# Random Mutations, Protein Mutability, and DNA Repair
## *Understanding protein tolerance to random amino acid changes through directed evolution*

Haiwei H. Guo

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2004

Program Authorized to Offer Degree: Pathology

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Haiwei H. Guo

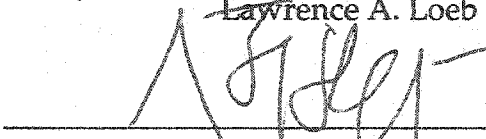
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

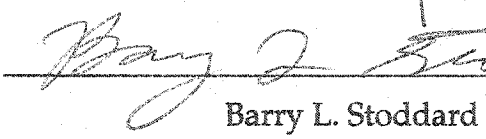
Lawrence A. Loeb

Reading Committee:

Lawrence A. Loeb



Raymond J. Monnat



Barry L. Stoddard

Date: 7/20/2004

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Hairuo H. Du

Date 7/20/2004

University of Washington

**Abstract**

Random Mutations, Protein Mutability, and DNA Repair:
*Understanding enzyme tolerance to random amino acid change through directed evolution*

Haiwei H. Guo

Chair of the Supervisory Committee:
Professor
Lawrence A. Loeb
Departments of Pathology and Biochemistry

In nature, evolution has given rise to the astonishing and wonderful diversity of organisms on this planet. In the laboratory, directed evolution can be a powerful technique to generate variants of a given protein with novel characteristics. The end products of the selection for desired traits from large combinatorial sets of mutants can also yield insight into protein structure and function. In both laboratory and nature, new mutations may be beneficial, but are often neutral or deleterious to overall protein function. One salient question that arises is the degree of tolerance of a protein to random amino acid change.

A major part of the corpus of this dissertation addresses this question of protein tolerance to random change. This basic question is quantitatively defined as calculating the probability that a random amino acid replacement will lead to a protein's functional inactivation. Using the human DNA repair enzyme 3-methyladenine DNA glycosylase (AAG), we develop a method to calculate inactivation probabilities from libraries of AAG mutants harboring random mutations throughout the gene. This analytical method is then applied to a

range of diverse proteins. Remarkably, inactivation probabilities were observed to be similar among many proteins. To delineate the nature of tolerated mutations, 244 surviving AAG mutants were sequenced. Over 920 tolerated mutations were assembled into "substitutability indices" of each amino acid position across the entire AAG gene and mapped onto secondary and tertiary structures. I discuss the general factors determining the tolerability of amino acid substitutions within the same chapter.

The next section of this dissertation describes the use of targeted random oligonucleotide cassette mutagenesis to study the AAG enzyme active site. Selection for altered properties yield novel human DNA glycosylases with altered active sites. This study also reveals the degree of plasticity of the human AAG substrate recognition pocket and highlights the essential residues for substrate recognition and catalysis.

This dissertation demonstrates a general approach to understanding proteins' tolerance to random mutations, and to creating novel DNA glycosylases. This work can serve as a stepping stone toward new enzymes remove select altered DNA bases for potential therapeutic and biotechnological applications.

## TABLE OF CONTENTS

Page

List of Figures ..... iii

List of Tables ..... iv

Chapter 1. DNA damage, repair, and directed evolution ..... 1
Introduction: DNA damage and repair ..... 2
The base excision repair pathway ..... 3
Uracil DNA Glycosylase (UDG) ..... 5
The helix-hairpin-helix (HhH) superfamily ..... 6
Alkyl-adenine DNA glycosylase (AAG) ..... 9
Directed Evolution ..... 12
PCR mutagenesis ..... 14
Random oligonucleotide cassette mutagenesis ..... 15
Questions: ..... 16
Sequence and degree of protein tolerance ..... 16
Crucial AAG active site residues ..... 17
Creating novel DNA glycosylases ..... 19

Chapter 2. Measuring protein tolerance to random amino acid substitutions . 29
Introduction ..... 30
Materials and methods ..... 31
Materials and strains ..... 31
Construction of PCR mutagenesis libraries ..... 31
Genetic selection for active enzymes, DNA sequencing ..... 33
AAG activity assay ..... 35
X-factor calculation and mutability visualization ..... 36
Results and discussion ..... 36
Protein tolerance to random amino acid substitutions ..... 36

The x-factor and the mutability of proteins ..... 39
Substitutability and AAG structure ..... 43
Some implications of the x-factor ..... 47

Chapter 3. Novel alkyl-DNA glycosylase active sites revealed by random mutagenesis and selection ..... 63
Introduction ..... 64
Materials and methods ..... 66
Materials and strains ..... 66
Construction of the mutant AAG libraries ..... 67
Construction of the YH library ..... 70
Genetic selection for mutants active against MMS lesions ..... 71
Genetic selection for mutator DNA glycosylases ..... 71
DNA sequence analysis ..... 72
Results and discussion ..... 72
AAG complementation of *E. coli* against MMS killing provides a stringent selection for active AAG mutants ..... 72
Targeted, "semi-rational" libraries of AAG substrate binding pocket mutants ..... 72
Selection of mutant AAGs that enhance survival after DNA alkylation damage ..... 74
Selected mutants reveal tolerated substitutions and variant glycosylase active sites ..... 74
a. Mutator DNA glycosylases ..... 76

Chapter 4. Concluding remarks ..... 85

Bibliography ..... 88

## LIST OF FIGURES

Figure Number Page

1.1 Schematic representation of principle steps of base excision repair ..... 22
1.2 Comparison of the AlkA and EndoIII active sites..... 23
1.3 Substrates of the human 3-methyladenine DNA glycosylase..... 24
1.4 Structural features of AAG ..... 25
1.5 Overview of random mutagenesis methods ..... 26
1.6 Representation of random oligonucleotide cassette..... 27
1.7 Schematic representation of protein sequence space ..... 28
2.1 Distribution of amino acid changes among unselected mutants..... 56
2.2 MV1932 dose response to the DNA alkylating drug MMS..... 57
2.3 Survival of the mutated libraries, normalized to that of wild-type..... 58
2.4 MV1932 survival at 0.2% MMS, normalized to wild-type AAG survival..... 59
2.5 Amino acid changes per mutant in the AAG PCR mutagenesis libraries..... 60
2.6 Tolerated amino acid changes along the AAG primary sequence ..... 61
2.7 Substitutability of AAG amino acid residues in relation to structure..... 62
3.1 Scheme of AAG MV1932 complementation..... 79
3.2 Unselected AAG random oligo mutagenesis library sequences..... 80
3.3 Sequences from selected AAG active site mutants ..... 81
3.4 Histograms of mutational loads, unselected and selected AAG mutants ..... 82
3.5 Mutant AAG increases mutation frequency in *E. coli*..... 83
3.6 The wild-type and Y127I, H136L double mutant active sites..... 84

## LIST OF TABLES


| Table Number | | Page |
|---|---|---|
| 1.1 | Mechanistic features of DNA glycosylases | 21 |
| 2.1 | DNA oligonucleotide sequences | 50 |
| 2.2 | PCR Mutagenesis conditions | 51 |
| 2.3 | Mutational biases of distinct mutagenic PCR protocols | 52 |
| 2.4 | Calculating the X-factor | 53 |
| 2.5 | X-values calculated from active site targeted cassette mutagenesis studies | 54 |
| 2.6 | Mean mutability indices of AAG motifs | 55 |
| 3.1 | MV1932 survival after MMS treatment | 78 |

## ACKNOWLEDGEMENTS

I would like to extend my heartfelt thanks to the following people who have made these works possible.

Larry Loeb has been my guiding mentor in our endeavors. His vision enabled the formulation of these projects and his encouragement and constant support have sustained them. His sense of optimism and scientific curiosity are inspirational. I would like to thank him for the great company in the lab and on the slopes, whether it be in the Washington Cascades or in the Swiss Alps. I would also like to thank Larry's wife, Phyllis, for the many happy get-togethers and her sumptuous food.

My committee members: Drs. Ray Monnat, Barry Stoddard, Charles Murry, and graduate school representative Albert La Spada. All of whom have also been my teachers in various medical and graduate school courses. They have challenged, encouraged, and provided kindly advice. Dr. Monnat has been a steady scientific consultant, providing critique, experimental insights, and good humor. I would also like to thank Dr. Stoddard for his advice, and his lab members Greg Ireton and Django Sussman for their generous help. Without them, the gorgeous structural and mutability visualizations in chapter 2 would have remained as figments of imagination.

Drs. Elinor Adman and Christophe Verlinde of the Molecular Structure Visualization Center have been invaluable for their analysis of mutants and structural discussions. Ellie has been truly an indispensable extended member of the lab. She has always graciously accommodated our request for her time and expertise.

My collaborator on the protein mutability project Juno Choe has been a source of challenge and expertise. His proficiency in computer programming is remarkable and has greatly facilitated the large amount of data analysis. Our friendship dates to our college days, and I wish him the best in his future studies and training.

Our collaborator Dr. Leona Samson at MIT, who has readily provided reagents, unpublished data, and valuable experimental insight.

Of course, these projects were propelled along with the help of present and past fellow lab members in the trenches of science. Eitan Glick has an almost uncanny ability to come up with a better or easier way, and seemingly endless patience for graduate students' cries for help. My thanks also to John Davidson, Jon Anderson, Premal Patel, Daisuke Umeno, James Shen, Ashwini Kamath-Loeb, Michael Fry, and Manel Camps for their expert advice. Lance Encell, Hisaya Kawate, Jessica Sneeden, Ern Loh, Jason Bielas, Jessica Hsu, Ranga Venkatesan, and honorary lab member Ronald "Rover" Cheung. Your friendships made the going easier and mean a lot to me. Your unique personalities made the lab an always more interesting place to be. My thanks to Ann Blank, who has been a voice of logic and careful editor of manuscripts; and to Cory Heindel and his legion of student helpers, for operating an efficient lab where it is a pleasure to do science.

# *Chapter 1. DNA damage, repair, and directed evolution*

**Chapter Summary:**

The genome is continuously subjected to insults by chemical and physical agents. Life has evolved numerous and overlapping DNA repair mechanisms to preserve the fidelity of the genetic material. Members of the base excision repair (BER) pathway remove diverse types of inappropriate base structures from DNA and restore the normal primary sequence. DNA glycosylases initiate this pathway by recognizing and removing a wide spectrum of lesions. Crystal structures of DNA glycosylases complexed to respective help shed light on this specific process. In this introductory overview, we examine prototypical members of this class of enzymes and focus on the structural basis for substrate recognition.

Members of the DNA glycosylase family provide facile experimental systems to study basic evolutionary questions, such as the tolerance of proteins and their active sites to random change. Directed evolution is general approach that can be used to address some of these questions. I will discuss the questions of interest and the approach to address them.

## INTRODUCTION: DNA damage and repair

Life carries and propagates its genetic information in the linear sequences of DNA. However, as strings of chemical compounds measuring nearly 2 meters long in an average human cell, DNA is susceptible to a variety of reactions that can alter its coding fidelity. Many agents, both endogenous and environmental, can attack DNA [1]. Among them include reactive oxygen species (ROS), generated from aerobic respiration and ionizing radiation, that can oxidize the complement of DNA bases [2]. S-adenosylmethionine, an intracellular molecule, can directly alkylate bases [3]. Lipid peroxidation products include acrolein and crotonaldehyde, which are metabolized to epoxides and can generate exocyclic etheno modifications of DNA bases. Mutagens from tobacco smoke and industrial processes, in addition to UV light exposure, are responsible for large numbers of DNA lesions that have been shown to produce cancer [4].

Damaged DNA bases may be cytotoxic, mutagenic, or both. These lesions are believed to have key roles in the pathobiology of cell senescence, aging, and cancer. Understandably, cells have evolved a variety of overlapping mechanisms that restore the fidelity of the normal primary DNA sequence after damage. Direct repair proteins, such as the human alkyl-guanine methyltransferase (AGT) remove the lesions in a single step reaction [5]. Enzymes of the Nucleotide Excision Repair (NER) pathway assemble into multi-unit complexes that recognize bulky, helix-distorting adducts, including dipyrimidine photoproducts, and replace a 24-32 base single strand oligonucleotide flanking the lesion [1]. Mismatch repair enzymes of the long

patch and short patch pathways recognize and remove specific base mispairing. Somatic recombination, overlapping of extensive excision repair tracts, ionizing radiation, and oxidative insults are all processes that produce double strand breaks. Proteins in the double-strand break repair pathway can initiate homologous recombination and non-homologous end-joining as means to repair the damage [6].

*The base excision repair pathway*

Base Excision Repair (BER) is another pathway that purges a large variety of damaged bases from the genome. Five enzymatic activities are required for complete BER. DNA glycosylases initiate BER by recognizing and removing inappropriate and corrupted bases. All DNA glycosylases cleave the C1′-*N* glycosylic bond between the target base and deoxyribose, thus freeing an aberrant base and leaving an apurinic/apyrimidinic (AP) site. 5′-AP endonuclease cleave the 5′ phosphate-deoxyribose bond and thus produces a single strand break. From this point on, the repair steps can proceed down one of two pathways. In the "short patch" process, the 5′-deoxyribo-phosphodiesterase activity of DNA polymerase β (pol β) removes the sugar-phosphate, leaving a gap in the DNA. Pol β then fills this gap, and the remaining DNA strand break is sealed by DNA ligase III. In the "long patch" pathway, polymerase δ or β extend the 3′OH end approximately 7 nucleotides, displacing the strand containing the 5′terminal deoxyribose phosphate. FEN1 (flap endonuclease or DNase IV), in association with Proliferating Cell Nuclear

Antigen (PCNA), excises the flap-like structure produced by polymerase displacement. The remaining nick is sealed by DNA ligase I [7]. Some DNA glycosylases possess additional AP lyase activity and are able to cleave the phosphodiester bond 3′ to the AP site, leaving the AP site with a 3′ nick. For schematic summary of BER steps and pathways, please see Figure 1.

The AP sites that result from DNA glycosylase activity are themselves mutagenic and cytotoxic. In the utilization of the BER pathway, cells subject themselves to the potential added toxicity of mismanaged intermediates. Indeed, *S. cerevisiae* deficient in the APN1 AP-endonuclease have an increased spontaneous mutation rate compared to wildtype [8]. This highlights that the activities of the BER pathway must be balanced for optimal protection against the biological consequences of damaged DNA bases. Until the last step of BER has been completed, damage is still present in the genome [9]. Parikh and workers, based on the observation that UDG binds AP sites, have proposed a mechanism in which UDG reduces the toxicity if the exposed AP site by binding and protecting it until downstream factors of the BER pathway HAP1, Pol β, and ligase, can come to the AP site and displace UDG [10]. Next we will examine prototypical members of the DNA glycosylase family of enzymes. For a summary of the key mechanistic features of these enzymes, see Table 1.

*Uracil DNA Glycosylase (UDG)*

Uracil is produced in the genome by cytosine deamination, leading to G:C to A:T transversion mutations. Uracil residues are also incorporated instead of thymine by the promiscuous use of deoxyuridine triphosphate (dUTP) as a precursor for DNA synthesis. Incorporation of U instead of T is not directly mutagenic but results in altered binding affinity for transcription regulatory proteins [11]. The major biological function of UDG is to excise uracil from DNA. UDG is the first known DNA glycosylase [12]. Its importance is highlighted by its conservation from bacteria to human and even some viruses [13]. In a recent series of exciting discoveries, uracil DNA glycosylase has been shown to play pivotal roles in both antibody diversification and targeted viral hypermutagenesis [14]. In somatic hypermutation of $F_V$ regions, activation induced deaminase (AID) converts cytosines into uracil. UDG acts directly down-stream to generate AP sites, which can miscode, or be converted into single strand breaks and be extended by error-prone DNA polymerases. Alternatively, in innate defenses against retroviral infection, CEM15/APOBEC3G deaminates cytosines of the first strand cDNA synthesized by viral reverse-transcriptase. The resulting uracils can then be targeted by UDG, or left to miscode [15].

The crystal structure of UDG bound to uracil containing DNA provided the first evidence of the nucleotide flipping mechanism hypothesized to be common among all DNA glycosylases. Flipped bases were first described for *Hha*I and *Hae*III methyltransferases [16] [17]. For UDG, it is observed that DNA

binds along a positively charged groove in the enzyme, which appears to progress along the minor groove until a uracil is encountered. The enzyme then kinks the DNA by compression of the flanking backbone of the same strand as the lesion and flips out the abnormal nucleoside residue into a specific pocket [13]. Extensive shape and electrostatic complementarity to uracil is observed within this pocket. Uracil stacks with Phe-158. Hydrogen bond pairs exist between Uracil N1 – His-268, U O2 – Gln-144, U O2 – His-268, U N3 – Asn-204, U O4 – Asn-204, and van der Waals interaction exist between U C5 and Tyr-147. The tyrosine ring excludes thymines due to steric clash with its 5-methyl groups. Purines are excluded by steric clashes with their bulky imidazole rings. Interestingly, the tight substrate recognition range of UDG can be altered by amino acid substitutions of key residues. The replacement of Asn-204 by aspartate or Tyr-147 by alanine, cysteine, or serine results in enzymes that have cytosine DNA glycosylase activity or thymine DNA glycosylase activity, respectively [18]. The cleavage of the C1′-*N* glycosilic bond is mediated by Asp-145, which activates a water molecule via deprotonation, which in turn breaks the C1′-*N* bond by nucleophilic attack.

***The Helix-hairpin-Helix superfamily***

Many DNA glycosylases belong to the helix-hairpin-helix (HhH) superfamily, including the *E. coli* MutY, endonuclease III, AlkA, and the human hOGG1. No overview of DNA glycosylase would be complete without a discussion of this family. Of greater relevance to this dissertation, is that it

appears nature, the ultimately molecular tinkerer, has experimented with using a common fold but altering the active site to accommodate different substrates. These related proteins adopt similar three-dimensional folds. The hallmarks of these enzymes is the presence of the named HhH motif and a Gly/Pro-rich stretch with nearby Asp (GPD) motif. The HhH motifs are involved in formation of a hydrophobic core that mimics the symmetry of the DNA double helix and enables strong DNA binding. Protein-DNA contacts involve a sugar-phosphate chain in a nonsequence-specific manner. Interestingly, the fusion of HhH motifs with *Taq* and *Pfu* DNA polymerses produced more processive and salt-resistant chimeric DNA polymerases [19].

Homologs of EndoIII have been found in gram-negative bacteria, gram-positive bacteria, Crenarchaeota, Euryarcaeota, fission yeast, budding yeast, invertebrates, and mammals [20]. It appears that the structure of family members contain an evolutionarily ancient protein fold that have been used many times to serve as the basis for nonspecific DNA binding and catalysis of glycosilic bond cleavage. To date, six distinct sets of substrates have been identified for members of this family, including: adenine mismatched with oxo-G (MutY), oxidized pyrimidines (EndoIII), oxidized guanine (hOGG1), alkylated purines (AlkA), pyrimidine dimer (*M. luteus* UV-endonuclease) and T/G mismatch (*Methanobacterium thermoautotrophicam* T/G mismatch glycosylase).

MutY is a DNA glycosylase that protects the genome against reactive oxygen species. As a member of the MutT, MutM (Ogg1, FPG), and MutY triad that synergistically protect *E. coli* against oxoG mutagenesis, MutY removes

Adenine mispaired opposite guanine and oxoG [21]. The crystal structure of MutY reveal a sterically tight adenine-mismatch pocket. Adenine mispaired with G and oxoG generally assume a *syn* orientation [22]. The steric clash of the 8-oxo group of oxoG with Glu-37, Leu-40, and Gln-42 of the minor groove reading motif, along with compression of the DNA phosphodiester backbone, may lead to nucleotide flipping of the mispaired adenine. Further specific interactions may come from a strictly conserved Gln-41 with the adenine, and the interaction of Arg-194 with the orphan guanine or oxoG [13].

In addition to reactive oxygen species, the genome is also continuously challenged with alkylation damage. 3-methyladenine (3-MeA) is one lesion, among many, that is known to block *in vitro* DNA synthesis by bacterial and viral DNA polymerases [23]. 3-MeA in the murine genome was shown to inhibit in vivo DNA replication and slow cell progression through the S-phase [24]. DNA glycosylases that remove 3-MeA are found in bacteria, yeast, plants, rodents, and humans [9]. *E. coli* posses two genes for the 3-MeA removal: Tag and AlkA. Tag is constitutively expressed and possesses high specificity for 3-MeA. AlkA is induced as a part of the *E. coli* adaptive response to alkylation damage [25]. In addition to removing 3-MeA, it has been reported that AlkA removes a large and diverse range of alkylated bases, oxidized bases, and even deaminated bases [26]. Although the preferred substrates are likely to be alkylated purines [27]. Interestingly, AlkA has even been shown to remove the normal A, T, C, G bases from intact DNA in vitro and induce spontaneous mutations when overexpressed in *E. coli* [28].

It is of interest to analyze the residues within these enzymes' active sites the examine the differences that account for the variations in substrate recognition. Two family members from which extensive crystal structure information is available are EndoIII and AlkA [7,27,29-31]. Endonuclease III from *E. coli* was originally isolated as DNA nicking activity after heavy UV irradiation [32]. It removes numerous forms of oxidized and ring fragmented pyrimidines [29]. Its 1.8 Å crystal structure was among the first of the DNA glycosylases to be elucidated [33]. AlkA has a recognition pocket rich in aromatic residues that may interact with electron-deficient substrates, such as alkylated purines, through $\pi$ donor / $\pi$ acceptor interactions. The open architecture of the AlkA active site help to account for its binding to bulky alkylation adducts, such as 7-(2-chloroethyl)-dG [34]. On the other hand, the cleft of Endonuclease III is lined with hydrophilic resides that would interact favorably with oxidatively damaged bases, such as thymine glycol. For visual comparisons of the two active sites, please see Figure 2.

