQQKYLVVTCNYEARKLGVKKLMNVRDAKEKCPQLVLVNGEDLTR
YREMSYKVTELLEEFSPVVERLGFDENFVDLTEMVEKRLQQLQSDELSAVTVSGHVYNNQSINLLDV
LHIRLLVGSQIAAEMREAMYNQLGLTGCAGVASNKLLAKLVSGVFKPNQQTVLLPESCQHLIHSLNH
IKEIPGIGYKTAKCLEALGINSVRDLQTFSPKILEKELGISVAQRIQKLSFGEDNSPVILSGPPQSF
SEEDSFKKCSSEVEAKNKIEELLASLLNRVCQDGRKPHTVRLIIRRYSSEKHYGRESRQCPIPSHVI
QKLGTGNYDVMTPMVDILMKLFRNMVNVKMPFHLTLLSVCFCNLKALNTAKKGLIDYYLMPSLSTTS
RSGKHSFKMKDTHMEDFPKDKETNRDFLPSGRIESTRTRESPLDTTNFSKEKDINEFPLCSLPEGVD
QEVFKQLPVDIQEEILSGKSREKFQGKGSVSCPLHASRGVLSFFSKKQMQDIPINPRDHLSSSKQVS
SVSPCEPGTSGFNSSSSSYMSSQKDYSYYLDNRLKDERISQGPKEPQGFHFTNSNPAVSAFHSFPNL
QSEQLFSRNHTTDSHKQTVATDSHEGLTENREPDSVDEKITFPSDIDPQVFYELPEAVQKELLAEWK
RAGSDFHIGHK

Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
>gi|6601457|gb|AAF18986.1|AF206019_1 REV1 protein [Homo sapiens]
MRRGGWRKRAENDGWETWGGYMAAKVQKLEEQFRSDAAMQKDGTSSTIFSGVAIYVNGYTDPSAEEL
RKLMMLHGGQYHVYYSRSKTTHIIATNLPNAKIKELKGEKVIRPEWIVESIKAGRLLSYIPYQLYTK
QSSVQKGLSFNPVCRPEDPLPGPSNIAKQLNNRVNHIVKKIETENEVKVNGMNSWNEEDENNDFSFV
DLEQTSPGRKQNGIPHPRGSTAIFNGHTPSSNGALKTQDCLVPMVNSVASRLSPAFSQEEDKAEKSS

TDFRDCTLQQLQQSTRNTDALRNPHRTNSFSLSPLHSNTKINGAHHSTVQGPSSTKSTSSVSTFSKA
APSVPSKPSDCNFISNFYSHSRLHHISMWKCELTEFVNTLQRQSNGIFPGREKLKKMKTGRSALVVT
DTGDMSVLNSPRHQSCIMHVDMDCFFVSVGIRNRPDLKGKPVAVTSNRGTGRAPLRPGANPQLEWQY
YQNKILKGKAADIPDSSLWENPDSAQANGIDSVLSRAEIASCSYEARQLGIKNGMFFGHAKQLCPNL
QAVPYDFHAYKEVAQTLYETLASYTHNIEAVSCDEALVDITEILAETKLTPDEFANAVRMEIKDQTK
CAASVGIGSNILLARMATRKAKPDGQYHLKPEEVDDFIRGQLVTNLPGVGHSMESKLASLGIKTCGD
LQYMTMAKLQKEFGPKTGQMLYRFCRGLDDRPVRTEKERKSVSAEINYGIRFTQPKEAEAFLLSLSE
EIQRRLEATGMKGKRLTLKIMVRKPGAPVETAKFGGHGICDNIARTVTLDQATDNAKIIGKAMLNMF
HTMKLNISDMRGVGIHVNQLVPTNLNPSTCPSRPSVQSSHFPSGSYSVRDVFQVQKAKKSTEEEHKE
VFRAAVDLEISSASRTCTFLPPFPAHLPTSPDTNKAESSGKWNGLHTPVSVQSRLNLSIEVPSPSQL
DQSVLEALPPDLREQVEQVCAVQQAESHGDKKKEPVNGCNTGILPQPVGTVLLQIPEPQESNSDAGI
NLIALPAFSQVDPEVFAALPAELQRELKAAYDQRQRQGENSTHQQSASASVPKNPLLHLKAAVKEKK
RNKKKKTIGSPKRIQSPLNNKLLNSPAKTLPGACGSPQKLIDGFLKHEGPPAEKPLEELSASTSGVP
GLSSLQSDPAGCVRPPAPNLAGAVEFNDVKTLLREWITTISDPMEEDILQVVKYCTDLIEEKDLEKL
DLVIKYMKRLMQQSVESVWNMAFDFILDNVQVVLQQTYGSTLKVT

