ABSTRACT

Title of Document: DEVELOPMENT AND APPLICATIONS OF CODON SCANNING MUTAGENESIS: A NOVEL MUTAGENESIS METHOD THAT FACILITATES IN-FRAME CODON MUTATIONS

Kelly Anne Daggett, Doctor of Philosophy, 2009

Directed By: Assistant Professor T. Ashton Cropp
Department of Chemistry and Biochemistry

The ability to create protein variants is a very valuable tool in biochemistry. Information about mechanistic roles of amino acid side chains, protein topology and binding can all be obtained. Methodologies to mutate proteins also allow for new catalytic activity to be achieved. While the routinely used methods to alter a protein sequence have proven to be useful, to some degree each of these methods requires some knowledge of protein structure to determine the site of mutation. Further, the routinely used methods also only allow for a specified site to be changed to a pre-determined residue (directed by oligonucleotides) or for multiple random sites to be changed to a non-specified residue. This dissertation focuses on the development of a method that allows for a new defined amino acid to replace a native amino acid at a random location within in the protein.

To introduce mutations at random locations within a protein coding sequence, three steps need to be accomplished. First, the coding sequence needs to be randomly digested on both strands; second, three nucleotides (a codon) at the digestion site need to be removed; and last, a new specified codon inserted. This process results in the replacement of a random codon with the new defined codon. To direct a mutation at a random location, the unique properties of a transposase/transposon are used to create both the double strand break and removal of three nucleotides. The insertion of the new defined codon is introduced using a linker sequence that when inserted in the correct reading frame a selectable phenotype is produced. This process has been termed Codon Scanning Mutagenesis (CSM).

The advantages of this method over current mutagenesis methods are (1) knowledge of structural information is not required, (2) oligonucleotides are not required to introduce the mutation and (3) the mutagenesis method allows for every amino acid to be mutated regardless of the DNA sequence. Further, this method allows for any natural and unnatural amino acid to be inserted at the mutation site, as well as the ability to create mutational mixtures or introduce multiple user defined mutations.

DEVELOPMENT AND APPLICATIONS OF CODON SCANNING MUTAGENESIS: A NOVEL MUTAGENESIS METHOD THAT FACILITATES IN-FRAME CODON MUTATIONS

By

Kelly Anne Daggett

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Assistant Professor T. Ashton Cropp, Chair
Professor Steven Rokita
Professor Philip DeShong
Assistant Professor Barbara Gerratana
Associate Professor John Fisher

Dedication

I dedicate this dissertation in loving memory
of my Dad, Michael, who taught me to always strive
for the best and to never give up on my dreams.

Acknowledgements

I would like to first thank my advisor Dr. Ashton Cropp, I am forever grateful for everything that he has done for me. Without your patience, continued support and excellence in teaching this dissertation would not have been possible. Thank you for all of the opportunities to grow as both researcher and teacher. You are a great advisor and I am thrilled that I am one of your first graduate students.

Thank you to all my committee members. You have all been very supportive and I am grateful to each of you for all of the advice and guidance that you have offered as well as the numerous letters of recommendations that you have written over the years. Dr. Rokita, thank you for helping me become a better teacher. Dr. Gerratana, thank you for taking the time to help me prepare for post-doc interviews. Dr. DeShong, thank you for being a wonderful educator and for all entertaining stories at happy hour. I would also like to thank Dr. Fisher for agreeing to serve as the Dean's representative.

Thanks to the entire Cropp lab, both past and present members. A special thanks to both Bryan Wilkins and Mark Layer. Bryan, it has been wonderful to work with you and this project would have never started without you (especially since it was yours in the beginning). It has been great to share the experience of being Ashton's first graduate students with you. I wish you much success in the future. Mark, thank you for all of the numerous mini-preps and PCRs' that you have performed during the infancy of this project, without you I would not have been able to work through the problems that arose efficiently. I also need to thank you for being a fantastic student and a fun person to work with. Thank you to Jia Liu, Zijun Chen

and Karina Herrera for making the lab an enjoyable work place and for all of your helpful comments. I also need to thank the past undergraduate students, Asher Page and Sam Marionni for making lab entertaining, I wish you both much success in your graduate careers. Thank you, Bryan and Jia for all of your editing and helpful comments.

To my fellow graduate students who started in 2004; Will Harrell, Brian Borak, Neil Campbell, Melissa Resto and Sara Lioi, for your friendship and all of the help that you have given me over the years. Thank you all for making this a memorable experience.

Thank you the members of the Gerratana and Rokita labs, for all the enzymes, reagents, etc. that I have "borrowed" over the years and insightful conversations.

To Joe Nemesh, for all of your support, patience (when I was crazy at times), and understanding throughout the 5 years. And for the invaluable help in formatting this dissertation. I also need to thank you for you friendship when we first met and helping me realize my potential, without which this dissertation would not have existed.

And last to my family, especially my parents, Mike and Barbara for instilling in me the values of getting an education at a young age. And thank you to my siblings; Mikie, Brittany and Kyle for all of the laughs over the years.

Table of Contents

List of Tables

List of Figures

List of Schemes

List of Equations

Abbreviations

2XYT	yeast extract tryptone medium
5-FU	5-fluorouracil
ATP	adenosine triphosphate
bp	base pairs
CSM	Codon Scanning Mutagenesis
DNA	deoxyribonucleic acid
DNAse	deoxyribonuclease
dNTP	deoxynucleotide triphosphate
EDTA	ethylenediaminetetraacetic acid
EIPCR	enzymatic inverse polymerase chain reaction
FLP	FLP recombinase
GFP	green fluorescent protein
GST	glutathione *S*-transferase
hGh	human growth hormone
kDa	kilodaltons
LB	Luria-Bertani broth
M9	minimal media with supplemented with M9 salts
MALDI-TOF	matrix assisted laser desorption ionization-time of flight
mphR	macrolide repressor protein
OD	optical density
P450	cytochrome P450
*p*Bpa	*p*-benzoylphenylalanine

PCR	polymerase chain reaction
PRTase	phosphoribosyltransferases
SDS-PAGE	sodium dodecyl sufate-polyacrylamind gel electrophoresis
SOC	super optimal catobilite-repression broth
VMA	VMA intein from *Saccharomyces cerevisiae*
UPRT	uracil phosphoribosyl transferase
UV	ultraviolet

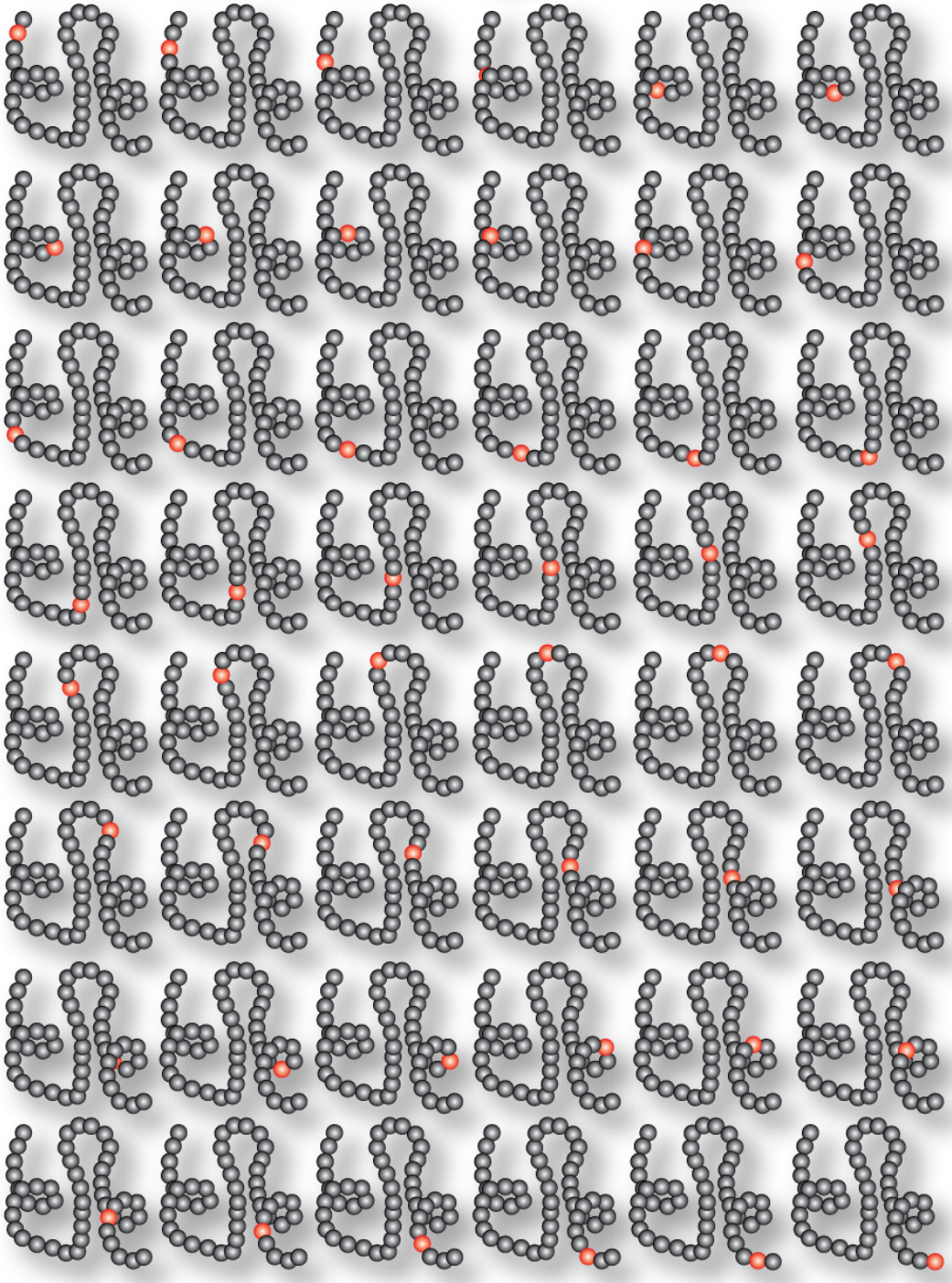
General Abstract

Proteins can be compared to a string of beads, where each bead is called an amino acid. In nature there are 20 different amino acids that can carry a charge (positive, negative or neutral) or contribute to structural aspects of the protein. These characteristics that distinguish one from the other are determined by what is called a functional group or side chain. The order of the amino acids dictates the function of the protein and is determined by a three letter code in the deoxyribonucleic acid (DNA). DNA is made up of four units, adenine (A), cytosine (C), guanine (G), and thymine (T). These four units are assembled into 64 three letter words called a codon. Proteins can be studied by altering a codon, resulting in the coding of a new amino acid.

Described in this thesis is a method that has been termed Codon Scanning Mutagenesis (CSM). This method allows for each three letter code to be randomly changed, resulting in the replacement of an amino acid. The replacement is done in three simple steps, first a circular piece of DNA is cut in a random place, second a codon is removed, third a different codon is added and the circle is closed. The recoding of the sequence will result in a new amino acid in place of the old. The result of subjecting a protein to CSM is depicted on the following page, where the red bead is scanned throughout the protein sequence, resulting in every variation.

Prior to the development of CSM, specific codons could be introduce at a specified location in the DNA sequence or mutations are incorporated a random location without the ability to specify which codon. CSM is a novel method in that a codon for a specific amino acid can be introduced at a random location.

Codon Scanning Mutagenesis



Chapter 1: Background and Significance

1.1 Information that can be gained from protein mutagenesis

The ability to change amino acid residues within a protein has had a profound impact in the biochemistry field. Mutating protein sequence allows for both mechanistic and structural information to be obtained, as well as the creation of proteins with improved or unique function. Protein mutagenesis allows for the identification of specific residues that are responsible for protein-protein interactions by mutating those residues that are believed to be responsible for binding. A classical example of such an experiment was performed by Wells and co-workers where the binding of the human growth hormone (hGH) receptor was investigated by introducing alanine mutations.[1] Alanine is chosen because all side chain interactions are eliminated and typically this amino acid does not alter the overall conformation of the protein. The dissociation constant of each mutant protein was then compared with the wild-type and the residues that are crucial for binding were revealed.

Differences in highly conserved sequences of protein homologs are often target sites to introduce mutations. For example, when the sequence of the highly conserved binding site of uracil phosphoribosyltransferase (UPRT) was compared with other phosphoribosyltransferases (PRTase) a proline residue was observed in the active site of UPRT where as other PRTases have an aspartic acid at the corresponding site.[2] To determine if this proline was responsible for activity, an aspartic acid mutation was introduced in place of the proline. Various kinetic and binding assays were performed on both the wild-type and P131D variant. A 54-59 fold in reduction of enzymatic activity

was observed for the P131D variant, indicating that the proline is essential for substrate binding.

Introducing amino acid mutations into a protein can also result in the creation of a protein with an improved function. Raines and co-workers demonstrated this by increasing the affinity of a protein to the cell wall by the introduction of a cationic patch.[3, 4] A cationic patch results from site-specifically replacing native residues located on the surface of the protein with arginine residues. The creation of this cationic patch increases the affinity of the protein to the cell wall and promotes internalization of the protein. It has been reported that the guanidinium groups of arginine interact with the lipids or heparin sulfate proteoglycans (HSPG)[5] and results in internalization.[6, 7] The addition of guanidinium groups increases the number of hydrogen bonds promoting sufficient cellular uptake.[8, 9]

As a proof of concept, Raines and coworkers proved that a protein can be internalized by the addition of a cationic patch. Protein internalization was demonstrated using green fluorescent protein (GFP) which has a net charge of -9 at neutral pH and has both acidic and basic residues on one face of the protein. Mutating five of the acidic residues on this face to arginine created a cationic patch on the surface of the protein. Incubation of the mutant protein with HeLa cells revealed that the cationic GFP was internalized.[3] Using the idea of altering the surface of a protein to have a higher affinity for the cell wall, arginine mutations were introduced in to ribonuclease (RNase A). It was found that the variants of RNase A that have the cationic surface had an increase of cytotoxicity and inhibited cell proliferation by 3-fold[4] as measured by the incorporation of [methyl-^{3}H]thymidine into cellular DNA.[10]

Structural information can also be determined by altering protein sequence. This is particularly useful for proteins in which it is difficult or impossible to obtain an X-ray crystal structure, such as transmembrane or multi-domain proteins. Topological information can be obtained by scanning cysteine accessibility mutagenesis (SCAM)[11] (Figure 1.1). Using sulfhydryl chemistry it is possible to locate amino acid positions that are both intracellular and extracellular. The positions of amino acids can be determined by labeling the whole protein with a membrane permeable sulfhydryl reagent that carries a biotin tag. Any positions that are not in the plane of the bilayer will be labeled. Only positions that are intracellular can be identified by incubating with a non-permeable reagent capable of blocking free thiols followed by a membrane permeable reagent that carries a biotin label. The only requirement is that the mutant contains one cysteine residue.[11] Additionally solvent exposed residues can be mapped by introducing cysteines and reacting with cyanylation reagent, such as 2-nitro-5-thiocyanobenzoic acid (NTCB) or 1-cyano-4-dimethylaminopyridinium tetrafluoroborate (CDAP). The reagent will only react with thiol groups which are exposed to solvent which then allows for peptide cleavage catalyzed by ammonia. Any mutant that has an extracellular cysteine will be cleaved at the site of the cysteine mutation.[12]

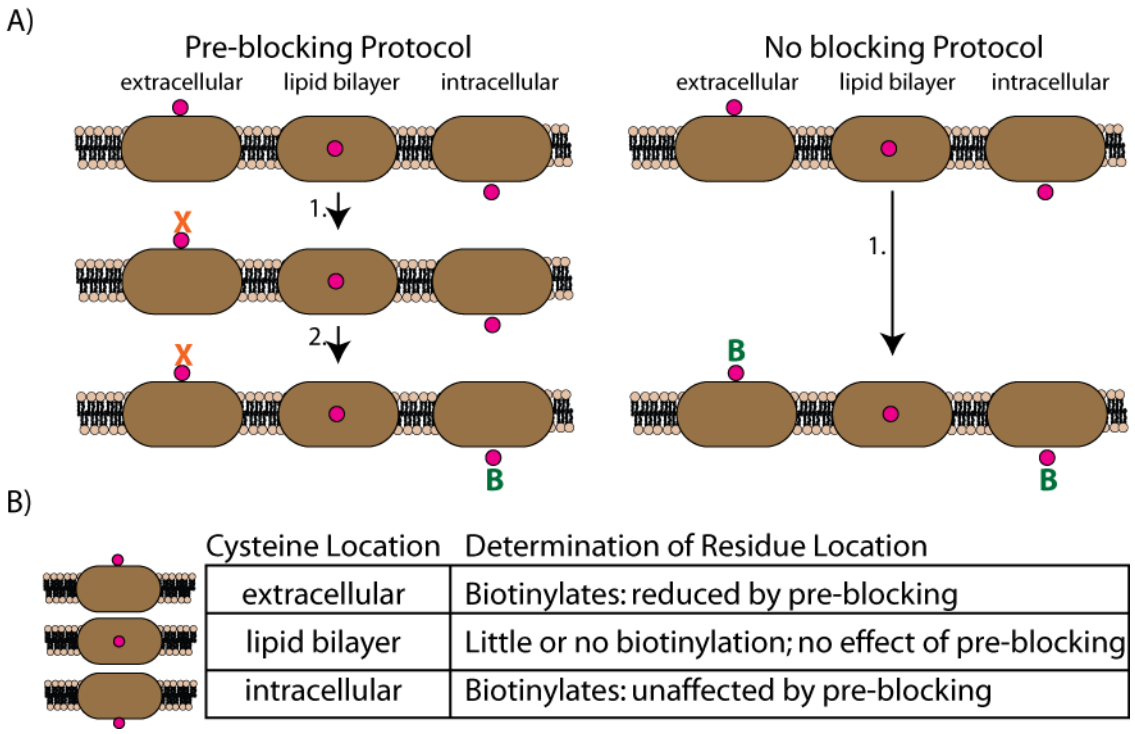


Figure 1.1: Determination of topology using SCAM.
SCAM allows for the location of cysteine residues in a transmembrane protein to be determined. (A) Cysteines (pink) within membrane proteins (brown) can occupy three different environments. On the left outlines the pre-blocking protocol, where a membrane-impermeable sulfhydryl reagent (X) reacts with cysteine residues which are extracellular. Both the blocked and unblocked pool is treated with a membrane-permeable regent (B) and introduces a biotin tag intracellular as well as to those which are extracellular in the unblocked pool. (B) Location of the residues can be determined by comparing the amount of biotin incorporated in both the blocked and non-blocked protocols.

Introducing protein mutations can also be valuable in understanding protein folding. By changing a side chain that may be involved in the formation of the tertiary structure to one of the other naturally occurring amino acids would alter the overall stability of that secondary structural element. The mutant can then be analyzed by absorbance, fluorescence or circular dichroism spectroscopy.[13] Altering the primary structure of a protein does not necessarily have to be detrimental to overall protein stability but rather can be beneficial by randomly introducing mutations to increase the rate at which the mature protein is formed. An example of a protein that has been evolved to fold at an increased rate as well as lower the ΔG of the folded protein, is super-folder

green fluorescent protein (sfGFP).[14] Waldo and co-worker randomly introduced mutations into GFP by fragmenting the gene and reassembling the random fragment multiple times. After each round of randomization the brightest clones were chosen. The kinetics and stability of sfGFP was then determined, it was found that the variant folds at a rate of 5.0×10^{-1} s^{-1} making it the fastest folding GFP variant. It was also found that sfGFP is folded and forms a mature chromophore even when fused to non-soluble protein domains. The mutations that were introduced into sfGFP are all solvent exposed except for one. The stability of the sfGFP can be attributed to increasing electrostatics, increasing H-bonding, and increasing non-polar interactions.[14]

Not only is protein mutagenesis a valuable tool to aid in the understanding of protein structure or function, but directed evolution has been shown to enhance protein function. Directed evolution can be used to create catalysts for organic reactions. This has been demonstrated using cytochrome P450s (P450) as a catalyst for difficult organic reactions, such as hydroxylation of linear alkanes[15] or enantioselective epoxidation.[16] The substrate specificity of enzymes can also be altered to create new function. It has been shown that non-native amino acids can be genetically encoded by introducing mutations into both the enzyme and substrate, in this case, an aminoacyl transferase and tRNA.[17] The introduction of non-native functional groups is further described in section 1.3. While several examples of the usefulness of protein mutagenesis have been described here, nearly all reports of studying proteins involve introducing mutations.

1.2 Traditional approaches to protein mutagenesis

A general method that allows for a desired change in a gene sequence to be made was first described by Smith in 1978.[18] The ability to both obtain the exact sequence of a

gene[19] and synthesize short segments of DNA (oligonucleotides) made it possible to introduce site specific mutations in the DNA sequence. Smith's invention of site-directed mutagenesis was further enhanced by the development of polymerase chain reaction (PCR) by Mullis.[20] Both site-directed mutagenesis and PCR have had a profound impact on the field of biochemistry and has facilitated numerous studies to be conducted on proteins.

The sequence of a protein is dictated by the nucleotides that comprise the codons of the gene. To create mutations in a protein sequence involves changing the DNA sequence. There are many ways in which a change in the DNA sequence can be performed. Site-directed approaches, section 1.2.1, allow specific changes in the DNA sequence to be made. Not only are site-directed approaches used in creating amino acid changes, but it can be a useful method in synthetic biology to generate silent mutations allowing for the removal of unwanted restriction endonuclease sites. Apart from site-directed approaches there are several random approaches that incorporate mutations at a non-specified site. Random mutagenesis approaches, discussed in 1.2.2, allow for gene libraries to be generated and leads to evolution of new protein function. Nearly all of the current methods that are used to introduce mutations, either site-specifically or random, are *in vitro* and involve the use of PCR and a unique set of oligonucleotides. Additionally, either chemical mutagens[21] or UV light[22] can promote DNA mutations which lead to protein variants.

1.2.1 Site-directed approaches to protein mutagenesis

Introducing a site-specific codon mutation is fairly simple with the use of oligonucleotides. While several different approaches have been developed, all utilize an

oligonucleotide that has been designed to contain a mutation. The methods described are based on Smith's early development of site-directed mutagenesis, which involved synthesizing a mutagenic oligonucleotide, annealing of the oligonucleotide to single stranded plasmid DNA, extending the new mutant strand with DNA polymerase I followed by ligation to yield a heteroduplex molecule (Figure 1.2 A).[18]

Kunkle and co-workers made great improvements on Smith's early site-directed approach. The method termed Kunkle mutagensis,[23] is in many ways similar to the original site-directed mutagenesis method. The Kunkle method involves first isolating plasmid DNA from an *E. coli* strain that lacks both dUTPase and uracil deglycosidase. Plasmid DNA that is isolated from this strain of E. coli will be single stranded and have uracil in place of thymine. The mutation is then inserted by a single mutagenic oligo, polymerase to synthesize the second strand, followed by ligation. The duplex DNA, where one strand lacks thymine and the other contains the wanted mutation is then transformed into *E. coli*, where through replication would remove the uracil containing strand, leaving only duplex DNA that has the specified mutation, as depicted in Figure 1.2 B.

While, Smith's and Kunkle's site-directed mutagenesis methods are useful, a disadvantage is the requirement that the plasmid containing the gene of interest must be single stranded. This requires using phage plasmid or isolation of the DNA from a particular strain and therefore can be a time consuming process. With PCR, generating mutations is faster and more reliable, than either of the two previously described methods, since the DNA strand containing the mutation is amplified exponentially. One method that utilizes PCR was developed by Stemmer in 1992, and is termed enzymatic

inverse PCR (EIPCR).[24] In this method, the entire plasmid containing the gene of interest is amplified. To incorporate the mutation a set of oligonucleotides is used, one of which has the mutation. In addition to the mutation in the oligonucleotide, unique restriction sites, typically type IIS endonuclease which cut outside of the recognition sequence, are designed such that after a successful amplification of the plasmid, the product can be digested. Removal of the restriction site, followed by successful ligation leaves the desired mutation, as shown in Figure 1.2 C.

One of the most widely used methods was developed by Stratagene, termed the "Quikchange" method. The one requirement for this method is that the plasmid DNA is isolated from an organism that methylates DNA. While this requirement seems similar to that of the Kunkle method, it differs in that nearly all routinely used *E. coli* strains carry the genes for DNA methylases. A set of unique oligonucleotides are required to change a specific codon in the protein coding sequence. These mutagenic oligonucleotides are designed so that either one or both contain the mutation of interest; additionally the oligonucleotides used are either partially or exact complements of each other. After successful PCR amplification using a high fidelity polymerase the methylated parent DNA can be removed by digesting with *Dpn*I, an endonuclease that recognizes only methylated DNA (DNA from a PCR reaction is not methylated), as shown in Figure 1.2 D. Not only can the described site-directed methods be utilized to change a codon, oligonucleotides can be designed such that either a deletion or insertion of nucleotides is achieved.