***Alkyl-Adenine DNA Glycosylase, AAG***

Similar to AlkA, the human 3-MeA DNA glycosylase removes a large range of alkylation adducted DNA bases, including 3-methyladenine, 7-methylguanine, and 3-methylguanine [9,35]. AAG also removes hypoxanthine, which derives from adenine deamination, and the lipid peroxidation derived base lesion 1,$N^6$-ethenoadenine (εA) [36]. For the chemical structures of the substrates removed by AAG, see Figure 3. The AAG cDNA was cloned by

complementation of alkylation repair deficient *E. coli* [37]. AAG does not belong to the HhH superfamily, as the crystal structure reveals a different fold from AlkA and other members of this family [38]. However, the active site of AAG is also rich in aromatic residues. Three critical residues, His-136, Tyr-159, and Tyr-127 appear poised for π-electron interactions with bases flipped out into the active site. Mouse knock-outs missing AAG appear healthy and do not exhibit an obvious phenotype [39]. However, AAG -/- cells are sensitive to alkylating agents when challenged [24]. Homologs of 3-methyladenine glycosylases are absent from known sequences of the drosophila genome [40]. In contrast, at least two AlkA and six Tag homologs are found in *Arabidopsis*, suggesting a greater need to repair alkylated DNA lesions among plants [41].

The crystal structure of AAG complexed with substrate containing oligonucleotide have provided the most detailed look at the enzyme's interactions with its substrate [38,42]. For detailed views, please refer to figure 4A-D. AAG contacts the phosphodiester backbone through a flat, positively charged DNA binding surface. A protruding β hairpin juts into the DNA minor groove. Tyr-162 is located at the apex of this hairpin and displaces the nucleotide targeted for base flipping into the active site pocket. Tyr-162 thus acts as a DNA intercalator, filling the gap left by the displaced nucleotide. The apparent ease of the nucleotide to being base-flipped provides one level of substrate discrimination. Lesions such as εA lose their Watson-Crick hydrogen bonding capability and hence are more easily rotated to an extra-helical conformation than normal basepairs.

Once within the active site pocket, the compatibility in hydrogen bonding character, aromaticity, and charge between the base lesion and those of the active site contribute further to substrate discrimination. The damaged base stacks between the aromatic side chains of Tyr-127 on one side and those of His-136 and Tyr-159 on the other. His-136 also contributes to the stability of the base-enzyme complex by hydrogen bonding with the 5′ phosphate of the flipped-out base and hydrogen bonding with Tyr-157 (Figure 4C).

The abundance of aromatic side chains within the AAG active site confers it with another degree of specificity toward alkylated purine adducts. The electron-withdrawing tendency of the adducted alkyl groups confers a positive charge on the bases. The positively charged bases form more favorable π-electron donor-acceptor stacking interactions with the enzyme aromatic side chains than neutral electron stacking interactions [43]. The electron withdrawing aberrant alkyl groups also weaken the glycosylic bond, making them good leaving groups with reduced need for catalytic assistance [28]. Leu-180 also makes contact with the flipped out base and contributes additional van der waals contact surface area.

Asn-169 is another critical residue in the AAG active site. The side-chain of Asn-169 sterically clashes with exocyclic $N^2$ amino group of guanines, helping to prevent normal guanines from being been gratuitously removed (Figure 4D). It is hypothesized the positive charge of 7-methylguanine and 3-methylguanine help to overcome this steric clash and enable their removal by AAG [42].

Finally, the cleavage of the C1′-*N* glycosylic bond is accomplished by an activated water molecule that has been deprotonated by Glu-125. Arg-182 and the main chain carbonyl of Val-262 further stabilize the water molecule through hydrogen bonding.

The DNA glycosylases provide fascinating case studies of protein-DNA interaction. Their conserved functions across kingdoms, as well as the example of convergent and divergent evolution, highlight their fundamental importance to the process of life. DNA glycosylases are small monomeric proteins that do not require cofactors for activity; and as such, are convenient models for studying interactions between damaged base and proteins. Many readily complement simple model organisms, such as *E. coli* or yeast, that are deficient in the corresponding biochemical activity. When expressed in the DNA alkylation repair deficient *E. coli* strain MV1932 (*alkA, ada*), AAG confers greater than 4 orders of magnitude in cell survival after challenge with the DNA alkylation drug methyl methanesulfonate (MMS). Thus providing a stringent selection for functional enzymes.

## DIRECTED EVOLUTION

Natural processes of evolution have enabled an astonishing diversity of life on earth. Through billions of years of exploration and alteration through continuous selection, organisms have developed capabilities that often surpass the imagination. Just as nature has used evolution to select for proteins that provide an advantage to the host, directed evolution in the laboratory uses

techniques to select or screen for mutant proteins that perform better than wild-type proteins under selected conditions. In contrast to "rational" protein engineering through site-directed mutagenesis, directed evolution techniques do not demand detailed understanding of the target protein in order to develop useful variants. Directed evolution however, does require a selection or screening step to enrich for mutants of interest.

"Semi-rational" random mutagenesis coupled with selection is also a powerful method in the study of protein structure-function relationships. In contrast to "rational" specific mutagenesis that examine the effect of one particular substitution at one particular position in the protein, random mutagenesis techniques can be used to generate upwards of millions of mutants that are then selected or screened for desired properties. Active mutants can reveal the degree of plasticity of the region of interest and highlight a range of tolerated substitutions that can confer novel activity to the target protein. Directed evolution can also often reveal unexpected biological insights into protein function. For example, the Loeb laboratory has used this approach to isolate mutant proteins involved in DNA replication and repair, such as low fidelity DNA polymerases, drug resistant thymidylate synthases, and more active herpes thymidine kinases [44-48].

A comprehensive review of all random mutagenesis techniques is beyond the scope of this dissertation. I will instead briefly focus upon two methods used in this work, PCR mutagenesis and random oligonucleotide cassette

mutagenesis. For a schematic overview of random mutagenesis methods, please see Figure 1.5.

*PCR Mutagenesis*

Polymerase chain reaction (PCR) amplification using the DNA polymerase *Taq*, which lacks a proofreading mechanism, can produce a library of mutated genes. As a gene or gene segment is amplified via PCR, mutations accumulate in the pool of amplified products, which can be cloned and analyzed for altered activity. Under normal conditions, wild-type *Taq* produces approximately one mutation in every 9000 bases that it synthesizes [49]. The mutation rate can be increased by several techniques. The use of manganese instead of magnesium as the divalent cation greatly increases the mutation frequency of DNA polymerases [50]. Biasing the ratios of natural deoxyribonucleotide triphosphate (dNTP) pools also increases the frequency of misincorporations and frameshifts [51]. PCR with various nucleoside analogs, including 8-oxo-dGTP and dPTP, have been successful in creating libraries of mutant genes [52,53]. Finally, the creation of mutant reduced-fidelity *Taq* polymerases have led to efficient ways of creating mutant libraries [54,55]. The region being mutated can be selected by the use of primers for a specific region of the gene. The number of mutations can be determined by altering the PCR reaction conditions, and by varying the number of rounds of amplification.

PCR mutagenesis, however, exhibit several disadvantages when compared with other methods. The ability of this technique to create high

mutation load within a specific, narrow region of the gene is limited. Furthermore, existing PCR mutagenesis protocols are limited by significant biases in the mutation spectrum, with *Taq* preferentially mutating at A:T basepairs [46,56,57]. These biases limit the ability to effectively search regions of sequence space. Significant work has been done within this body of work to overcome this bias, and will be further discussed in chapter 2. Even with these limitations, PCR mutagenesis is a relatively straightforward and efficient technique to produce mutations across the entirety of a protein.

***Random oligonucleotide cassette mutagenesis***

Random oligonucleotide incorporation is an efficient method to generate mutations within a limited region of a gene. Degenerate oligonucleotides are used to replace a portion of the gene. This creates an instant library of mutant genes that contain a random assortment of substitutions over the region of interest. The randomness of the library can be carefully titrated by adjusting the fraction of the wild-type nucleotides at a position compared to the other three degenerate nucleotides [58]. For a schematic of a random oligonucleotide, see Figure 1.6.

The advantages of random oligo mutagenesis include the ability to target a specific region of a gene and randomly mutate the gene at precisely controlled levels. By controlling the degree of degeneracy of base substitution in the synthesis of the DNA oligo, the mutational bias can also be accounted for when analyzing any resulting mutants. Numerous DNA oligonucleotide vendors are

available. Proteins created by this method can then be selected or screened for activity.

One disadvantage of this technique is the fact that only a small region of the gene can be mutated at a given time. Although with the advent of custom oligo synthesis reaching the length of tens of kilobases, this limitation is being overcome. More importantly, this method also requires a fair amount of information be already known about the protein of interest, so that a "semi-rational" decision can be made on where to mutate the enzyme before the commencement of mutagenesis. Useful information are usually those from structure studies or conservation alignments that highlight catalytic domains or recognition pockets. Although the requirement for the level of *a priori* knowledge is great, it is less daunting than the information needed for precise "rationally" designed variant created using the method of site-directed mutagenesis.

## QUESTIONS

### *Sequence space and degree of protein tolerance*

One underlying idea in directed evolution and protein evolution is the concept of a "sequence space." A protein's sequence space is defined as the connected network of all possible sequences stemming from original sequence [59]. The sequence space of any protein is astronomically large. For example, the sequence space for a protein of 300 amino acids would contain $20^{300}$ points. Clearly it is not possible to explore all possible points.

Within an enzyme's local multi-dimensional sequence space, it is possible to explore surrounding maxima and minima of activity toward a defined substrate. This sequence search can be visualized as navigating an artificial landscape of peaks and valleys. The topology of the landscape is determine by complex factors, including the enzyme's affinity toward the substrate, its ability to stabilize the transition state, and release of the reaction products after catalysis. More fundamentally, mutations that affect the stability of the overall protein structure would also impact the local topology.

Broadly, new mutations can be beneficial to protein function, or be neutral or deleterious. It has been asserted that proteins are "highly plastic" to random mutations. Such assertions were often based on the isolation of active mutants harboring multiple mutations from random mutagenesis libraries, and from the wide sequence divergence of homologous enzymes that perform similar functions from diverse organisms. To date, though, there lacks a quantitative estimate of the *degree* of tolerance a protein can exhibit toward random mutations at a random site in its sequence. A significant portion of this dissertation is aimed at developing a general method to calculate this degree of tolerance, or conversely, the degree of sensitivity to random inactivating mutations. This method is applied to the study of the human AAG and other proteins and will be discussed in greater detail in chapter 2.

*Crucial AAG active site residues*

DNA glycosylase present an interesting paradox. First, there are "specialist" DNA glycosylases that show exquisite selectivity toward their substrates, even in the presence of overwhelming ratios of closely related structural relatives, as in the examples of UDG, hOGG1, and MutY. In contrast, "generalist" enzymes such as EndoIII, AlkA, and AAG recognize a wide range of adducts. The unifying theme of the first group of enzymes, gleamed from their crystal structures, are that strategic positioning of amino acid residues that closely discriminate between substrate and non-substrate through steric hindrance can accomplish the high specificity demanded. The generation of cytosine and thymine DNA glycosylases from UDG through single amino acid substitutions that alter the binding site geometry corroborate these observations [18]. On the other hand, the wide range of adducts recognized by EndoIII, AlkA, and AAG seem to point the way toward a different type of substrate recognition mechanism. For AlkA, a strategy of recognizing electron-deficient alkylated bases via $\pi$-electron stacking interactions has been proposed [27,31]. Instead of being targeted for excision by their abnormal shapes, alkylated purines might be recognized for their electron deficient rings and positive charge. Despite the lack of similarity between the overall three-dimensional structure of AAG and AlkA, it has been proposed that the two proteins appear to use a similar $\pi$-electron strategy to recognize their diverse array of substrate bases [38]. This remains to be tested in greater detail.

Prior studies with Motif A and Motif B of DNA polymerases reveal that although residues are very conserved throughout evolution, many of these sites, in fact, tolerate a large range of amino acid substitutions and still yield functional enzymes [44,45,60]. This raises the unexpected possibility that one cannot easily predict the requirement for a specific residue by simply looking at its degree of conservation. These observations also raise the intriguing proposition that enzymes are highly plastic and can tolerate a variety of substitutions. We further extend these studies to the active site of AAG, which contains numerous evolutionary conserved residues. The findings from these studies will be further described in chapter 3.

*Creating novel DNA glycosylases*

The goal of directed evolution is often to create enzyme variants with altered activities. We have selected for variant DNA glycosylases that confer increased mutation frequency in cells. The goal is to show that DNA repair mechanisms, when damaged or dysregulated, can paradoxically produce mutations. It has been proposed the spontaneous mutation rate in somatic cells is insufficient to account for the multiple mutations observed in human cancers, and that cancer cells are genetically unstable [61-63]. Defects within the BER pathways may produce increased mutation rates consistent with that of a "mutator phenotype" [64]. The results from these investigations will be further detailed in chapter 3.

The 3-methyladenine DNA glycosylase family is also of potential interest in cancer therapy, as that mammalian members appear to possess some activity in removing adducts formed by nitrogen mustards, a class of chemotherapeutic drug, but at much reduced rates than the bacterial AlkA protein [65].

At the clinical level, many chemotherapeutic drugs derive their anti-tumor effects by damaging DNA. However, patients are subjected to toxicities due to nonspecific damage to their normal cells. It may be possible to develop novel DNA glycosylases that repair the damage inflicted by these drugs by altering the substrate recognition sites of existing glycosylases. In order to alleviate negative side-effects, genes coding for such novel enzymes can be transduced into selected cells by targeted gene therapy.

**Table 1.** Mechanistic features of DNA glycosylases

| Protein | Typical Substrate | Base Recognition | Nucleophile |
|---|---|---|---|
| Udg | **U** | π–stacking, steric clash, highly specific H-bonds | $H_20$ |
| Mug | **U**:G | π–stacking, Non specific H-bonds | $H_20$ |
| hOGG1 | **oxoG**:C | π–stacking, highly specific H-bonds | Lys249 |
| MutY | **A**:oxoG | highly specific H-bonds | $H_20$ |
| Endo III | **TG** | H-bonds | Lys120 |
| AlkA | **3-mA**, **7-mG** | strong π–stacking | Asp238 |
| AAG | **3-mA**, **εA**, **Hx** | strong π–stacking, steric clash | $H_20$ |

Abbreviations: oxoG, 7,8 dihydro-8-oxoguanine; TG, thymine glycol; 3-mA, 3-methyladenine; 7-mG, 7-methylguanine; εA, 1,$N^6$-ethenoadenine. Adapted from [66].

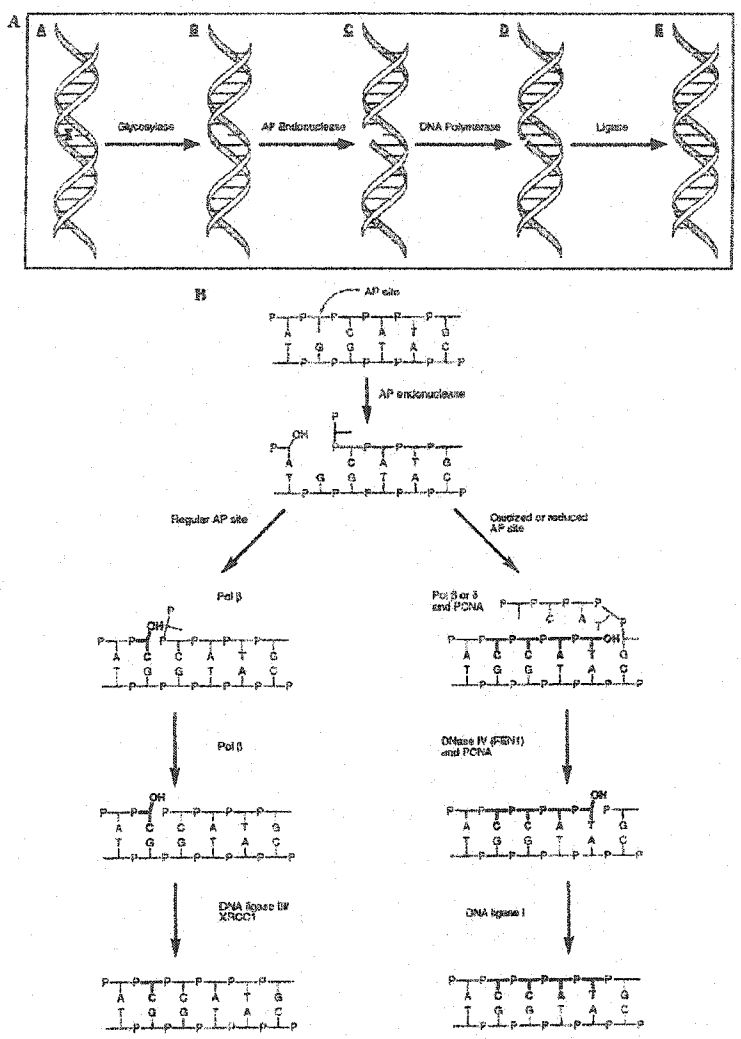



Figure 1.1

**Schematic representation of principle steps of base excision repair**
A). Base modifications are initially recognized and removed by DNA glycosylases. The DNA strand is cleaved, either by the lyase activity present in some DNA glycosylases or AP endonuclease. A DNA polymerase synthesizes the replacement nucleotide and the phosphodiester backbone is sealed by a ligase.
B). In the short-patch pathway (left), polymerase β synthesizes one nucleotide and removes the 5′-terminal baseless sugar via its lyase activity, supported by the scaffolding protein XRCC1. DNA ligase III seals the remaining nick. In the long-patch pathway (right), pol δ/ε, supported by the replication factor PCNA, synthesizes a repair tract of 2-6 bases. FEN1 excises the overhanging flap, and DNA ligase I seals the remaining nick. Until the final ligation step is completed, toxic and mutagenic DNA repair intermediates, such as AP sites and single-stranded breaks, are still present in the genome. Figure adapted from [7].

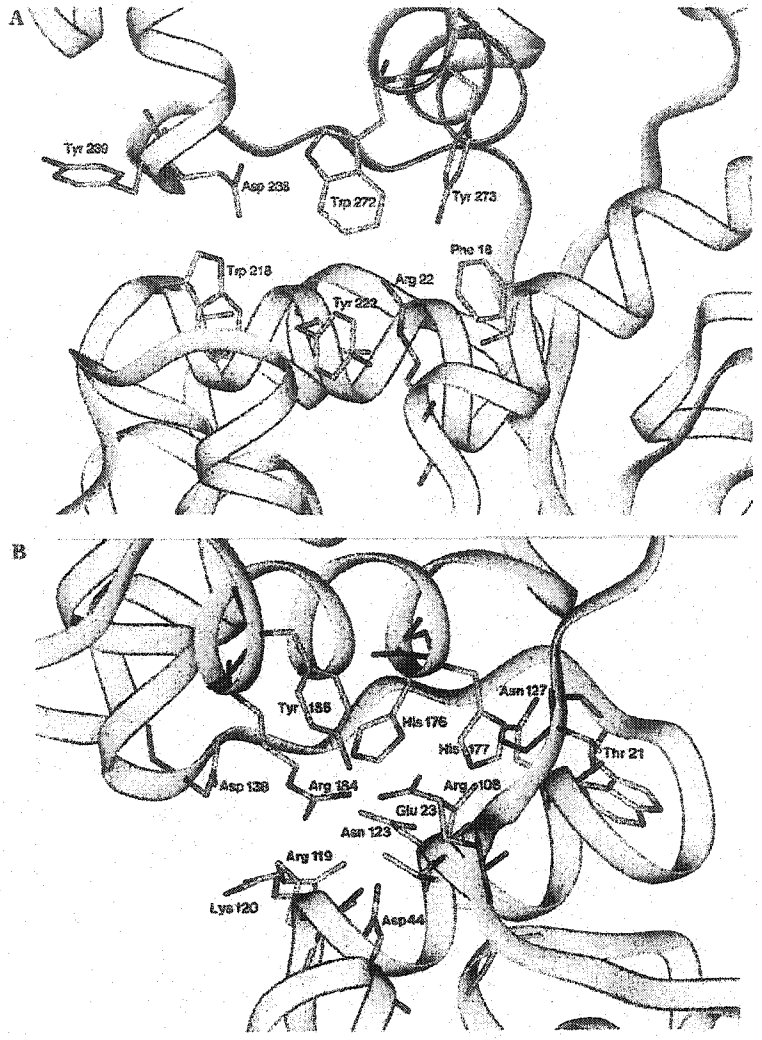



Figure 1.2

**Comparison of the AlkA and EndoIII active sites**
Both AlkA and EndoIII are members of the helix-hairpin-helix (HhH) superfamily of DNA glycosylases. A). Note the abundance of aromatic sidechains in the AlkA active site, which recognizes slightly positively charged alkylated bases.
B). Note the richness of basic residues in the EndoIII active site, which recognizes negatively charged DNA lesions stemming from oxidative damage.
Adapted from [7].