Saccharomyces cerevisiae
Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
Saccharomyces.
>gi|2507536|sp|P12689.2|REV1_YEAST DNA repair protein REV1
MGEHGGLVDLLDSDLEYSINRETPDKNNCLSQQSVNDSHLTAKTGGLNARSFLSTLSDDSLIEYVNQ
LSQTNKNNSNPTAGTLRFTTKNISCDELHADLGGGEDSPIARSVIEIQESDSNGDDVKKNTVYTREA
YFHEKAHGQTLQDQILKDQYKDQISSQSSKIFKNCVIYINGYTKPGRLQLHEMIVLHGGKFLHYLSS
KKTVTHIVASNLPLKKRIEFANYKVVSPDWIVDSVKEARLLPWQNYSLTSKLDEQQKKLDNCKTVNS
IPLPSETSLHKGSKCVGSALLPVEQQSPVNLNNLEAKRIVACDDPDFLTSYFAHSRLHHLSAWKANL
KDKFLNENIHKYTKITDKDTYIIFHIDFDCFFATVAYLCRSSSFSACDFKRDPIVVCHGTKNSDIAS
CNYVARSYGIKNGMWVSQAEKMLPNGIKLISLPYTFEQFQLKSEAFYSTLKRLNIFNLILPISIDEA
VCVRIIPDNIHNTNTLNARLCEEIRQEIFQGTNGCTVSIGCSDSLVLARLALKMAKPNGYNITFKSN
LSEEFWSSFKLDDLPGVGHSTLSRLESTFDSPHSLNDLRKRYTLDALKASVGSKLGMKIHLALQGQD
DEESLKILYDPKEVLQRKSLSIDINWGIRFKNITQVDLFIERGCQYLLEKLNEINKTTSQITLKLMR
RCKDAPIEPPKYMGMGRCDSFSRSSRLGIPTNEFGIIATEMKSLYRTLGCPPMELRGLALQFNKLVD
VGPDNNQLKLRLPFKTIVTNRAFEALPEDVKNDINNEFEKRNYKRKESGLTSNSLSSKKKGFAISRL
EVNDLPSTMEEQFMNELPTQIRAEVRHDLRIQKKIQQTKLGNLQEKIKRREESLQNEKNHFMGQNSI
FQPIKFQNLTRFKKICQLVKQWVAETLGDGGPHEKDVKLFVKYLIKLCDSNRVHLVLHLSNLISREL
NLCAFLNQDHSGFQTWERILLNDIIPLLNRNKHTYQTVRKLDMDFEV

Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
Saccharomycetales; Saccharomycetaceae; Saccharomyces.
>gi|6320627|ref|NP_010707.1| DNA polymerase eta
MSKFTWKELIQLGSPSKAYESSLACIAHIDMNAFFAQVEQMRCGLSKEDPVVCVQWNSIIAVSYAAR
KYGISRMDTIQEALKKCSNLIPIHTAVFKKGEDFWQYHDGCGSWVQDPAKQISVEDHKVSLEPYRRE
SRKALKIFKSACDLVERASIDEVFLDLGRICFNMLMFDNEYELTGDLKLKDALSNIREAFIGGNYDI
NSHLPLIPEKIKSLKFEGDVFNPEGRDLITDWDDVILALGSQVCKGIRDSIKDILGYTTSCGLSSTK
NVCKLASNYKKPDAQTIVKNDCLLDFLDCGKFEITSFWTLGGVLGKELIDVLDLPHENSIKHIRETW
PDNAGQLKEFLDAKVKQSDYDRSTSNIDPLKTADLAEKLFKLSRGRYGLPLSSRPVVKSMMSNKNLR
GKSCNSIVDCISWLEVFCAELTSRIQDLEQEYNKIVIPRTVSISLKTKSYEVYRKSGPVAYKGINFQ
SHELLKVGIKFVTDLDIKGKNKSYYPLTKLSMTITNFDIIDLQKTVVDMFGNQVHTFKSSAGKEDEE
KTTSSKADEKTPKLECCKYQVTFTDQKALQEHADYHLALKLSEGLNGAEESSKNLSFGEKRLLFSRK
RPNSQHTATPQKKQVTSSKNILSFFTRKK

OTHER THAN Y FAMILY

```
>gi|17366980|sp|Q9NP87.1|DPOLM_HUMAN DNA polymerase mu (Pol Mu)
MLPKRRRARVGSPSGDAASSTPPSTRFPGVAIYLVEPRMGRSRRAFLTGLARSKGFRVLDACSSEAT
HVVMEETSAEEAVSWQERRMAAAPPGCTPPALLDISWLTESLGAGQPVPVECRHRLEVAGPRKGPLS
PAWMPAYACQRPTPLTHHNTGLSEALEILAEAAGFEGSEGRLLTFCRAASVLKALPSPVTTLSQLQG
LPHFGEHSSRVVQELLEHGVCEEVERVRRSERYQTMKLFTQIFGVGVKTADRWYREGLRTLDDLREQ
PQKLTQQQKAGLQHHQDLSTPVLRSDVDALQQVVEEAVGQALPGATVTLTGGFRRGKLQGHDVDFLI
THPKEGQEAGLLPRVMCRLQDQGLILYHQHQHSCCESPTRLAQQSHMDAFERSFCIFRLPQPPGAAV
GGSTRPCPSWKAVRVDLVVAPVSQFPFALLGWTGSKLFQRELRRFSRKEKGLWLNSHGLFDPEQKTF
FQAASEEDIFRHLGLEYLPPEQRNA

>gi|6687796|emb|CAB65074.1| DNA polymerase lambda [Homo sapiens]
MDPRGILKAFPKRQKIHADASSKVLAKIPRREEGEEAEEWLSSLRAHVVRTGIGRARAELFEKQIVQ
HGGQLCPAQGPGVTHIVVDEGMDYERALRLLRLPQLPPGAQLVKSAWLSLCLQERRLVDVAGFSIFI
PSRYLDHPQPSKAEQDASIPPGTHEALLQTALSPPPPPTRPVSPPQKAKEAPNTQAQPISDDEASDG
EETQVSAADLEALISGHYPTSLEGDCEPSPAPAVLDKWVCAQPSSQKATNHNLHITEKLEVLAKAYS
VQGDKWRALGYAKAINALKSFHKPVTSYQEACSIPGIGKRMAEKIIEILESGHLRKLDHISESVPVL
ELFSNIWGAGTKTAQMWYQQGFRSLEDIRSQASLTTQQAIGLKHYSDFLERMPREEATEIEQTVQKA
AQAFNSGLLCVACGSYRRGKATCGDVDVLITHPDGRSHRGIFSRLLDSLRQEGFLTDDLVSQEENGQ
QQKYLGVCRLPGPGRRHRRLDIIVVPYSEFACALLYFTGSAHFNRSMRALAKTKGMSLSEHALSTAV
VRNTHGCKVGPGRVLPTPTEKDVFRLLGLPYREPAERDW
```