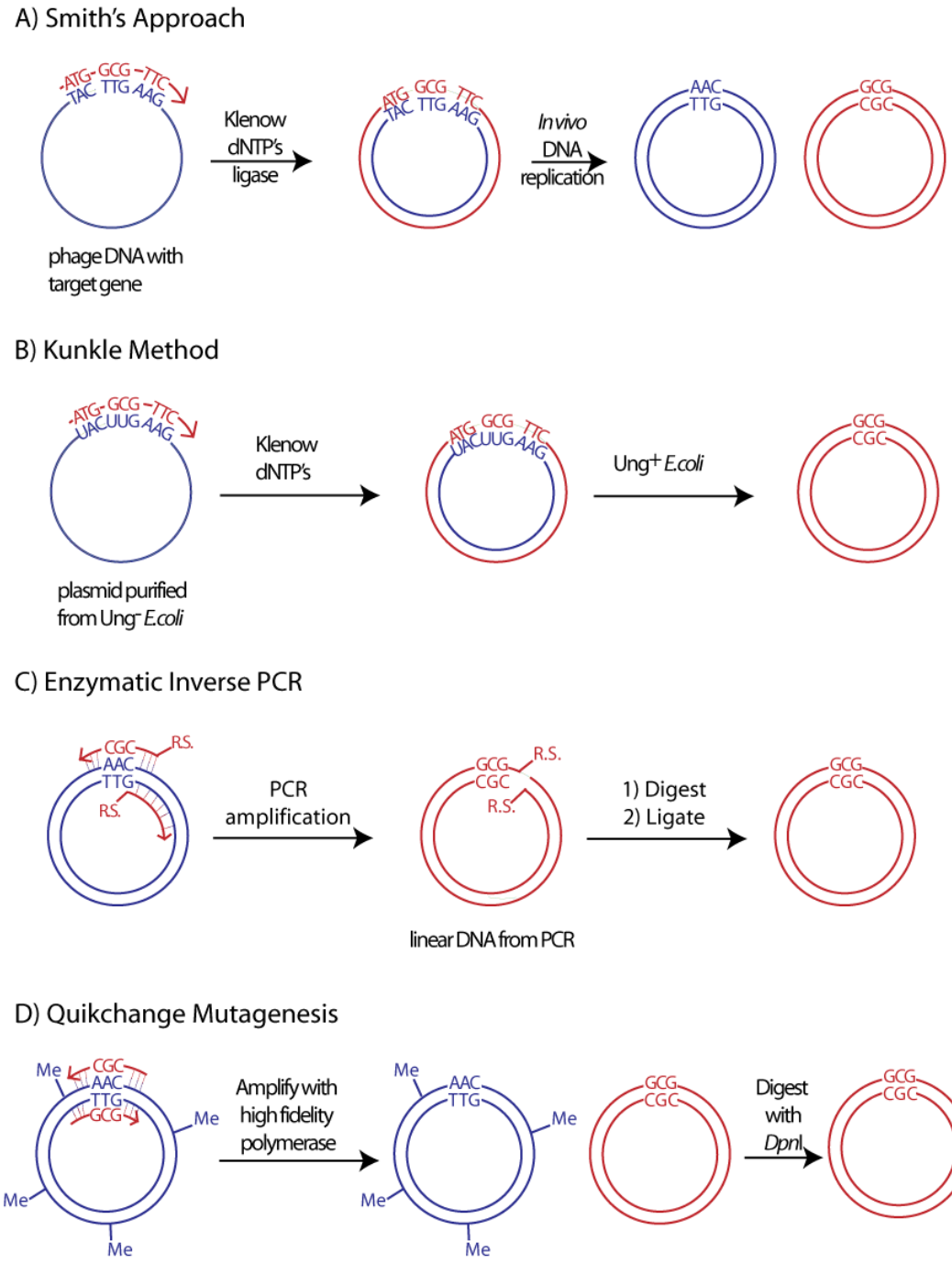


Figure 1.2: Comparison of site-directed mutagenesis methods.
All 4 methods depict creating an Asn to Ala mutation by mutating AAC to GCG. (A) In Smith's approach of site-directed mutagenesis, the parental DNA must be isolated form phage so that it is single stranded. This method resulted in obtaining both the parental DNA as well as the mutant. (B) In the Kunkle approach plasmid DNA is isolated from Ung- *E. coli*, the mutation of interest is incorporated site specifically and the parental DNA is destroyed when transformed into an Ung+ strain. (C) EIPCR, does not involve isolating the parental DNA from a specific strain. Both strands are amplified in a traditional PCR reaction, to contain the mutation of choice and a unique restriction site (R.S.) The mutation is then introduced by digestionand re-ligation. In the last example (D) Quikchange mutagenesis, the additional step processing step in EIPCR is eliminated by using overlapping oligonucleotides and removing parental DNA by *Dpn*I digestion.

1.2.2 Random approaches to protein diversification

Several approaches to randomly diversify proteins have been developed. The directed evolution of a protein requires that several non-specific mutations are incorporated into the DNA sequence followed by a high throughput screen or genetic selection that facilitates the isolation of a desired phenotype. One widely used approach is error-prone PCR. This method is based on traditional PCR that amplifies DNA, however by changing the concentration of either Mn^{2+} or Mg^{2+} in the reaction buffer, the fidelity of *Taq* polymerase can be altered. This results in mismatched bases during the amplification process, as depicted in Figure 1.3 A.[25]

A method that does not change the DNA sequence, but instead recombines the mutations from different strands of sequence in segments, is DNA shuffling.[26] This method is most commonly combined with error prone PCR and is comprised of fragmentation with DNase I and the desired length of the randomized fragments is isolated, smaller fragments favor more recombination. The isolated fragments are then reassembled by PCR without terminal primers. Since fragments will overlap, the fragments themselves act as primers, annealing to other fragments in the mixture. The recombined gene is then amplified with terminal primers and cloned into an expression vector and screened for desired phenotype, as shown in Figure 1.3 B. After those mutants which have the desired function are combined, multiple rounds of shuffling can be preformed. This method also allows for similar genes from different organisms to be recombined, however a constraint is that there is a limitation on the fragmented genes that can be recombined due to sequence homology.

In an effort to create a method that is not dependent on sequence homology, non-homologus random recombination has been proposed.[27] Similar to DNA shuffling, the starting DNA pieces are digested with DNAse to create random fragments of a desired length. Unlike DNA shuffling, the fragments are then joined with ligase and capped on the ends to control the gene length with hairpins. DNA hairpins are single stranded DNA that has a complementary sequence on either end. Within the hairpin, there is an internal restriction site and oligonucleotides can anneal to the hairpins to amplify the recombined gene, see Figure 1.3 B.

There are many other ways to randomly create protein variants. One method is circular permutation. This technique utilizes the randomness of DNAse to introduce a single double strand break within a circular gene resulting in fusion of the N- and C-termini. The gene encoding the protein of interest is amplified with oligonucleotides that contain DNA coding for a protein region of flexibility, typically this linker segment would have a high percentage of glycine residues. Ligation of the PCR product to itself results in a double stranded circular gene. The circular gene is then treated with DNAse in the presence of Mn^{2+}. The randomly linear DNA segment is then inserted into an expression plasmid and assayed for desired function.[28] Specific insertions can also be incorporated into a protein, using commercially available kits. This method is controlled by using mobile genetic elements called transposons, described in 3.2.1, and have been optimized to target DNA once. The random insertion of a linker sequence is useful in determining gene function or identifying regions of flexibility in a protein. One kit is the Mutation Generation System (MGS) kit (Finnzymes). This kit utilizes a transposon to randomly insert a selectable marker into DNA. This mobile gene is removed by digestion

with an endonuclease, followed by ligation leaving a "scar" in the protein coding sequence. This scar is a 15 bp insertion or a five amino acid linker that has been randomly inserted, as shown in Figure 1.3 C. By scanning a five amino acid linker throughout the gene encoding β-galactosidase, Savilahti found that every mutant contained a 15 bp insertion and the site of insertion was randomly disturbed. Using the MGS kit allows for regions within the protein that are tolerant of insertions to be indentified.[29]

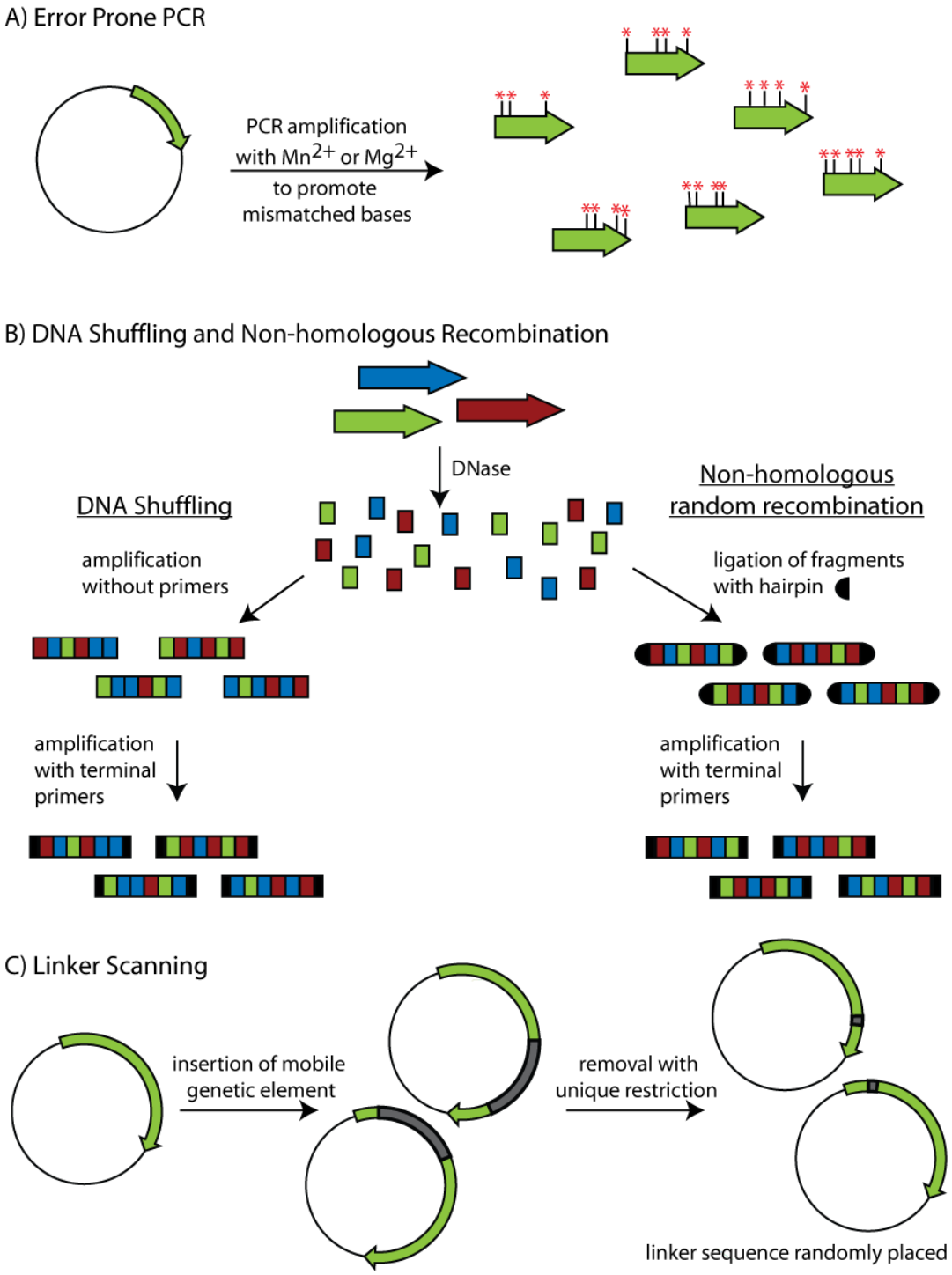


Figure 1.3: Comparison of irrational approaches to protein diversification.
Depiction of irrational protein diversification approaches. (A) Error prone PCR, lowering the fidelity of *Taq* polymerase results in random nucleotide mis-matches, represented by *. (B) In recombination genes of interest; red, blue and green, are randomly fragmented. In DNA shuffling the fragments are combined by PCR without the use of external oligonucleotides. The desired length is then amplified with external oligonucleotides. This method requires that there is homology between different fragments. On the right shows non-homologus recombination, where T4 ligase is used to form the desired length gene. (C) Insertion of a linker sequence (grey) by insertion of a mobile genetic element followed by removal.

An all *in vivo* method that is used to diversify a gene sequence is to use a strain of *E. coli* that has the DNA repair pathways deleted. The three DNA repair pathways that have been deleted are mutS, which repairs mismatches; mutD, responsible for 3'- to 5' exonuclease of DNA polymerase III and mutT, making the strain unable to hydrolyze 8-oxo-dGTP. Transforming plasmid DNA into a commercially available strain from Stratagene, XL1-red, which lacks these DNA repair pathways, will promote mis-matches in DNA during the replication process. Additionally, a temperature sensitive plasmid carrying the mutD5 gene[30], can be used in place of using the mutator strain XL1-red. Random mutations are introduced into the target DNA by co-transformation of the mutD5 plasmid and target plasmid. By adjusting the growing temperature conditions the *E. coli* can be cured of the mutD5 carrying plasmid. This method creates a temporary mutator strain, resolving one of the disadvantages to using the XL1-red strain. While the plasmid carrying the target gene is replicated mutations are also incorporated into the chromosome making the strain non-viable after long growth times.

1.3 *In vivo* incorporation of unnatural amino acids

With recent developments introducing a mutation into a protein is not limited to only the 20 natural amino acids. Schultz and coworkers have been able to expand the genetic code to incorporate non-native amino acids *in vivo*. The introduction of these unique amino acids into a protein sequence allows for new chemical experiments to be performed that would otherwise be difficult or impossible. For instance none of the 20 natural amino acids contain a keto group. With the expanded genetic code technology, it is possible to incorporate *p*-acetylphenylalanine, whose carbonyl group under physiological conditions functions as a chemical handle for labeling proteins with

fluorescent dyes or biotin.[31] Two other chemical handles that can be incorporated include azides[32] or alkynes.[33] These functional groups can be further modified by copper catalyzed cycloaddition or Staudinger ligation. This methodology also enables the control of protein function with light. The binding affinity of a protein can be controlled by the addition of *p*-phenylalanine-4'-azobenzene,[34] which has a photo-isomerizable double bond changing both structure and dipole. Protein activity can also be regulated by incorporating photo-removable protecting groups, which can block a biological reaction from occurring until the group has been removed.[35] Structural information can also be probed by incorporating a photo-cross-linking amino acid, *p*-benzoylphenylalanine,[36] which allows the isolation and characterization of protein-protein, protein-ligand[37] or protein-DNA interactions.[38]

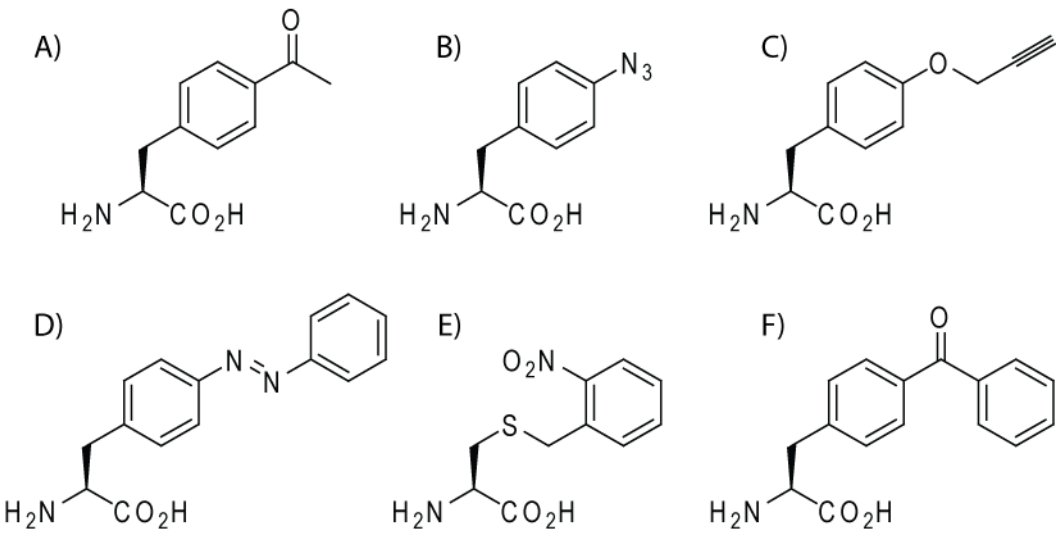


Figure 1.4: Examples of unnatural amino acids used to incorporate unique function *in vivo*.
A) *p*-acetylphenylalanine, B) *p*-azidophenylalanine, C) *p*-propargyloxyphenylalanine, D) *p*-azophenylphenylalanine, E) *o*-nitrobenzylcysteine, F) *p*-benzoylphenylalanine.

These unique functional groups are inserted by using the traditional approaches to site-directed mutagenesis. A specific codon is mutated to the amber codon, TAG, using

mutagenic oligonucleotides. The amber codon, TAG is a unique codon and therefore can code for the incorporation of a unique functional group. The introduction of unnatural amino acids into a protein is by suppression of the nonsense amber codon, TAG.[39] Additionally single or multiple unnatural amino acids can also be incorporated using an *in vitro* approach in response to a 4 base code.[40-42] The *in vivo* approach is what is used in the Cropp lab and the technique used in this thesis, so the details of the *in vitro* approach will not be discussed.

In the genetic code there are 3 stop codons, the amber codon, TAG, is the least abundant in all kingdoms of life. To synthesize protein *in vivo* that contains an unnatural amino acid, a unique codon (TAG), an orthogonal tRNA and a cognate aminoacyl-tRNA synthetase are required. To incorporate unnatural amino acids into *E. coli* the tRNA/synthetase pair from the archaeal bacteria *Methanococcus jannaschii*[43] is used and for eukaryotic organisms the tRNA/synthetase pair is derived from *E. coli*.[44] Structural constraints of the tRNA/synthetase pair from an orthogonal organism does not allow for any cross-talk with the endogenous system to occur.[45] Since in all kingdoms of life the amber stop codon, TAG is blank and does not have a tRNA to recognize it, the anticodon in the orthogonal tRNA is mutated to CUA to base-pair to UAG, followed by randomization of nucleotides to allow for specific incorporation of the unnatural amino acid.[46] The residues in the active site of the orthogonal synthetase were then mutated by directed evolution to accept the unnatural amino acid and charge the tRNA.[47] A series of *in vivo* positive and negative selections are then performed on the protein variants to remove those that are not orthogonal to the system (negative selection) and enhance those that are orthogonal (positive selection).

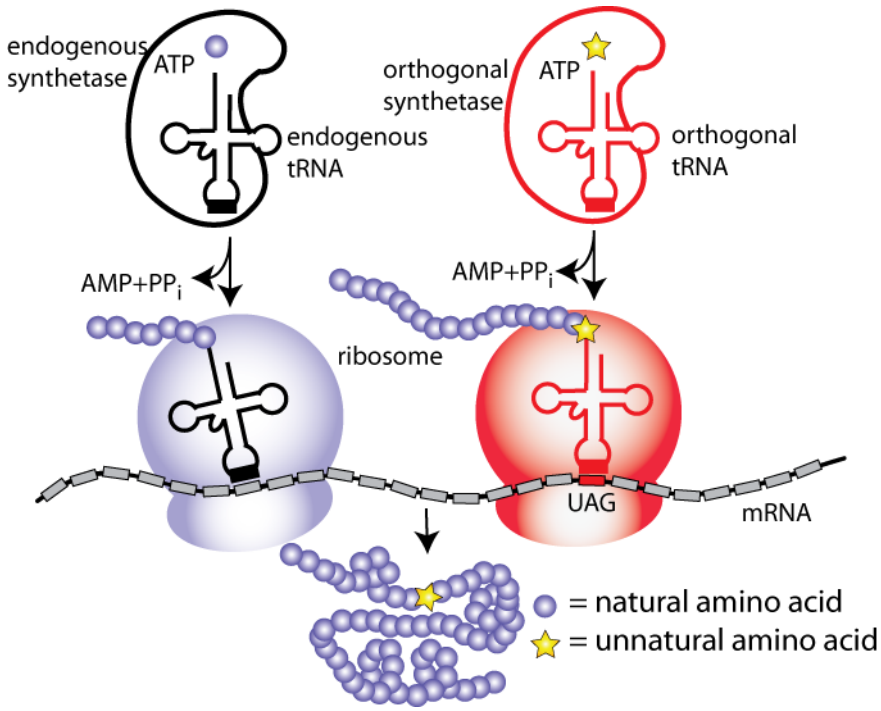


Figure 1.5: General method of *in vivo* incorporation of an unnatural amino acid.
An unnatural amino acid is inserted as a response to the amber stop codon UAG. An orthogonal synthetase charges the mutant orthogonal tRNA with the unnatural amino acid.

1.4 Disadvantages of traditional protein mutagenesis methods

The ability to mutate proteins in order to study them has shown to be extremely valuable, whether it is to improve protein function in the case of directed evolution, or site specifically change codons to alter the chemical reactivity of residues resulting in enhanced or deleterious protein function, as well as introducing unique function with unnatural amino acids. All the traditional methods to achieve these mutations either incorporate multiple random mutations, which might only occur at the DNA level and not the protein level, or the ability to site-specifically mutate the coding sequence. There are a few disadvantages to these methods, first is that site-directed approaches allow for only one mutation to be created at a time and requires a pair of oligonucleotides per desired mutation. It would therefore, not be feasible to use this approach in mutating every amino acid position throughout the entire protein codon sequence. For example, if one wanted to

identify solvent exposed residues on a protein using thiol protection assays and there was little or no information about the tertiary structure of the protein of interest, then the codon for cysteine, TGC, would be introduced at every codon position. If the protein of interest is 200 amino acids in length, then 400 unique oligonucleotides, 200 PCR reactions followed by screening and sequencing multiple clones of the desired mutation would be required, as shown in Figure 1.6. So the traditional site-directed approaches are only useful if there is previous knowledge of protein structure.

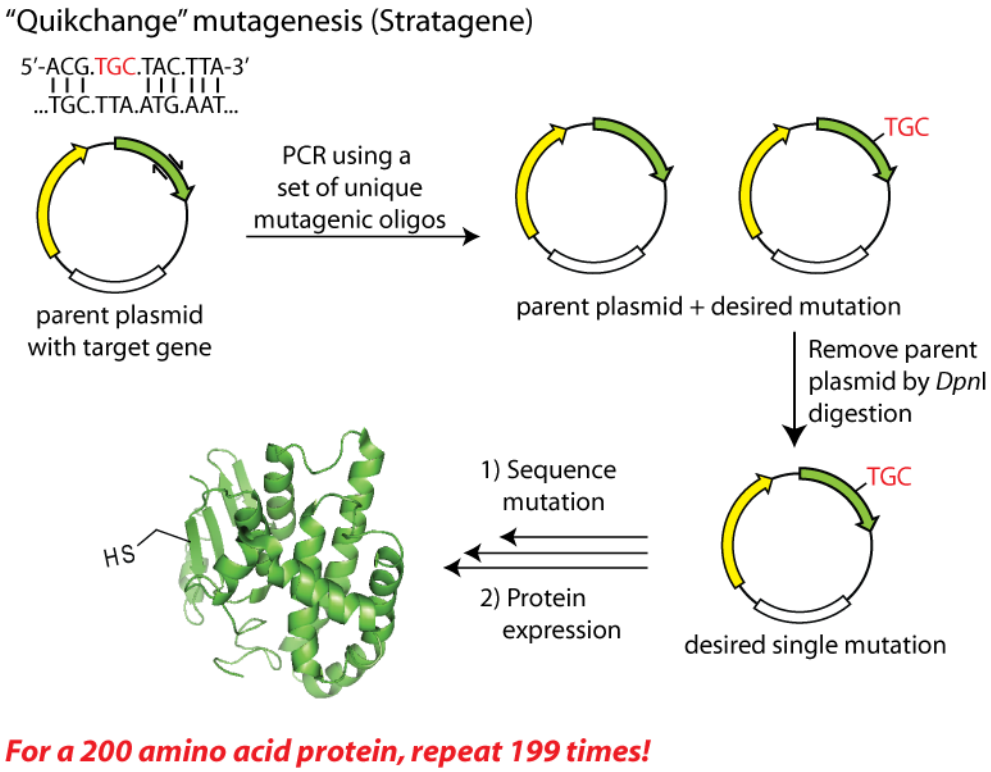


Figure 1.6: Traditional Quikchange mutagenesis depicting incorporating a cysteine. If one would want to perform a cysteine scan on the entire protein, 200 individual mutagenesis reactions would need to be performed followed by sequencing to verify the mutation.

There are many disadvantages associated with the random approaches, one is there is no control over the mutation that is achieved, and since there is redundancy in the genetic code most mutations will be silent (a wobble base mutation) or limited to amino

acids that are in the same row or column of the standard genetic code, Table 1.1. Second, several of the random approaches described doe not have control over reading frame, which could cause early termination or frame shifts. This dissertation will focus on the development of a novel method that incorporates the best aspects of each, randomness and the ability for the researcher to choose the mutant codon as well as have a random method that allows for a defined number of mutations.

Table 1.1: Standard genetic code

1st nucleotide	2nd nucleotide: T	2nd nucleotide: C	2nd nucleotide: A	2nd nucleotide: G
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys
	TTA Leu	TCA Ser	TAA STOP	TGA STOP
	TTG Leu	TCG Ser	TAG STOP	TGG Trp
C	CTT Leu	CCT Pro	CAT His	CGT Arg
	CTC Leu	CCC Pro	CAC His	CGC Arg
	CTA Leu	CCA Pro	CAA Gln	CGA Arg
	CTG Leu	CCG Pro	CAG Gln	CGG Arg
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser
	ATC Ile	ACC Thr	AAC Asn	AGC Ser
	ATA Ile	ACA Thr	AAA Lys	AGA Arg
	ATG Met	ACG Thr	AAG Lys	AGG Arg
G	GTT Val	GCT Ala	GAT Asp	GGT Gly
	GTC Val	GCC Ala	GAC Asp	GGC Gly
	GTA Val	GCA Ala	GAA Glu	GGA Gly
	GTG Val	GCG Ala	GAG Glu	GGG Gly

1.5 Specific Aims

The aim of this dissertation was to develop a universal random mutagenesis method that incorporates a user specified mutation, termed Codon Scanning Mutagenesis (CSM). The rational for developing CSM arose while working on incorporating an isotopic photoaffinity label, [D_{11}]-*p*-benzoylphenylalanine (*p*Bpa). Incorporating [D_{11}]-*p*Bpa allows for *in vivo* cross-linking between two monomers and the protein-protein

interaction to be identified by a unique fingerprint in the mass spectra. The isotopic label would be useful in the identification of unknown interactions; however this would require generating many TAG mutations throughout the gene sequence. It became clear that a method was needed which incorporates a specific mutation randomly. An ideal CSM, would allow for any mutation to be created, regardless of the nucleotide sequence and specifically introduce the mutation in the correct reading frame. Additionally, the use of the developed method would allow for the introduction of multiple defined mutations as well as, create mutant libraries with a specified percentage of various codons.

This dissertation is assembled in the order that experiments were conducted and the influence that the results provided on future experiments. Described first is the incorporation of [D_{11}]-*p*Bpa the conclusions of which inspired the development of a novel mutagenesis method that facilitates in-frame codon mutations. The details of the components needed to accomplish such a task are then described followed by two proof of principal experiments using well characterized targets, 1) glutathione *S*-transferase (GST) for the random incorporation of a photoaffinity label and 2) uracil phosphoribosyl transferase (UPRT) in an alanine scan. The specific details of CSM were also optimized in the second proof of principle experiment.