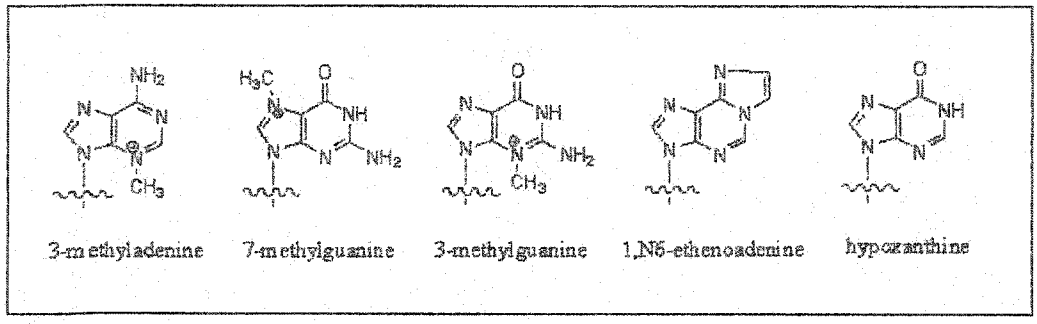



Figure 1.3

**Substrates of the human 3-methyladenine DNA glycosylase**

3-mA, 7-mG, and 3-mG are lesion deriving from methyl alkylation damage. εA derives from lipid peroxidation aldehydes. Hypoxanthine is an oxidative deamination product of adenine.

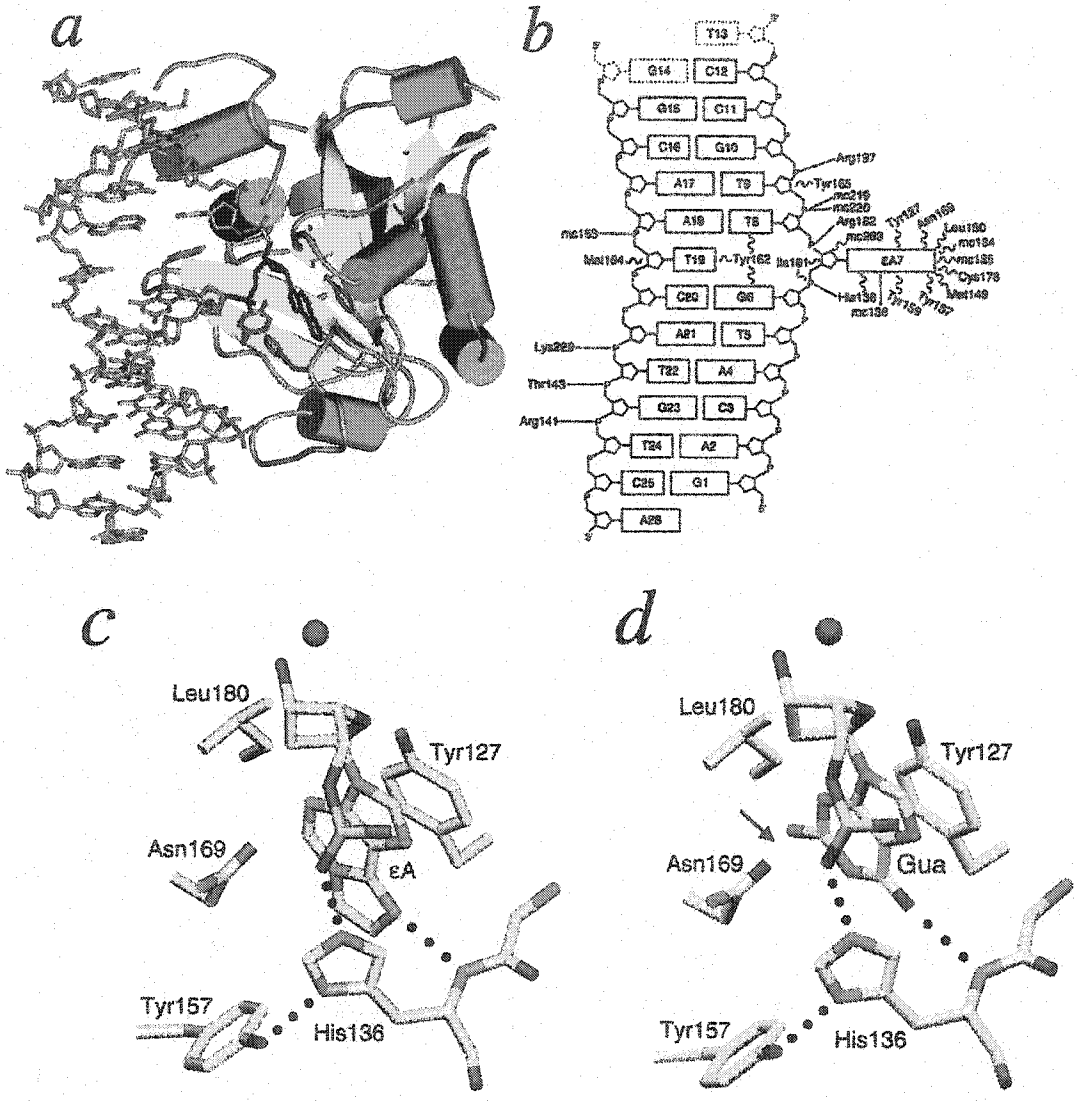



Figure 1.4

**Structural features of AAG**
See text for discussion of salient features.
Adapted from [42].

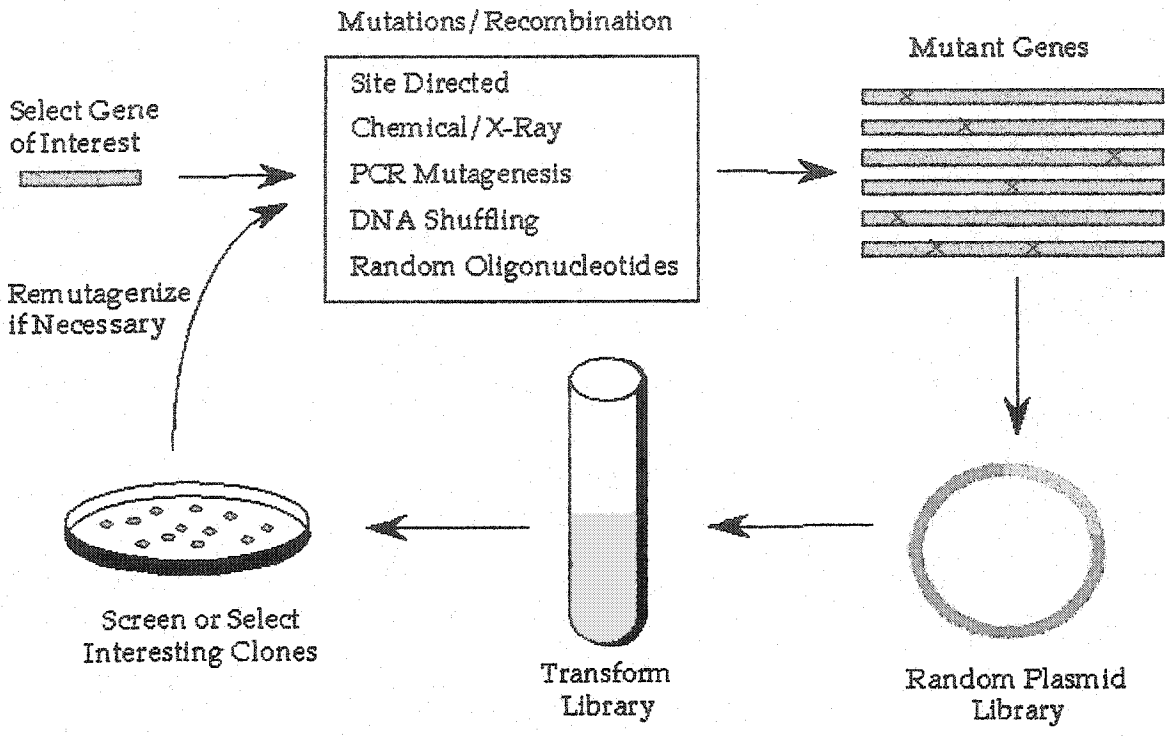



Figure 1.5

**Overview of random mutagenesis methods**
The gene of interest is mutagenized by one of several methods and cloned into an expression vector, creating a mutant gene library. Genes from the library are transformed into bacteria, which produce the enzyme variants that lead to a discernable phenotype. The bacteria expressing the mutant enzymes are selected or screened for interesting characteristics. The mutant gene of interest can be further mutated if necessary.

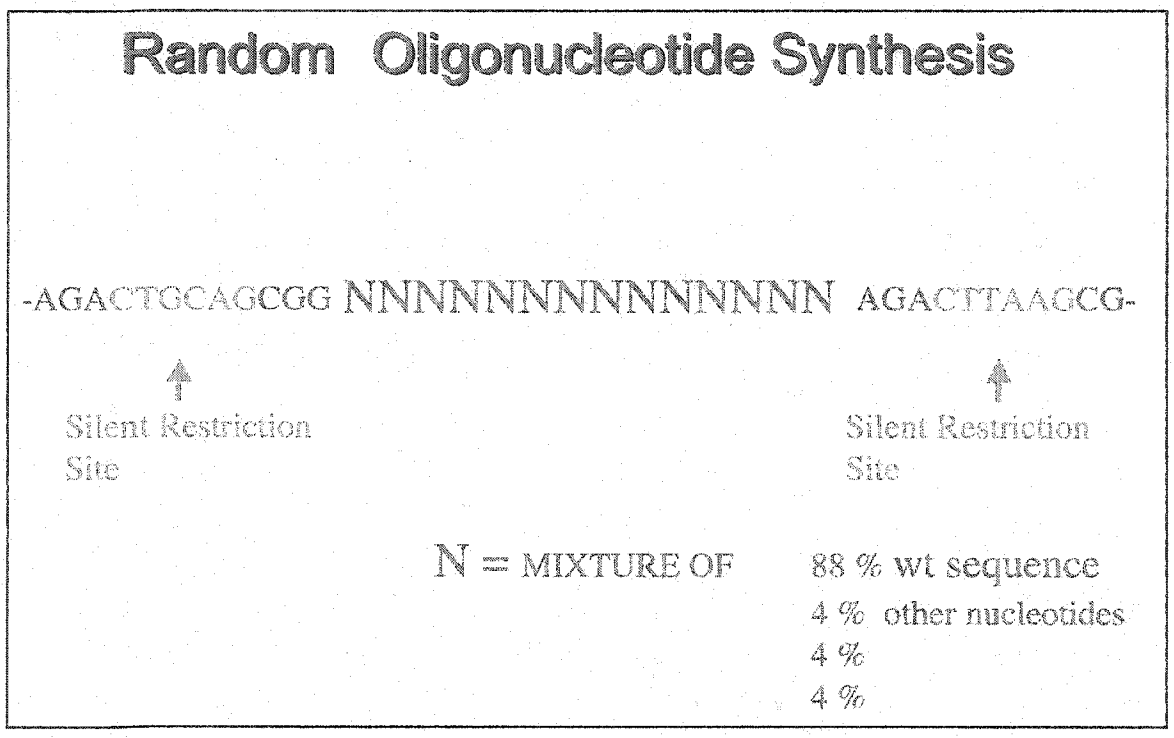



Figure 1.6

**Representation of random oligonucleotide cassette**
The mutation load at the N sites can be carefully adjusted, simply by varying the molar ratios of wild-type versus doping nucleotides. Silent restriction site facilitate the replacement of the wild-type sequence by the randomized fragment.

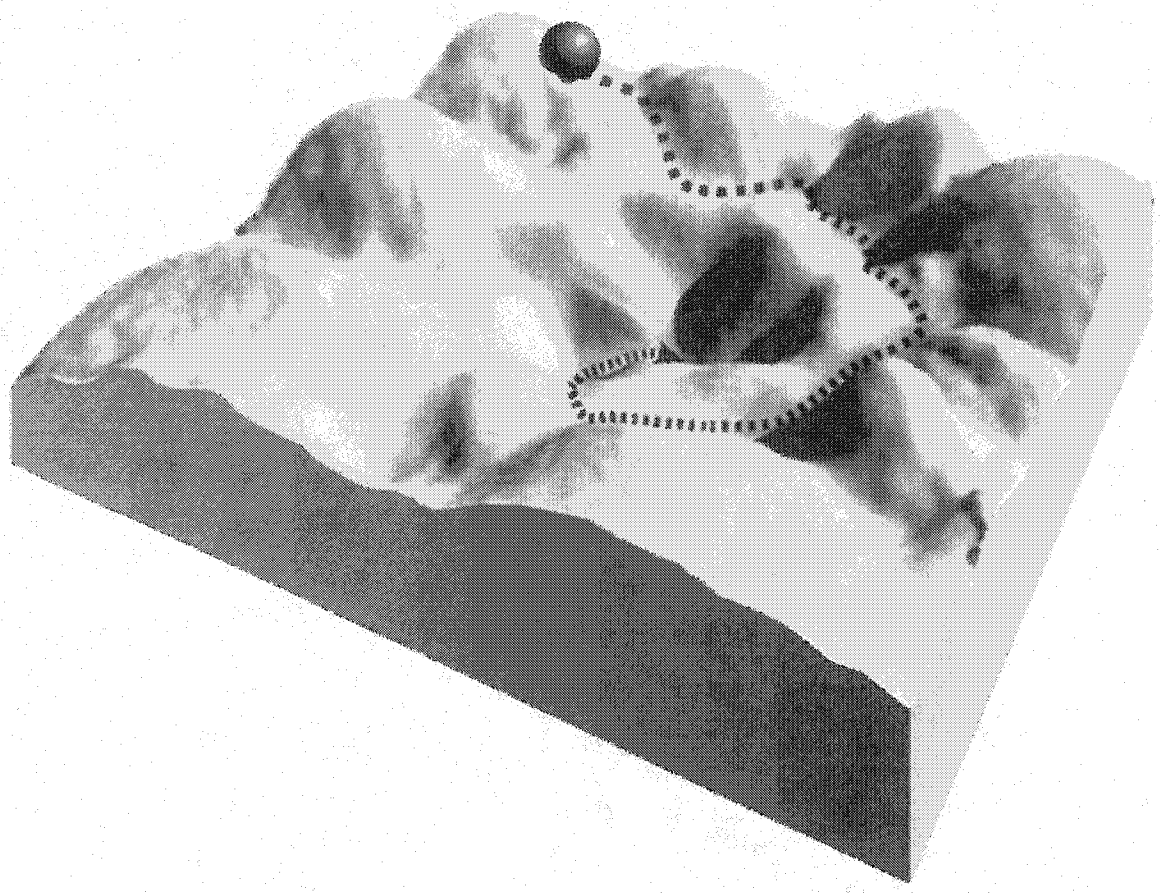


Figure 1.7

**Schematic representation of protein sequence space**
This artificial topology shows possible variant amino acid sequences for an enzyme, with the height of the peaks representing the fitness of the enzyme under some set conditions.

# *Chapter 2. Measuring protein tolerance to random amino acid substitutions*

**Chapter Summary:**

Mutagenesis of protein encoding sequences occurs ubiquitously; it enables evolution, accumulates during aging, and is associated with disease. Many biotechnological methods exploit random mutations to evolve novel proteins. In order to quantitate protein tolerance to random change, it is vital to understand the probability that a random amino acid replacement will lead to a protein's functional inactivation. We define this probability as the "x-factor." Here, we develop a broadly applicable approach to calculate x-factors and demonstrate this method using the human DNA repair enzyme 3-methyladenine DNA glycosylase (AAG). Three gene-wide mutagenesis libraries were created, each with $10^5$ diversity, and averaging 2.2, 4.6, and 6.2 random amino acid changes per mutant. After determining the percentage of functional mutants in each library using high stringency selection (>19,000-fold), the x-factor was found to be 34% ± 6%. Remarkably, reanalysis of data from studies of diverse proteins reveals similar inactivation probabilities. To delineate the nature of tolerated amino acid substitutions, we sequenced 244 surviving AAG mutants. The 920 tolerated substitutions were characterized by substitutability index and mapped onto the AAG primary, secondary, and known tertiary structures. Evolutionarily conserved residues show low substitutability indices. In AAG, β-strands are on average less substitutable than α-helices; and surface loops that are not involved in DNA binding are the most substitutable. Our results are relevant to such diverse topics as applied molecular evolution, the rate of introduction of deleterious alleles into genomes in evolutionary history, and organisms' tolerance of mutational burden.

## INTRODUCTION

A fundamental aspect of evolution is that mutations generate novel alleles that are then favored by selection. However, new coding mutations can be deleterious, neutral, or beneficial. Mutations can result from environmental and endogenous damage to DNA, and from errors during DNA synthetic processes. In humans, random mutations produce inherited diseases and accumulate with aging and cancer [61,67]. Conversely, targeted hypermutagenesis by immune defenses helps to generate antibody diversity and was recently shown to inactivate retroviral genomes [15]. John Maynard Smith proposed more than thirty years ago that the occurrence of functional mutant proteins that differ from the wild-type by one residue is likely frequent for evolution to be possible [59]. Since then, numerous evolutionary and mutagenesis studies have led to the assertion that proteins are highly plastic in tolerating amino acid substitutions [68-70]. However, to date we lack a quantitative measure of the degree of proteins' tolerance for random amino acid changes that occur at a random position in the protein. If a rigorous measure of proteins' degree of tolerance of random amino acid changes can be defined, then such fundamental calculations as the steepness of protein fitness landscapes, or the rate of introduction of deleterious mutations into coding genomes can be more clearly delineated. Further understanding of the nature of tolerated amino acid substitutions can also lend insight into protein folding and design.

Here, we develop the concept of the probability of inactivating a protein with a random codon replacement producing amino acid change at a random

location along its sequence. For conciseness, this concept is named the x-factor. We describe an analytical method for calculating the x-factor of proteins from randomly mutated libraries, and demonstrate the method using the human DNA repair enzyme 3-methyladenine DNA glycosylase (AAG, MPG, ANPG). Nine hundred and twenty tolerated amino acid substitutions in active mutant enzymes were identified and substitutions were mapped to the available x-ray crystal structure of AAG. We examine the applicability of the x-factor concept to diverse proteins by reanalyzing results from prior studies. These findings reveal a similar range of inactivation probabilities.

## MATERIALS AND METHODS

### *Materials and strains*

*E. coli* strain MV1932 (AB1157 *ada alkA1*) [71] was derived from strain AB1157 (*argE3 hisG4 leuB6 proA2 thr-1 ara-14 galK2 lacY1 mtl-1 xyl-1thi-1 rpsL31 supE44 tsx-33*). Chemicals were from Sigma-Aldrich and enzymes were from NEB unless otherwise indicated. DNA oligonucleotides were purchased from IDT (Coralville, IA) unless stated otherwise.

### *Construction of PCR mutagenesis libraries*

For each of the low, medium, and high mutation frequency libraries, two separate error prone PCR mutagenesis protocols that generate complementary mutational spectra were performed in series. Mutazyme in the GeneMorph kit (Stratagene, La Jolla, CA) preferentially mutates at G:C (72% at G:C, 26% at A:T,

2% inserts or deletions, abbreviated as indels). *Taq* with 0.5 mM $Mn^{++}$ and dNTP bias (1 mM dCTP, 1 mM dTTP, 0.2 mM dATP, 0.1 mM dGTP) favors mutations at A:T basepairs (18% at G:C, 77% at A:T, 5% indels) [46]. The two protocols were used in series to yield approximately equal mutational frequencies at G:C and A:T basepairs. Mutation frequencies were titrated by varying the amount of template material. (For primer sequences and detailed PCR conditions, see Tables 1 and 2). Briefly, each library was amplified first with Mutazyme using primer pair 1 and 2. Mutated products were gel purified and subsequently amplified with *Taq* according to the protocol for mutagenic reaction condition 5 in the Diversify PCR Mutagenesis Kit (Clontech) using primer pair 3 and 4. Finally, to generate greater yield and blunt ends for cloning, libraries were fidelity amplified using *Pfu* (Stratagene) using nested 5'-phosphorylated primers that hybridized close to the boundaries of the AAG gene (primer pair 5 and 6). For comparisons of mutational spectra between *Taq*, Mutazyme, and the combination protocols described here, see Table 3.

PCR Mutagenesis libraries were gel purified and blunt-end cloned into the *EcoR*V site of pGRFP2, which were previously linearized with *EcoR*V and dephosphorylated. pGRFP2 is pUC derived cloning vector that expresses the cloned protein N-terminally fused with green fluorescent protein (GFP) from the lac promoter. The GFP-AAG fusion protein complemented MV1932 over non-expressors by over 19,000 fold (Figure 4, Table 4). DH10B (Invitrogen) cells were transformed with the libraries and grown overnight on LB-glucose-carbenicillin (carb) plates. Aliquots of transformed cells were diluted and plated to calculate

library sizes. The plasmid libraries were harvested using HiSpeed Maxi kits (Qiagen). Inserts cloned in the reverse orientation generate a unique *BamHI* site; and were eliminated by cutting with *BamHI*. The remaining circular DNA was gel purified and transformed into *E. coli* MV1932 cells.

20, 18, and 28, totaling 66 mutant AAGs were randomly picked from the unselected low, medium, and high mutational frequency libraries before methyl methanesulfonate (MMS) selection and were sequenced. All mutant AAGs were in the forward direction and distinct from each other. Mutations were distributed did not exhibit apparent hotspots (Figure 1).

*Genetic Selection for Active Enzymes, DNA sequencing*

MV1932 cells were transformed with pGRFP2-AAG, low, medium, high libraries; and with empty pGRFP2 vector and grown to confluence, diluted 1:100, and grown to mid-log phase in LB-carb at 37 °C. Cultures were treated with 0.2% MMS for one hour and drug washed away. Pre-treated and post-treated cultures were serially diluted and plated on LB-carb in triplicate to calculate survival means and standard deviations. The fractions of surviving clones in libraries were normalized to wild-type survival. The standard deviations of library survivals after normalization were calculated from the formula:

$$\mathrm{Var}(H) / H^2 = \{\mathrm{Var}(X) / X^2\} + \{\mathrm{Var}(Y) / Y^2\}$$

Where $H = X / Y$ and $\mathrm{Var} = (\mathrm{Stand.\ Dev.})^2$

The dose of MMS used was within a range of drug in which the library and control populations were proportionally affected. MMS sensitivity assays were also performed at 0.15% and 0.25% MMS; and the percent library survivals relative to controls and each other were similar at these MMS concentrations (Figures 2 and 3).