Chapter 2: *In vivo* incorporation an isotopic label and the limitations of traditional mutagenesis methods

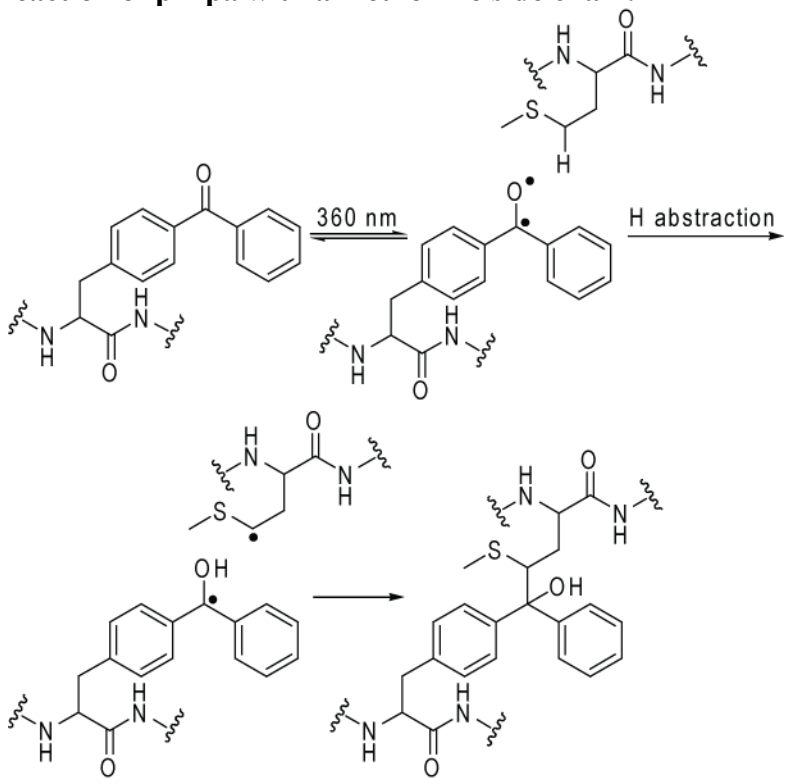
2.1 Introduction

As stated in Chapter 1, one of the most useful functions of performing unnatural amino acid mutagenesis is the ability to capture two proteins at the interface or trap a ligand to the active site. The use of photo-cross-linking amino acids coupled with mass spectrometry allows for identification of such interactions to be determined.[48-50] However, this process is limited by the ability to assign an unknown covalent interaction in the mass spectra. It has been shown that by incorporating 1:1 mixtures of isotopically labeled cross-linking reagents that the indicative isotopic pattern results in the mass spectra.[48] In an attempt to eliminate this delay in data analysis, it was thought that a genetically encoded isotopic label could be introduced into a protein, cross-linking initiated with light followed by treating the cross-linked peptides with trypsin and analyzing the digested peptides by mass spectrometry for a unique fingerprint.

The unique fingerprint could arise from the incorporation of an isotopically labeled amino acid. The unnatural amino acid *p*-benzoylphenylalanine has 11 aromatic hydrogens. Replacement of all the hydrogens with deuterium would result in an amino acid that is 11 atomic mass units larger. The hypothesis being that if incorporated in a 1:1 mixture with unlabeled, there would be a peak for M and M+11. To test this hypothesis the [D_{11}]-*p*Bpa was synthesized and provided by Bryan Wilkins.[51] Glutathione *S*-transferase was used to test both the incorporation of the labeled amino acid by the evolved *p*Bpa synthetase and the ability to identify cross-linked fragments by the unique M and M+11 fingerprint. This protein was chosen due to the precedence of cross-linking the dimers both specifically[36] and non-specifically.[52] Additionally it is known that there is

a phenylalanine located at the dimeric interface.[53] The advantage of performing cross-linking experiments by genetically encoding benzophenone is that the side chain of an excited benzophenone reacts with adjacent alkyl groups within 4Å. The excited triplet state is also short lived and therefore will result in specific cross-linking events to occur. Another advantage is that cross-linking is promoted by irradiating the sample with 360 nm light which does not damage proteins.[54] Specific cross-links were also seen in our experiments with a genetically encoded isotopically labeled benzophenone. Analysis of the mass spectra showed that only the predicted peptides were able to cross-link.

Scheme 2.1: Reaction of p-Bpa with a methonine side chain.



2.2 Results and Discussion

In order to incorporate the unnatural amino acid, the gene encoding GST was cloned into the expression vector pBADmycHisA, adding a C-terminal polyhistidine tag.

The F51TAG mutation was introduced site-specifically and co-transformed with pSUP-*p*Bpa,[55] a plasmid carrying the orthogonal tRNA/synthathase pair into chemically competent *E. coli.* Expression of both wild-type and F51TAG GST in the presence and absence of [D_0] and [D_{11}]-*p*Bpa were performed. After a five hour expression, protein samples were isolated by Ni-affinity chromatograph and analyzed by SDS-PAGE. Cross-linking of monomers was initiated by irradiating protein samples in 15 and 30 min intervals, Figure 2.1.

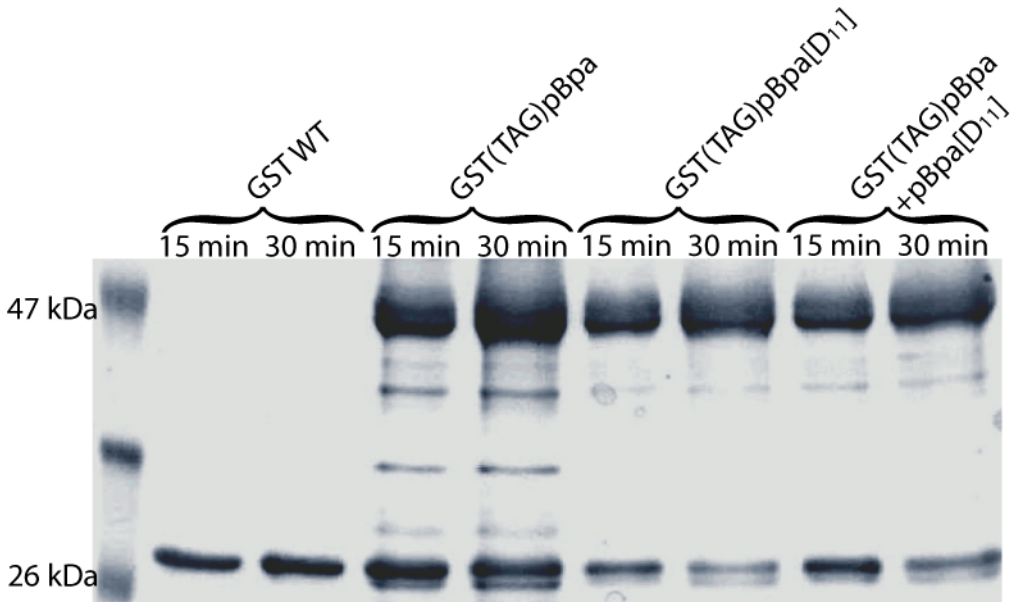


Figure 2.1: SDS-PAGE analysis of cross-linking with [D_0] and [D_{11}]-*p*Bpa
Purified protein samples were irradiated with 360 nm light for both 15 and 30 min. Wild-type GST did not result in any cross-linking, where as those samples which were expressed in the presence of [D_0] and [D_{11}]-*p*Bpa show ~50-60% cross-linking.

Samples of the purified products, both cross-linked and un-cross-linked were separated by SDS-PAGE. As it can be seen in Figure 2.1, there is a band for the cross linked product when either the labeled or unlabeled *p*Bpa added to the media. This indicates that the [D_{11}]-*p*Bpa has the same chemical properties and is a substrate for the previously evolved aminoacyl-tRNA synthetase. Both monomer and dimer bands were isolated and treated with trypsin. The peptides resulting from the digest were then

analyzed by MALDI-TOF mass spectrometry. Indeed when analyzing the data obtained by mass spectrometry the expected, M and M+11 fingerprint could be identified, Figure 2.2. It was predicated that the fragment carrying the *p*Bpa would give rise to a peak at m/z = 2333.2 which corresponds to the fragment 45-FELGLE(*p*Bpa)PNLPYYIDGDVK-63. However, a peak at m/z = 2460.5 was observed corresponding to the mis-cleaved fragment with an additional lysine residue, 44-KFELGLE(*p*Bpa)PNLPYYIDGDVK-63. It was also observed in the expression performed in media supplemented with a 1:1 mixture of [D_0] and [D_{11}]-*p*Bpa, a peak at 2460.5 and 2471.4, confirming the hypothesis that a unique fingerprint would arise from genetically incorporating an isotopic label. Upon, analyzing the mass spectra of the samples which were cross-linked, the unique M and M+11 fingerprint was clear, 3028.8 and 3039.5 was observed. The mass difference of 695.6 (3028.8 – the original peptide, 2333.2) is the exact mass of the peptide of the interacting monomer, 131-MREDR-134, see Figure 2.2.

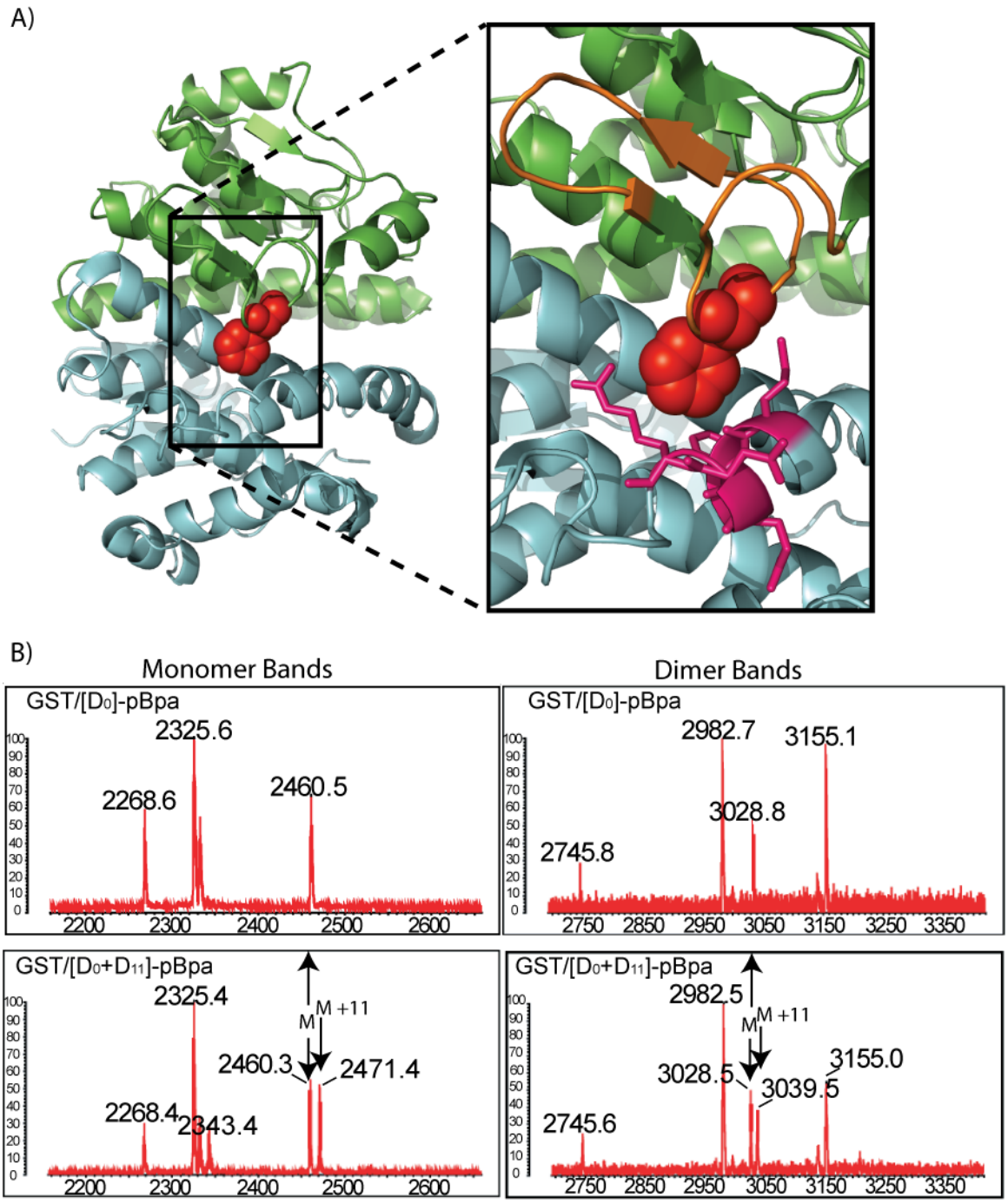


Figure 2.2: MALDI-TOF spectra of GST with [D_0] and [D_{11}]-*p*Bpa
A) Crystal structure of GST (PDB entry 1y6e) with the F51 highlighted in red. This position resides on the fragment shown in orange and interacts with the peptide shown in pink (MREDR). B) Each spectra is a zoom of the area with the assigned peak. For both monomer (left panel) and dimer (right panel) a double appears with two peaks 11 mass units in difference when expressions are done in the presence of a 1:1 mix of [D_0] and [D_{11}]-*p*Bpa. MALDI-TOF experiments were done by Bryan Wilkins.

2.3 Conclusions

Incorporating [D_{11}]-*p*Bpa shows an M and M+11 fingerprint as expected. Further no additional peaks with the signature M and M+11 fingerprint were observed. Incorporating genetically encoded isotopic labels should decrease the time commitment associated with interpreting mass spectral data. One can envision characterizing protein-protein or protein-ligand interactions in an unknown system by incorporating a photoactive isotopically labeled amino acid. While this appears to be a great advance in determining unknown interactions, we are currently limited by methodology used in creating the amber codon mutants. If the site of interaction is unknown then multiple residues would need to be mutated in order to obtain any significant data. Therefore, a general method that incorporates the amber codon, TAG at random codon positions is needed. The ability to incorporation of an isotopic label when added to the codon mutagenesis method described in this dissertation would be an extremely valuable tool in obtaining structural information without the need for NMR or X-ray crystallography.

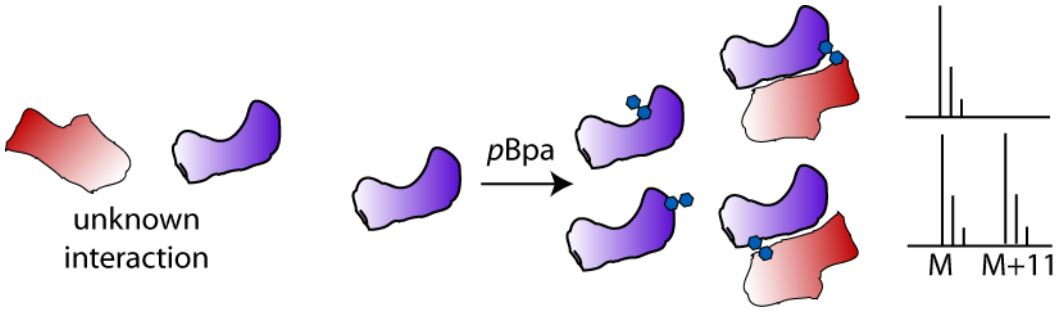


Figure 2.3: Hypothetical representation of indentifying an unknown interaction using [D_{11}]-*p*Bpa
The topology of an unknown interaction can be identified by the unique M and M+11 fingerprint. Incorporation of [D_{11}]-*p*Bpa at multiple sites of one protein (purple) and cross-linking with the red protein would allow for identification of those residues at the interface.

2.4 Materials and Methods

Materials. Plasmids pGEX-T4-1 and pBADmycHisA were purchased from GE HealthCare and Invitrogen. The plasmid encoding the tRNA/synthethase pair was obtained from Peter Schultz (Scripps). The non-labeled amino acid was purchased from Bachem and the labeled amino acid was synthesized by Bryan Wilkins.[51] All enzymes were purchased from New England Biolabs (Ipswich, MA). All reagents used were molecular biology grade.

General Methods. Plasmid DNA was isolated from *E. coli* using Fermentas GeneJet Kit according to the manufactures recommendations. All PCR reactions were carried out using a PTC-200 Peltier Thermo cycler. PCR products were purified using Qiagen PB buffer. Digested DNA was purified by running DNA on a 1% agarose TAE gel, bands excised and DNA purified using Qiagen QG buffer.

Cloning of GST and introducing a TAG mutations.

Table 2.1: Oligonucleotides used to incorporate a photo-affinity label

CL192	CTAGGATCCCCTATACTAGGTTATTGG
CL310	CCAGTCGACGCCTCTAGAAACCAGATCCGATTT
CL206	TTGGGTTTGGAGTAGCCCAATCTTCCT
CL207	AGGAAGATTGGGCTACTCCAAACCCAA

As a preliminary test, the small homodimer, glutathione-S-transferase (GST) was chosen. It is known that when Phe51, located at the dimer interface, is mutated to *p*Bpa the protein can cross link with the other monomer when irritated with 360 nm light. The gene encoding GST was amplified from pGEX-T4-1 (GE Healthcare) using CL192 and CL310. The product was purified and digested with *Bam*HI and *Sal*I and cloned into the *Bgl*II and *Sal*I sites of pBAD-mycHisA (Invitrogen). When expressed, glutathione is fused to an N-terminal MDPSSR leader peptide and C-terminal hexa-histidine tag. To

incorporate the TAG mutation at the Phe51 position standard Quikchange PCR was conducted using oligonucleotides CL206 and CL207. The mutation was verified by both the lack of the presence of a band on an SDS-PAGE gel as well as sequencing.

Expression to introduce [D_0] and [D_{11}]-*p*Bpa. Expressions were performed by inoculating 50 mL of LB medium containing 100 μg mL^{-1} ampicillin and 35 μg mL^{-1} chloramphenicol and grown the presences of 1 mM *p*Bpa (Bachem) or 1 mM *p*Bpa[D_{11}], or 0.5 mM *p*Bpa/0.5 mM *p*Bpa[D_{11}] and grown to an OD_{600} = 0.8. The cultures were induced with 0.2% arabinose and expressed for 5 hours.

Purification and cross-linking. Proteins were then isolated by binding to Probond Purification resin (Invitrogen) according to the manufacturer's protocol for native isolation using binding buffer containing imidazole (50 mM NaH_2PO_4, pH 8.0, 10 mM imidazole, 0.5 M NaCl). Purified protein (100 μL) was then irradiated with 350 nm hand-held 100W Black Ray lamp from a distance of 5 cm at room temperature for both 15 and 30 min intervals. The cross linked samples were then analyzed by SDS-PAGE.

Tryptic digests and MALDI-TOF.[51] Experiments were performed by Bryan Wilkins following an in-gel tryptic digest protocol. Digested samples were then analyzed by MALDI using internal standards insulin and insulin oxB. MALDI spectra were recorded on an Axima-CFR spectrometer (Shimadzu).

Chapter 3: A mutagenesis method that is not dependent on mutagenic oligonucleotides

3.1 Introduction

The previous experiment of incorporating [D_{11}]-*p*Bpa at the dimer interface of GST proved that it is possible to characterize transient protein-protein interactions using mass spectrometry. It is not feasible to identify unknown interactions by identifying the M and M+11 fingerprint. However, if an unknown interaction is too be determined this would require the independent generation of many amber codon mutations throughout the protein coding sequence with traditional mutagenesis methods, and is therefore a major limitation of using the genetically encoded isotopic label. The development of a method that can randomly incorporate a specific codon mutation that is not dependent on oligonucleotides would be an invaluable tool in creating protein variants and facilitating this research. To create random mutations that are codon specific, it was envisioned that a target plasmid with the gene of interest could be randomly digested once to create a double strand break followed by the removal of three nucleotides and subsequent insertion of the specified mutant codon. This approach would lead to a general method that allows for scanning a specific codon mutation, see Figure 3.1. Ideally, the codon mutations would also only occur in the correct reading frame.

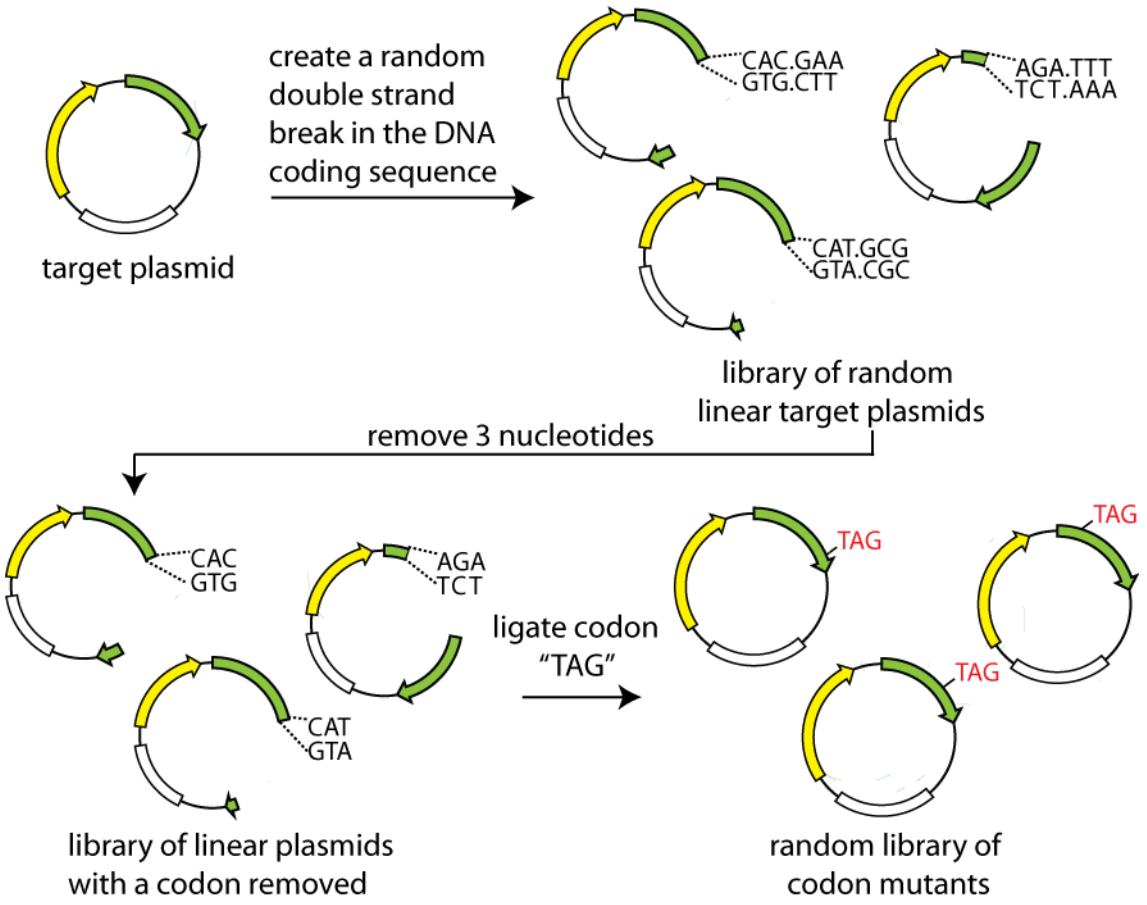


Figure 3.1: Process to randomly mutating a codon position on a library scale.
To introduce random codon mutations throughout a gene of interest the target plasmid would be randomly linearized, followed by the removal of three sequential nucleotides from the coding sequence and ligation of the new mutant codon, in this case TAG. This process would create a random single mutation without any net addition or deletion of nucleotides.

3.2 Creating a double strand break

In order to create a codon mutation in the DNA coding sequence, each strand of the duplex DNA needs to be broken. An endonuclease that is capable of randomly digesting DNA is DNase. As stated earlier DNase has proven to be useful in a variety of protein evolution experiments. DNase is used to create multiple random fragments for recombining genes as well as to create a single random break in a protein coding sequence for circular permutation.

The use of DNase to introduce the random double stand break was considered. Controlling the reaction to only cut once was fairly difficult even when following previous reports of using ~1 unit of DNase per mg of DNA and performing the reaction at low temperatures, 16 ^{o}C for short periods of time. It is also reported that besides for a single random cut there are also truncations of the gene as well as nicks in the DNA strand.[28] In designing a random codon scanning method the use of DNase did not seem to be the most reliable option since in many cases unwanted deletions can occur and controlling the reaction was difficult. Recently a useful method in targeting DNA once is the use of mobile genetic elements, called transposons.

3.2.1 Transposon Mutagenesis

Mobile genes were first discovered in maize by Barbara McClintock and are responsible for phenotypic differences in the kernels.[56] These mobile genetic elements have shown to be useful in many areas of molecular biology and genetics, the most widely used and best understood is bacteriophage Mu.[57] Transposons have been used in mapping techniques, DNA sequencing, and to analyze protein-DNA complexes. While these mobile genes are found in nature, and therefore is an *in vivo* process, a protocol has been developed that allows for the transposition reaction to occur *in vitro* with high efficiency. The *in vivo* Mu transposition reaction is fairly complex and requires cofactors and accessory proteins.[58] The *in vitro* process has been modified to require the minimum components. However, the transposition reaction was not very efficient, since the mobile gene was part of a plasmid and required the transposase to cut and paste this element.[59] The most recent *in vitro* process eliminates the need for the transposase to cut the transposon, thereby enhancing the rate of the reaction. This is achieved by using a linear

double stranded transposon, where on either end there are recognition sites for the transposase including a selectable maker. It is necessary that the ends of the transposon are pre-digested with *Bgl*II.[60] The mechanism of MuA transposon incorporating into target DNA is outlined in Figure 3.2. The use of a transposon to introduce random codon mutations appears to be a more reliable option, when compared with DNase. There is also evidence that this reaction is indeed randomly distributed throughout the target DNA and only incorporates once per target.[29] Further, there are several commercially available kits that use the MuA transposon/transposase or a transposase from the RNase super family Tn5, where the reaction conditions have been optimized and are reproducible.

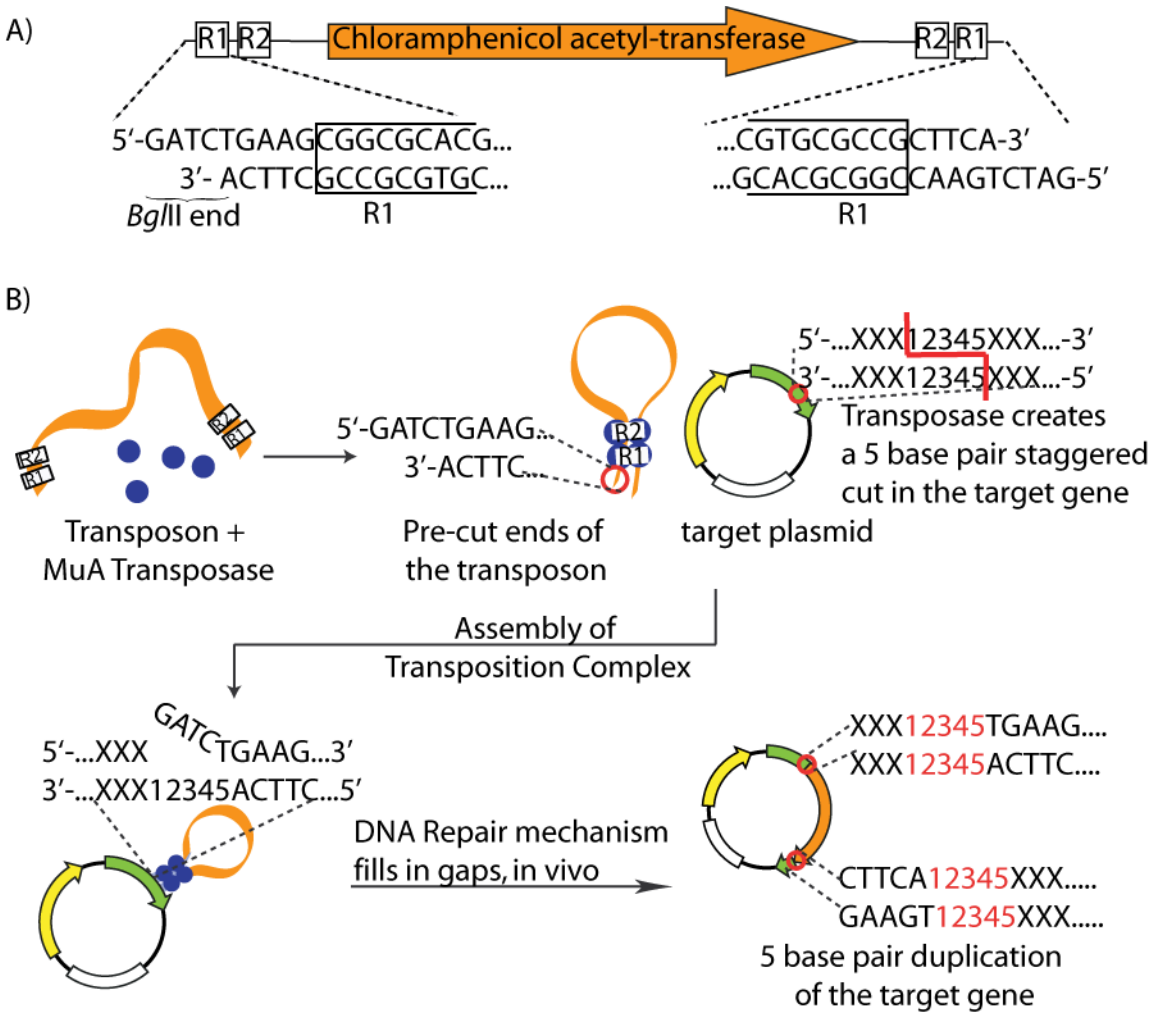


Figure 3.2: MuA transposon and incorporation into target DNA.
A) The ends of the transposon are pre-cut by BglII. On either end the recognition sequences are identical and there are four nucleotides that do not alter the efficiency of the transposition reaction. B) For the transposon to incorporate into target DNA, the transposon is mixed with Hyper MuA transposase. The transposase recognizes the target DNA and creates a five base pair staggered cut, at this time there is assembly of the transposition complex followed incorporation of the transposon. *In vivo* DNA repair mechanisms will fill in the five base pair gap. The overall process results in a five base pair duplication of the target gene at the site of transposon incorporation.

3.2.2 Deletion of a triplet nucleotide from a protein coding sequence

There are four base pairs on either end of the MuA transposon that are not recognized by the MuA transposase but are incorporated into the final DNA target. This sequence can be altered to contain an endonuclease restriction site allowing for the transposon to be removed. The ability to add a restriction site has enabled transposon

mutagenesis to be a useful methodology for pentapeptide scanning. Pentapeptide scanning is accomplished by modifying the transposon to have *Not*I restriction sites which are eight bp in length, GCGGCCGC.[29] As shown in Figure 3.2, during transposon insertion, five base pairs in the target sequence are duplicated,[61, 62] as well as the base on either end of the transposon that is left over from the pre-cut ends. Transposons have also been modified to create libraries of C-terminal deletions, by introducing a series of stop codons at the transposon ends.[63]

Building on the ability to change the end sequence of a MuA transposon, it was shown by Jones that by using a restriction enzyme that cuts outside its recognition sequence, type IIS, the result would be a net removal of nucleotides.[64] The type IIS restriction enzyme that was used is *Mly*I. The recognition sequence is not a palindrome so the direction of which it cuts can be controlled and cuts both strands five base pairs downstream of the recognition site. Due to the duplication of base pairs and the positioning of the restriction sites, when the transposon is removed from the target DNA the result will be a triplet nucleotide removal from the target gene. Triplet nucleotide removal is further described in Figure 3.3.

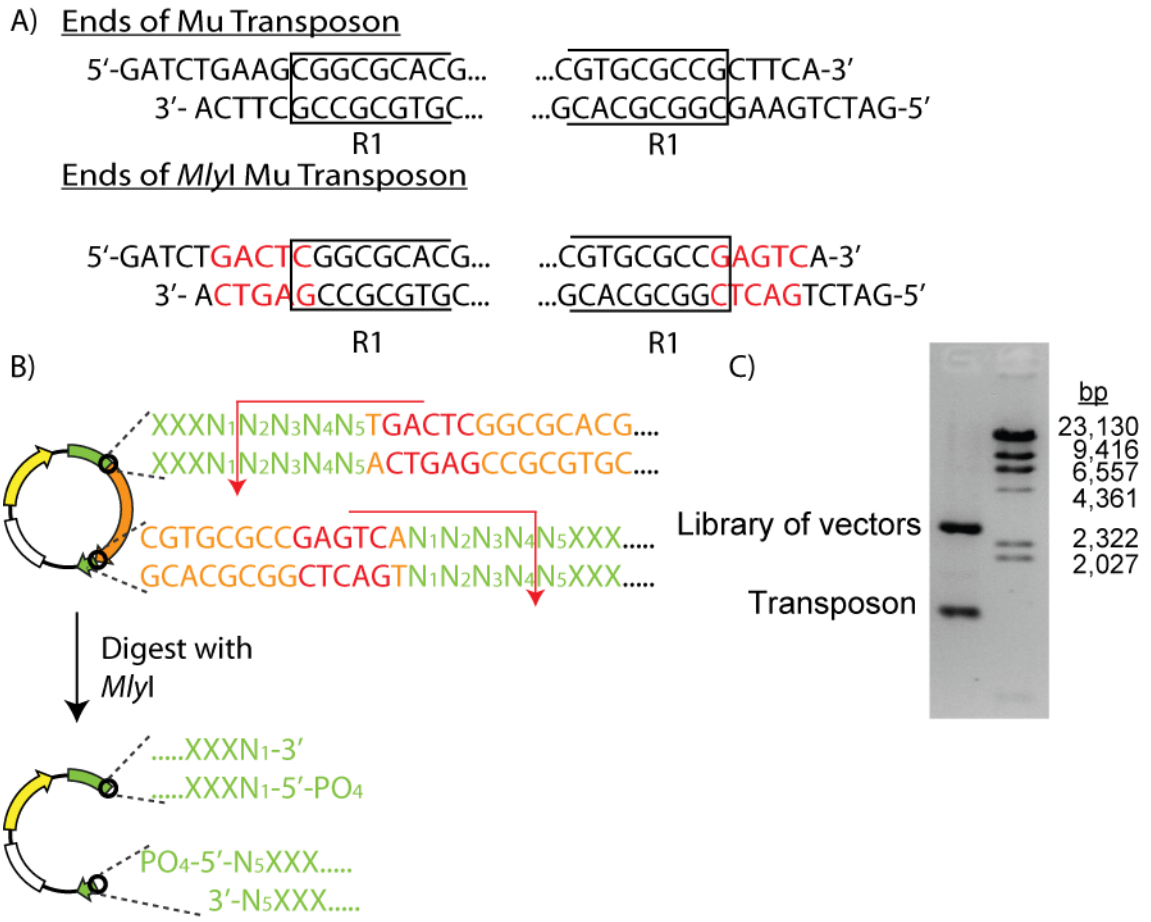


Figure 3.3: *Mly*I-Mu transposon to remove three nucleotides.
A) The difference in the ends of the transposon are shown in red, the *Mly*I recognition sequence is also highlighted in red. The mechanism of incorporation of the transposon is the same as in Figure 3.2. B) The recognition sites for *Mly*I are placed so that up to N_1 and N_5 will be cut resulting in deletion of $N_2N_3N_4$. The gel on the right shows an *Mly*I digestion of a pool of library clones.

3.3 Insertion of a new codon

The use of a modified transposon allows for the removal of three nucleotides and sets the stage for replacement of a new codon. Due to the instability of three base pair duplex DNA, the mutational codon would need to be part of a selectable phenotype. Three nucleotide bases is not enough to remain duplex DNA and therefore T4 DNA ligase would not be able to ligate in the codon. The simplest approach to insert a new codon would be to use a linker that codes for a selectable phenotype, such as β-lactamase.

The linker could then have the mutational codon as well as *Mly*I restriction sites appropriately placed on each end.

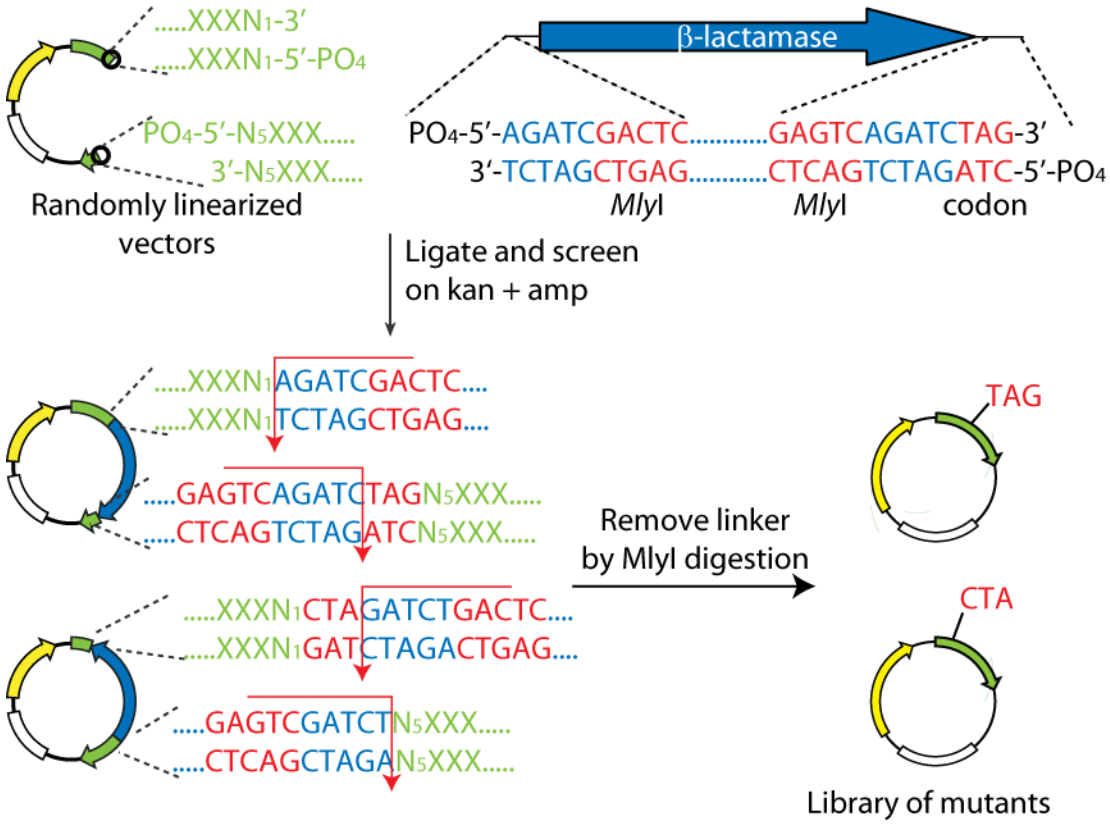


Figure 3.4: A simple approach to inserting a new codon.
The pool of randomly linearized $N_2N_3N_4$ deletion vectors can be ligated with a linker sequence that carries the codon to be inserted, TAG. Similar to the MlyI-Mu transposon, the restriction sites are placed so that there will be a clean removal of the linker only leaving behind the codon. In this process there would also be CTA mutations, as a result of the linker inserting in the reverse direction.

Recently, the triplet nucleotides NNN[65] and TAG[66] have been randomly introduced into a gene using this simple approach. However, only one-sixth of the library generated would have the mutation in-frame and the specific codon that is wanted (Figure 3.5). This is due to the fact that when the transposon is removed, the ends would be blunt thus any insert that is ligated can go in both the forward or reverse direction, scanning both TAG and CTA. Secondly, the transposon reaction is random and the insertion of a new codon can occur in any of the three reading frames. To develop a universal scanning

method that only facilitates in-frame mutations the linker insertion would have to produce a selectable phenotype only when inserted in the correct frame.

5′-ATG TCG ACT CGG AGT ATT TTA CCC GGT-3′ Met Ser Thr Arg Ser Ile Leu Pro Gly	original sequence
5′-ATG TCG ACT TAG AGT ATT TTA CCC GGT-3′ Met Ser Thr Amb Ser Ile Leu Pro Gly	in-frame mutation
5′-ATG TCG ACT CTA GGT ATT TTA CCC GGT-3′ Met Ser Thr Leu Gly Ile Leu Pro Gly	out of frame mutation (+1)
5′-ATG TCG ACT CGT AGT ATT TTA CCC GGT-3′ Met Ser Thr Arg Ser Ile Leu Pro Gly	out of frame mutation (+2)
5′-ATG TCG ACT CTA AGT ATT TTA CCC GGT-3′ Met Ser Thr Leu Ser Ile Leu Pro Gly	in-frame, reverse
5′-ATG TCG ACT CCT AGT ATT TTA CCC GGT-3′ Met Ser Thr Pro Ser Ile Leu Pro Gly	out of frame (+1), reverse
5′-ATG TCG ACT CGC TAT ATT TTA CCC GGT-3′ Met Ser Thr Arg Tyr Ile Leu Pro Gly	out of frame (+2), reverse

Figure 3.5: Insertions of TAG in all 6 frames.
Only 1 out of 6 insertions will give the correct mutation. The mutational codon, TAG is highlighted in red as well as the amino acid mutations that would occur.

3.4 Development of a reading frame selectable linker

The most straight forward way to develop a reading frame selectable linker would be to not include the promoter in the linker sequence as well as removing the start codon from the selectable gene. When the linker is inserted there would be C-terminal fusions.[67, 68] When designing CSM, the linker would be best if when inserted in-frame selection can occur in the presence of an antibiotic selection, such as β-lactamase in the presence of ampicillin. A disadvantage of this simple reading frame approach is that the possibility that within the target gene there could be several ATG start codons next to potential ribosome binding sites.[69] The linker would not necessarily have to insert into the correct reading frame to give a selectable response and would lead to false positives resulting in unwanted out-of-frame mutations. To ensure that mutations would only occur in the correct reading frame, a second layer of selection within the system is needed.

The first generation of a reading frame selectable linker was based on a reading frame selectable plasmid that was developed by Benkovic and co-workers for directed evolution experiments.[70]

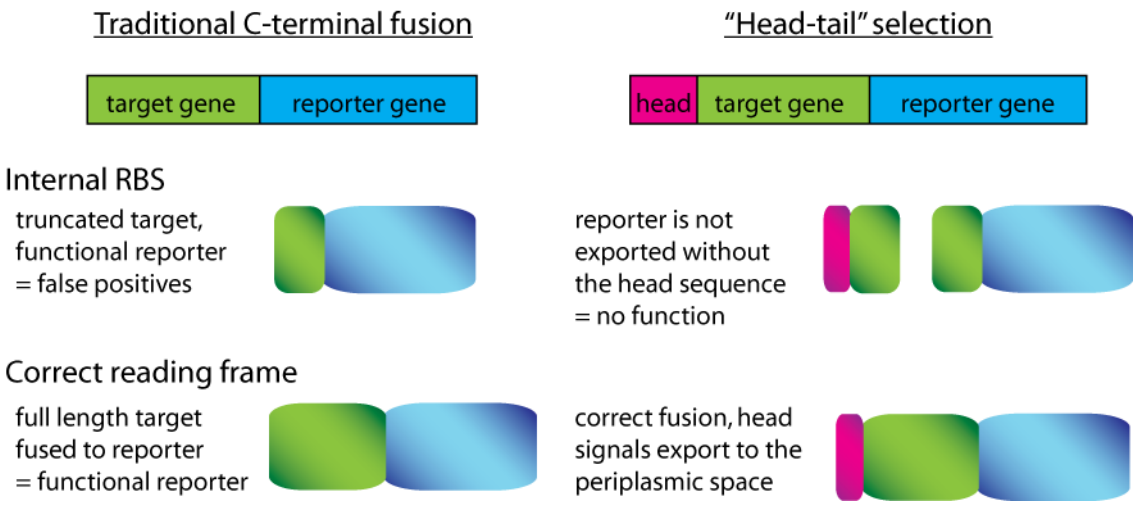


Figure 3.6: Comparison of C-terminal fusion selection with a "head-tail" selection system.
The left panel displays the traditional C-terminal fusion selection approach. This approach leads to many false positives due to possible internal RBS. On the right, the head-tail system requires that a sequence both on the N and C termini of the target protein are translated for proper function. Any internal RBS, will separate these two components eliminating false positives.

This selection system puts the mutant gene of interest in between a head and tail sequence, see figure 3.6. Translation of both the signaling peptide (head sequence) and β-lactamase (tail sequence) is required for exporting to the periplasm. If any N-terminal truncations, as a result of an internal RBS, were to occur then exportation would not occur and result in ampicillin sensitivity.

To test the feasibility of using a head-tail system for the reading frame selection the signaling peptide was fused to the target gene. A linker sequence was then created by amplifying β-lactamase, without the signaling peptide and start codon. A small scale library was created using the Mu-*Mly*I transposon following the manufactures recommendations for the reaction. The transposon was then removed by *Mly*I digestion

followed by insertion of the linker and selection on ampicillin. After removal of the linker sequence several clones were sent to be sequenced. Sequencing results revealed that all mutations did in fact occur in the correct reading frame, however all were clustered toward the N-terminus of the target gene, Figure 3.6. The bias of the mutations favoring the N-terminus of the target gene is caused by the fusion of β-lactamase to the target gene and that each protein resulting from the gene fusion would be different, as shown in Figure 3.8.

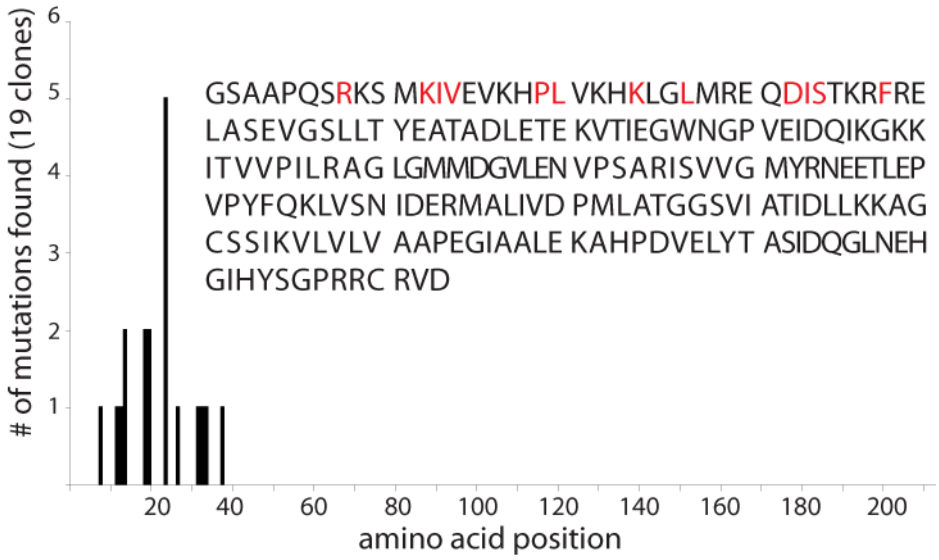


Figure 3.7: Position of insertion using a "head-tail" reading frame selection.
19 individual clones were sequenced, giving 12 different mutation sites. However in a 213 amino acids protein, no mutations were found past position 38.

3.4.1 Incorporating an intein into the selection system

Incorporating and intein into the selection system would allow for all selectable proteins to be the same regardless of where the linker is inserted into the gene. An intein is a self splicing protein, similar to introns and exons but occurs after protein translation. Inteins excise themselves by either a cleavage or splicing mechanism.[71, 72] Two separate linkers that include inteins were developed for use in the CSM method.

Generation 2: a non-biased method. The second generation reading frame selectable linker that was developed incorporates the use of an intein to minimize the bias that was observed with the head-tail selection system. This selection system is based off of pInSALect,[73] a selection plasmid that addressed the issues that were present in the original pSALect[70] system. It was found that by simply adding the second layer of selection pressure of requiring fusion of the target gene to both a head and tail section that is required for proper protein activity was not reliable. A requirement of the selection system is that the "head-target protein-tail" fusion is soluble. Protein fusions which are mis-folded[74] or form protein aggregates would be selected against.[73]

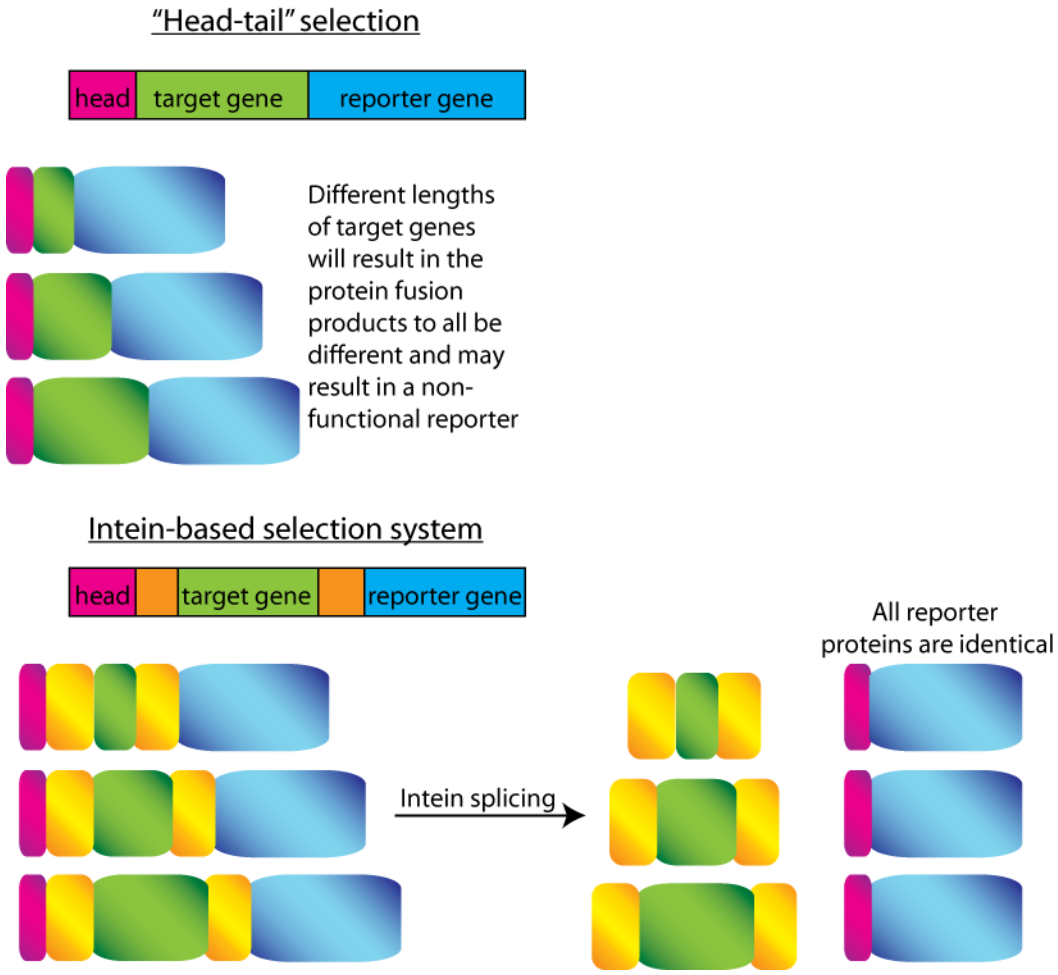


Figure 3.8: Comparison of an intein selection vs. a non-intein selection.
Top panel: Without an intein present, all fusion products will be different lengths. The various gene segments that would be introduced between the "head" and "tail" genes could result in proteins that are non-functional. Bottom panel: The addition of intein segments on either end of the target gene allow for all reporter proteins to be identical.

In the CSM method, the head sequence, tatSS-Nterm-VMA intein is fused to the target gene in the target plasmid. The linker sequence includes the C-terminal portion of the intein as well as β-lactamase. The restriction sites were placed in such a manner that would allow for read through of the fusion gene if inserted in the correct reading frame. The initial test of this linker system showed that the bias towards the N-terminus of the gene is no longer present. Additionally to test if the reading frame system would still work if inserted at the C-terminus of the gene, a double strand break was site-specifically

introduced and the linker ligated and selected for on carbenicillin. Initially selections were done in the presences of ampicillin, however the libraries appeared to be contaminated with out-of-frame mutants. This was also the case in the reported pInSALect plasmid, where it was hypothesized that ampicillin hydrolysis was the cause of obtaining false-positives. Carbenicillin is a more stable analog than ampicillin. Selections using the intein system were tested at both 30 and 37 ^{o}C and based on counting colonies it was confirmed that lower growth temperature resulted in more colonies. Lower growth temperatures were also observed in the initial report of pInSALect by Lutz.[73]

Figure 3.9 shows the location of the linker as the amino acid position that would be mutated. Nine unique locations of the linker were observed out of 12 individual colonies. Incorporating the intein into the selection system proved to remove the bias that was observed previously. While at first glance the N-terminal bias still seems to be present, further sequencing of library of GCG mutants introduced into UPRT, a 217 amino acid protein, showed mutations to be evenly dispersed as well as a mutation at position 204, shown in Figure 5.2.