Surviving colonies were grown overnight and picked. Plasmids were amplified using TempliPhi (Amersham, Piscataway, NJ) according to manufacturer's protocol. Reactions were then treated with shrimp alkaline phosphatase (Amersham), and enzymes were heat inactivated at 95 °C for 10 min. Aliquots were used for BigDye3.1 (ABI) sequencing of the AAG gene in entirety with 3 uni-directional overlapping sequence reads with primers 7, 8, and 9. Minimal accepted sequencing quality threshold, as assessed by Phred [72], was 25. Briefly, Phred quality values are log transformed error probabilities that have been calibrated against real data sets. A score of 30 indicates one error in every 1,000 bases, and a score of 40 indicates one error in every 10,000 bases. The actual mean Phred score for all sequencing data used in our analysis is 45.5. To rule out mutagenesis in *E. coli* as a confounding factor, we sequenced the AAG coding region from sixteen wild-type AAG expressors and the multiple cloning site and flanking regions from sixteen empty vectors in non-AAG expressors after MMS selection. No mutations were observed in more than 10 kb of sequence from both populations (data not shown).

*AAG activity assay*

Crude lysate from MV1932 exhibits no detectable glycosylase activity toward a double stranded DNA substrate containing a single ethenoadenine (εA) lesion, which is recognized and removed by human AAG. Crude lysates were prepared from MV1932 harboring wild-type, mutant, or empty vector grown to identical $A_{600}$ readings. Cells were resuspended in 100uL of glycosylase buffer (20mM Tri-HCl pH 7.8, 100mM KCl, 5 mM EDTA, 1mM EGTA, and 10mM 2-mercaptoethanol) containing lysozyme (0.4mg/mL), frozen, and thawed on ice for 2 hours. Cell debris was removed by centrifugation. Glycerol was added to the supernatant to 10% final concentration. Protein content was estimated by Bradford assay (Bio-Rad), and aliquots were frozen.

The AAG DNA glycosylase assay was adapted from [73] with modification. An εA containing oligonucleotide (Midland CRC, Midland, TX) was 5′ radio-labeled with polynucleotide kinase and annealed to 2-fold excess of complementary oligomer (primers 10 and 11). 100 fmoles of the resulting substrate were incubated with 3.5 ug crude lysate in 20 uL glycosylase buffer at 37 °C for 30 min. NaOH was added to 0.2 M and samples heated to 70 °C for 30 min to convert abasic sites to DNA strand breaks. Tris-HCl (pH 7.5) was added to 0.25M and DNA fragments were separated by denaturing polyacrylamide gel electrophoresis. Bands were visualized and quantified by phosphorimaging screen exposure.

*X-factor calculation and Substitutability visualization*

$X_T$ variables were calculated by finding best-fit values from equation #1 after iteratively stepping through all possible values from 0 to 1 at .001 increments. To visualize individual residue substitutability, the B-factor value for each residue in the AAG-εA co-crystal PDB file (PDB ID code 1f4r) [42] was replaced with the substitutability index score to color the molecular representation. Figure 2 was generated with the program PyMOL (Delano Scientific, San Carlos, CA). A web based x-factor calculator applicable to any protein of interest and the altered AAG PDB file enabling visualization by structure and substitutability are both accessible at:

http://depts.washington.edu/loeblabs/protocols/protocols.htm.

## RESULTS AND DISCUSSION

*Calculating protein tolerance to random amino acid substitutions*

The probability of protein inactivation with one random amino acid substitution, the x-factor ($x_{sub}$), can be calculated from the fractions of mutants ($f_n$) with (n) number of amino acid changes within a gene-wide randomly mutated library, and from the proportion of mutants that survive functional selection (S). For example, $f_0$ denotes the fraction of the unselected library with 0 amino acid change, $f_1$ denotes the fraction with 1 aa change, and so on:

$$f_0(1-x_T)^0 + f_1(1-x_T)^1 + f_2(1-x_T)^2 + f_n(1-x_T)^n + \ldots = S \quad \text{or} \qquad \text{(equation \#1)}$$

$$\sum_{n=0} f_n (1-x_T)^n = S$$

where $x_T$ is the total protein inactivation probability with random amino acid change, including frameshifts (indels). $x_T$ can be solved after experimental determination of the $f_n$, S, and (i) values. Indels are found at low percentages in random mutagenesis libraries, but invariably produce protein inactivation. To determine the true x-factor ($x_{sub}$) resulting only from a random codon substitution (missense or nonsense mutation), the indel fraction in the total mutational pool (i) is subtracted from $x_T$ to obtain $x_{sub}$.

$$x_{sub} = x_T - i \qquad \text{(equation \#2)}$$

In order to measure the probability of inactivation by random amino acid substitutions we used the gene encoding the human 3-methyladenine glycosylase (AAG). AAG protects cells against DNA alkylation damage by excising alkylated base lesions including 3-methyladenine, 7-methylguanine, and 1,$N^6$-ethenoadenine (εA) [9]. The 894 base pair AAG cDNA encodes a 298 amino acid, 33 kD monomeric protein that complements the DNA alkylation repair deficient strain MV1932 (*ada alkA1*) [71] against toxicity induced by the alkylating drug methyl methanesulfonate (MMS) [37]. Under MMS challenge, MV1932 cells expressing AAG from our pUC based vector exhibit greater than 19,000-fold survival advantage over non-AAG expressing MV1932 controls (Figure 4), thus providing a stringent and specific selection for active mutant AAG enzymes.

The crystal structure of the catalytically competent ΔN79 (residues 80-298) AAG protein complexed with εA substrate oligo reveals that the enzyme binds to DNA via a flat positively charged face. A β-hairpin extends into the DNA minor groove and flips the targeted nucleotide into the enzyme active site [38,42]. A water molecule is deprotonated by Glu-125 to form a hydroxyl nucleophile that cleaves the glycosilic bond between the damaged base and the sugar. The resulting abasic site is later cleaved and replaced with a normal nucleotide by the subsequent actions of an endonuclease, a DNA polymerase, and a DNA ligase [9].

We used PCR mutagenesis to generate low, medium, and highly mutated AAG cDNA libraries averaging 2.2, 4.6, and 6.2 amino acid changes per gene (a change is defined as a missense, nonsense, or indel). Sequencing of 20, 18, and 28 mutants from each unselected library revealed the ($f_n$) and indel (i) values of each library (Figure 5, table 4). Expression of AAG and AAG mutant libraries protected MV1932 cells against MMS induced cell death. The fractional survival of each library relative to wild-type yielded the survival (S) values (Figure 4, Table 4). Solving for $X_{sub}$ using equations 1 and 2 yielded the x-factors ($x_{sub}$) of the low, medium and high libraries at: 39%±4%, 30%±5%, and 33%±3% (mean ± standard deviation), respectively. The x-factors from the three libraries are within the 95% confidence interval of each other. The average x-factor is 34% ± 6%. Thus, the overall probability of inactivating AAG with a single random amino acid change occurring randomly in the protein is approximately 34%, or one-third (Table 4).

*The x-factor and the substitutability of proteins*

Using three different libraries we obtained a consistent value for the probability that a random amino acid change will inactivate AAG. Our findings beg the question of whether a similar x-factor is seen in other proteins. It may be argued that the wide range of protein functions should demand drastically different mutabilities of various proteins. On the other hand, proteins face essentially similar requirements, such as the need to properly fold into soluble globular structures necessary for function [74]. General types of changes leading to unfolding would inactivate various proteins. To address these questions, we reanalyzed data from diverse published studies and calculated inactivation probabilities. First, we examined random oligonucleotide mutagenesis studies in which mutations were targeted to the catalytic center of enzymes, and from which ($f_n$) and (S) data are available [44,45,60,75-77]. We reasoned that these critical segments are expected to tolerate few substitutions. The results from human, bacterial, and viral enzymes are shown in Table 5. Despite the different enzymes and selection systems used, inactivation probabilities within these sensitive regions range from 44% to as high as 81%, averaging ~60%, thus supporting our hypothesis. Second, Markiewicz and coworkers previously examined 12 or 13 different amino acid substitutions at each residue across 90% of the *E. coli* lac repressor protein using amber codon suppressor strains, which often corresponded to two or three nucleotide changes per codon [78]. In our reanalysis of their data, we counted close to 1,380 single mutants that were

inactive, approximately 20% of which were temperature sensitive, out of a total of 4,049 examined. This yielded a x-factor for the lac repressor gene of 34%, which correlates well with our results for human AAG. Third, the x-factor of a protein is conceptually similar to the proportion of new deleterious alleles that arise during the evolution of the source organism. Eyre-Walker and Keightly calculated the percentage of deleterious substitution mutations that were eliminated from the human lineage by purifying selection. They examined synonymous and nonsynonymous substitution rates from coding regions of 46 homologous proteins from humans and chimpanzees [79]. Interestingly, they conclude that at least 38% of spontaneous mutations in the human lineage were sufficiently deleterious to have been eliminated by selection [79]. Together, these findings from multiple and independent experimental approaches suggest a range of similar x-factors over the length of diverse proteins.

Enzyme inactivation can result from indirect structure disrupting mutations or from direct alterations of the catalytic mechanism. The AAG functional assay is sensitive to both modalities. Minimal AAG activity necessary for complementation was assessed by measuring initial reaction rates under saturating substrate conditions in lysates of ten random surviving clones. The results indicated that approximately 5-10% of wild-type activity is necessary for survival at the MMS dose used (data not shown). Thus, the minimal level of necessary AAG activity in our experiment is functionally defined. However, in natural evolution, the minimal activity requirement threshold may be much more variable, and likely depends upon the selective environment and

interactions with other genetic components. Because selection acts at the level of the organism, certain decreases in activity may be well tolerated if masked by compensatory mechanisms, or by a tolerant environment. The minimal level of required activity for various gene products therefore are likely to vary.

Hydrophobic/hydrophilic properties appear to be crucial overall determinants of protein structure [70,74]. The buried core is sensitive to non-hydrophobic changes and those that disrupt packing, whereas the solvent-accessible surface is generally more tolerant of change. Residue size, charge, hydrogen-bonding characteristics, and bond angle flexibility are other folding factors that may be perturbed by random substitutions.

AAG is a simple monomeric protein. Larger proteins with multiple functional domains and multiple interacting partners may exhibit more complex inactivation dynamics. The x-factor calculation assumes that the effects of multiple mutations are independent, in that the effects of mutations on protein function are largely additive. This is supported by findings on the λ repressor [80]. However, at higher mutational loads effects of mutation may interact in more complex ways, with increased possibility of compensatory or synergistic effects. These results with AAG may slightly underestimate the x-factor, since the N-terminal 79 amino acids of AAG are not required for enzymatic activity. Tolerated substitutions are slightly elevated in this N-terminal region. Nevertheless, protein-folding principles apply, and mutations in this region that cause overall misfolding or aggregation will produce inactivation.

There likely are variations in the substitutability of different proteins. The hydrophobic core is generally less tolerant of change than the solvent accessible exterior [70]. Therefore, x-factors may also be influenced by proteins' sizes and surface-to-volume ratios. Axe and coworkers found that 5% of single amino acid substitutions lead to an inactivated barnase enzyme [81]. Rennell et. al. found that ~16% of amino acid substitutions in T4 lysozyme caused inactivation [82]. The difference from the above findings may be attributed to barnase and T4 lysozyme's small sizes, which are 110 amino acids and 164 amino acids, respectively. Highly conserved proteins such as histones are likely to be relatively intolerant to mutation, whereas protein domains such as $F_v$ regions of antibodies may exhibit increased tolerance against misfolding. Residues that are post-translationally modified are also expected to be intolerant of change.

The x-factor is calculated for amino acid replacements, and can include the generation of stop codons. The frequencies of stop codons in the low, medium, and high libraries are 4%, 9%, and 7.5%, respectively. The x-factor can be converted for single nucleotide substitutions. Largely due to degeneracy at the third position, the nucleotide-x-factor is expected to be less than the amino-acid-x-factor. Multiplying the amino-acid-x-factor of ~34% by the probability of non-synonymous codon change accessible by one nucleotide (415/549) yields the nucleotide-x-factor of ~26%.

Our mutagenesis scheme of creating predominantly random single nucleotide substitutions mimics the generation of natural diversity. Three naturally occurring human SNPs arose in our database of tolerated AAG

substitutions: P64L, T199A, and A258V. These variations did not exhibit appreciable effects on MV1932 complementation when individually assayed (data not shown).

*Substitutability and AAG structure*

Previously, we have focused on the probability of amino acid changes being inactivating. We have also examined situations in which amino acid substitutions are tolerated. In order to analyze the nature of tolerated substitutions, we sequenced 244 mutant AAG cDNAs from the highly mutated library that complemented MV1932. This yielded a total of 920 tolerated amino acid changes. Figure 1 maps the mutations along the AAG primary sequence. The types of tolerated amino acid substitutions at each position are indicated. Residues without bars reflect zero identified substitutions.

A residue's "substitutability index" is defined as the percent sequenced clones with a substitution at that residue. Many positions that are evolutionarily conserved are also essential for activity [38,42] and did not tolerate changes in our assay. Examples include Glu-125, Arg-182, and Val-262, each of which interacts with the activated water molecule that hydrolyzes the sugar-base glycosylic bond. Other non-substituted amino acid residues include Tyr-162, which projects from a surface β-hairpin and acts as a "nucleotide flipper." Met-164 and Tyr-165 assist in this base flipping mechanism by destabilizing the base pair adjacent to the flipped nucleotide. Y162A, M164A, Y165A single substitution mutants were generated by Lau et al. and assayed using a genetic

complementation system [42]. The Y162A mutant exhibited large impairment of glycosylase activity, while M164A and Y165A showed only moderate impairment [42]. Correspondingly, in our study, no substitutions were observed at Tyr-162, whereas positions Met-164 and Tyr-165 showed moderate substitutability, allowing Ile and Arg, and Phe substitutions, respectively (Fig. 1). Within the substrate-binding pocket, the flipped out base stacks between the aromatic side-chains of Tyr-127, His-136, and Tyr-159. Y127F, H136Q, and Y159F mutants were also generated previously [42]. Y127F exhibited the most profound decrease in activity, whereas Y159F was the least affected [42]. In our data set, Tyr-127 was concordantly unsubstituted, and His-136 tolerated only one Tyr replacement. Tyr-159 was substituted by both Phe and Asn.

There are positions in AAG that are not evolutionarily conserved, but did not exhibit any tolerated changes. The individual spatial arrangements of these interactions are likely unique to AAG. Although some of these positions may display substitutions if even more mutants are sequenced, the structural basis for lack of substitutions at many of these positions highlights three general mechanisms: specific hydrogen bonding interactions, unique hydrophobic packing, and ion binding. For example, specific hydrogen bonding requirements are emphasized by Glu-116's interaction with Arg-118, which, in turn, interacts with Glu-188 and Glu-245 in 3-way interaction. Arg-261 provides a hydrogen pair partner to the evolutionary conserved and unsubstituted Asp-132. This pair packs adjacent to Tyr-127, which forms part of the active site pocket. Hydrophobic packing constraints are observed at Gly-119, which is at the core of

a β–strand, less than 4.5 Å away from Leu-184. No other side-chains can fit in this tight space. Similar packing constraints are observed at Leu-184, which is less than 4.5 Å from the unsubstituted Leu-225. Cys-167 is buried less than 4.5 Å from Ile-227 and is close to the Cα of Cys-222. Mutations of buried residues may require concomitant mutations of other closely packed residues in order to maintain optimal packing. Interestingly, at least one mutant in our study appears to demonstrate this principle. It contains I170V and L181M substitutions that pack adjacently in the hydrophobic core. The conversion of Leu-181 to the slightly bulkier methionine is found to coexist with the conversion of Ile-170 to the smaller valine. Lastly, lack of substitutions at Ser-171 highlights the role of ion binding. Ser-171's side-chain oxygen binds to a $Na^+$ ion, which has been postulated to enhance the structural stability of the active site floor [42].

In contrast, certain regions in AAG appear highly substitutable. Examples include the first 79 N-terminal residues that have been shown previously to be unnecessary for *in vitro* enzyme activity and DNA binding specificity [36,83]. Residues 80-81, 200-207, 249-254, and 296-298 are also highly substitutable (Fig. 1). In accord, they display low electron density in x-ray crystallography and were inferred to be disordered loops [38].

In Figure 2, the relative substitutability indices of residues are mapped onto the available crystal structures of the N79Δ AAG mutant. Dark blue residues are the least substitutable, and red residues are the most tolerant of change. Figure 2a-b shows surface residues, and Figure 2c-d facilitates views into the protein core. One striking feature is the general immutability of the

DNA interacting face and specifically, the nucleotide-flipper Tyr-162 (Fig. 2a). A surface region distant from the DNA binding face (Fig. 2b) was also observed to have low substitutability scores; Glu-188, Arg-118, Glu-245, Glu-116, and Arg-110 participate in a network of charged contacts that likely contribute to protein stability. In the protein interior, a conspicuous pattern of alternating unsubstituted and substitutable sites is seen in the β4 (165-171) strand (Figs. 1, 2c-d). Cys-167, Asn-169, and Ser-171 are relatively unsubstituted, because their side-chains face toward the active site and are involved in substrate recognition or $Na^+$ binding [42]. In contrast, Met-168 and Ile-170 tolerate hydrophobic substitutions, because their side-chains face the opposite direction and pack into the hydrophobic core. Solvent accessible surfaces generally exhibit higher substitutability compared with buried residues. This is evident in Figure 2c-d, where the exposed exterior sides of several α-helices exhibit greater substitutability than their interior-facing sides.

Averages of substitutability indices in different structural motifs are presented in Table 6. In AAG, evolutionarily conserved and catalytically crucial residues are significantly less substitutable than the rest of the protein. Nonconserved residues adjacent to conserved residues in the primary sequence are generally less substitutable than other nonconserved residues, reflecting their involvement in functionally important regions. This observation suggests that they may also be fruitful targets for directed evolution studies. β-strand residues, as a group, are less tolerant of substitution than are α-helices. This may be explained in part by the fact that in this α–β protein, the β sheets are generally

less solvent accessible and therefore possess fewer surface residues that are more likely to tolerate substitutions. Loops and turns, expectedly, are the most substitutable.

*Some Implications of the x-factor*

We observed that various residues of a protein are differentially sensitive to substitutions, and that tolerance of the entire protein to random change can be defined by the x-factor. The x-factor is a description of an intrinsic property of individual proteins and protein motifs, and can be a guiding parameter in the study of natural and artificial evolutionary processes. For example, using the estimated inactivation probability of approximately 34% and assuming mutually independent effects on inactivation probability by multiple mutations, the isolation of active mutants harboring many mutations from large random mutagenesis libraries ($>10^5$) is not surprising [53]. In contrast, a single, non-3bp indel event almost certainly leads to inactivation ($x \approx 1$). Therefore, indel frequencies should be minimized in efforts to evolve novel proteins from high mutation load libraries. Retroviruses, such as HIV, may be susceptible to increased mutational burden, and lethal mutagenesis of viral genomes by introducing mutations through the use of nucleoside and ribonucleotide analogs has been proposed [84]. Given our findings, such efforts may be further enhanced by the use of analogs that efficiently induce frameshift mutations. Viral genomes that encode multiple proteins as different reading frames of the

same genetic sequence may be particularly sensitive to agents that generate frameshifts.

It is estimated that the human mutation rate per coding diploid genome per generation is 3.2, including base substitutions, indels, and larger changes [85]. Multiplying this number by the general x-factor of ~34%, then the rate of introducing deleterious coding alleles by random substitution is approximately 1.0 per diploid genome per sexual generation. This is likely an underestimate, since indels inactivate coding regions much more efficiently than base substitution mutations. Dominant negative mutations may also more efficiently produce a deleterious phenotype, though the frequency of mutations that act in a dominant negative manner is largely unknown. Interestingly, our deleterious coding allele rate calculation of 1.0 is congruent with the estimate of 1.6 independently calculated by Eyre-Walker and Keightly, which was based on the assumption of 60,000 genes in the human genome [79].

Overall, our method of gene wide random mutagenesis and sequencing highlights the relative importance of specific residues to enzyme structure and function through the numbers and types of tolerated substitutions. This work validates and extends from previous structural studies. Interestingly, the substitutability indices of individual residues can be obtained independently of conservation or structural information, and are generally consistent with both. The extensive database of tolerated amino acid substitutions is obtained from a more expedient form of gene wide study than previous techniques. In comparison with the traditional method of gene-wide alanine scanning, in which

alanine substitution constructs need to be made for each amino acid and each individual mutant screened, our method of functional selection followed by sequencing of active mutants is less complex, and able to test larger areas of protein sequence-space. Particularly that automated sequencing has become even more robust, accessible, and cost-effective. In comparison with the use of codon suppressor methods of testing, our method is again more readily accessible. For example, in the LacI work that was cited [78], over 4,000 individual constructs had to be made and assayed in 12-13 amber codon suppressor strains, representing a heroic effort. The database of tolerated substitutions can provide a valuable resource for predicting the effects of mutations on protein function, which has been a focus of recent investigations [86,87].