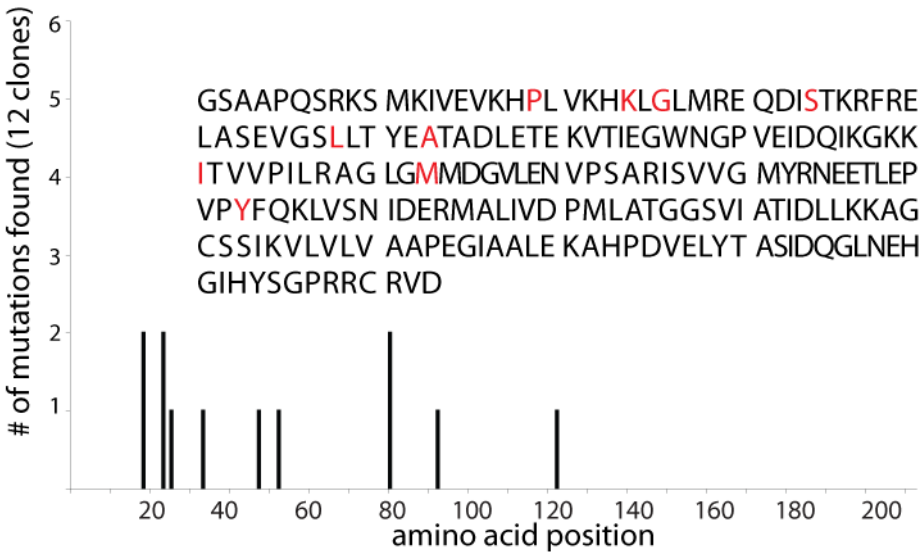


Figure 3.9: Position of insertion using the intein selection system.
12 individual clones were sequenced and 9 different codon mutations were observed. With the intein system the mutations are relatively dispersed throughout the entire sequence when compared with the non-intein selection system (Figure 3.7).

Generation 3: thymidylate synthase linker. In the initial applications of the CSM described in Chapters 4 and 5, fusing the N-terminus of the intein directly to the target gene was not ideal. An attractive method would eliminate the need to sub-clone the genes which have a single codon mutation into an expression vector. It was also thought that CSM should allow for multiple specific mutations to be incorporated. The most recent attempts at a reading frame selectable linker is based on a selection system that uses an intein thymidylate synthase reporter.[75] It has been shown for an intein to catalyze excision the N-terminus is not required. Rather then having N-intein-gene-C-intein-reporter, with a C-terminal cleavage the construct can be assembled gene-C-intein-reporter.[76] To select for ligation events that occur in-frame a thymine deficient strain of *E. coli* was made using the gene knockout method of Wanner, Figure 3.10.[77] Through homologus recombination, the gene encoding thyA was replaced by a selectable maker, which was then removed by λRed recombinase. The ΔthyA *E. coli* strain is unable to

survive without media supplemented with thymine or a plasmid expressing thyA, see Figure 3.14 in experimental.

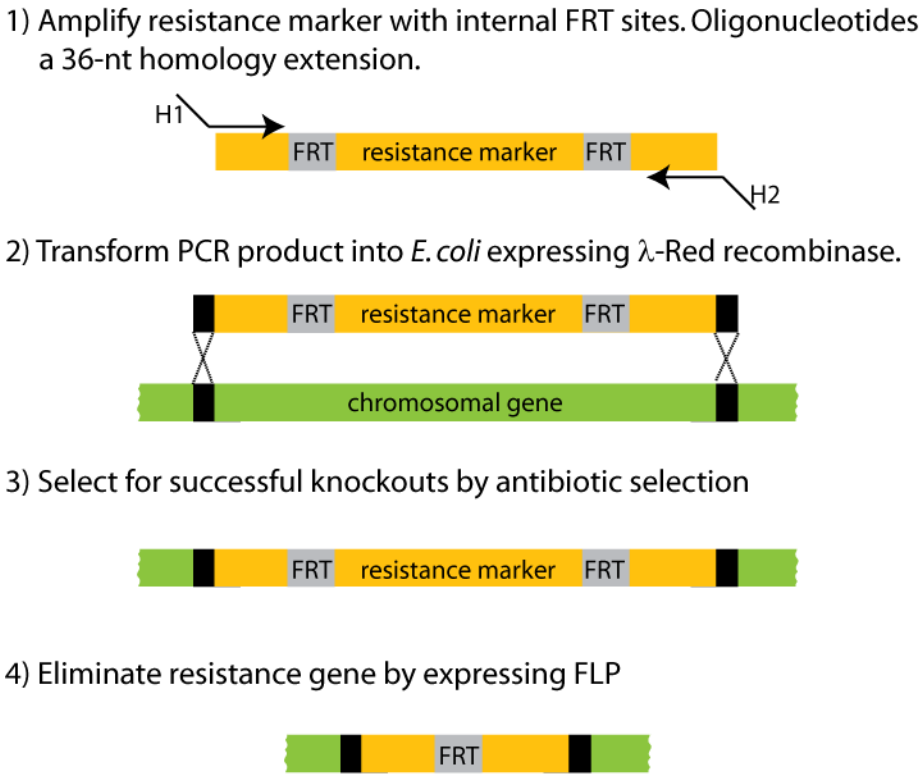


Figure 3.10: PCR mediated chromosomal gene knockout.
A resistance marker is amplified with oligonucleotides that have regions of homology (H1 and H2). This segment also carries the FRT sites. The PCR product is transformed with a plasmid carrying the gene for λRed recombinase. Through homologus cross over the resistance maker replaces the chromosomal gene. The resistance marker can be removed by FLP recombinase that recognizes the FRT sites.

The linker was made by amplifying the intein and thyA from pPPV.[75] After many attempts of inserting the linker into both a specific site and a library, the ability to survive on media not supplemented with thymine could not be recovered. It was hypothesized that a sufficient amount of thymidylate synthetase was not produced to survive on media lacking thymine which resulted in the cell death. This was tested by first growing ΔthyA cells transformed with a library of plasmids with the thyA linker inserted on media supplemented with thymine. All colonies resulting from this transformation where then pooled and plated on media lacking thymine. As expected where the linker inserted in-

frame the cells were able to survive. Since, the ligation of linker and library is blunt and an intermolecular ligation event would be rare, a majority of the clones on the plates supplemented with thymine would not contain the linker sequence. To remove any members of the library that do not contain the linker the initial screen for positive insertions would be required. The thyA linker was redesigned to also carry β-lactamase, where ampicillin resistance is not dependent on inserting into the correct reading frame. The linker can first be ligated into the library, transformed into ΔThyA cells and grown on M9 agar supplemented with thymine and ampicillin. Any colonies that survive on ampicillin would have the linker inserted. The colonies can then either be pooled and dilutions plated or replica plated on M9 agar without thymine.

This linker eliminates the need to sub-clone the mutant genes into an expression plasmid and the only requirements are that the restriction site for *Mly*I is not present. While this linker follows the simple approach to a reading frame selectable linker described in 2.4, the possibility of insoluble protein fusions cannot occur due to the presences of the intein. The thyA based linker simplifies the CSM process but it is not clear if only in-frame mutations will be obtained. In a test, where the linker was site-specifically placed in-frame and out-of-frame, it was found that the out-of-frame insertion was able to survive on media lacking thymine. Upon analyzing the sequence prior to the insertion site it was found that there is a start codon and possible ribosome binding site, shown in Figure 3.11.

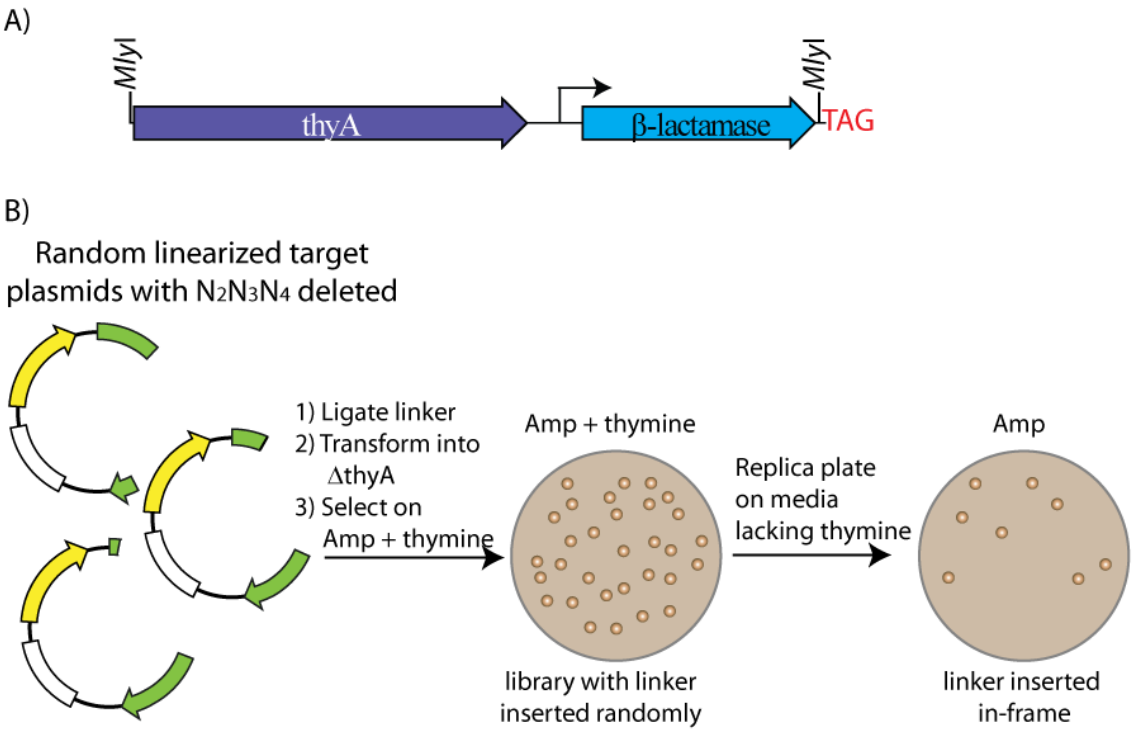


Figure 3.11: Selection using the thyA linker
A) The thyA linker also carries the gene for ampicillin resistance. Resistance to ampicillin is not dependent on inserting in-frame. B) The linker sequence can be ligated with the codon deletion linear library. First the library is selected for on media supplemented with both thymine and ampicillin. Next the resulting colonies can be replica plated onto media lacking thymine. Only where the linker inserted in-frame will be able to survive. C) Sequence of the linker site-specifically inserted in-frame and out-of-frame. The additional base for the out-of-frame is highlighted. Although inserted out-of-frame +1, expression of thyA proceeded.

When performing CSM the linker of choice would depend on how clean the library of interest needs to be. For a library that contains a greater number of in-frame

mutations, the intein-β-lactamase system over the thyA system may be a better choice. While fusing the protein of interest to the N-terminus of the intein may render the protein inactive, in some cases the protein of interest would still be functional. In the proof of concept experiments using CSM, it was found that uracil phosphoribosyl transferase (UPRT) was functional when fused to the N-terminus of the VMA intein, as described in Chapter 5. While in the first experiment using CSM, sub-cloning of the mutant library was required, described in Chapter 4, and in this case the thyA system may have been a better choice. The thyA based system may prove to be useful in performing multiple scans, as well as, rapidly introduce mutations and screen for desired function by performing CSM on an expression plasmid. However, if creating libraries of protein variants where the assay for protein function is not performed in *E. coil*, then sub-cloning would be required, regardless of which method is used.

3.5 Codon Scanning Mutagenesis using a NotI-Mu-transposon

When working out the specific details of CSM, it was originally thought that performing a blunt ligation on library scale would be very inefficient and difficult. And in fact, it was found that performing blunt ligations was difficult to obtain the number of clones required to cover all possible sites of insertion. Later in the development it was found that the inefficiency of the ligation was caused by (1) removing 5'-phosphates from the library to inhibit intramolecular ligations and (2) reaction conditions, both the ratio of library to linker and the amount of ATP. It was found that the optimal concentration of ATP for blunt ligations was 0.5 mM as opposed to the standard 1 mM. However, prior to the optimization of the blunt library ligation, details and preliminary tests of the method using a commercially available modified transposon were designed

and performed. The modified transposon has the restriction sites for *Not*I and is marketed for linker scanning mutagenesis. It was originally thought that using the *Not*I modified transposon would be best because a cohesive end ligation would be done rather then blunt end ligation in the case of *Mly*I. The major difference between the two modified transposons is that the *Not*I transposon will leave behind a total of 15 bases, whereas the *Mly*I transposon removes three bases. If using the *Not*I transposon, the linker needs to be designed to remove 18 bases as well as leave behind three for the new codon. A series of typeIIS restriction sites are used and would require two digestion-blunting-ligation steps. The details of this linker are depicted in Figure 3.12 Either of the reading frame selection systems described in section 3.4.1 could also be adapted for the *Not*I system.

While using a cohesive end ligation for the linker resulted in more ligation events the limitation of processing the linker proved to be problematic. One was digesting with *Bsa*XI. It was later found that this enzyme requires two sites to digest DNA efficiently. This method also required more processing of the DNA which could lead to unwanted mutations from all of the digestion as well as increases the risk of loosing members during the various purification steps. It was during this time that the blunt ligation was optimized and all future experiments were performed using the *Mly*I transposon.

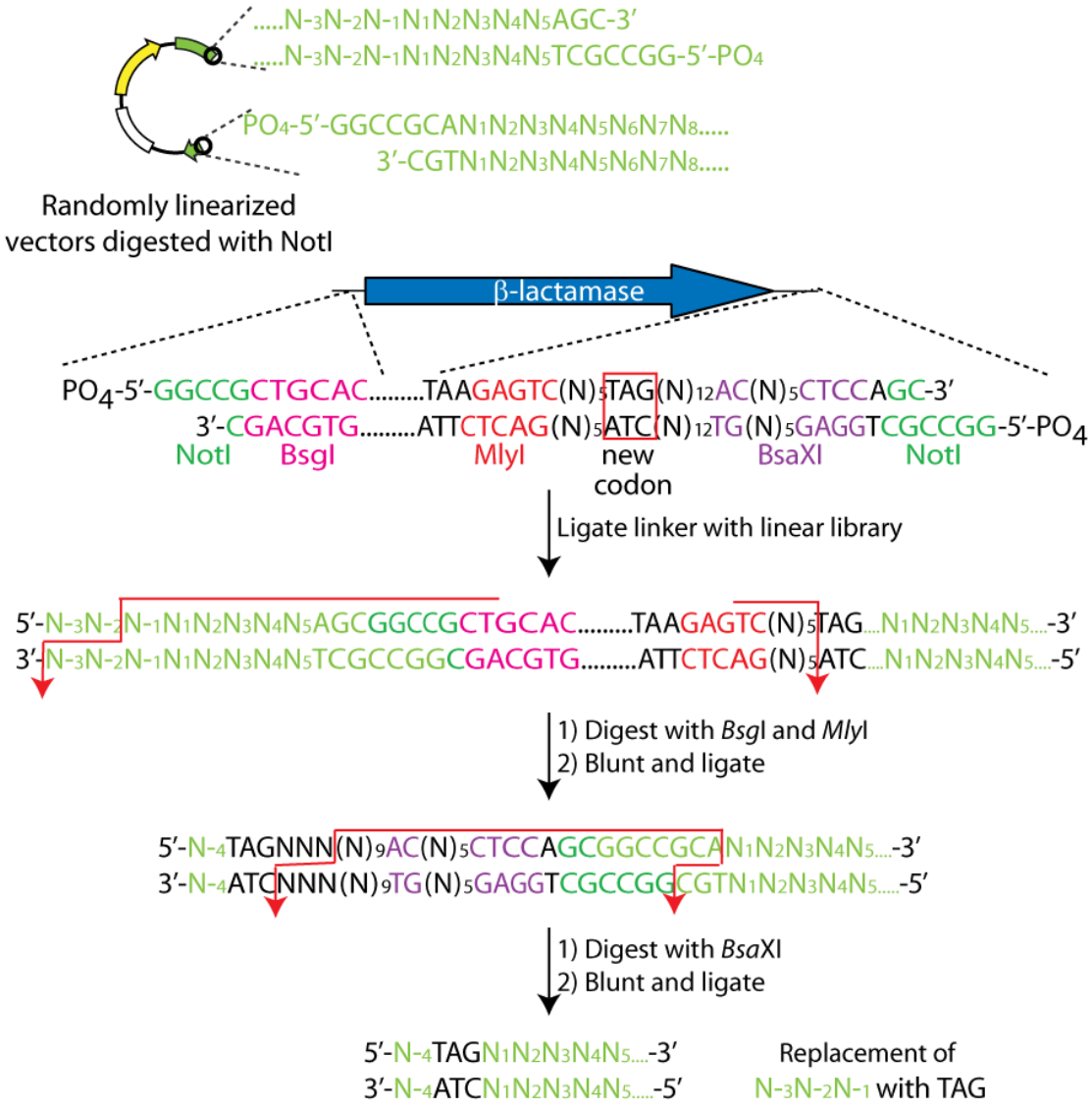


Figure 3.12: CSM using a transposon modified with *Not*I restriction sites.
The linear library of vectors is generated by using the commercially available MGS kit from Finnzymes. The linker is digested with *Not*I and ligated with the linker library. To process the mutation, two digestion steps are required. The first set, *Mly*I and *Bsg*I removes the five base pair duplication and replaces a codon with the mutation codon (TAG). The second step, *Bsa*XI removes the *Not*I restriction site.

3.6 Development of a transposon that can create random in-frame codon mutations

While developing the reading frame selectable linker, it was thought that the transposon can be used to randomly delete in-frame-codons. If the modified MuA transposon can delete three nucleotides at random and in-frame mutations can be chosen based on a selectable response. Then the DNA sequence encoding the selectable

response can be the inserted between the two recognition sites of the transposon. This was worked out using the intein-β-lactamase system and the pIT-target. Using CL826 and CL827 the C-terminus VMA intein and β-lactamase were amplified from pIntSALect. The oligonucleotides were designed with the five base pair duplication of the reaction in mind and when inserted in-frame there would be translation throughout the recognition sequence. The transposition reaction conditions were performed according to the manufactures recommendations. Transposon insertions were selected for on LB agar that was supplemented with kanamycin and carbenicillin and compared to media only supplemented with kanamycin to determine transposition frequency. By sequencing individual clones, it was found that the transposon did in fact insert in-frame and removal of the transposon results in an in-frame codon deletion. The properties of this in-frame deletion transposon are currently being used by Jia Liu, who is expanding on the developed CSM method.

3.7 Overall process of CSM and the determination of sufficient library coverage

In the following experiments described in Chapters 4 and 5 the reading frame selection system that was used for CSM is based off of the intein based "head-tail" system described by Lutz and coworkers.[73] An outline of the method is depicted in Figure 3.13. The following equation for estimating the number of clones needed to determine if sufficient library coverage was obtained was used:[78]

Equation 3.1: $L = -V \ln(1 - F)$

Where L is the number of clones in the library, V is the number of variants and F is the percent confidence. In the example in Figure 3.13, the number of clones to ensure 99% library coverage in the transposition reaction there are 2,100 possible insertion sites. The

size of the resistance marker and origin of replication are subtracted from the overall size of the plasmid. Approximately 10,000 clones would be required for complete library coverage.

$$L = -2100 \ln(1 - 0.99) =\sim 10{,}000$$

The number of clones required for ligation of the linker followed by selection on carbenicillin for the correct reading frame is far less. If the target gene is 600 bp then there are 200 possible codons to mutate. Using equation 3.1, approximately 930 clones would be required to ensure that complete library diversity is maintained.

$$L = -200 \ln(1 - 0.99) =\sim 930$$

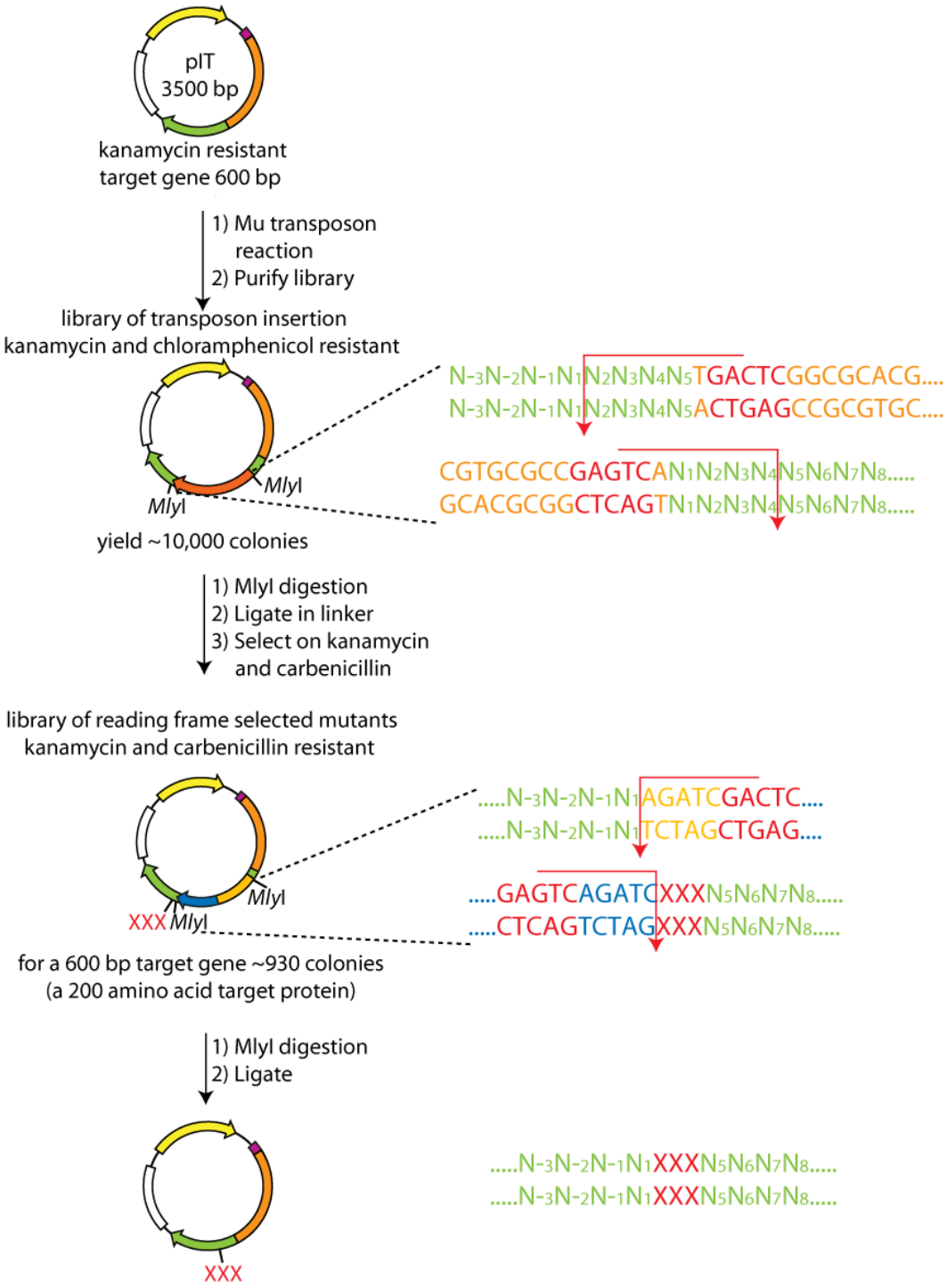


Figure 3.13: Overall process of CSM.
pIT target plasmid: tatSS (pink) and N-term intein (orange) is fused to the target gene (green). Insertion of the transposon should yield ~10,000 colonies. This library can be further purified by removing the target gene with the transposon inserted and ligating with the original plasmid backbone. Removal of the transposon by *Mly*I removes $N_2N_3N_4$. Ligation of the linker with C-term intein (light orange) and β-lactames (blue) and selection on carbenicillin should yield ~930 colonies. Digestion with MlyI leaves the mutation codon XXX in place of $N_2N_3N_4$.

3.8 Materials and Methods

Materials. All restriction enzymes, Phusion Polymerase and T4 DNA ligase were purchased from either New England Biolabs (Ipswich, MA) or Fermentas (Glen Burnie, MD). MuA transposase was purchased from Finnzymes (Finland). Oligonucleotides were synthesized by both Integrated DNA Technologies, Inc. (Coralville, IA) and Invitrogen (Carlsbad, CA). The plasmid pInSALect was obtained from Stefan Lutz, Emory University and pPPV from Michael Hecht, Princeton University. DNA sequencing was done by University of Michigan DNA Sequencing Core. Antibiotics; kanamycin, ampicillin, carbenicillin were purchased from Fisher. Chemical competent Gene Hogs were used in all experiments except were noted.

General Methods. Plasmid DNA was isolated from *E. coli* using Fermentas GeneJet Kit, Zymo Research Zyppy Plasmid Miniprep kit, or Qiagen Plasmid Midi prep kit according to the manufactures recommendations. All PCR reactions were carried out using a PTC-200 Pelteir Thermo cycler. PCR products were purified using Qiagen PB buffer. Digested DNA was purified by running DNA on a 1% agarose TAE gel, bands excised and DNA purified using either Qiagen QG buffer or via electroelution followed by isopropanol precipitation.

Table 3.1: Oligonucleotides used for constructing the components used for CSM

CL266	GCTTAGATCTGACTCGGCGCACGAAAAACGCGAAAG
CL861	GGATCGACTCTCCTGGGTATTCGCAATAATCTTAATACTGAG
CL863	CTAGATCTGACTCAATTACCAATGCTTAATCAGTGAGGCACCTA
CL1064	AAGAGTCGAACAGCGAGAAGAGCAAATTCGAGCTCGAACAAC
CL1065	AAGAGTCTGCAACGCTGAAGAGCTTCTAGATACGTAAGATCTTTACA CCGCCATC
CL1066	AAACCAGATCTGACGAAAGGGCCTCG
CL1067	AACCATCTAGAGACAGTTACCAATGC

***Mly*I-Mu transposon.** The *Mly*I transposon was created by PCR amplifying the Entranceposon (M1-CamR) (Finnzymes) with CL266 to incorporate *Mly*I restriction sites. The 1254 bp product was phenol-chlorofom extracted followed by ethanol precipitation. The purified product was then digested with *Bgl*II for 3 hours, phenol-chloroform extracted and ethanol precipitated. The DNA pellet was re-suspended in 50 μL of water and quantified by gel electrophoresis. Later in the development of the method the transposon was cloned into the *Sma*I site of pUC18 to allow for higher quantities of the pre-cut transposon to be produced.