We advance the concept of the x-factor as a measure of protein tolerance to random substitutions. The x-factor may also be useful in measuring genomic robustness against mutations. It has been hypothesized that evolvability, or the ability to generate heritable variation, may be favored in certain environments [88]. Genomes experiencing high mutational burden may face selective pressure to evolve proteins that are tolerant of change, in which case the observed x-factors are expected to be less than x-factors of homologous proteins from more faithfully propagated genomes. It may be of particular interest to examine x-factors from various protein families and from diverse organisms.

Table 2.1 DNA oligonucleotide Sequences

| Primer | Oligonucleotide sequences (5'->3') |
|---|---|
| 1 | CGTATTACCGCCTTTGAGTGAGCTGAT |
| 2 | GCGACACGGAAATGTTGAATACTCATACTC |
| 3 | GCAGCGAGTCAGTGAGCGAGGAA |
| 4 | TTATTGAAGCATTTATCAGGGTTATTGTCTCAT |
| 5 | GCCGCGGCCGCGAT |
| 6 | CCGCGGCGCGCTCGAGTC |
| 7 | TCGAAAGATCCCAACGAAAAGAGA |
| 8 | CCCTCGCCCTCGCCCTCGAT |
| 9 | CGCAGCCCAGGCACCTGC |
| 10 | TTGGCAGCAGAATATTGCTeAGCGGGAATTCGGCGCG |
| 11 | CGCGCCGAATTCCCGCTAGCAATATTCTGCTGCCAA |

Sequences of primers used for PCR mutagenesis library construction, DNA sequencing, and AAG activity assays.

Table 2.2 PCR Mutagenesis conditions

| | | **Mutazyme mutagenesis** | | ***Taq* w/ $Mn^{++}$ mutagenesis** | |
|---|---|---|---|---|---|
| | | Template Quantity | Amplification cycles | Template Quantity | Amplification cycles |
| low | | 40 ng | 35 | 204 ng | 27 |
| medium | | 232 pg | 35 | 62 ng | 27 |
| high | round 1 | 6 ng | 35 | 14 ng | 27 |
| | round 2 | 880 pg | 35 | | |

Mutazyme and error prone *Taq* PCR amplification protocols were used sequentially to achieve similar mutational frequencies at G:C and A:T basepairs.

**Table 2.3 Mutational biases of distinct mutagenic PCR protocols**

| | *Taq* (Mutagenic Buffer) | Mutazyme | Low | Medium | High |
|---|---|---|---|---|---|
| A->N, T->N | 77% | 25.6% | 52% | 50% | 40% |
| G->N, C->N | 18% | 72.5% | 42% | 40% | 55% |
| insertions | 0% | 0.8% | 0% | 2% | 0.3% |
| deletions | 5% | 1.1% | 6% | 8% | 5.2% |

Side-by-side comparisons of the mutation frequencies at A:T and G:C basepairs achieved by *Taq* under mutagenic condition [46], Mutazyme I from Stratagene, and the protocols used for the low, medium, and highly mutated AAG libraries. The mutant AAG libraries described in this work exhibit mutational frequencies at A:T and G:C basepairs that are closer to 50:50 than achieved either *Taq* or Mutazyme alone, though at slightly higher indel frequencies then *Taq* or Mutazyme alone.

## Table 2.4 Calculating the x-factor

| | % of library with (n) number of amino acid changes ($f_n$x100) | | | | | | | | | | | | | Average | Library | % indels | % survival (Sx100) | x-factor ($x_{sub}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Mut. Freq. | Size | (i x 100) | | |
| wt-AAG | 100 | — | — | — | — | — | — | — | — | — | — | — | — | | | | 100 ± 8.6 | |
| low | 5 | 20 | 40 | 20 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.2 | $2x10^5$ | 6.1 | 32.7 ± 3.5 | 0.39 ± 0.038 |
| medium | 5.6 | 5.6 | 5.6 | 11.1 | 5.6 | 33.3 | 22.2 | 5.6 | 0 | 5.6 | 0 | 0 | 0 | 4.6 | $1x10^5$ | 9.9 | 18.2 ± 3.3 | 0.30 ± 0.052 |
| high | 0 | 3.6 | 7.1 | 10.7 | 3.6 | 10.7 | 14.3 | 10.7 | 17.9 | 14.3 | 3.6 | 3.6 | 0 | 6.2 | $0.9x10^5$ | 5.5 | 10.7 ± 2.3 | 0.33 ± 0.034 |
| vector only | — | — | — | — | — | — | — | — | — | — | — | — | — | | | | 0.0051 ± 0.00017 | |
| | | | | | | | | | | | | | | | | | Average x-factor | 0.34 ± 0.06 |

Distribution of amino acid mutation load frequencies ($f_n$), library survival (S), and x-factors ($x_{sub}$) in the low, medium, and highly mutated libraries. Indels (i) are expected to produce nearly 100% inactivation and are thus subtracted from the unadjusted x-factors ($x_T$) to yield x-factors due to amino acid substitutions ($x_{sub}$). As expected, increasing average mutation load results in lower percentage of active enzymes.

**Table 2.5 X-values calculated from active site targeted cassette mutagenesis studies**

| Protein | Organism | Protein region (aa #) | Ave. Sub Freq. | Survival fraction (S) | X-value | Reference |
|---|---|---|---|---|---|---|
| DNA Polymerase η | *Homo sapiens* | 52-73 | 3.8 | 0.02 | 0.81 | 17 |
| DNA polymerase I | *Thermus aquaticus* | 605-617 (Motif A) | 3.8 | 0.04 | 0.59 | 18 |
| Thymidylate Synthase | *Homo sapiens* | 196-199, 204-212 | 4.2 | 0.1 | 0.6 | 19 |
| Reverse Transcriptase | HIV | 67-78 (β3 β4 loop) | 4.1 | 0.11 | 0.59 | 20 |
| DNA polymerase I | *Thermus aquaticus* | 659-671 ( O helix) | 2.7 | 0.11 | 0.8 | 21 |
| Thymidine kinase | *Herpes Simplex-1* | 155, 161-165 | 2.4 | 0.32 | 0.44 | 22 |

Available $f_n$ and S values were used to derive the inactivation probabilities, which are greater than 34% due to the concentration of mutations near the enzyme active sites.

**Table 2.6 Mean mutability indices of AAG motifs**

| | Motif Residue mutability | | | Non-motif residues | | | t-test |
|---|---|---|---|---|---|---|---|
| | Mean | ± | Stand. Dev. | Mean | ± | Stand. Dev. | p-value |
| Entire protein | 1.38 | ± | 1.10 | | | | |
| Evolutionarily Conserved | 0.75 | ± | 0.85 | 1.53 | ± | 1.10 | 6.37E-08 |
| Nonconserved adjacent to conserved | 1.25 | ± | 0.84 | 1.60 | ± | 1.16 | 0.017 |
| Alpha helices | 1.19 | ± | 0.97 | 1.41 | ± | 1.12 | 0.19 |
| Beta strands | 0.73 | ± | 0.76 | 1.52 | ± | 1.12 | 1.01E-08 |
| Turns and Loops | 1.57 | ± | 1.12 | 0.93 | ± | 0.89 | 3.46E-07 |
| Functionally important residues | 0.56 | ± | 0.60 | 1.41 | ± | 1.11 | 1.04E-04 |

The mutability index of individual residues (number of observed amino acid changes divided by the number of active mutants sequenced at that position) is expressed as a percentage (X100), categorized by motifs, and averaged. T-test is performed against indices of motif non-members for differences in mean mutability indices, indicative of differing importance to enzyme function.

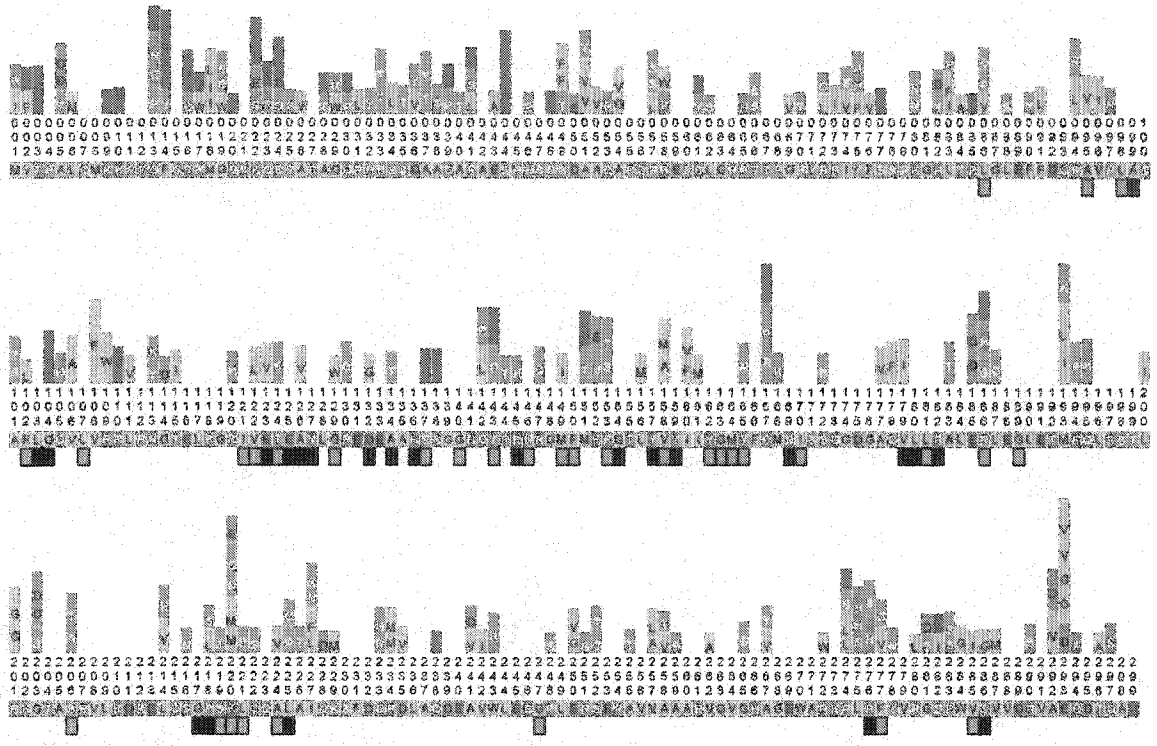


Figure 2.1

**Distribution of amino acid changes among unselected mutants**
66 unselected mutants from the low, medium, and high libraries were sequenced. Mutations producing amino acid changes are shown here.

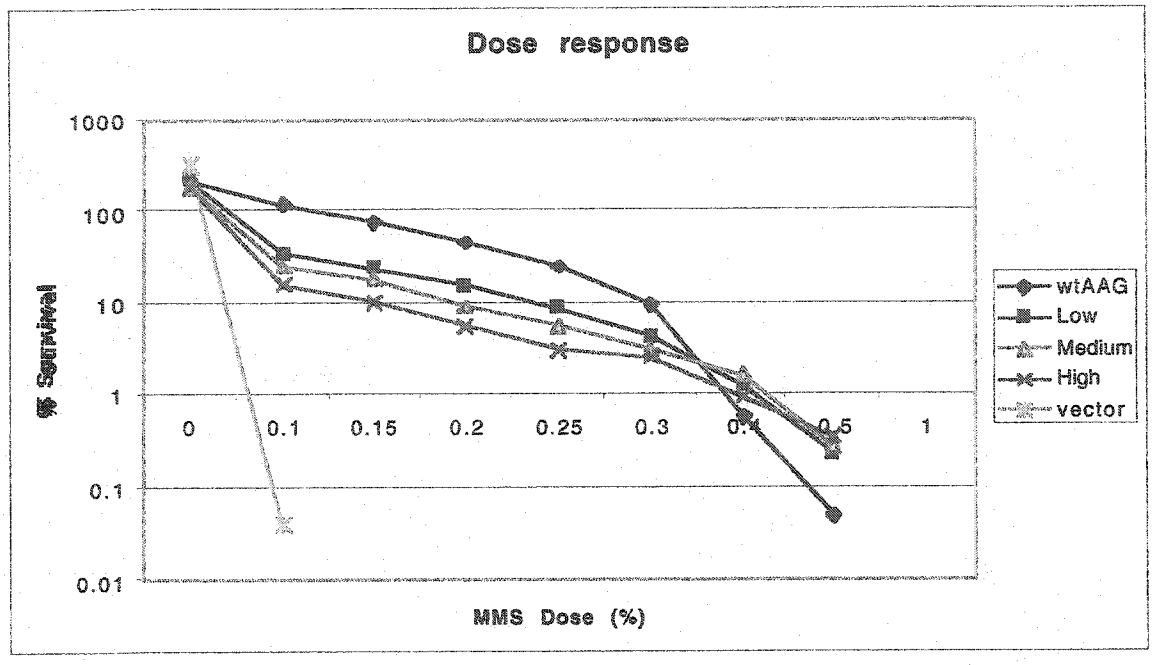



Figure 2.2

**MV1932 expressing wild-type AAG; low, medium, and highly mutated AAG libraries; and empty vector control survival in response to increasing dose of the DNA alkylating drug MMS.**

One potential source of variability in the S-value is the stringency of selection used. To control for this variability, the library survivals were normalized to the survival of the wild-type population. In addition, a dose of MMS (0.2%) was used in which the survivals of the wild-type and of the libraries are proportionally affected, such that the S-value, is largely independent of drug dosage and hence of selection stringency. The dose response ranges are shown over a broad range of MMS concentration. The absolute survivals show fairly graded response over a broad range. Interestingly, at higher MMS dosages (0.4%-0.5%), the fractional survival of the mutant libraries are greater than that of the wild-type AAG expressing population, suggesting that subset mutant AAG populations within the libraries may complement MV1932 against MMS toxicity more effectively than wild-type AAG.

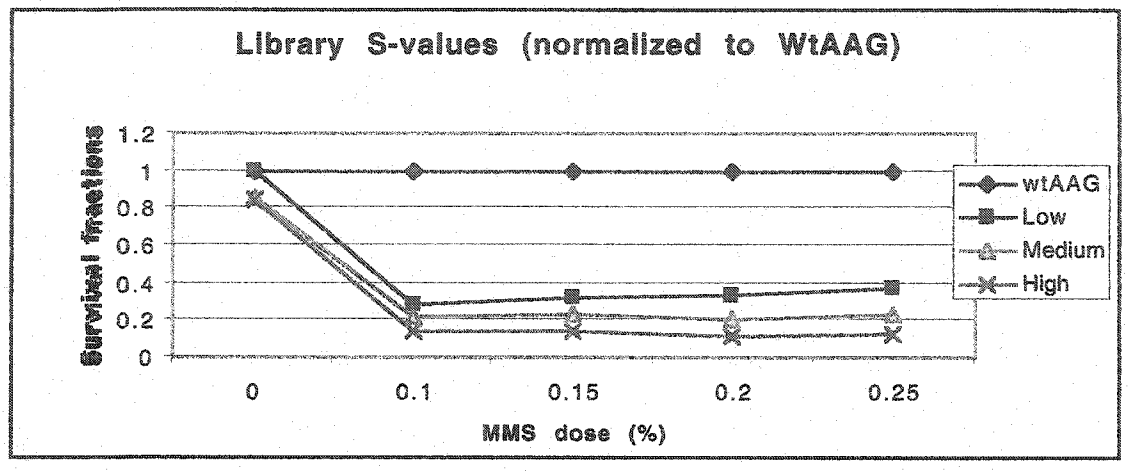



Figure 2.3

**Fractional survival of the low, medium, and highly mutated libraries, normalized to wild-type AAG survival.**
Between the dose of 0.15 and 025% MMS, the fractional survival of the libraries normalized to the wild-type AAG expressing population are largely flat, hence the choice of 0.2% MMS for the actual survival experiments.

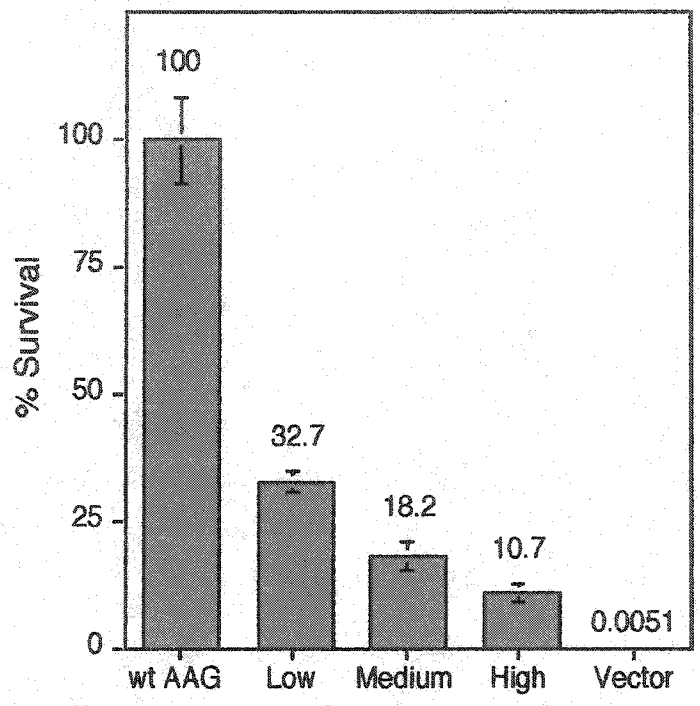



Figure 2.4

**Percent survival of MV1932 expressing wild-type AAG; low, medium, and highly mutated AAG libraries; and empty vector control, normalized to wild-type AAG survival.**
Survival percentages were measured at 0.2% MMS. As expected, higher mutational load produce reduced fractions of active enzymes.

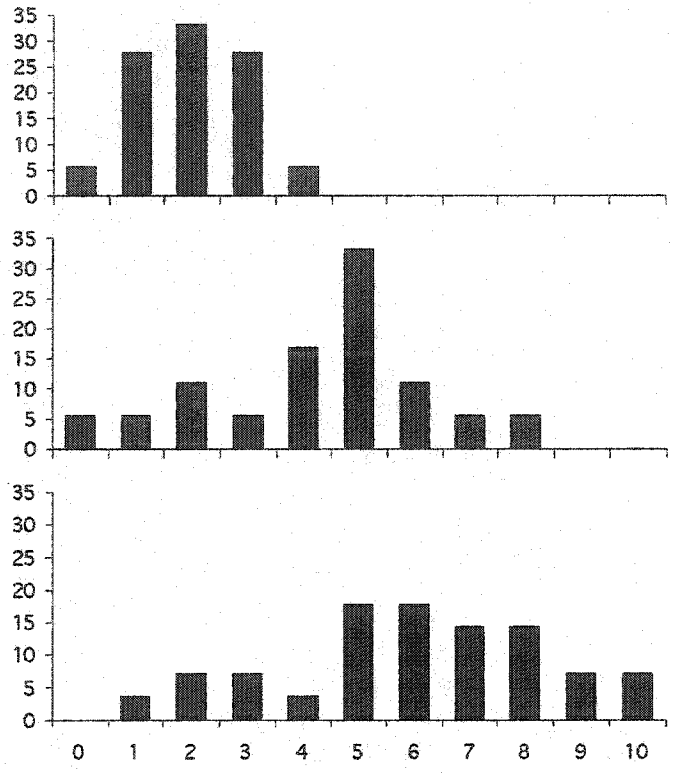



Figure 2.5

**Histograms of amino acid changes per mutant in the low, medium, and highly mutated AAG PCR mutagenesis libraries**

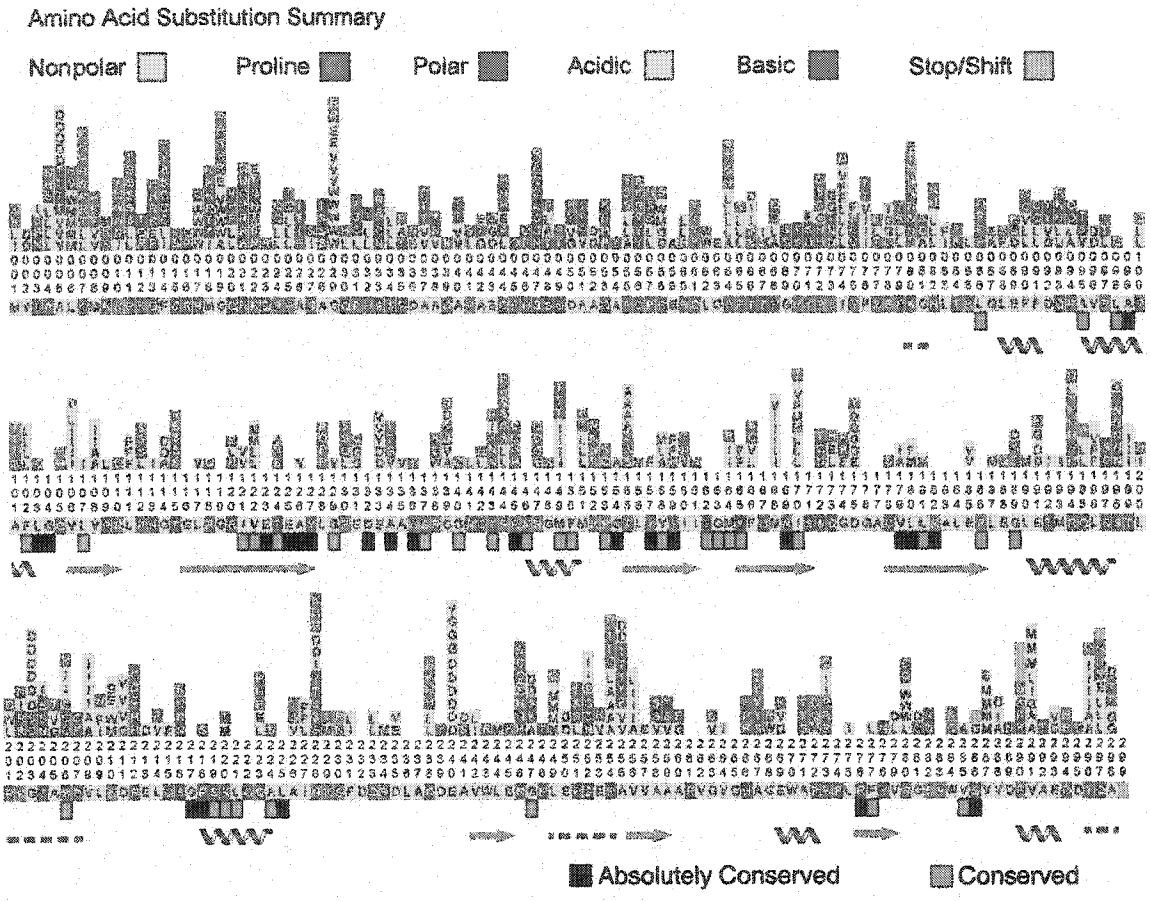



Figure 2.6

**Tolerated amino acid changes along the AAG primary sequence**
Mutations are shown with evolutionary conservation and secondary structures. 244 active AAG mutants were sequenced and observed amino acid substitutions are shown above the wild-type sequence. Colored bars indicate general categories of amino acids. Numbers are read vertically and indicate residue position. Below the wild-type sequence, evolutionarily invariant residues are marked by black boxes, and conserved residues by gray boxes. α-helices (helix), β-strands (arrows) and disordered regions (dashed lines) are indicated. Homologous sequences (from human, mouse, rat, *B. burgdorferi, B. subtilis, A. thaliana, and M. tuberculosis*) were identified with PSI-BLAST [38], and secondary structure calling was performed with MOE (Molecular Operating Environment).