ThyA gene knockout. To delete the gene encoding thymidylate synthase from the genome of *E. coli* the PCR based gene knockout method that was described by Datsenko and Wanner.[77] Oligonucleotides CL50 and CL51 were designed to amplify the kanamycin resistance gene from pKD13.[77] Flanking the kanamycin resistance gene is a 36 nucleotide homology with the sequence that flanks the sequence on either side of the genomic thyA gene. PCR gave a 1400 bp band which was purified by ethanol precipitation. The amplification of the resistance gene also carries FLP recognition target sites. Gene Hogs were then transformed with pKD46 which carries the gene for λ-Red recombinase. 4 mL of LB supplemented with 100 μg mL^{-1} ampicillin was added after a one hour recovery. Once the culture reached saturation the 5 mL's was added to 250 mL 2XYT supplemented with 100 μg mL^{-1} ampicillin and grown at 30 ^{o}C until the culture reached OD_{600} = 0.3. At that time 10 mM of L-arabinose was added to begin expression of λRed recombinase. The culture was grown to OD600 = 0.6 then washed with cold water and electroporated with the kanamycin PCR product. 1 mL 2XYT was added to the transformation for recovery at 37 ^{o}C. After 1 hour a 250 μL aliquot was removed and

plated, the remainder 750 μL was recovered for 7 hours then plated on M9 media supplemented with thymine and kanamycin. Individual clones were then picked and replica plated on M9 media supplemented with thymine and trimethoprim. Selection on trimethoprim eliminates any clones which have functional ThyA.[79] A single clone which survives the selection is then grown in media supplemented with thymine. The now kanamycin resistant, thyA deletion strain is made competent and transformed with pCP20 a plasmid carries FLP recombinase and recognizes the FRT-flanked resistance gene.[80] Colonies which are sensitive to all antibiotics and can survive on media supplemented with thymine and thymine and trimethoprim were colony PCR amplified and the product sequenced to show that the expected scar remains. Figure 3.14 compares Gene Hogs and the ΔthyA deletion strain carrying pUC18 an ampicillin resistant plasmid. The plates show that the strain can survive on M9 media in the presence of thymine (D) and trimethoprim, (C) whereas Gene Hogs are not able to survive on trimethoprim (B and C). Additionally there is zero growth of the thyA deletion strain on media that lacks thymine (A and B).

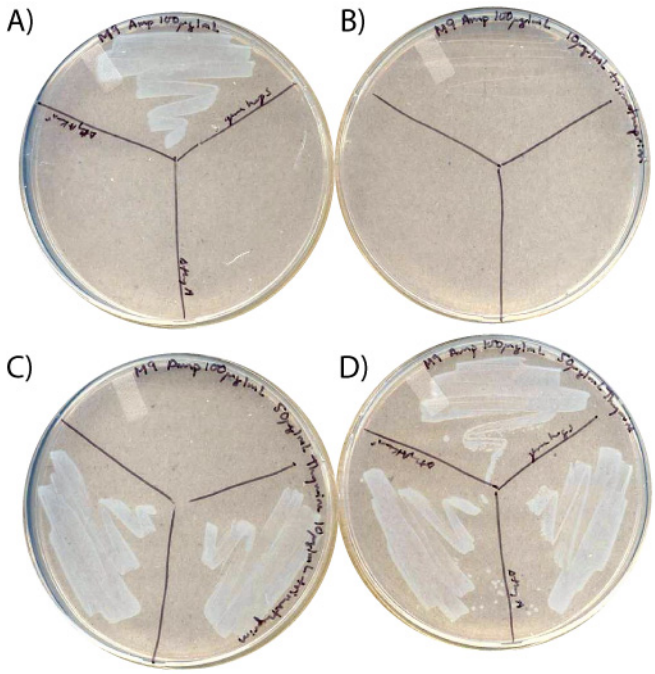


Figure 3.14: Growth of the ΔthyA strain.
All plates are M9 media supplemented with 100 μg mL^{-1} ampicillin (A) no additional additives, (B) 10 μg mL^{-1} trimethoprim, (C) 50 μg mL^{-1} thymine and 10 μg mL^{-1} trimethoprim, (D) 50 μg mL^{-1} thymine.

Chapter 4: Scanning photoaffinity mutagenesis

4.1 Introduction

The first protein that was chosen to test the developed CSM method was glutathione S-transferase (GST) from *Schistosoma japonicum*. This protein was picked because it is a small protein, 219 amino acids and is a well characterized homo-dimer.[52, 53] Another unique feature is that when a photo-reactive amino acid is placed at the interface cross-linking can occur.[36] This then shows that CSM is useful in scanning unnatural amino acids throughout a protein coding sequence. From preliminary tests with the linkers it was decided at the time to use the intein-β-lactamase linker and target plasmid that were both derived from pInSALect.[73] The gene encoding GST was amplified and cloned into pIT, so that GST would be fused to the N-terminus of the intein.

4.2 Results and Discussion

4.2.1 Generation of an amber codon-scanned library of the gene encoding glutathione *S*-transferase

Using the *Mly*I-Mu transposon a library of $>10^4$ independent colonies was generated.[81] This library was further purified by digesting to pool all GST target genes containing the transposon and religating with the vector. Removal of the transposon with *Mly*I results in deleting three nucleotides. The linker that was used has a TAG codon and when inserted in the correct reading frame a selectable phenotype is produced. When this library was plated in both the presence and absence of carbenicillin, it was observed that ~90% of the library was removed. This is consistent with wanting to remove 5/6th of the

library (83.3%). Removal of the linker leaves TAG in place of the three nucleotides that were removed when the transposon was excised.

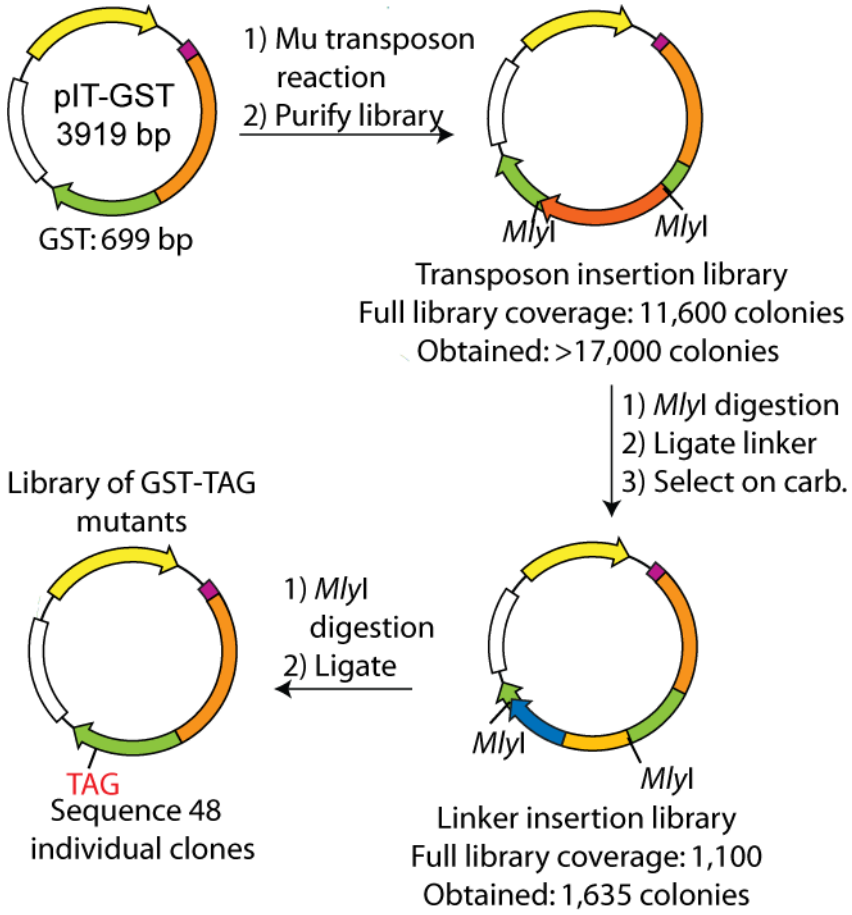


Figure 4.1: Process of using CSM to create random TAG mutants in GST.
The target plasmid, pIT-GST is used in a transposition reaction where successful insertions are screened on kanamycin (resistance of target plasmid) and chloramphenicol (resistance of transposon). Based on the size of the target plasmid it was estimated that 11,600 clones are required for 99% library coverage (using equation 3.1). The gene encoding GST is 699 bp, there are only 233 positions where the linker can insert in the correct reading frame and for 99% confidence 1,100 clones are required. In both cases the clones obtained were sufficient for full library covrage.

While processing the library an individual clone from the linker ligation was removed and processed separately. Based on the insertion site of the linker the sequence of the transposon clone was determined. Digestion with *Mly*I followed by ligation yielded the expected TAG mutation, Figure 4.2. The sequencing represents a single event that is occurring in a library of thousands.

Figure 4.2: Sequence of a single clone through each step of CSM.
In the wild-type sequence the ACA that will be mutated is highlighted. The removal of the ACA codon is shown by the site of the *Mly*I-transposon (no sequence data available). Next is the insertion site of the *Mly*I linker, and cut sites are indicated by red arrows. Last is the final sequence showing the TAG mutation in place of ACA.

4.2.2 Expression and purification of random TAG GST mutants

Removal of the linker generated thousands of clones, where each member presumably would contain the expected mutation seen in Figure 4.2. Ten representative TAG-GST mutants were chosen at random and cloned into the expression plasmid pBADmycHisA and transformed into *E. coli* expressing the translational components for inserting the unnatural amino acid *p*Bpa. Each of the 10 representative clones was

individually grown and expressed with or without *p*Bpa present. Resulting mutant proteins were isolated by Ni^{2+} resin purification and analyzed by SDS-PAGE, Figure 4.3.

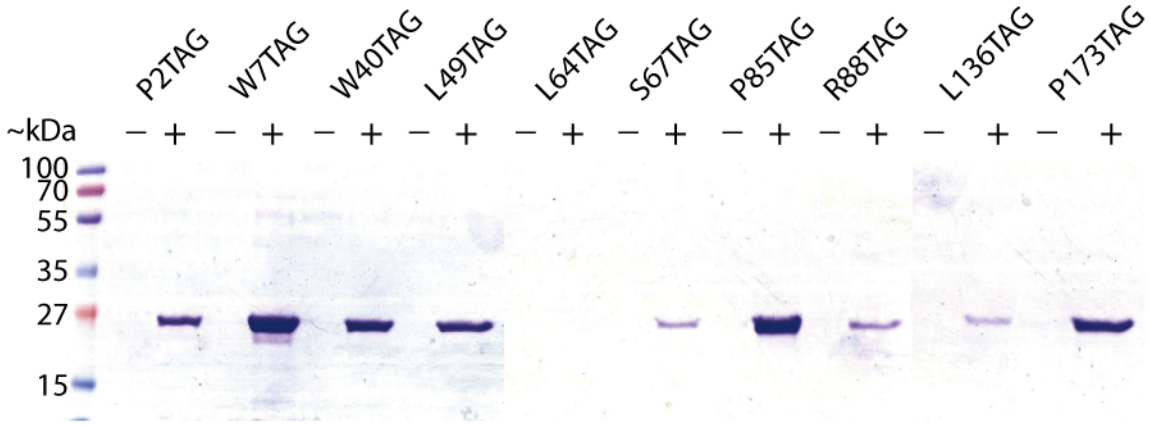


Figure 4.3: SDS-PAGE analysis of 10 amber mutants.
Expressions were done in the absence (-) and presence (+) of pBpa. All mutants expressed, except for L64TAG.

4.3 Conclusions

As it can be seen in Figure 4.3, only when the culture was supplemented with unnatural amino acid the protein was expressed and purified. Not all mutants expressed the same level of protein, and one clone (L64TAG) did not produce any protein even after multiple attempts. We believe that the inconsistency in the amount of protein produced is based on the context effects of amber suppression. Only mutants which showed good expression with *p*Bpa were irradiated with 365 nm light and loaded onto an SDS-PAGE gel to analyze if a covalent bond had formed across the dimeric interface, Figure 4.4. The only mutant that showed minor cross-linking is P85TAG. When the positions of the mutants are referenced on the crystal structure of GST, it can be seen that P85 is at the dimer interface and would be expected to cross link.

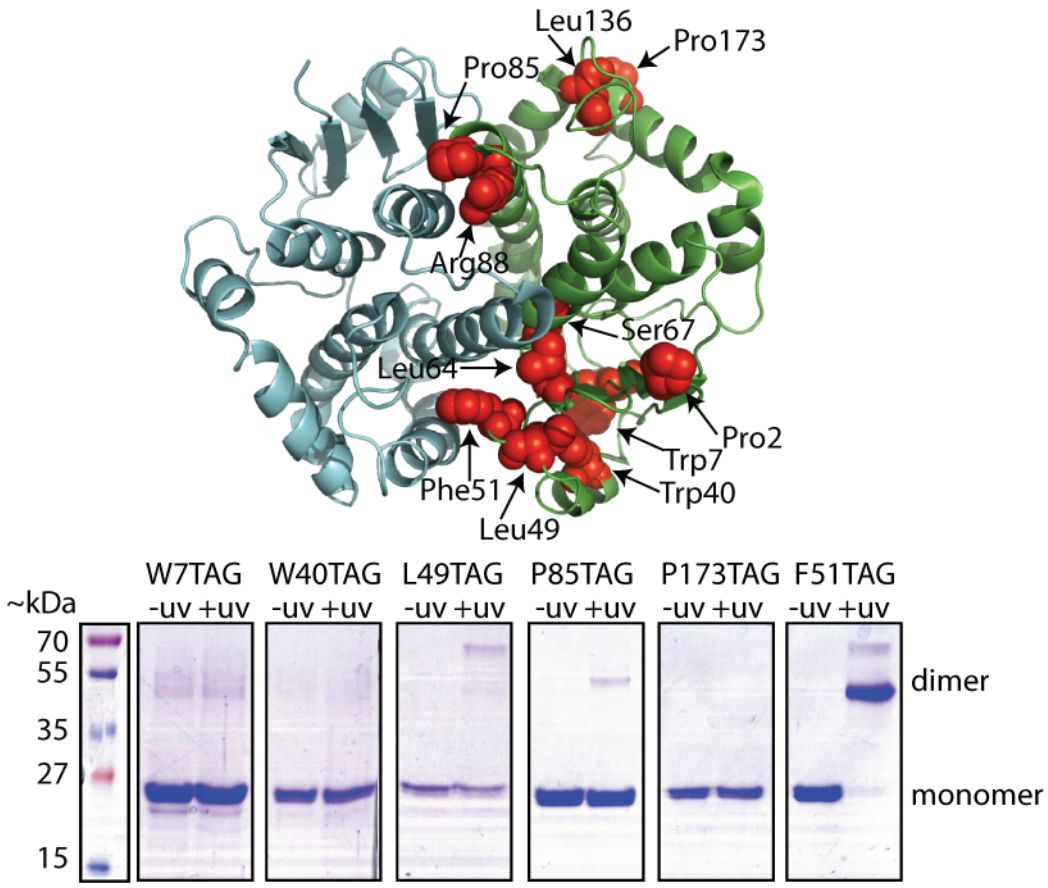


Figure 4.4: Location and photo-activity of amber mutants
The location of each of the mutations (red) was identified in the crystal structure (PDB 1Y6E). Mutants which produced the greatest amount of protein were irradiated with 360 nm light to promote cross-linking. The known cross-linking mutation F51TAG was used as a control.

Although only 10 individual clones were analyzed in the initial test of CSM, those 10 only represented a small fraction of the clones obtained. The use of the method of scanning a photo-cross-linking amino acid throughout a protein would allow for unknown interactions to be found. It would be possible to express the pool of mutants with the [D_{11}]-*p*Bpa, described in Chapter 2 and analyzed by mass spectrometry or individuals can be expressed spatially separated and purified in a high-throughput manner and analyzed for successful cross-links using a colorimetric assay.

4.3.1 Unexpected mutations and possible causes

In this first experiment using the developed CSM method, 48 TAG mutants were sequenced. Of the 48 sequenced it was found that only 15 different clones had the correct TAG mutation. Other mutations included either an addition or deletion of a codon before the mutation, see Figure 4.5. There were also unexpected mutations where an additional base was removed, either part of the coding sequence or the T and G of the scanned TAG codon. It was also found that some did not have a TAG mutation but instead were missing between 2 and 6 in-frame nucleotides. After analyzing the sequence data it was determined that majority of the unexpected mutations were caused by mis-cutting of *Mly*I, which could have been caused by either long reaction times or inappropriate reaction conditions. It is not believed this is caused by star activity, *Mly*I is a dimer and proper cutting requires that both dimers reach exactly 5 bases (NEB, personal conversation). The mis-cutting of *Mly*I, can cause the addition or deletion of a codon before the mutation, as well as the removal of the first base of the mutational codon as well as the 2 to 6 in-frame deletions, shown in Figure 4.5. The result of the last base missing from the mutational codon is caused by oligonucleotide synthesis. Oligonucleotides are synthesized from 3' to 5' and with standard desalting ~30% that are either lacking the 5' base or a deletion within the oligonucleotide (IDT, FAQ website, and personal conversation with representatives). It was found that even with additional purification the last base was missing from the mutational codon. These unexpected mutations are addressed in Chapter 5.

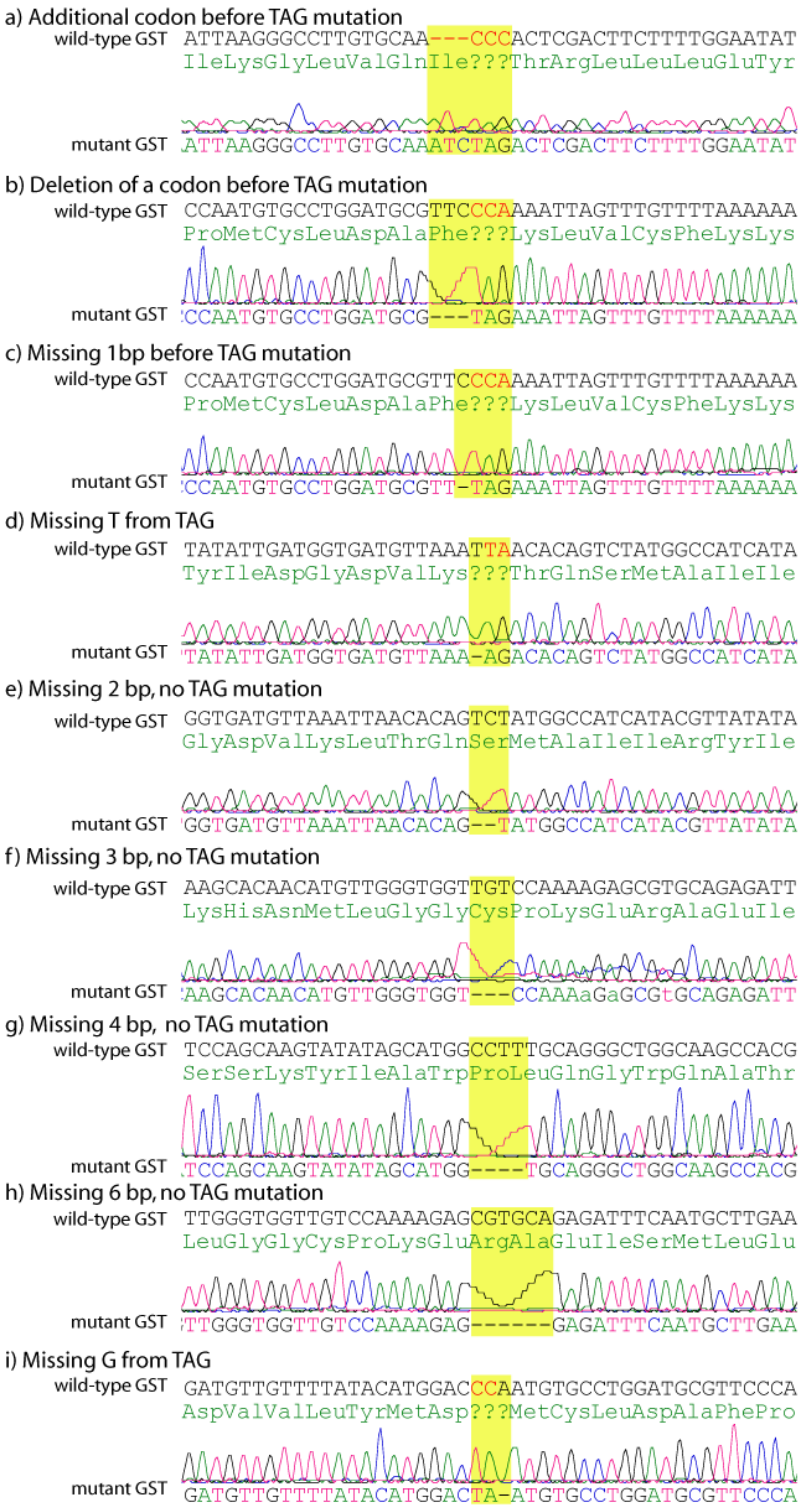


Figure 4.5: Unexpected mutations from CSM

The unexpected mutations that were identified from sequencing include (a) an additional codon or (b) deletion of a codon. As well those which had a TAG, but were missing and additional base (c and d). And several which did not have a TAG present but were missing (e) 2 bp, (f) 3 bp, (g) 4 bp and (h) 6 bp. There were also mutations that were (i) missing the G from the TAG codon.

4.4 Materials and Methods

Materials. All restriction enzymes, Phusion Polymerase and T4 DNA ligase were purchased from either New England Biolabs (Ipswich, MA) or Fermentas (Glen Burnie, MD). MuA transposase was purchased from Finnzymes (Finland). Oligonucleotides were synthesized by both Integrated DNA Technologies, Inc. (Coralville, IA) and Invitrogen (Carlsbad, CA). The plasmid pInSALect was obtained from Stefan Lutz, Emory University. Plasmids pGEX-T4-1 and pBADmycHisA were purchased from GE HealthCare Life Sciences and Invitrogen. DNA sequencing was done by University of Michigan DNA Sequencing Core. All media and antibiotics were molecular biology grade. Chemical competent Gene Hogs (Invitrogen) were used in all experiments.

General Procedures. Plasmid DNA was isolated from *E. coli* using Fermentas GeneJet Kit, Zymo Research Zyppy Plasmid Miniprep kit, or Qiagen Plasmid Midi prep kit according to the manufactures recommendations. All PCR reactions were carried out using a PTC-200 Pelteir Thermo cycler. PCR products were purified using Qiagen PB buffer. Digested DNA was purified by running DNA on a 1% agarose TAE gel, bands excised and DNA purified using either Qiagen QG buffer or via electroelution followed by isopropanol precipitation.

Table 4.1: Oligonucleotides used in photo-affinity mutagenesis

CL502	GTCTAGATCTCCCCGCGCGTTGGCCG
CL227	ATGTATATCTCCTTGAAATTGTTATCCGCTCAC
CL273	TTTCAAGGAGATATACATATGAACAATAACGATCTCTTTCAGGACTCA CGTCGGCGTTTT
CL276	CTCGAGGCCTCTAGAGAATTCGGATCCGGCCATGGTACCTCTGACAAC TTTAGAGTACAA
CL192	CTAGGATCCCCTATACTAGGTTATTGG
CL310	CCAGTCGACGCCTCTAGAAACCAGATCCGATTT
CL195	TATCGTCTTGAATCCAACCCGGTA
CL196	TACCGGGTTGGATTCAAGACGATA

CL197	ATGTTGGACGAGTAGGAATCGCAGAC
CL198	GTCTGCGATTCCTACTCGTCCAACAT
CL665	GTCGTTCAGAAAAGTCAGCACAGAGCCCACAAAAGTGATTCAAGTCG
CL666	CGACTTGAATCACTTTTGTGGGCTCTGTGCTGACTTTTCTGAACGAC

pIT-GST construction. Plasmid pIT-GST contains that Tat-signal sequence, the N-terminal portion of the cis-splicing VMA intein and glutathione S-transferase from *Schistosoma japonicum*. All three genes are in-frame fusions expressed from a lac^P promoter. pIT-GST was generated by amplifying the lac^P promoter from pUC19 using oligonucleotides CL502 and CL227 and gel purifying the 200 bp product. The Tat-signal sequence and N-terminal portion of the VMA intein was amplified from pInSALect using oligonucleotides CL273 and CL276, producing a 998 bp product which was then gel purified. Each product was then used in an overlap PCR with CL502 and CL276. The 1141 bp product was then digested with *Bgl*II and *Eco*RI and inserted into the *Bam*HI and *Eco*RI sites of pKQ. The GST gene was then amplified from pGEX-T4-1 using oligonucleotides CL192 and CL310 resulting in a 675 bp product which was gel purified and digested with *Bam*HI and *Sal*I and inserted into the *Bam*HI and *Sal*I sites of pIT. Several restriction sites were then removed by Quikchange mutagenesis. One located in the origin of replication, which was removed using CL195 and CL196, one in the kanamycin resistance marker using CL197 and CL198 and another in the intein using CL665 and CL666.

Deleting a triplet nucleotide. To create the randomly TAG codon scanned library, the target plasmid, pIT-GST was used in an in vitro transposon mutagenesis reaction. In a 20 μl reaction, 400 ng of target plasmid was mixed with a 1.3 molar excess of the *Mly*I transposon with pre-cut *Bgl*II ends, prepared as described in 2.2.2. To the

DNA 2 μL of 10X HyperMu reaction buffer (1.5 M potassium acetate, 0.5 M Tris-acetate (pH 7.5), 0.1 M magnesium acetate and 40 mM spermidine) and 1 μL of Mu transposase (Epicenter) was added. The reaction was then incubated at 30 ^{o}C for 4 hours and stopped by adding SDS to a final concentration of 0.1% (w/v) and heating the reaction to 75 ^{o}C for 10 min. The reaction was then placed on ice and 4 μL was transformed into 200 μL of chemically competent *E. coli* and recovered in 1 mL SOC for 1 h at 37 ^{o}C. Freshly transformed *E. coli* were then plated on LB agar supplemented with 50 μg mL^{-1} kanamycin and 10 μg mL^{-1} chloramphenicol and grown at 37 ^{o}C. The next day there were ~17,000 chloramphenicol resistant colonies and it was determined that that the efficiency of the transposition reaction was ~2.5% when compared with plates that only contain kanamycin.