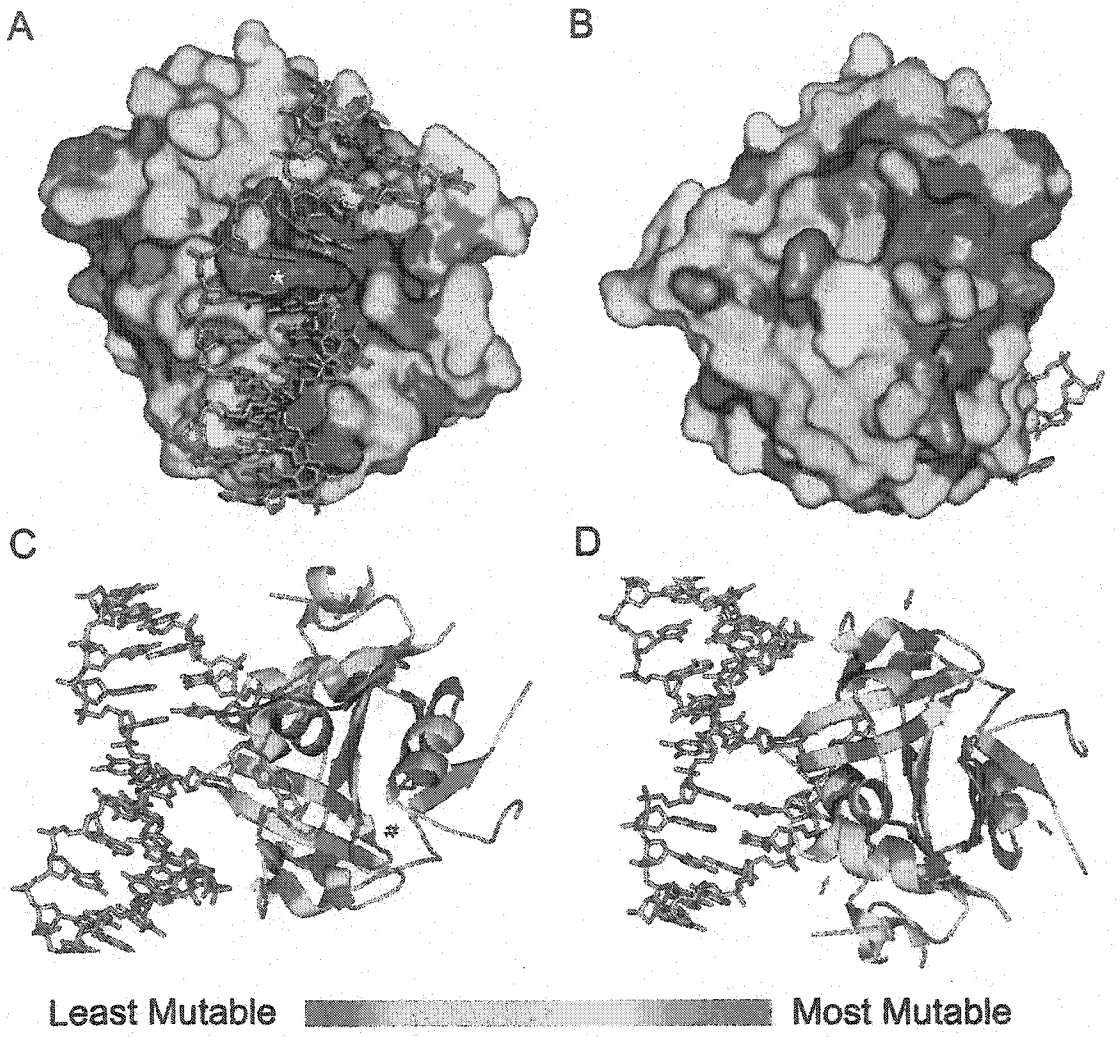



Figure 2.7

**Substitutability of AAG amino acid residues and structure.**
Individual residues' substitutability scores are indicated by their color in the spectrum, with red being the most mutable and dark blue being the least. A). The DNA interacting face of AAG. The DNA binding face and intercalating Tyr-162 (*) are largely intolerant of substitutions, whereas distant loops are generally tolerant of change. B). Rotation of figure A by 180°, showing the opposite side of the AAG surface. C-D). The substitutability of AAG resides shown by secondary and tertiary structure representation. Views are rotated 180° relative to each other. Residues near the active site of AAG, adjacent to the extra-helical εA DNA lesion, are generally intolerant of change. The β4 (165-171) strand is indicated by (#). Arrows point toward α-helices with solvent accessible faces that exhibit greater substitutability than their buried sides.

# *Chapter 3. Novel alkyl-DNA glycosylase active sites revealed by random mutagenesis and selection*

**Chapter Summary:**

Alkyladenine DNA glycosylases initiate the base excision repair pathway by recognizing and removing a variety of alkylated DNA bases. In order to study the substrate recognition pocket of alkyl-DNA glycosylases, we targeted active site residues of the human DNA repair enzyme Alkyl-DNA glycosylase (AAG) for random mutagenesis by random oligonucleotide cassette mutagenesis. More 4.5 million AAG active pocket variants were generated. More than 10,000 individual survivors were obtained after genetic selection with the DNA alkylating drug Methyl Methane-sulfonate (MMS) for *E. coli* expressing active human AAG. DNA sequence analysis of 120 human AAG mutants reveal a diverse range of alternate alkylated DNA base binding sites. Some mutants being able to protect *E. coli* against alkylation damage more effectively than the wildtype. This is the largest collection of mutant DNA glycosylases (repair enzymes) reported to date. This study reveals the degree of plasticity of the human AAG substrate recognition pocket and highlights the essential residues for substrate recognition and catalysis. This study also demonstrates a general approach for creating novel DNA glycosylases that remove select altered DNA bases for potential therapeutic and biotechnology applications.

## INTRODUCTION

Diverse DNA repair mechanisms, including the base excision repair pathway, have evolved to protect genomes against environmental and endogenous agents [11]. Alkylation of DNA bases produces a variety of lesions that threaten genomic fidelity and cell viability. DNA glycosylases initiate base excision repair by recognizing and removing inappropriate bases from DNA by hydrolyzing the *N*-glycosylic bond attaching the target base to the sugar-phosphate DNA backbone. The subsequent abasic site deoxyribose sugar is removed by an apyrimidinic/apurinic (AP) endonuclease alone or an AP endonuclease in combination with the AP lyase activity of some DNA glycosylases. A DNA polymerase fills in the resulting gap in DNA, and DNA ligase finally seals the repaired strand [66]. Remarkably, the human 3-methyladenine DNA glycosylase (AAG, MPG, APNG) is the only human DNA glycosylase identified to date that removes alkylated base lesions [9]. AAG in specific and DNA glycosylases in general face the challenge of recognizing appropriate damaged targets amid a vast background of normal bases, and the substrate binding pocket of these enzymes play a crucial role. Our interest in the structure of the alkylated-base recognition pocket led us to apply the method of targeted random mutagenesis followed by selection to study the active site of the human 3-methyladenine DNA glycosylase.

Random oligonucleotide cassette mutagenesis coupled with selection is a powerful method in the study of protein structure-function relationships. In contrast to specific "rational" mutagenesis that examine the effect of one

particular substitution at one particular position in the protein, random oligonucleotide replacement "semi-rationally" target groups of amino acid residues of interest. Random mutagenesis techniques can be used to generate upwards of millions of mutants that are then selected or screened for desired properties. Active mutants reveal the degree of plasticity of the region of interest, highlight a range of tolerated substitutions that can confer novel activity to the protein of interest, and often reveal unexpected biological insights into protein function [89].

The human AAG has been reported to recognize a range of alkylation damaged bases, including 3-methyladenine, 7-methylguanine, and 3-methylguanine [35]. In addition, AAG also removes hypoxanthine, which derives from adenine deamination; and the lipid peroxidation derived base lesion 1,$N^6$-ethenoadenine (εA) [36]. Co-crystals of AAG complexed with DNA reveal that it, akin to other DNA glycosylases, uses a nucleotide "flipping" mechanism where a damaged base is flipped out of the double helix and bound into an enzyme active site [38,42]. A "nucleotide flipper" consisting of a tyrosine projecting from a β-hairpin on the surface of AAG inserts into the DNA minor groove, displacing the targeted nucleotide into the enzyme active site. Once in the substrate recognition pocket, the damaged base stacks between aromatic side chains and makes van der Waals contacts with additional residues. The cleavage of the C1'-*N* glycosylic bond is accomplished by an activated water molecule that has been deprotonated by Glu-125. Therefore, it is suggested that the hydrogen-

bonding characteristics, shape, aromaticity, and charge of the damaged base all contribute to recognition of the damaged base by AAG [42,90]

To evaluate the mutability of the AAG substrate recognition pocket and to generate novel DNA glycosylases with altered activities, we have created more than 4.5 million mutants of AAG harboring random substitutions in the enzyme active site. Using stringent selection in the DNA alkylation repair deficit *E. coli* strain MV1932, we have isolated more than 10,000 clones that confer protection against killing by the DNA alkylating agent methyl methane-sulfonate (MMS). The range of tolerated substitutions provides insight into the plasticity of the substrate binding pocket, and highlight a general approach for the generation of DNA glycosylases with new substrates specificities.

## MATERIALS & METHODS

### *Materials and strains*

*E. coli* strain MV1932 (AB1157 *ada alkA1*) was derived from strain AB1157 [71]. Chemicals were from Sigma-Aldrich (St. Louis, MO), enzymes were from NEB (Beverly, MA), DNA oligonucleotides were purchased from IDT (Coralville, IA) unless stated otherwise.

*Construction of the mutant AAG libraries*

*Creating the pUC-AAG expression vector*

The full length human AAG cDNA was generously provided by Leona Samson [37]. A 6X polyhistidine tag was introduced to the N-terminus of the AAG ORF by PCR amplification with the forward primer 5'GCA CGC GGA TCC ATG GGT GGT TCT CAT CAC CAT CAC CAT CAC GGT GGG ATC GAA GGT CGT GCT GTC ACC CCC GCT TTG CAG ATG AAG AAA CCA AAG CAG3' and the reverse primer 5'GAT GCG GCC GCG GAG TTC TGT GCC ATT AGG AAG TCG CCG3'. The amplified product was gel purified and cut with *BamH*I and *Sac*II. The pUC-AAG plasmid was generated by the cloning of this fragment into pUC19 using the *BamH*I and *Sac*II restriction sites.

*Creation of the pUC-AAG-SK and KX dummy constructs*

Silent restriction sites were created in the AAG coding frame in order to facilitate random oligonucleotide cassette insertion. To avoid contamination of the random mutagenesis libraries with recircularized vector containing wt-AAG, dummy constructs with internal deletions of the AAG gene were made prior to library creation. First, The pre-existing *Sac*II site in the pUC19 multiple cloning site was destroyed by *Sac*II digestion, followed by mung bean exonuclease end blunting and religation. To generate silent restriction sites within the AAG reading frame, site-directed mutagenesis using the QuickChange kit (Stratagene) was performed on pUC-AAG, introducing the silent mutations A354C to create a

*Sac*II site; G462T to create a *Kpn*I site; and A549T, G552A to create a *Xba*I site (numbers are in reference to the wildtype AAG reading frame).

To generate the internal deletion inserts for SK and KX dummy construction, the sense oligo 5'CCG CGG CCT GCA CGT GGT CGG TAC CTG TGC TAG CAA GTA TCT AGA3' was annealed to the antisense oligo 5'CTT AGA TAC TTG CTA GCA CAG GTA CCG ACC ACG TGC AGG CCG C3' to generate a DNA fragment with *Sac*II and *Xba*I sticky ends and internal *Pml*I, *Kpn*I, and *Nhe*I restriction sites. The fragment was digested with *Kpn*I and the fragments ligated into pUC-AAG that was pre-cut with *Sac*II and *Kpn*I in one reaction to generate the pUC-AAG-SK dummy, and ligated with pUC-AAG that were pre-cut with *Kpn*I and *Xba*I in another reaction to generate the pUC-AAG-KX dummy. Ligations were incubated with T4 DNA ligase (Invitrogen), and all constructs were subsequently restriction digestion mapped and sequenced. For conciseness, library constructs are designated SK or KX, which refer to the *Sac*II and *Kpn*I, and *Kpn*I and *Xba*I silent restriction sites used for random oligo cassette inserts in the respective libraries.

The SK random oligonucleotide cassette was constructed by annealing two single stranded DNA oligonucleotides:

1). SK sense: 5'- ACA GAA CTC CGC GGC CGC ATC GTG GAG ACC GAG **GCA TAC CTG GGT** CCA GAG GAT **GAA GCC GCC CAC TCA** *CGC GGT GGC CGG CAG ACG* -3' with

2). SK antisense: 5'- CAC GTA CAG GGT ACC CGG CTT CAT GAA CAT GCC TCG GTT GCG AGG *CGT CTG CCG GCC ACC GCG* -3'

The KX random oligonucleotide cassette was constructed by annealing the two single stranded DNA oligonucleotides:

3). KX sense: 5'-G AAG CCG GGT ACC CTG TAC **GTG TAC ATC** ATT TAC GGC ATG TAC TTC **TGC ATG AAC ATC** *TCC AGC CAG GGC GAC GGG* -3'

with 4). KX antisense: 5'- CAG CGG CTC TAG AGC TCG **CAG CAA GAC** GCA AGC *CCC GTC GCC CTG GCT GGA* -3'

The **bold, underlined** nucleotide positions were synthesized to constitute 88% wild-type sequence and 4% of each of the other three nucleotides. The 18 basepair region for hybridization are *italicized*. 150 pmoles of each pair of sense and antisense oligos were annealed at their complementary 3' ends in equimolar proportions in 1X Klenow buffer. 25 U of Klenow DNA polymerase and dNTPs were added to copy the template containing the randomized regions and to make fully duplex DNA. The reaction mix was heat inactivated and washed using Microcon YM-30 (Millipore, Bellerica, MA). SK random oligos were restriction enzyme digested with *Sac*II and *Kpn*I. KX random oligos were restriction enzyme digested with *Kpn*I and *Xba*I. The resulting reaction mixtures were heat inactivated, purified by Microcon protein removal column, and washed using Microcon YM-30. The SK and KX random oligonucleotide cassettes were ligated into the corresponding pre-cut pUC-AAG-SK-dummy, and pUC-AAG-KX-dummy constructs using T4 DNA ligase. Ligation products were

Microcon washed, and cut with restriction enzymes (*Pml*I in SK-dummy and *Nhe*I in KX-dummy) that linearize recircularized dummy constructs, but not mutant AAG constructs with successful random oligo inserts. The reactions were Microcon washed, and were used to transform XL1-Blue cells by electroporation. Sequencing of unselected and selected mutants did not reveal any surviving dummy constructs.

After plasmids containing the random libraries were electroporated into *E. coli* XL1-Blue, the sizes of each of the libraries were determined by plating an aliquot onto LB-carbenicillin (carb) agar plates. The remainder of the libraries was amplified by growing the transformed LX1-Blue cells in 1 liter of LB-carb for 12 hrs at 37°C. The pUC-AAG-SK and pUC-AAG-KX mutant AAG library plasmids were harvested by Maxi-prep (Qiagen, Valencia, CA).

***Construction of the YH library***

Positions Tyr-127 and His-136 were targeted for detailed analysis by the generation of libraries containing combinations of completely degenerate mutants at these two positions. The library is named YH in reference to Y-127 and H-136 positions. Essentially identical procedures were used as in the construction of the SK and KX libraries, except that the random YH sense oligo: 5′-ACA GAA CTC CGC GGC CGC ATC GTG GAG ACC GAG GCA **NNN** CTG GGT CCA GAG GAT GAA GCC GCC **NNN** TCA *CGC GGT GGC CGG CAG ACG*-3′ was used in place of the 1). SK sense oligo.

The **bold, underlined** nucleotide positions were synthesized to constitute a completely degenerate mixture consisting of 25% of each of the four nucleotides.

*Genetic selection for mutants active against MMS lesions*

The alkylated DNA repair deficient *E. coli* strain MV1932 (*alkA ada*) [71] was obtained from the Samson laboratory. MV1932 were transformed with pUC-AAG-SK and pUC-AAG-KX libraries and grown to mid-log phase in LB-carb. Prior to MMS selection, library aliquots were plated on LB-carb, and unselected clones were isolated and sequenced in order to characterize the composition of the unselected SK and KX libraries. 1 mL of MV1932 expressing wt pUC-AAG, SK library, KX library, and SK Dummy were treated to a final 0.2% (w/v) MMS for exactly one Hour. Afterwards, cells were serial diluted into M9 minimal media and spread onto LB-carb plates and incubated for 24 hours at 37°C. Survival curves were plotted from counting colonies.

*Genetic selection for mutator DNA glycosylases*

YH libraries were transformed into BW528 (*xth nfol::kan*) cells, grown overnight, and plate on LB-carb-rifampicin (100ug/ml). After incubation at 37 °C for 36 hours. Colonies were picked and plasmids from cells isolated by mini-prep. Plasmid containing putative mutator glycosylases were retransformed into fresh MV1932 cells and reassayed for rifampicin resistance mutation efficiency.

*DNA sequence analysis*

Plasmids from 72 unselected and 120 selected mutants of the pUC-AAG library were isolated by miniprep (Qiagen). The 198 bp *Sac*II-*Xba*I region of AAG was sequenced using the primer 5'-ACTGGGGTTGGAGTTCTTCG-3' and using BigDye 3.0 (ABI). Mutant Sequences were aligned and identified with Sequencher 3.0 (Gene Codes Corporation, Ann Arbor, MI).

## RESULTS AND DISCUSSION

### *AAG complementation of* E. coli *against MMS killing provides a stringent selection for active AAG mutants*

The Expression of wildtype AAG in the DNA alkylation repair deficient strain MV1932 confers as much as 54,000 fold protection against MMS killing (Table1), compensating for the strain's lack of endogenous DNA alkylation repair enzymes. This selective window provides a highly specific selection for active AAG mutants. Indeed, there were no inactive mutants with frameshift or nonsense mutations in the final selected pool, the only mutants found were those with amino acid substitutions.

### *Targeted, "semi-rational" libraries of AAG substrate binding pocket mutants*

Many evolutionarily conserved residues in AAG cluster around the active site. The crystal structure of AAG complexed with DNA containing an etheno-Adenine adduct reveals the residues that constitute the substrate binding pocket [42]. The residues Glu-125, Tyr-127, Ala-134, Ala-135, His-136, Tyr-159, Cys-167,

Asn-169 and Leu-180 are arranged in close proximity to the flipped-out alkylated base, with many making van der Waals contacts. In order to assess the degree of plasticity of residues that form the substrate binding pocket, we targeted most of the constituent residues and their flanking neighbors for random mutagenesis. Two libraries were created, designated SK (aa residues numbers 126-129, 133-137), and KX (aa residue numbers 158-160, 167-170, and 179-181) that contain amino acid substitutions at each of the numbered positions. Oligonucleotides containing 12% random substitutions at nucleotides encoding for Ala-126, Tyr-127, Leu-128, Gly-129; Glu-133, Ala-134, Ala-135, His-136, Ser-137 were used to replace the wild-type sequence in the SK library. Similarly constructed oligonucleotides were used to replace amino acid residues Val-158, Tyr-159, Iso-160; Cys-167, Met-168, Asn-169, Iso-170; Val-179, Leu-180, and Leu-181 in the KX library. The SK library was designed to average 3.2 nucleotide substitutions per mutant, which translates to approximately 2.4 amino acid substitutions. The KX library was designed to average 3.6 nucleotide substitutions per mutant, which translates to approximately 2.6 amino acid substitutions. Because of its crucial role in deprotonating a bound water and forming the hydroxyl nucleophile necessary for glycosylic bond cleavage, Glu125 was not mutated. The unselected pUC-AAG-SK and pUC-AAG-KX libraries each contain 2.8 $X10^6$ and 1.7$X10^6$ individual members, respectively.