Ten individual colonies were picked and analyzed by restriction digest. It was determined that the reaction was indeed random, however due to the large size of the plasmid a large percentage of the library contained the transposon not in the target gene, GST. To purify the library, all colonies were pooled and plasmid DNA isolated. Members of the library that contain the transposon in GST were isolated by digestion with *Bam*HI and *Sal*I, which gave 4 bands: pIT backbone + transposon, pIT backbone, GST + transposon and GST. The bands for the pIT backbone and GST + transposon were isolated by gel electrophoresis, relegated and transformed into chemically competent *E. coli*, recovered in 4 mL SOC for 1 hour and the plated on LB agar supplemented with 50 μg mL^{-1} kanamycin and 10 μg ml^{-1} chloramphenicol and grown at 37 ^{o}C. This purification step yielded ~30,000 colonies, of which 10 were chosen and analyzed by

restriction digest showing that all members have the transposon in the GST gene. All clones were then pooled and plasmid DNA isolated.

Inserting the TAG-intein-β-lactamase linker. The transposon library was then digested with *Mly*I to remove the transposon and random three nucleotides. The randomly linearized library was then ligated overnight with the second generation reading frame selectable linker (see 2.4.2) that contains the TAG codon. The ligation was performed in a 1:5 library to linker ratio, with a decrease in ATP concentration from 1 mM to 0.5 mM as recommended for blunt ligations. All 20 μL of the ligation was then transformed by adding 800 μL of chemically competent *E. coli* and recovering in 4 mL SOC for 1 hour at 37 ^{o}C. The transformed ligation was then plated on LB agar supplemented with 50 μg mL^{-1} kanamycin and 40 μg mL^{-1} carbenicillin. At this state the plates were grown at 30 ^{o}C, which was found to be a critical step for intein-mediated splicing. The library resulted in 1635 kanamycin and carbenicillin resistant colonies. The library was then pooled and plasmid DNA isolated and the linker removed by *Mly*I digestion to leave behind the TAG codon scar. The digested library was gel-extracted and ligated at low concentrations at room temperature for 2 h. The mutant library ligation was then transformed by mixing 5 μL of the ligation reaction with 200 μL of chemically competent *E. coli*, recovered in 1 mL SOC for 1 h and plated on LB agar supplemented with 50 μg mL^{-1} kanamycin and grown at 37 ^{o}C, which resulted in >40,000 individual colonies. From the library, 48 individual colonies were chosen to be sequenced, of which 10 were chosen as representatives of the library and expressed in the presence of the unnatural amino acid p-benzoylphenylalanine.

Expression and Purification. The ten representative clones were each digested with *Bam*HI and *Sal*I to remove the TAG mutant gene and inserted into the *Bgl*II and *Sal*I sites of the expression plasmid pBADmycHisA (Invitrogen). Each of the ligations was then transformed into chemically competent Gene Hogs that contained the plasmid pSUP-*p*Bpa. All expressions were done on a 50 mL sale in LB media, 50 μg mL^{-1} ampicillin, 30 μg mL^{-1} chloramphenicol and supplemented with 20 mM *p*Bpa (racemic). All individual clones were expressed in both the presence and absence of amino acid. All cultures were grown to OD600 = ~0.5-0.6 then induced with 0.04% arabinose and grown for an additional 6 hours at 37 ^{o}C.

Mutant proteins with or without *p*Bpa were then purified by centrifuging each culture to pellet cells and resuspending in 2 mL of native binding buffer (100 mM HEPES, 10 mM imidazole, 1 mM PMSF). Cells were lysed by sonication (3 min, 10 s on, 10 sec off) and then centrifuged at 13,000 rpm for 20 min to clear the lysate. To the cleared lysate 100 μL of 50% slurry mix of Promega HisLink resin was added. The resin and lysate were incubated on ice for 30 min with mixing. The resin was washed twice with 1.5 mL of wash buffer (100 mM HEPES, 50 mM imidazole) followed by elution with 200 μL of elution buffer (100 mM HEPES, 500 mM imidazole). A fraction of each of the purified mutants there were expressed in the presences of *p*Bpa were then irradiated with 365 nm hand-held UV to promote photo-cross-linking for 15 min.

Chapter 5: Codon Scanning Mutagenesis to identify residues essential for catalysis

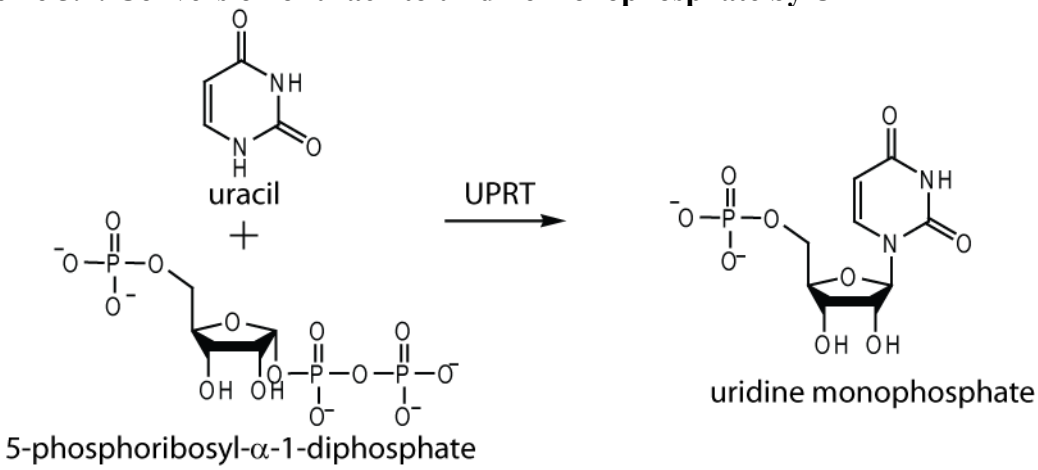
5.1 Introduction

Several unwanted and unexpected mutations were observed after scanning *p*-benzoylphenylalanine through glutathione *S*-transferase using the developed CSM method. It was thought that to address these issues, a well characterized protein that has a simple assay would be ideal. The protein uracil phosphoribosyl transferase (UPRT) was chosen due to the size of the protein, availability of a crystal structure (Lokanath, N.K., unpublished) and the ease of the assay to select for either functional or non-functional mutants. To prove the versatility of the developed CSM method, a linker carrying the alanine codon GCG will be used to generate a library of UPRT having randomly placed in-frame single GCG codon mutations. In many examples of determining active site residues, alanine is a common choice, due to the size and un-reactivity of the side chain. Previous work on introducing a mutation in the active site of UPRT resulted in a variant with a decreased activity. It was expected that the results of an alanine scan on UPRT, followed by a high-throughput assay and sequencing would show that any residue either located in the active site or essential for folding would decrease the catalytic activity.

UPRT is involved in the salvage pathway for nucleotide synthesis. All phosphoribosyltransferases catalyze the transfer of the 5-phosphribosyl from 5'-phosphoribosyl-α-1-diphosphate to a nitrogenous base. The salvage pathway for nucleotide biosynthesis recycles free bases from the breakdown of nucleic acids. UPRT transfers the 5-phosphoribosyl moiety to uracil forming uridine monophosphate, which

then is converted to dUMP. dUMP is the precursor to dTMP which is catalyzed by thymidylate synthase.

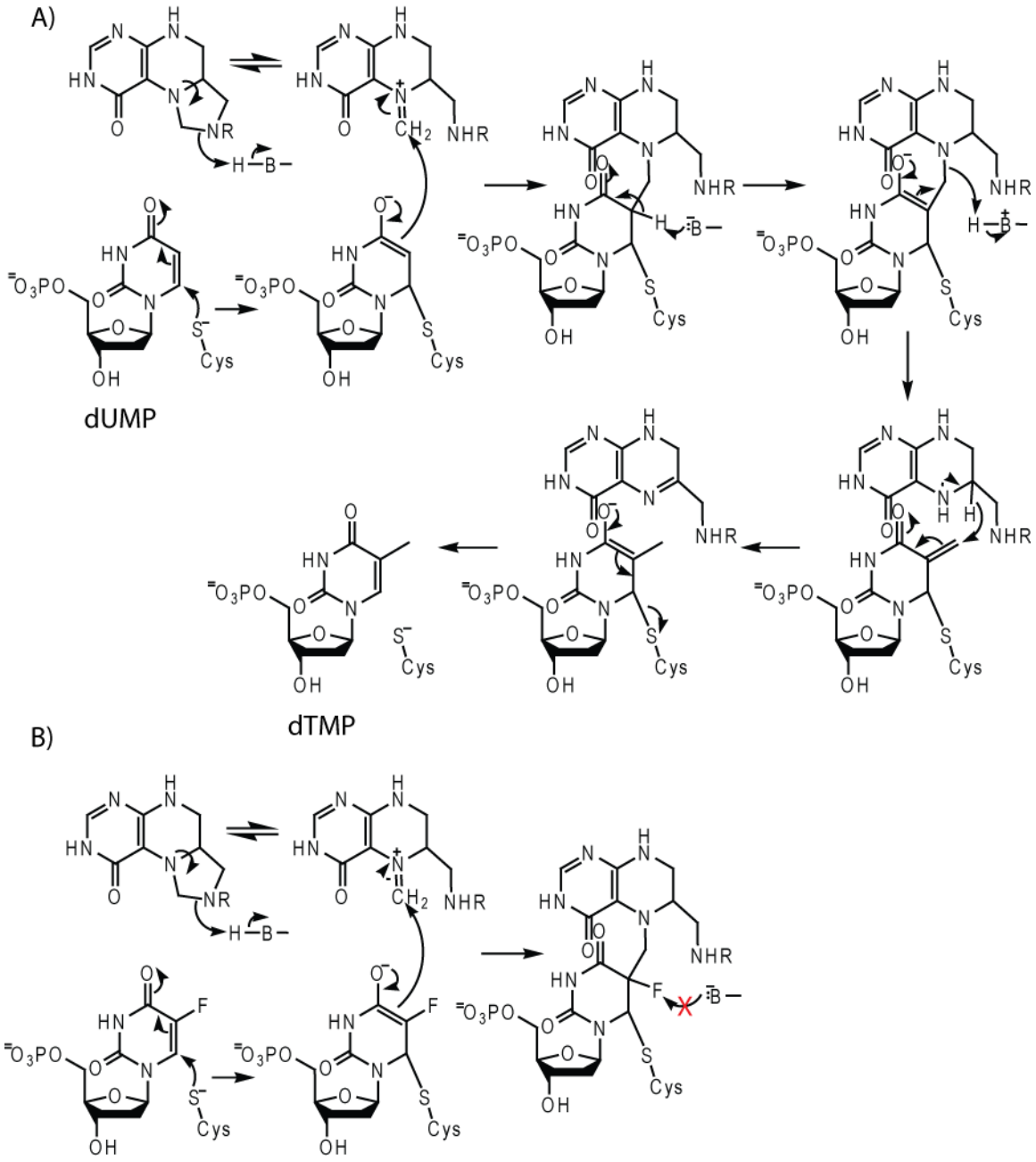
Scheme 5.1: Conversion of uracil to uridine monophosphate by UPRT



To assay the mutant UPRT the pro-toxin 5-flurouracil is used. With functional UPRTase, 5-flurouracil is converted to 5-fluoro-dUMP, which is a substrate for thymidylate synthase. 5-fluoro-dUMP inhibits and inactivates thymidylate synthase by forming covalent complex to both the enzyme and tetrahydrofolate which would inhibit thymine synthesis and result in cell death. Since UPRT is endogenous to *E. coli* a strain that has a deleted open reading frame of genomic UPRT (GH371) is used.

Alanine mutations were introduced into *E. coli* UPRT using CSM. The mutant library was transformed into the GH371 strain of *E. coli*. When plated in the presence of 5-flurouracil any mutants that are active should not be able to survive. Those which are able to survive would have either a mutation that does not allow for proper folding or a mutation in the active site that would not allow for the conversion of 5-flurouracil to 5-fluro-dUMP, the thymidylate synthase inhibitor.

Scheme 5.2: Conversion of dUMP to dTMP.



A) Conversion of dUMP to dTMP by thymidylate synthase. B) 5-fluoro-dUMP inhibits is an irreversible inhibitor of thymydylate synthase by remaining covalently bound to both the Cys residue in the active site of ThyA as well as tetrahyrdofolate.

5.2 Results and Discussion

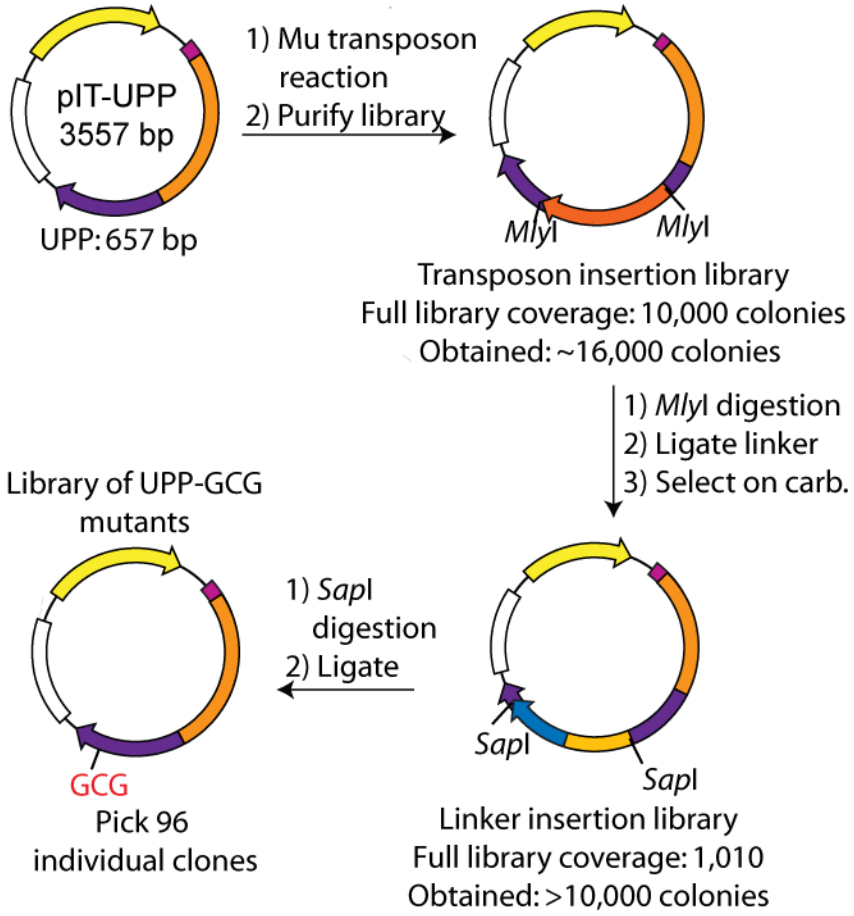


Figure 5.1: Process of using CSM to create random GCG mutations in UPP.
The target plasmid pIT-UPP (purple) is used in a transposition reaction where ~16,000 colonies were obtained where 10,000 were required for 99% library coverage. Removal of the transposon and insertion of the linker followed by selection on carbenicillin yielded greater then 10 times the number required for 99% confidence that all possible mutations were created. After the selection for reading the linker was removed and 96 individual clones containing a GCG codon mutation were picked, assayed for function and sequenced.

5.2.1 Addressing the unexpected mutations

Use of a high fidelity *Mly*I. While there were many unexpected mutations that were observed all but one example were due to the mis-cutting of *Mly*I. Suppliers of restriction enzymes are now offering high-fidelity versions of many of the enzymes. The fast-digest *Mly*I (*Sch*I) from Fermentas claims to have buffer optimization that results in cleavage at the correct site. After sequencing individual clones that have been digested

with *Mly*I, it was found that both the addition and deletion of additional nucleotides was no longer a problem and that as expected only three nucleotides were removed. Using a high fidelity enzyme eliminated the unexpected mutations that were observed previously in Figure 4.5 a-h.

Redesigning the linker. To address the problem were the last nucleobase of the mutational codon was missing, the linker was redesigned such that the blunt ends are generated by releasing the linker from a plasmid by *Mly*I digestion. This ensures that the linker is blunt, contains all nucleobases for the mutational codon as well has being phosphorylated. Since *Mly*I is needed to release the linker from a plasmid used for replication, a different restriction endonuclease is needed to leave behind the mutational codon in the target gene. Like *Mly*I the restriction enzyme used to release the linker and leave the mutation would need to be a type II restriction endonuclease. It would have been ideal to have an enzyme that cuts blunt, however *Mly*I is the only known enzyme that has this activity. Other type IIS endonucleases cut out side the recognition sequence but will leave overhangs, which would then require another step in the process to blunt the ends. The appeal of *Sap*I, is that the enzyme leaves a three base pair overhang. The new linker was designed so that on both ends there is the mutational codon, upon digestion with *Sap*I, the two three base pair sticky ends will be complementary to each other and will be the mutational codon. This method also elevates the possibility of *Sap*I mis-cutting and removing extra nucleotides, if the two ends are not identical then a ligation event will not take place. The limitation of the *Sap*I linker is the ability to scan the amber codon TAG throughout a gene is slightly more complicated and would require that the reading frame selection is done either in a suppressor strain or supplemented with

a plasmid that has the orthogonal tRNA-synthetase pair but incorporates an endogenous amino acid.

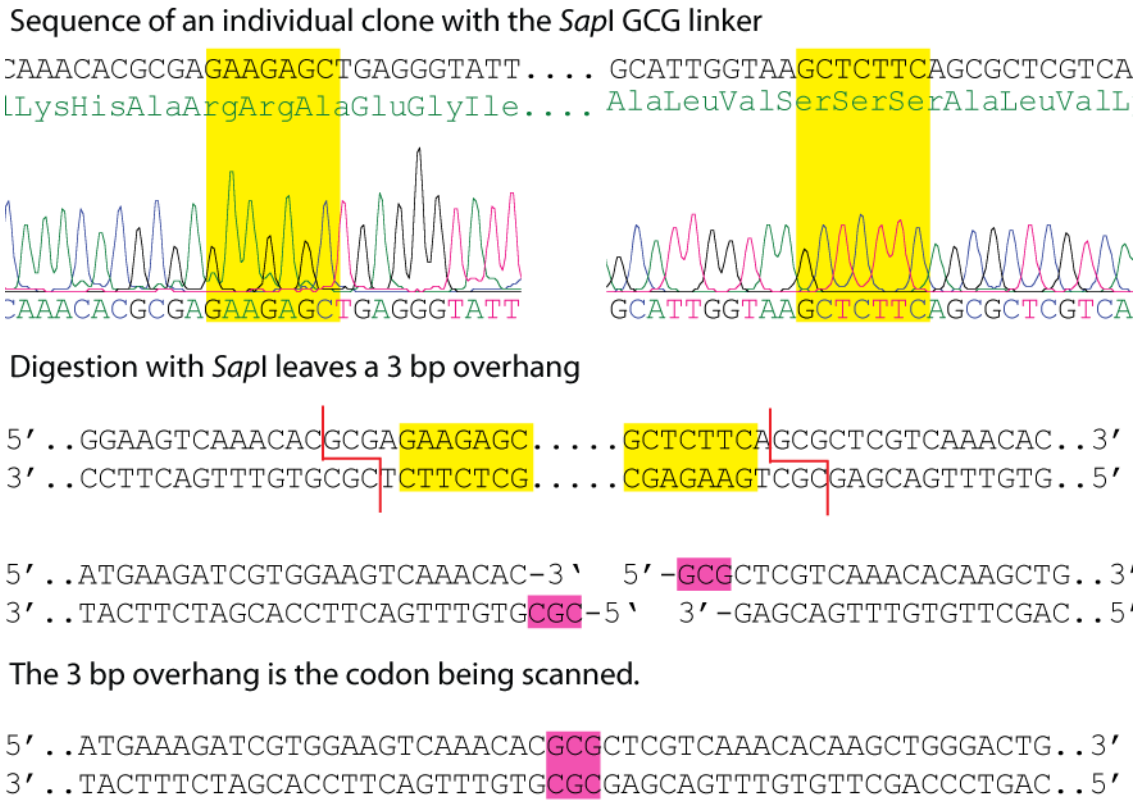


Figure 5.2: Reading frame linker with *Sap*I sites.
Sequencing of an individual clone shows that it is inserted in-frame and has the codon GCG on either end. Digestion of SapI leaves a three base pair overhang. The overhang is the mutational codon, GCG.

5.2.2 Introducing random GCG mutations into the gene encoding UPRT

CSM was performed as previously described in Chapter 4, using the redesigned SapI linker. The reaction with the *Mly*I transposon gave >16,000 clones, consistent with the TAG-GST library. Through each step of the process, linker insertion and removal, the number of clones required to ensure that every position has been targeted was obtained. The linker insertion resulted in thousands of colonies and removal to generate the GCG mutation gave ~4000 colonies, where for each step ~1000 colonies are ideal. 96 mutants

were chosen to be analyzed for functional UPRT in the presence of 5-FU. The results of the selection can be seen in Figure 5.2 and Table 5.1.

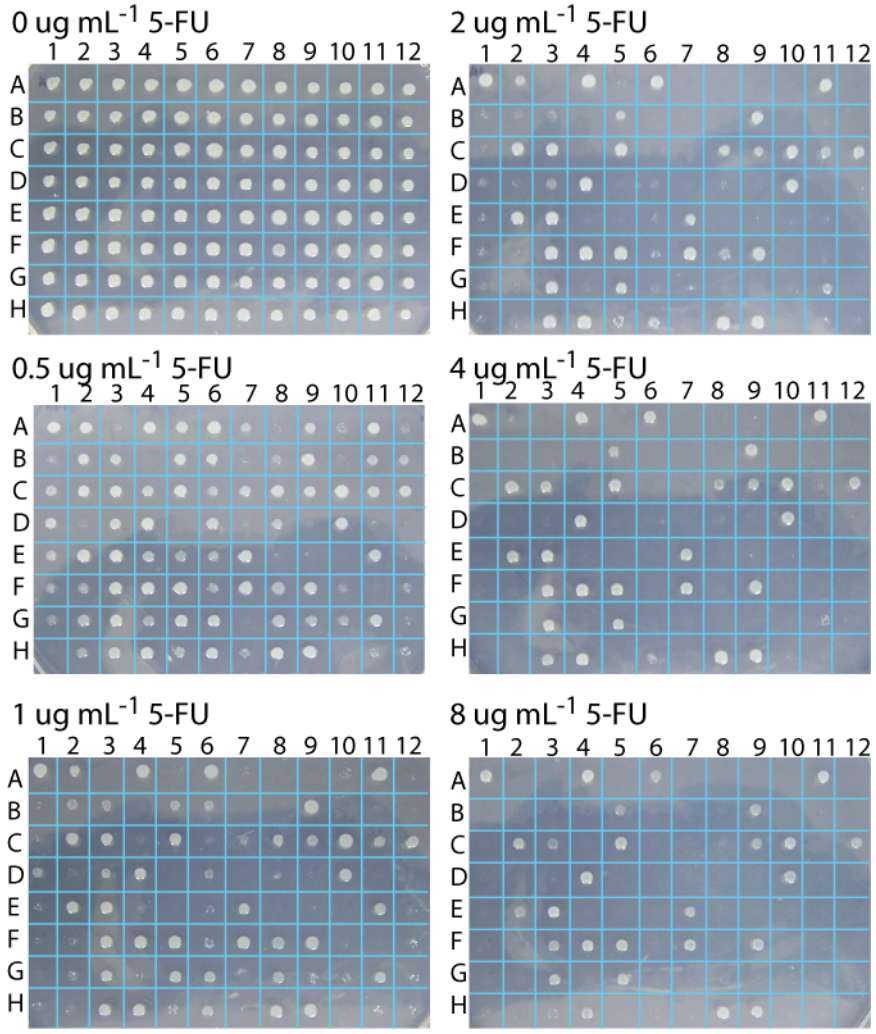


Figure 5.3: Assay of Ala UPRT mutants in the presence of 5 FU.
Those which survive on 5-FU have an Ala mutation that has altered the active site or overall protein structure.