*Selection of mutant AAGs that enhance survival after DNA alkylation damage*

Nineteen codons in five separate regions of the human AAG cDNA were randomized by replacing two cDNA segments corresponding to parts of the active site with partially random nucleotide sequences. The two libraries, totaling approximately 4.5 million individual variant human AAG cDNAs, were transformed into DNA glycosylase deficient *E. coli*. The randomized regions from 37 unselected members of the SK library were sequenced and multiple substitutions observed, averaging 2.2 amino acid changes (Figure 2) per mutant cDNA. The randomized regions from 35 unselected members of the KX library were also sequenced. There were an average of 2.3 amino acid changes (Figure 2) per mutant cDNA. A range of amino acid changes were seen at each codon (figure 2). Thus, the libraries encompass the expected diversity.

The two libraries were subjected to selection with 0.2% MMS and plated. The surviving fractions for Library A was 3.6 X $10^{-3}$%, and for library B was 2.2 X $10^{-2}$%, indicating that most of the mutants were unable to confer resistance to MMS at the dosage used, which was expected. Surviving colonies were picked for DNA sequencing.

*Selected mutants reveal tolerated substitutions and variant glycosylase active sites*

The degree of plasticity of active site residues generally followed the extent of evolutionary conservation, though with some unexpected exceptions. Three absolutely conserved residues Tyr-127, His-136, and Asn-169, did not

tolerate any single substitutions. The aromatic, planar Tyr-127 and His-136 residues form π-electron stacking interactions with the flipped out base and stabilizes it in the active site [38,42]. Substitutions at these positions invariably weaken substrate binding. The Asn-169 side chain has been proposed to help exclude normal guanines by sterically clashing with its exocyclic amino group ($N^2$) [42]. Perturbations at these three positions render AAG insufficient for the requirements of the selection. These results confirm the essential roles of these three residues in AAG function.

ii). The nonconserved residues Leu-128, Glu-133, Iso-160, and Met-168 tolerate a large and diverse range of substitutions (Figure 3). Given that none of these residues make direct contacts with the substrate, these results are not surprising. However, another evolutionarily nonconserved residue that does not make direct contact with the substrate is revealed to be immutable. Cys-167 may provide crucial structural support to maintain the shape of the active site pocket.

iii). The semi-conserved residues Gly-129, Ser-137, Val-158, and Iso-170, and Leu-181 tolerated a small limited region of substitutions.

iv). Strikingly, large numbers of mutants were recovered in which Leu-180 was substituted by phenylalanine and only phenylalanine (Figure 3). Mutants containing Leu-180 to Phe substitutions were enriched for by more than 10 fold after selection, even greater than the degree of enrichment of the wild-type AAG. It has been proposed that the specificity of alkylated DNA glycosylases, such as AAG and the E. coli AlkA [91] for alkylated bases is conferred by the enzymes' aromatic residue rich active pockets. The crystal structure of these enzymes

confirmed in large part by aromatic stacking interaction between Tyr-127, His-136, and Tyr-159. Leu-180 makes a hydrophobic contact with the C3-methyl group of 3-methyladenine modeled in the active site. Substitution of a phenylalanine ring for leucine at position 180 enhances aromatic π-electron interaction with the methylated base.

*Mutator DNA glycosylases*

By selecting for mutators from a library of altered AAGs, it may be possible to obtain variant AAGs that increase the endogenous mutation frequency. The rifampicin resistance assay was used in order to test this hypothesis. Rifampicin is an antibiotic that binds and inhibits RNA polymerase II at the active site. Resistance to rifampicin can arise from point mutations in the *rpoII* gene leading to altered RNA polymerase IIs that no longer bind rifampicin [92]. Spontaneous rifampicin resistance is an established assay for bacterial chromosomal mutation frequency.

The crystal structure of AAG and our data has shown Tyr-127 and His-136 of AAG to be crucial in substrate recognition (Figure 6). Mutations at these residues may produce mutator proteins. Therefore, a library (designated as YH) was created in which all possible amino acid substitutions at positions 127 and 136 are represented. Mutators were selected for by the rifampicin assay among the members of this library. A double mutant Y127I, H136L was isolated. This mutant reproducibly elevates the endogenous mutation frequency by as much

1,000 fold when expressed in *E. coli* (Figure 5). The active site of this double mutant is visualized in conjunction with that of the wild-type AAG (Figure 6).

In this section, we describe a general approach to evolve novel DNA glycosylases by targeted random substitutions in the enzyme active site. This points to the feasibility of reengineering DNA glycosylases to remove novel substrates from DNA, for biomedical, or biotechnology applications, or generating experimental reagents.

**Table 3.1 MV1932 survival after MMS treatment**

| | Survival after MMS treatment | |
|---|---|---|
| | Percent Survival | Ratio over dummy |
| Wild type AAG | $2.5 \times 10^{0}$ | 54,000 |
| AAG library A | $3.6 \times 10^{-3}$ | 80 |
| AAG library B | $2.2 \times 10^{-2}$ | 480 |
| Dummy | $4.6 \times 10^{-5}$ | 1 |

Cultures of MV1932 cells expressing either wild-type or inactivated (Dummy) human AAG, and randomized AAG libraries SK and KX were treated with 0.2% MMS. Cultures were washed, diluted, and plated, and the number of surviving colonies counted after incubation at 37°C for 24 hours. Plasmids from randomized libraries' survivors were sequenced.

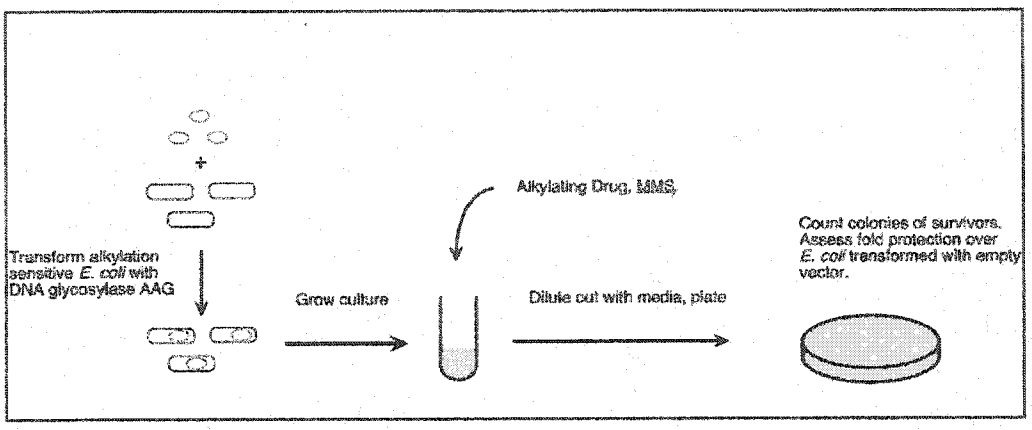



Figure 3.1

**Scheme of AAG MV1932 complementation**

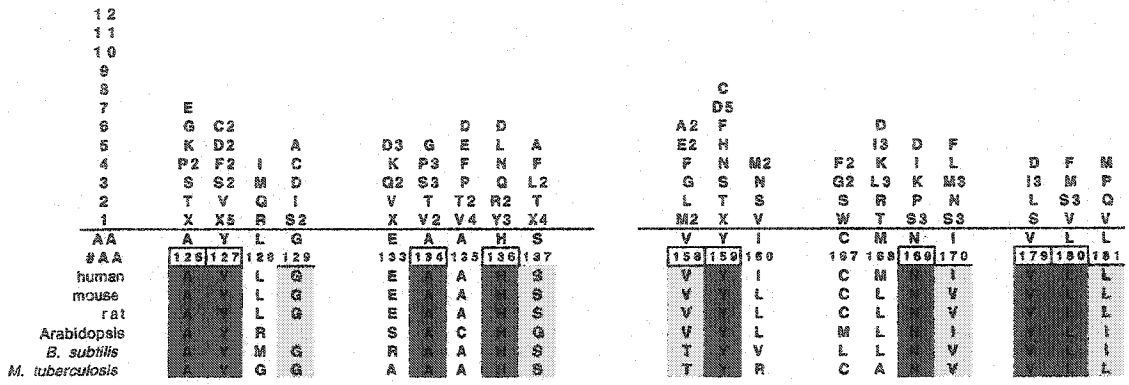



Figure 3.2

**Unselected AAG random oligo mutagenesis library sequences**
All sites subjected to random oligonucleotide insertional mutagenesis were mutated. Below the human primary sequence are aligned sequence of AAG homolog from diverse organisms. Purple indicates absolute conservation, also indicated by rectangles. Yellow indicates conservative changes.

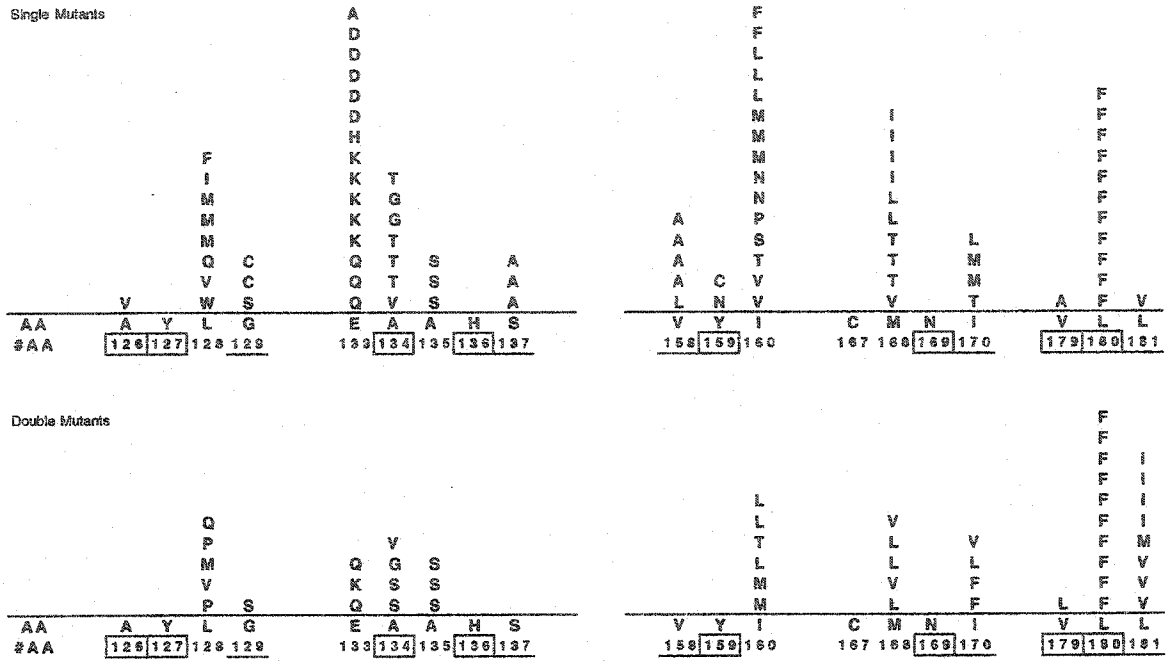



Figure 3.3

**Sequences from selected AAG active site mutants**
Amino acid substitutions in selected AAG mutants. The wild-type AAG sequence is shown below the solid line, and the substitutions observed at each position and their frequency are indicated above the solid line. Phylogenetically invariant residues are boxed, and positions with conservative residue substitutions are underlined.

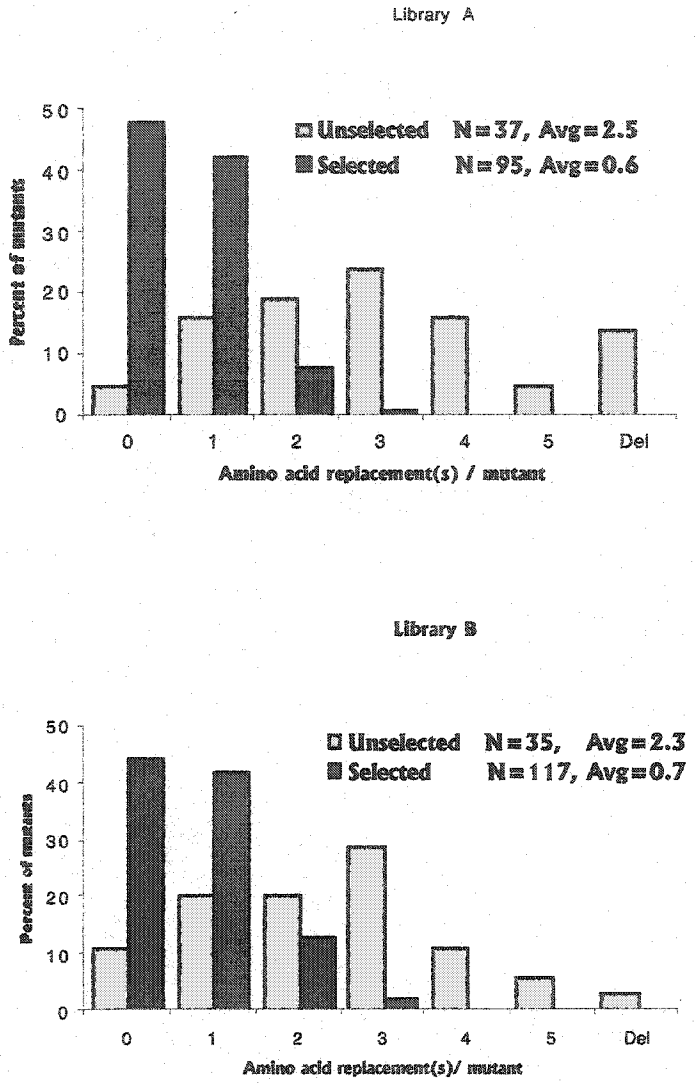



Figure 3.4

**Histograms of mutational loads among unselected and selected AAG mutants**
The SK library is shown at the top. The KX library is shown at the bottom. Note the "left-ward shift" toward lower average mutational load when the unselected mutants are compared with the selected mutants.

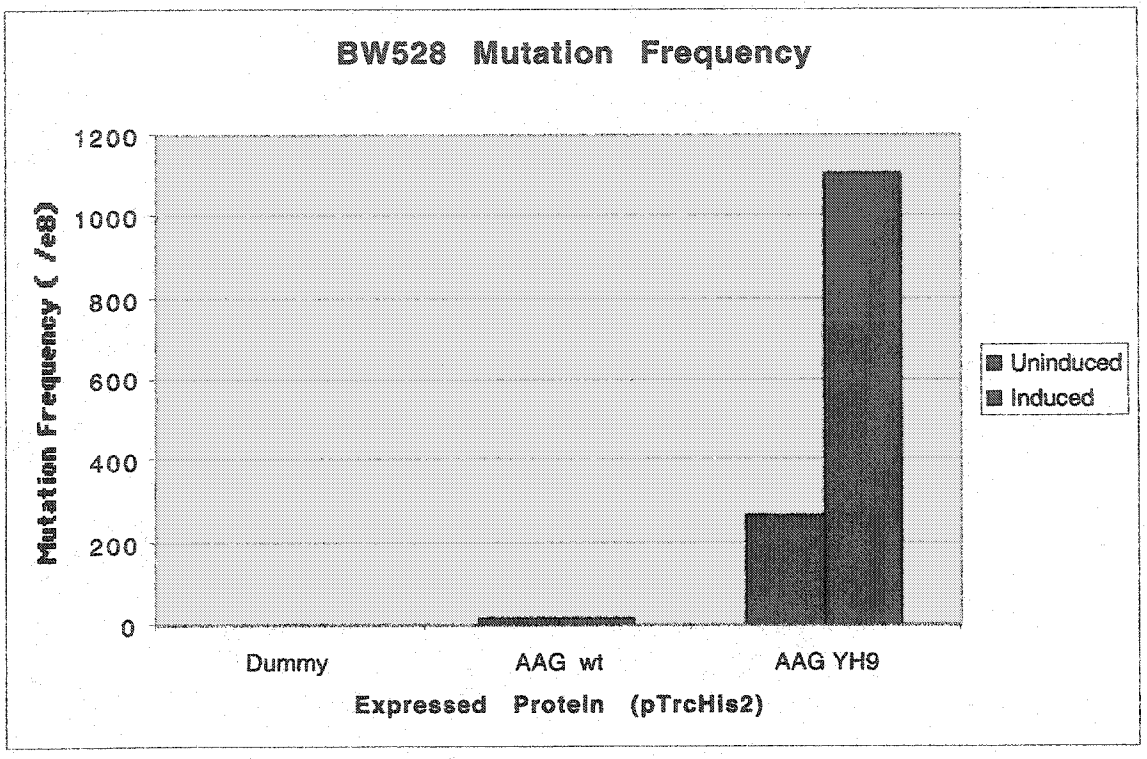



Figure 3.5

**Mutant AAG increases mutation frequency in *E. coli***
The double mutant Y127I, H136L increases mutation frequency by as much as 1,100 fold when expressed in *E. coli*.

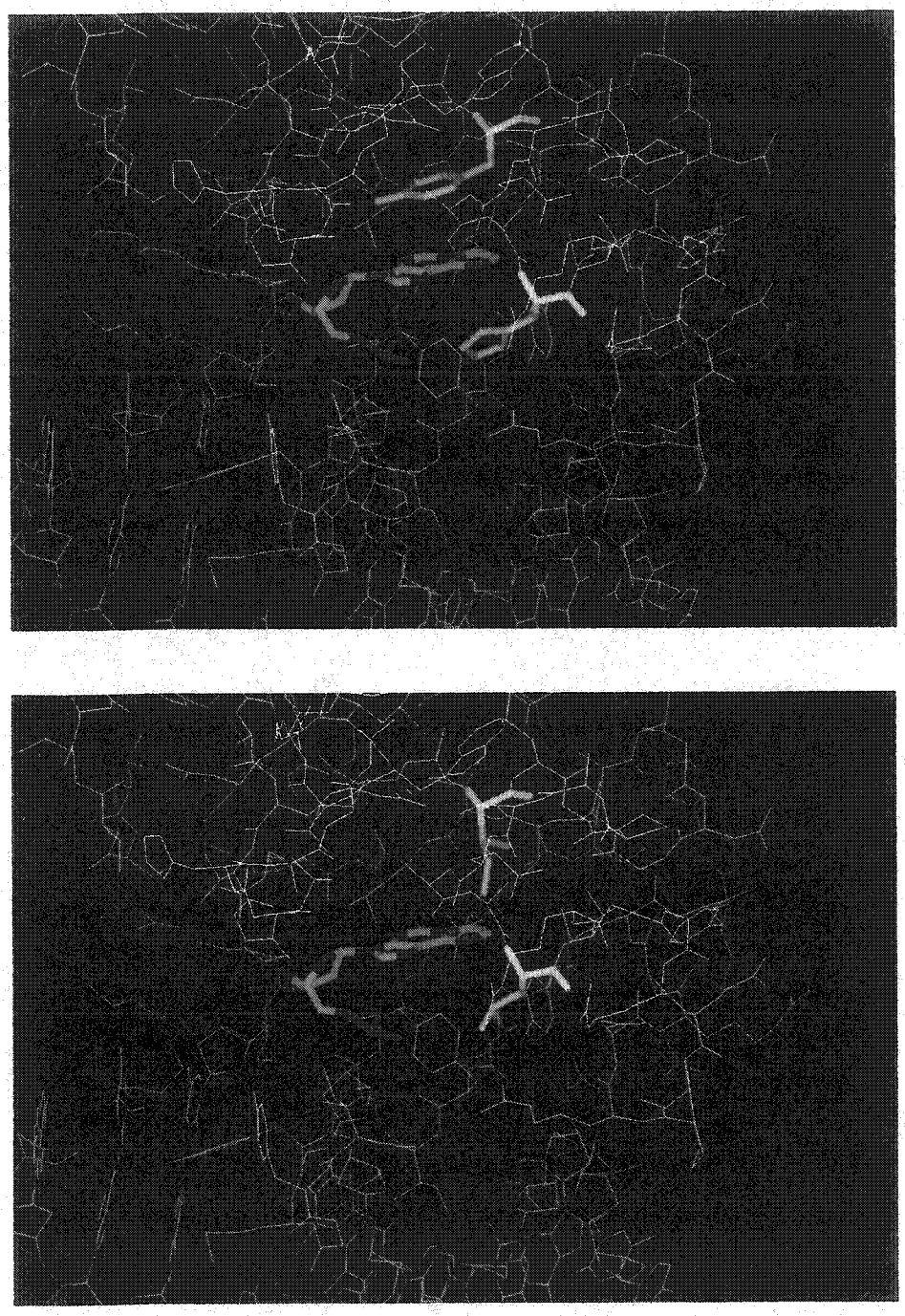


Figure 3.6

**The wild-type (top) and Y127I, H136L double mutant (bottom) active sites**

## *Chapter 4. Concluding remarks*

"You get what you select for" is a common saying in the directed evolution field. In this work, I have used AAG as a model protein to explore the degree of tolerance that a protein exhibits toward random amino acid changes. I have also isolated mutant human 3-methyladenine DNA glycosylase repair proteins that remove DNA alkylation lesions. I have targeted the active site of AAG for mutagenesis in order to obtain variant active site mutants. I have also isolated a variant DNA glycosylase that cause elevated mutation frequencies in *E. coli*. In the future, mutant DNA glycosylases that are active against other DNA substrates may be isolated, if sufficiently efficient selection or screening strategies can be devised.

Directed evolution is a semi-rational method to develop novel proteins. Directed evolution is often necessary because our understanding of protein structure and folding properties are not sufficiently robust to allow us to design enzymes *de novo*. However, given the advent of high resolution protein structures from crystallographic and NMR studies, increasing computational power, and more sophisticated protein structure prediction algorithms, protein design is steadily becoming an eventual reality. To rationally design mutant AAGs with altered substrate specificities, new DNA adducts may be modeled into the active site, and the amino acid constituents of the active site can be re-designed in order to accommodate the new substrate. However, this strategy

relies on a static "snap-shot" view of the enzyme, and does not directly account for the other dynamic steps of lesion detection, base-flipping, and enzymatic catalysis. In the choice of the new target substrate, these other factors also need to be taken into account.

As our ability to design proteins becomes more sophisticated, directed evolution will continue to play a role in the optimization of novel proteins. Directed evolution can aid in searching distant sequence space to identify mutations far from the active site. Kuhlman and coworkers were able to design a small protein from scratch with a structure verified by x-ray crystallography [93]. Also using computer-based rational design, Dwyer *et al* successfully converted the catalytically inert ribose-binding protein into an enzyme that is active as a triose phosphate isomerase (TIM) [94]. The designed TIM mutant contained thirteen amino acid replacements and exhibited an activity above background. To improve catalytic activity, directed evolution in the form of error-prone PCR was used to generate and isolate mutants with $10^5$-$10^6$ fold activity above background [94]. Enhanced mutants exhibited changes at protein surface residues, away from the active site. This illustrates the ability of gene-wide mutagenesis, followed by selection, to search sequence space of distant residues to complement rational design strategies.

The creation of large numbers of variant AAGs with altered active sites after selection by MMS may highlight the creation of more specific, "tailored" 3-methyladenine DNA glycosylases. Recent works by O'Brien and Ellenberger have shown that the wild-type AAG is most proficient at removing the

hypoxanthine base lesion, with a $10^8$-fold enhancement. Hypoxanthine, an oxidative deamination product of adenine, can mispair with cytosine during replication. This led the authors to hypothesize that the primary substrate target of AAG *in vivo* is hypoxanthine. Then our stringent selection against 3-methyladenine lesions, which is the predominant toxic lesion created by MMS [95], would select for mutant AAGs with redirected selectivity toward 3-methyladenine. It is not clear if these mutants, such as L180F, would exhibit decreased activity toward hypoxanthine or ethanoadenine. This remains to be tested by *in vitro* studies.

The use of DNA damaging agents continues to be a mainstay of current cancer therapy [96]. Alkylating agents such as cyclophosphamide, chlorambucil, and mechloroethamine make bulky alkyl-base adducts, most frequently at the N7 position of guanines. Due to their toxicity to rapidly dividing cells, the side-effects of their use include GI tract epithelial ablation, allopecia, and bone marrow suppression, which is often the dose-limiting effect. Gene therapy of bone marrow cells with engineered drug resistance genes has been proposed [89]. Continuing advances in understanding gene transduction processes and specific cell targeting suggest that the creation of DNA glycosylases that specifically remove chemotherapy-induced base lesions may be clinically useful. This dissertation lays the foundation for further work to advance into this aim: to further combine understanding of DNA glycosylase mechanisms with directed evolution and rational design to create more therapies to alleviate suffering from cancer and cancer therapy.

## *Bibliography*


1. Hoeijmakers, J.H. (2001) Genome maintenance mechanisms for preventing cancer. *Nature*, **411**, 366-74.

2. Bruner, S.D., Norman, D.P. and Verdine, G.L. (2000) Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature*, **403**, 859-66.

3. Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709-15.

4. Lindahl, T. and Wood, R.D. (1999) Quality Control by DNA Repair. *Science*, **286**, 1897-1905.

5. Encell, L.P., Coates, M.M. and Loeb, L.A. (1998) Engineering human DNA alkyltransferases for gene therapy using random sequence mutagenesis. *Cancer Res*, **58**, 1013-20.

6. Yu, Z., Chen, J., Ford, B.N., Brackley, M.E. and Glickman, B.W. (1999) Human DNA repair systems: an overview. *Environ Mol Mutagen*, **33**, 3-20.

7. David, S.S. and Williams, S.D. (1998) Chemistry of Glycosylases and Endonucleases Involved in Base-Excision Repair. *Chem Rev*, **98**, 1221-1262.

8. Xiao, W. and Samson, L. (1993) In vivo evidence for endogenous DNA alkylation damage as a source of spontaneous mutation in eukaryotic cells. *Proc Natl Acad Sci U S A*, **90**, 2117-21.

9. Wyatt, M.D., Allan, J.M., Lau, A.Y., Ellenberger, T.E. and Samson, L.D. (1999) 3-methyladenine DNA glycosylases: structure, function, and biological importance. *Bioessays*, **21**, 668-76.

10. Parikh, S.S., Mol, C.D., Slupphaug, G., Bharati, S., Krokan, H.E. and Tainer, J.A. (1998) Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. *Embo J*, **17**, 5214-26.

11. Friedberg, E.C., Walker, G.C. and Siede, W. (1995) *DNA repair and mutagenesis*. ASM Press, Washington, D.C.

12. Lindahl, T. and Nyberg, B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, **13**, 3405-10.

13. Guan, Y., Manuel, R.C., Arvai, A.S., Parikh, S.S., Mol, C.D., Miller, J.H., Lloyd, S. and Tainer, J.A. (1998) MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nat Struct Biol*, **5**, 1058-64.

14. Harris, R.S., Sheehy, A.M., Craig, H.M., Malim, M.H. and Neuberger, M.S. (2003) DNA deamination: not just a trigger for antibody diversification but also a mechanism for defense against retroviruses. *Nat Immunol*, **4**, 641-3.

15. Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S. and Malim, M.H. (2003) DNA deamination mediates innate immunity to retroviral infection. *Cell*, **113**, 803-9.

16. Klimasauskas, S., Kumar, S., Roberts, R.J. and Cheng, X. (1994) HhaI methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357-69.

17. Reinisch, K.M., Chen, L., Verdine, G.L. and Lipscomb, W.N. (1995) The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell*, **82**, 143-53.

18. Kavli, B., Slupphaug, G., Mol, C.D., Arvai, A.S., Peterson, S.B., Tainer, J.A. and Krokan, H.E. (1996) Excision of cytosine and thymine from DNA by mutants of human uracil-DNA glycosylase. *Embo J*, **15**, 3442-7.

19. Pavlov, A.R., Belova, G.I., Kozyavkin, S.A. and Slesarev, A.I. (2002) Helix-hairpin-helix motifs confer salt resistance and processivity on chimeric DNA polymerases. *Proc Natl Acad Sci U S A*, **99**, 13510-5.

20. Cunningham, R.P. (1997) DNA glycosylases. *Mutat Res*, **383**, 189-96.

21. Michaels, M.L. and Miller, J.H. (1992) The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J Bacteriol*, **174**, 6321-5.

22. McAuley-Hecht, K.E., Leonard, G.A., Gibson, N.J., Thomson, J.B., Watson, W.P., Hunter, W.N. and Brown, T. (1994) Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs. *Biochemistry*, **33**, 10266-70.

23. Larson, K., Sahm, J., Shenkar, R. and Strauss, B. (1985) Methylation-induced blocks to in vitro DNA replication. *Mutat Res*, **150**, 77-84.

24. Allan, J.M., Engelward, B.P., Dreslin, A.J., Wyatt, M.D., Tomasz, M. and Samson, L.D. (1998) Mammalian 3-methyladenine DNA glycosylase protects against the toxicity and clastogenicity of certain chemotherapeutic DNA cross-linking agents. *Cancer Res*, **58**, 3965-73.

25. Thomas, L., Yang, C.H. and Goldthwait, D.A. (1982) Two DNA glycosylases in Escherichia coli which release primarily 3-methyladenine. *Biochemistry*, **21**, 1162-9.

26. Krokan, H.E., Standal, R. and Slupphaug, G. (1997) DNA glycosylases in the base excision repair of DNA. *Biochem J*, **325 ( Pt 1)**, 1-16.

27. Labahn, J., Scharer, O.D., Long, A., Ezaz-Nikpay, K., Verdine, G.L. and Ellenberger, T.E. (1996) Structural basis for the excision repair of alkylation-damaged DNA. *Cell*, **86**, 321-9.

28. Berdal, K.G., Johansen, R.F. and Seeberg, E. (1998) Release of normal bases from intact DNA by a native DNA repair enzyme. *Embo J*, **17**, 363-7.

29. Hatahet, Z., Kow, Y.W., Purmal, A.A., Cunningham, R.P. and Wallace, S.S. (1994) New substrates for old enzymes. 5-Hydroxy-2'-deoxycytidine and 5-hydroxy-2'-deoxyuridine are substrates for Escherichia coli endonuclease III and formamidopyrimidine DNA N-glycosylase, while 5-hydroxy-2'-deoxyuridine is a substrate for uracil DNA N-glycosylase. *J Biol Chem*, **269**, 18814-20.

30. Hollis, T., Ichikawa, Y. and Ellenberger, T. (2000) DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, Escherichia coli AlkA. *Embo J*, **19**, 758-66.

31. Yamagata, Y., Kato, M., Odawara, K., Tokuno, Y., Nakashima, Y., Matsushima, N., Yasumura, K., Tomita, K., Ihara, K., Fujii, Y., Nakabeppu, Y., Sekiguchi, M. and Fujii, S. (1996) Three-dimensional structure of a DNA repair enzyme, 3-methyladenine DNA glycosylase II, from Escherichia coli. *Cell*, **86**, 311-9.

32. Radman, M. (1976) An endonuclease from Escherichia coli that introduces single polynucleotide chain scissions in ultraviolet-irradiated DNA. *J Biol Chem*, **251**, 1438-45.

33. Thayer, M.M., Ahern, H., Xing, D., Cunningham, R.P. and Tainer, J.A. (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *Embo J*, **14**, 4108-20.

34. Singer, B. and Hang, B. (1997) What structural features determine repair enzyme specificity and mechanism in chemically modified DNA? *Chem Res Toxicol*, **10**, 713-32.

35. Kartalou, M., Samson, L.D. and Essigmann, J.M. (2000) Cisplatin adducts inhibit 1,N(6)-ethenoadenine repair by interacting with the human 3-methyladenine DNA glycosylase. *Biochemistry*, **39**, 8032-8.

36. O'Connor, T.R. (1993) Purification and characterization of human 3-methyladenine-DNA glycosylase. *Nucleic Acids Res*, **21**, 5561-9.

37. Samson, L., Derfler, B., Boosalis, M. and Call, K. (1991) Cloning and characterization of a 3-methyladenine DNA glycosylase cDNA from human cells whose gene maps to chromosome 16. *Proc Natl Acad Sci U S A*, **88**, 9127-31.

38. Lau, A.Y., Scharer, O.D., Samson, L., Verdine, G.L. and Ellenberger, T. (1998) Cystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanisms for Nucleotide Flipping and Base Excision. *Cell*, **95**, 249-258.

39. Engelward, B.P., Weeda, G., Wyatt, M.D., Broekhof, J.L., de Wit, J., Donker, I., Allan, J.M., Gold, B., Hoeijmakers, J.H. and Samson, L.D. (1997) Base excision repair deficient mice lacking the Aag alkyladenine DNA glycosylase. *Proc Natl Acad Sci U S A*, **94**, 13087-92.

40. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. and al, e. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185-95.

41. Arabidopsis-Genome-Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796-815.

42. Lau, A.Y., Wyatt, M.D., Glassner, B.J., Samson, L.D. and Ellenberger, T. (2000) Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. *Proc Natl Acad Sci U S A*, **97**, 13573-8.

43. Gallivan, J.P. and Dougherty, D.A. (1999) Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A*, **96**, 9459-64.

44. Patel, P.H. and Loeb, L.A. (2000) DNA polymerase active site is highly mutable: evolutionary consequences. *Proc Natl Acad Sci U S A*, **97**, 5095-100.

45. Glick, E., Vigna, K.L. and Loeb, L.A. (2001) Mutations in human DNA polymerase eta motif II alter bypass of DNA lesions. *Embo J*, **20**, 7303-12.

46. Kawate, H., Landis, D.M. and Loeb, L.A. (2002) Distribution of mutations in human thymidylate synthase yielding resistance to 5-fluorodeoxyuridine. *J Biol Chem*, **277**, 36304-11.

47. Landis, D.M., Heindel, C.C. and Loeb, L.A. (2001) Creation and characterization of 5-fluorodeoxyuridine-resistant Arg50 loop mutants of human thymidylate synthase. *Cancer Res*, **61**, 666-72.

48. Kokoris, M.S., Sabo, P. and Black, M.E. (2000) In vitro evaluation of mutant HSV-1 thymidine kinases for suicide gene therapy. *Anticancer Res*, **20**, 959-63.

49. Tindall, K.R. and Kunkel, T.A. (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry*, **27**, 6008-6013.

50. Beckman, R.A., Mildvan, A.S. and Loeb, L.A. (1985) On the fidelity of DNA replication: manganese mutagenesis *in vitro*. *Biochemistry*, **24**, 5810-5817.

51. Bebenek, K., Roberts, J.D. and Kunkel, T.A. (1992) The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J Biol Chem*, **267**, 3589-96.

52. Pavlov, Y.I., D.T., M., Izuta, S. and Kunkel, T.A. (1994) DNA replication fidelity and 8-oxodeoxyguanosine triphosphate. *Biochemistry*, **33**, 4685-4701.

53. Zaccolo, M. and Gherardi, E. (1999) The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 beta-lactamase. *J Mol Biol*, **285**, 775-83.

54. Suzuki, M., Avicola, A.K., Hood, L. and Loeb, L.A. (1997) Low fidelity mutants in the O-helix of *Thermus aquaticus* DNA polymerase I. *J. Biol. Chem.*, **272**, 11228-11235.

55. Patel, P.H., Kawate, H., Adman, E., Ashbach, M. and Loeb, L.A. (2001) A single highly mutable catalytic site amino acid is critical for DNA polymerase fidelity. *J Biol Chem*, **276**, 5044-51.

56. Vartanian, J.P., Henry, M. and Wain-Hobson, S. (1996) Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Res*, **24**, 2627-31.

57. Shafikhani, S., Siegel, R.A., Ferrari, E. and Schellenberger, V. (1997) Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotechniques*, **23**, 304-10.

58. Christians, F.C. and Loeb, L.A. (1996) Novel human DNA alkyltransferases obtained by random substitution and genetic selection in bacteria. *Proc Natl Acad Sci U S A*, **93**, 6124-8.

59. Smith, J.M. (1970) Natural Selection and the Concept of a Protein Space. *Nature*, **225**, 563-564.

60. Suzuki, M., Baskin, D., Hood, L. and Loeb, L.A. (1996) Random mutagenesis of Thermus aquaticus DNA polymerase I: concordance of immutable sites in vivo with the crystal structure. *Proc Natl Acad Sci U S A*, **93**, 9670-5.

61. Loeb, L.A., Loeb, K.R. and Anderson, J.P. (2003) Multiple mutations and cancer. *Proc Natl Acad Sci U S A*, **100**, 776-81.

62. Davidson, J.F., Guo, H.H. and Loeb, L.A. (2002) Endogenous mutagenesis and cancer. *Mutat Res*, **509**, 17-21.

63. Suzuki, M., Christians, F.C., Kim, B., Skandalis, A., Black, M.E. and Loeb, L.A. (1996) Tolerance of different proteins for amino acid diversity. *Mol Divers*, **2**, 111-8.

64. Guo, H.H. and Loeb, L.A. (2003) Tumbling down a different pathway to genetic instability. *J Clin Invest*, **112**, 1793-5.

65. Mattes, W.B., Lee, C.S., Laval, J. and O'Connor, T.R. (1996) Excision of DNA adducts of nitrogen mustards by bacterial and mammalian 3-methyladenine-DNA glycosylases. *Carcinogenesis*, **17**, 643-8.

66. Scharer, O.D. and Jiricny, J. (2001) Recent progress in the biology, chemistry and structural biology of DNA glycosylases. *Bioessays*, **23**, 270-81.

67. Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.

68. Creighton, T.E. (1993) *Proteins*. W.H. Freeman, New York.

69. Bashford, D., Chothia, C. and Lesk, A.M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol*, **196**, 199-216.

70. Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A. and Sauer, R.T. (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306-10.

71. Posnick, L.M. and Samson, L.D. (1999) Imbalanced base excision repair increases spontaneous mutation and alkylation sensitivity in Escherichia coli. *J Bacteriol*, **181**, 6763-71.

72. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, **8**, 186-94.

73. Wyatt, M.D. and Samson, L.D. (2000) Influence of DNA structure on hypoxanthine and 1,N(6)-ethenoadenine removal by murine 3-methyladenine DNA glycosylase. *Carcinogenesis*, **21**, 901-8.

74. Beasley, J.R. and Hecht, M.H. (1997) Protein design: the choice of de novo sequences. *J Biol Chem*, **272**, 2031-4.

75. Landis, D.M. and Loeb, L.A. (1998) Random sequence mutagenesis and resistance to 5-fluorouridine in human thymidylate synthases. *J Biol Chem*, **273**, 25809-17.

76. Kim, B., Hathaway, T.R. and Loeb, L.A. (1996) Human immunodeficiency virus reverse transcriptase. Functional mutants obtained by random mutagenesis coupled with genetic selection in Escherichia coli. *J Biol Chem*, **271**, 4872-8.

77. Black, M.E. and Loeb, L.A. (1993) Identification of important residues within the putative nucleoside binding site of HSV-1 thymidine kinase by random sequence selection: analysis of selected mutants *in vitro*. *Biochemistry*, **32**, 11618-11626.

78. Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S. and Miller, J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol*, **240**, 421-33.

79. Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344-7.

80. Gregoret, L.M. and Sauer, R.T. (1993) Additivity of mutant effects assessed by binomial mutagenesis. *Proc Natl Acad Sci U S A*, **90**, 4246-50.

81. Axe, D.D., Foster, N.W. and Fersht, A.R. (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry*, **37**, 7157-66.

82. Rennell, D., Bouvier, S.E., Hardy, L.W. and Poteete, A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol*, **222**, 67-88.

83. Roy, R., Biswas, T., Hazra, T.K., Roy, G., Grabowski, D.T., Izumi, T., Srinivasan, G. and Mitra, S. (1998) Specific interaction of wild-type and truncated mouse N-methylpurine-DNA glycosylase with ethenoadenine-containing DNA. *Biochemistry*, **37**, 580-9.

84. Loeb, L.A., Essigmann, J.M., Kazazi, F., Zhang, J., Rose, K.D. and Mullins, J.I. (1999) Lethal mutagenesis of HIV with mutagenic nucleoside analogs. *Proc Natl Acad Sci U S A*, **96**, 1492-7.

85. Drake, J.W., Charlesworth, B., Charlesworth, D. and Crow, J.F. (1998) Rates of spontaneous mutation. *Genetics*, **148**, 1667-86.

86. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*, **11**, 863-74.

87. Saunders, C.T. and Baker, D. (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol*, **322**, 891-901.

88. Radman, M., Matic, I. and Taddei, F. (1999) Evolution of evolvability. *Ann N Y Acad Sci*, **870**, 146-55.

89. Encell, L.P., Landis, D.M. and Loeb, L.A. (1999) Improving enzymes for cancer gene therapy. *Nat Biotechnol*, **17**, 143-7.

90. O'Brien, P.J. and Ellenberger, T. (2004) Dissecting the broad substrate specificity of human 3-methyladenine-DNA glycosylase. *J Biol Chem*, **279**, 9750-7.

91. Hollis, T., Lau, A. and Ellenberger, T. (2000) Structural studies of human alkyladenine glycosylase and E. coli 3-methyladenine glycosylase. *Mutat Res*, **460**, 201-10.

92. Ovchinnikov, Y.A., Monastyrskaya, G.S., Guriev, S.O., Kalinina, N.F., Sverdlov, E.D., Gragerov, A.I., Bass, I.A., Kiver, I.F., Moiseyeva, E.P., Igumnov, V.N., Mindlin, S.Z., Nikiforov, V.G. and Khesin, R.B. (1983) RNA polymerase

rifampicin resistance mutations in Escherichia coli: sequence changes and dominance. *Mol Gen Genet*, **190**, 344-8.

93. Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364-8.

94. Dwyer, M.A., Looger, L.L. and Hellinga, H.W. (2004) Computational Design of a Biologically Active Enzyme. *Science*, **304**, 1967-1971.

95. Sedgwick, B. (2004) Repairing DNA-methylation damage. *Nat Rev Mol Cell Biol*, **5**, 148-57.

96. Goodman, L.S., Hardman, J.G., Limbird, L.E. and Gilman, A.G. (2001) *Goodman & Gilman's the pharmacological basis of therapeutics*. McGraw-Hill, New York.

# VITA

Haiwei Guo was born in Harbin, China in 1974 to parents who were instructors at the local technical university. His father in electrical engineering, his mother in chemistry. His family immigrated to the United States and settled in the San Francisco Bay Area in 1984. He attended high school at Lynbrook High School in San Jose. Eager to experience life on the East Coast, he left for M.I.T. for his undergraduate studies, in Cambridge, Massachusetts. There, he did research at the Whitehead Institute, rowing in crew, and resided in the Chi Phi fraternity in backbay Boston. He graduated in 1997 with a B.S. in biology and after a stint travelling, started the M.D./Ph.D. program at the University of Washington. He has been researching directed evolution and DNA repair in the laboratory of Dr. Lawrence A. Loeb for the past five years.