Table 5.1: Survival of alanine UPRT mutants on 5-FU

	Survival on 5-FU ug mL^{-1}					
	0	0.5	1	2	4	8
Pro18						
A12	+	-	-	-	-	-
B12	+	+ / -	-	-	-	-
C4	+	+	-	-	-	-
D6	+	+	- /+	-	-	-
E1	+	+ / -	-	-	-	-
E4	+	+	-	-	-	-
G8	+	+	+ / -	-	-	-
Leu19						
B5	+	+	- / +	- / +	- / +	- / +
C9	+	+	+	+	+ / -	+ / -
C10	+	+	+	+	+	+
E7	+	+	+	+	+	+ / -
G3	+	+	+	+	+	+
G5	+	+	+	+	+	+
Leu 24						
B3	+	+	- / +	-	-	-
Gly 25						
H6	+	+	+	- / +	-	-
Arg28						
C2	+	+	+	+	+	+
D10	+	+	+	+	+	+
Gln30						
H5	+	+ / -	-	-	-	-
H7	+	- /+	-	-	-	-
Thr 34						
D2	+	-	-	-	-	-
Gly45						
H12	+	-	-	-	-	-
Pro84						
F8	+	+	+	+ / -	-	-
G6	+	+	+	-	-	-
Leu97						
F4	+	+	+	+	+	+
Ser102						
B1	+	-	-	-	-	-
Arg112						
B9	+	+	+	+	+	+
C3	+	+	+	+	+	- / +
E2	+	+	+	+	+	+ / -
F3	+	+	+	+	+	+ / -
H4	+	+	+	+	+	+
Leu117						
B2	+	+	- / +	-	-	-
B6	+	+	- / +	-	-	-

	Survival on 5-FU ug mL^{-1}					
	0	0.5	1	2	4	8
Leu119						
A6	+	+	+	+	+	+ / -
A8	+	-	-	-	-	-
A10	+	-	-	-	-	-
B11	+	+ / -	-	-	-	-
C1	+	+	-	-	-	-
C6	+	+ / -	-	-	-	-
D9	+	-	-	-	-	-
E5	+	+	-	-	-	-
E6	+	+ / -	-	-	-	-
F2	+	+ / -	-	-	-	-
F12	+	-	-	-	-	-
G12	+	-	-	-	-	-
H3	+	+	+	+	+	-
H9	+	+	+	+	+	+
Pro121						
D7	+	-	-	-	-	-
Pro140						
A9	+	+	-	-	-	-
C11	+	+	+	+	-	-
G11	+	+	+	- / +	-	-
Leu142						
C5	+	+	+	+	+	+
C12	+	+	+	+	+	+
F5	+	+	+	+	+	+
F7	+	+	+	+	+	+
F9	+	+	+	+	+	+
Pro172						
C8	+	+	+	+	+ / -	-
D8	+	+ / -	-	-	-	-
Pro183						
G2	+	+	-	-	-	-
Ala190						
A7	+	+ / -	-	-	-	-
D11	+	-	-	-	-	-
Gln194						
A2	+	+	+	+/ -	-	-
F10	+	-	-	-	-	-
Pro204						
B8	+	- / +	-	-	-	-
D3	+	+	+	-	-	-
W.T.						
E12	+	-	-	-	-	-
F6	+	+ / -	-	-	-	-
G10	+	+ / -	-	-	-	-
H2	+	+	-	-	-	-

+ growth, - no growth, -/+ slight growth, +/- slightly more growth

5.2.3 Analysis of alanine mutations

Sequencing of the 96 alanine scanned UPRT clones confirmed that the undesired mutations that were observed in the GST library were caused by both *Mly*I mis-cutting and the purity of the oligonucleotides. Of the 96 individual clones sent for sequencing only 27 were unreadable. Of the 69, only 1 was missing the last G of the GCG codon. There were 22 different mutations found out of 64 clones, the other 4 clones were wild-type UPRT, there are nine GCG codons in the wild-type sequence. The mutations were then corresponded to the ability to survive on 5-FU and location in the protein crystal structure.

Majority of the detrimental alanine mutations make sense based on the location within the crystal structure. An interesting result is the Pro131Ala, in previous studies it was determined that this residue was necessary for enzymatic activity. When Pro131 was mutated to Asp there was a decrease in activity by 50-60 fold.[2] However in the assay performed on the randomly selected mutants, one of which has the catalytic proline mutated to alanine and an assay on 5-FU confirms that this mutant is still functional. This may also explain why there were not many detrimental mutations found in the library. Even if small amounts of 5-FU is being converted to 5-F-dUMP that is still enough to bind and inhibit thymidylate synthase. Of the clones that were able to survive on high concentrations of 5-FU, two were arginines, R28A and R112A. Each of these could be responsible for multiple hydrogen bond interactions and would contribute the overall tertiary structure. The other three were leucine mutations, L19A, L97A, L142A. L142A is located in the active site of UPRT, the others are located in hydrophobic regions, L19A may be interrupting interactions between the two monomers, shown in purple and yellow in Figure 5.2. The decrease in size of the side chain between Leu and Ala may allow for

water molecules to be present disrupting the other structure. It was surprising to find that all of the clones with a Pro to Ala were unable to survive on concentrations greater then 1 μg mL^{-1} 5-FU. One would have thought that these mutations would have been detrimental to protein folding, rendering the enzyme inactive. However, to conclusively say that particular residues are essential for enzyme activity, further purification and assays would need to be conducted. The 5-FU assay that was performed, Figure 5.1, appears to not be consistent for all mutants. For instance the P119A variant was found 14 times, see Table 5.1, three of which survive up to 8 μg mL^{-1} 5-FU and the others between 0 and 0.5 μg mL^{-1} 5-FU. It is possible that those mutants which can survive on drastically different amounts of 5-FU, are producing protein at different rates. To determine if this is the case, the mutants should be expressed from an inducible promoter and measure the amount of protein produced by SDS-PAGE. Interestingly, Q194A, had little effect on UPRT, whereas arginine mutations seem to inactive UPRT.

5.3 Conclusions

In conclusion the changes made to CSM, using a high fidelity *Mly*I and redesigning the linker appear to have minimized unexpected mutations. Ideally further experiments need to be done on mutants obtained from CSM. A similar experiment that was performed by Jensen that identified the P131 residue as being essential in activity[2] could be done all the mutants obtained that were able to survive on high concentrations of 5-FU. However, we were more interested in optimizing the method then obtaining data on UPRT. This proof of concept experiment has also proved that CSM will be useful in obtaining preliminary enzymatic data. A slight disadvantage to this developed method is that it is most effective if there is a high-throughput screen for activity.

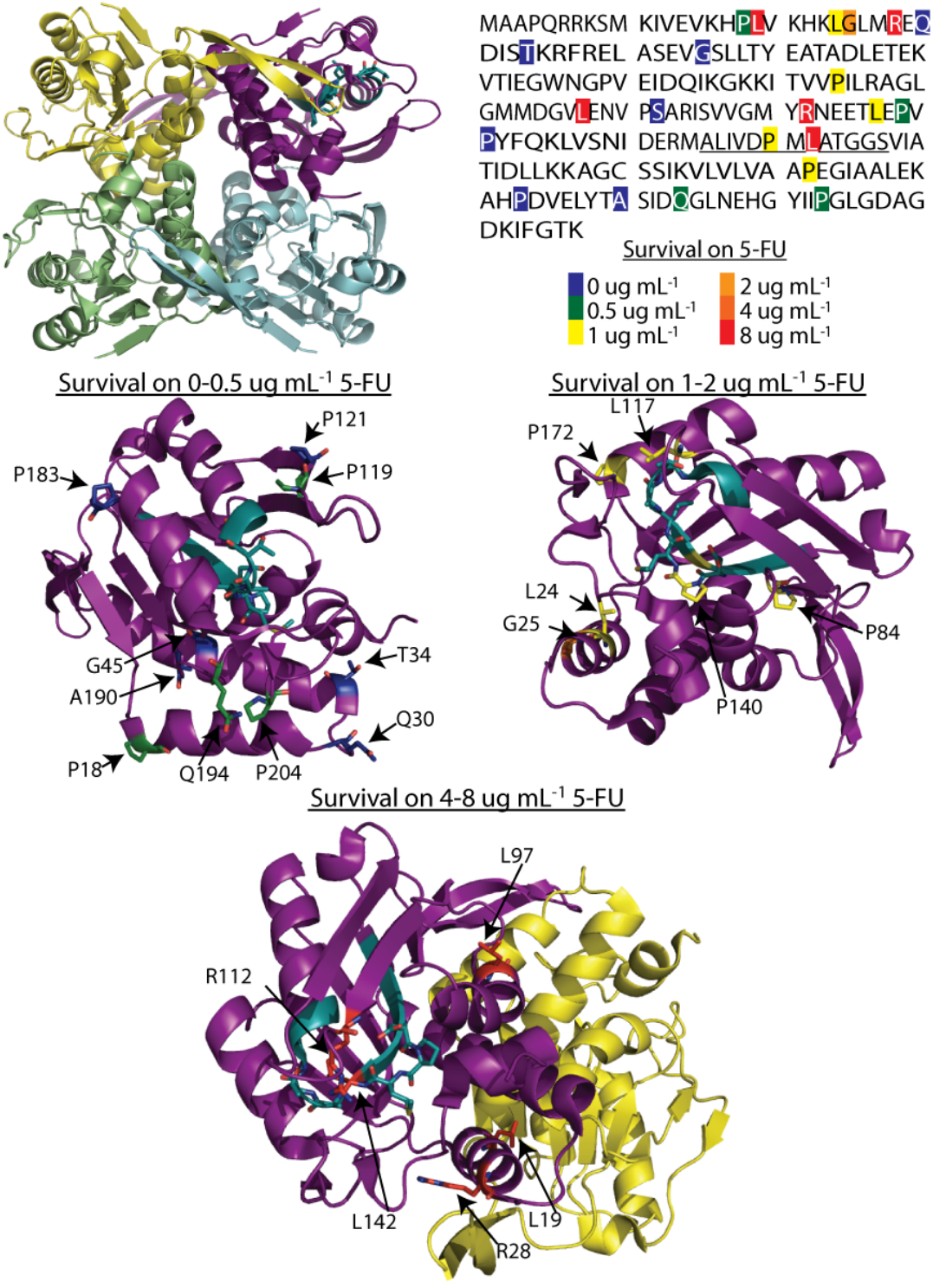


Figure 5.4: Location of Ala mutants in UPRT.
Highlighted in teal is the active site residues of UPRT. Top right panel highlights the position of mutation and ability to survive on 5-FU. PDB structure 2EHJ.

5.4 Materials and Methods

Materials. All restriction enzymes, Phusion Polymerase and T4 DNA ligase were purchased from either New England Biolabs (Ipswich, MA) or Fermentas (Glen Burnie, MD). MuA transposase was purchased from Finnzymes (Finland). Oligonucleotides were synthesized by both Integrated DNA Technologies, Inc. (Coralville, IA) and Invitrogen (Carlsbad, CA). The plasmid pInSALect was obtained from Stefan Lutz, Emory University. Plasmids pGEX-T4-1 and pBADmycHisA were purchased from GE HealthCare Life Sciences and Invitrogen. DNA sequencing was done by University of Michigan DNA Sequencing Core. All media and antibiotics were molecular biology grade. Chemical competent Gene Hogs were used in all experiments.

General Procedures. Plasmid DNA was isolated from *E. coli* using Fermentas GeneJet Kit, Zymo Research Zyppy Plasmid Miniprep kit, or Qiagen Plasmid Midi prep kit according to the manufactures recommendations. All PCR reactions were carried out using a PTC-200 Pelteir Thermo cycler. PCR products were purified using Qiagen PB buffer. Digested DNA was purified by running DNA on a 1% agarose TAE gel, bands excised and DNA purified using either Qiagen QG buffer or via electroelution followed by isopropanol precipitation.

Table 5.2: Oligonucleotides used in CSM to identify critical residues in UPRT

CL1008	AAAGGATCCCATATGAACAGTTTTGGCAACTTG
CL834	AAAAGAATTCTGCT TTTCTTCGCGAATTAATTCC
CL761	CTGGATCCGCTGCCCCTCAACGCAGAAAG
CL196	TACCGGGTTGGATTCAAGACGATA
CL1047	AAGAGTCGAACAGCGAGAAGAGCTGAGGGTATTCGC
CL1048	AAGAGTCTGCAACGCTGAAGAGCTTACCAATGCTT
CL283	CACGACGGGGAGCCAGGCAACTAT
CL284	ATAGTTGCCTGGCTCCCCGTCGTG

pIT-upp construction. In the early stages of development of the CSM method, it was first thought that the transposon reaction could be performed, followed by ligation of the linker and would result in only mutations in the gene of interest. However, in the first scan of GST, it was found that the intein was still functional if inserted into the N-terminus of the VMA intein near the fusion. This led to the step of purifying the library by separating the members of the library that have the transposon inserted into the target gene from those where the transposon inserted else ware in the plasmid. To eliminate the purification step the N-terminus of the intein was truncated by using an individual clone containing a linker in the upp sequence and by inverse PCR removing 4, 8, 12, and 16 amino acids. It was found that all were functional when plated on carbenicillin. It was unclear if it was due to where the linker was inserted into the gene or if the truncated intein where 16 amino acids was removed was actually functional. To verify if the intein can in fact be truncated at the fusion between the N terminus and the target gene, experiments would need to be done where an insoluble protein is inserted. If the splicing mechanism is not occurring then the *E. coli* would be sensitive in the presence of carbenicillin.

When constructing the target plasmid for UPRT, only 4 amino acids were truncated. This was done by amplifying the pIT portion from pIT-GST using CL1008 and CL834, the 2899 bp product was purified. The gene encoding UPRT was amplified from pKQ*upp* with CL761 and CL196 this resulted in a 1410 bp product which was then purified. Both pIT and *upp* were digested with *Bam*HI and *Eco*RI, the upp digestion of the *upp* amplified product resulted in two bands, *Eco*RI was a restriction site that was in pKQ*upp*. After successful ligation of the two pieces, the plasmid was transformed into

GH371 a *upp* knock out strain obtained from Jason Chin (MRC). In the presence of 1 μg mL^{-1} 5-fluorouracil it was found that UPRT, although fused to the VMA intein was still functional. This confirmed that following the scan, subcloning of the mutant genes would not be necessary.

***Sap*I-GCG-intein-β-lactamase linker construction.** The C-terminus and β-lactamase were amplified from pIntSALect using CL1047 and CL1048 and Phusion polymerase. The 1391 bp product was purified and ligated into the *Eco*RV site of pLac-upp. This plasmid was created by removing the N-terminus of the intein from pIT-upp and is used a general plasmid to clone blunt ended fragments. Selection of plasmids containing the insert are screened by transforming into GH371 and growing in the presences of 5-flurouracil. After successful ligation of the linker, the *Mly*I site in the β-lactamase gene was removed using CL283 and CL284. The new linker is then removed from the plasmid by *Mly*I digestion, the 1391 bp band for the linker is purified by electroelution, followed by ethanol precipitation. This linker is now blunt and phosphorylated and can be ligated into a library.

Deleting a triplet nucleotide. To delete a triplet nucleotide with the *Mly*I transposon, reaction conditions are identical to that that was previously described in Chapter 4. The only difference is that pIT-upp was used as the target plasmid. Removal of the transposon was performed with FD-*Mly*I, for 5 min.

Inserting the *Sap*I-GCG-intein-β-lactamase linker. The ligation of the linker with the library of linearized vector was performed as described in Chapter 4. Rather then removing the linker with *Mly*I, *Sap*I is used to leave behind the GCG codon scar. The digested library was gel-extracted and ligated at low concentrations at room temperature

for 2 h. The mutant library ligation was then transformed by mixing 5 μL of the ligation reaction with 200 μL of chemically competent GH371 (upp deletion strain), recovered in 1 mL SOC for 1 h and plated on LB agar supplemented with 50 μg mL^{-1} kanamycin and grown at 37 ^{o}C, which resulted in ~1000 individual colonies. From the library, 96 individual colonies were picked into LB medium supplemented with 50 μg mL^{-1} kanamycin in a 96 well block. The block was incubated with shaking at 37 ^{o}C until grown to saturation.

Sequencing and selection of alanine mutants. The 96 well block was then replica plated on LB agar and M9 agar containing 0, 0,5, 1, 2, 4, and 8 μg mL^{-1} 5-fluorouracil. Colony PCR was done to show the plasmids were correct by amplifying the gene encoding UPRT with CL311 and CL196, the product of the PCR should be ~2000bp and show that none of the plasmids have the linker still inserted. Plasmid DNA was then isolated by miniprepping the remainder of the liquid culture. All 96 individual plasmids were then sequenced at Yale Sequencing. Replica plates of the 96 well block were grown at both 30 ^{o}C and 37 ^{o}C, although in each condition the growth was identical. Any clones that are able to survive on 5-flurouracil would have a mutation that inactivates the enzyme.

Chapter 6: Conclusions, Optimization and Future Applications

6.1 Summary of Codon Scanning Mutagenesis

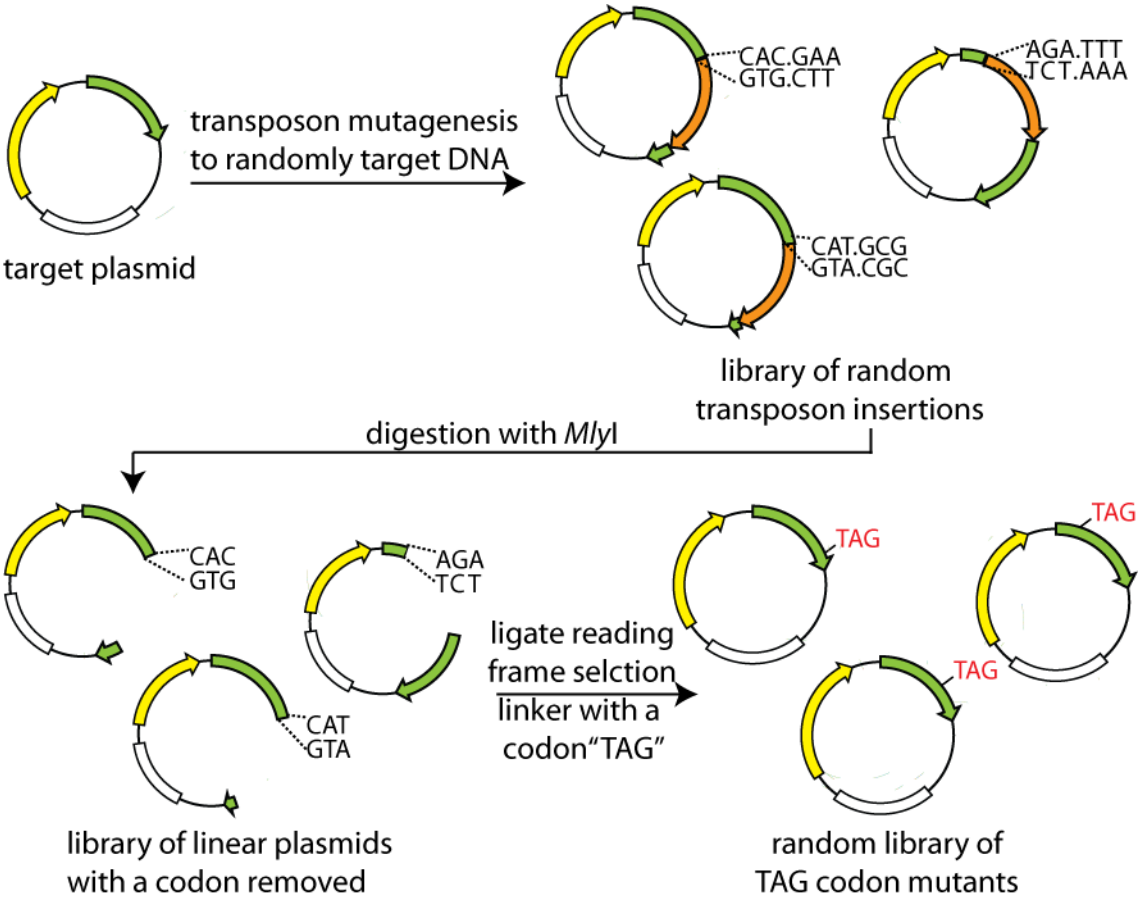


Figure 6.1: Process of using CSM to create a random codon mutations.
CSM requires four steps. 1) Transposon mutagenesis to insert unique restriction sites that allow for the double strand break. 2) Removal of the transposon and three nucleotides. 3) Ligation of a reading frame selectable maker carrying the mutation codon (TAG). 4) Removal of the linker to replace a native codon with the codon of choice, in this case a TAG.

To date the developed CSM method is the only method that allows for generating defined, random, in-frame codon mutations. Two selection systems have been developed to obtain the in-frame mutations. It was found that the most efficient way of selecting for the correct reading frame is incorporating an intein into the selection limiting any protein folding problems that could be associated with the fusion of the target protein to the

selectable marker. The use of an intein allows for the selectable protein to be identical regardless of where the linker has inserted. The first selection method was based off of a non- biased reading frame selection plasmid that separates a head and tail region by the VMA intein. The second selection method was a thymidylate synthase selection and uses a self cleavage intein. This second system does not have a leader sequence that is required for protein function and this allows for possible out-of-frame mutations to occur, although very minimal, due to possible ribosome binding sites within the target gene.

Each selection system has its advantages and disadvantages. The β-lactamase system would ensure that all mutations are in-frame by eliminating the synthesis of functional protein from an internal ribosome binding site. The disadvantage to this method is that some target proteins may not be functional when fused to the N terminus of the intein and therefore would require sub-cloning of the scanned mutant gene. Within the pIT plasmid two *Nde*I sites were incorporated on either end of the intein to allow for easy removal, although this adds an additional step to the method an intramolecular ligation is far more efficient then an intermolecular ligation. The advantage of the thymidylate synthetase selection is that scanning and assaying the mutant proteins can be done in all one plasmid but this library could be contaminated by out-of-frame mutations.

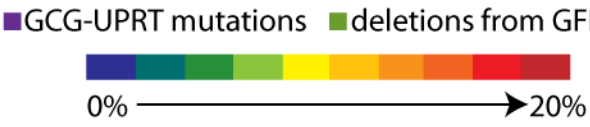
During the development of CSM, a transposon was developed that allows for random in-frame codon deletions. This deletion transposon can be used to quickly and efficiently make protein truncations. While this has been developed using the β-lactamase system and requires the pIT plasmid, a transposon that uses the thymidylate selection can easily be created and is expected to work in the same fashion.

6.1.1 Bias observed

In all the sequencing that has been performed on the mutants obtained from using the MuA transposase to introduce mutations it was found that mainly codons that have a high GC content were mutated. It was initially thought that it was due to the reading frame selection. However, after analyzing sequencing data that was obtained in the Cropp lab by Jia Liu, where three nucleotides were randomly deleted without a reading frame selection, it was found that the three nucleotides deleted were mainly GC rich. Shown in Table 6.1 is the codon that was targeted and the $N_2N_3N_4$ from the duplication of 5 nucleotides during the transposition reaction. Results are from 118 individual clones.

Table 6.1: Frequency of mutations observed from CSM and random deletions.

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser 2.5%	UAA Stop 3.4%	UGA Stop 6.8%
UUG Leu 0.8%	UCG Ser	UAG Stop	UGG Trp 4.2%
CUU Leu	CCU Pro	CAU His	CGU Arg 5.9%
CUC Leu 5.1%	CCC Pro	CAC His	CGC Arg 0.8%
CUA Leu	CCA Pro 12.7%	CAA Gln 1.7%	CGA Arg
CUG Leu 8.5%	CCG Pro 18.6%	CAG Gln 1.7%	CGG Arg 0.8%
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr 0.8%	AAC Asn	AGC Ser 0.8%
AUA Ile	ACA Thr 0.8%	AAA Lys	AGA Arg 0.8%
AUG Met	ACG Thr	AAG Lys	AGG Arg 1.7%
GUU Val	GCU Ala 0.8%	GAU Asp	GGU Gly 0.8%
GUC Val	GCC Ala	GAC Asp 0.8%	GGC Gly
GUA Val	GCA Ala 1.7%	GAA Glu 1.7%	GGA Gly 11%
GUG Val	GCG Ala 3.4%	GAG Glu	GGG Gly 0.8%



Literature precedence for a bias in the insertion of the MuA transposon. After sequencing both GCG-UPRT library and deletions from Jia Liu a trend toward GC rich codons was observed. Since the site of mutation is determined by the site of transposon insertion, there must be a bias of the transposase for GC rich regions of DNA.

Although the transposon kits are marketed as being completely random, the removal of $N_2N_3N_4$ from the initial five base pair duplication were analyzed from three recent reports of using the transposon,[29, 64, 66] results shown in Table 6.2. The frequency of mutation observed is out of 238 clones. Naturally, the number of unique insertion sites was increased with the number of clones sequenced. However a bias towards codons with a high GC content was still observed. This bias can easily be seen when comparing the each row. The second row, where each codon begins with C, has a higher frequency of being targeted by the transposon. Both the second and fourth column, where C and G are the second base is also common sites for transposon insertion.

Table 6.2: Frequency of $N_2N_3N_4$ at the five base pair duplication site.

UUU Phe	UCU Ser 1.2%	UAU Tyr	UGU Cys 0.4%
UUC Phe	UCC Ser 0.8%	UAC Tyr 0.4%	UGC Cys 1.6%
UUA Leu 1.2%	UCA Ser 7.5%	UAA Stop 0.8%	UGA Stop 3.4%
UUG Leu 2.1%	UCG Ser 2.9%	UAG Stop	UGG Trp 8.8%
CUU Leu	CCU Pro 1.2%	CAU His 0.4%	CGU Arg 1.3%
CUC Leu 0.8%	CCC Pro 2.1%	CAC His 0.4%	CGC Arg 5.5%
CUA Leu 0.8%	CCA Pro 7.9%	CAA Gln 0.8%	CGA Arg 1.3%
CUG Leu 5.9%	CCG Pro 3.9%	CAG Gln 3.8%	CGG Arg 5.0%
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn 0.4%	AGC Ser
AUA Ile 3.7%	ACA Thr 0.8%	AAA Lys 0.4%	AGA Arg 0.4%
AUG Met 1.7%	ACG Thr 2.1%	AAG Lys 0.8%	AGG Arg 1.2%
GUU Val	GCU Ala	GAU Asp	GGU Gly 1.2%
GUC Val	GCC Ala	GAC Asp	GGC Gly 1.7%
GUA Val 0.4%	GCA Ala 2.1%	GAA Glu 0.8%	GGA Gly 1.7%
GUG Val 0.8%	GCG Ala 4.6%	GAG Glu 0.8%	GGG Gly 1.2%

Triplet Nucleotide Removal TriNex Pentapeptide Scanning

0% → 10%

6.1.2 Deep sequencing to confirm observed bias

Based on the sequencing that was obtained in both scans using CSM as well as those which were performed by Jia Liu which uses the MuA transposase but for a different application, insertions at GC rich areas of DNA is preferred. This was also

observed in mutations that were found in the literature, shown in Tables 6.1 and 6.2. It would be interesting to determine if the AT rich codons are also being mutated but in a low frequency. A deep sequencing experiment[82] could be performed on either of the libraries to validate that there is indeed a bias. Results of the deep sequencing experiment should show that all codons are mutated and the frequency at which the mutation occurs. As stated earlier the bias that we are observing may be that we have not sequenced enough clones.

6.2 Optimization of Codon Scanning Mutagenesis

Incorporating a defined set of multiple adjacent mutations. One limitation of this method is that only one mutation can be made at a time. The method could be adapted to scan a pre-chosen peptide linker, allowing for more then one codon to be mutated but would require that the codons are adjacent to one another. The methodology of deleting greater then 1 codon at a time is currently being developed by Jia Liu. Using the intein-β-lactamase selection system, he generated an asymmetrical transposon that is capable of selecting for the correct reading frame. The difference in the sequence on either end of the transposon allows for EIPCR to be performed to specifically place the restriction site for *Bsg*I. The placement of the *Bsg*I restriction site determines the number of nucleotides that can be removed. A linker segment with a defined number and sequence can then be inserted into the deletion plasmid, similar to a method termed transposon-directed base-exchanged mutagenesis (TDEM),[83] which uses the same concepts as described in Chapter 3.5.

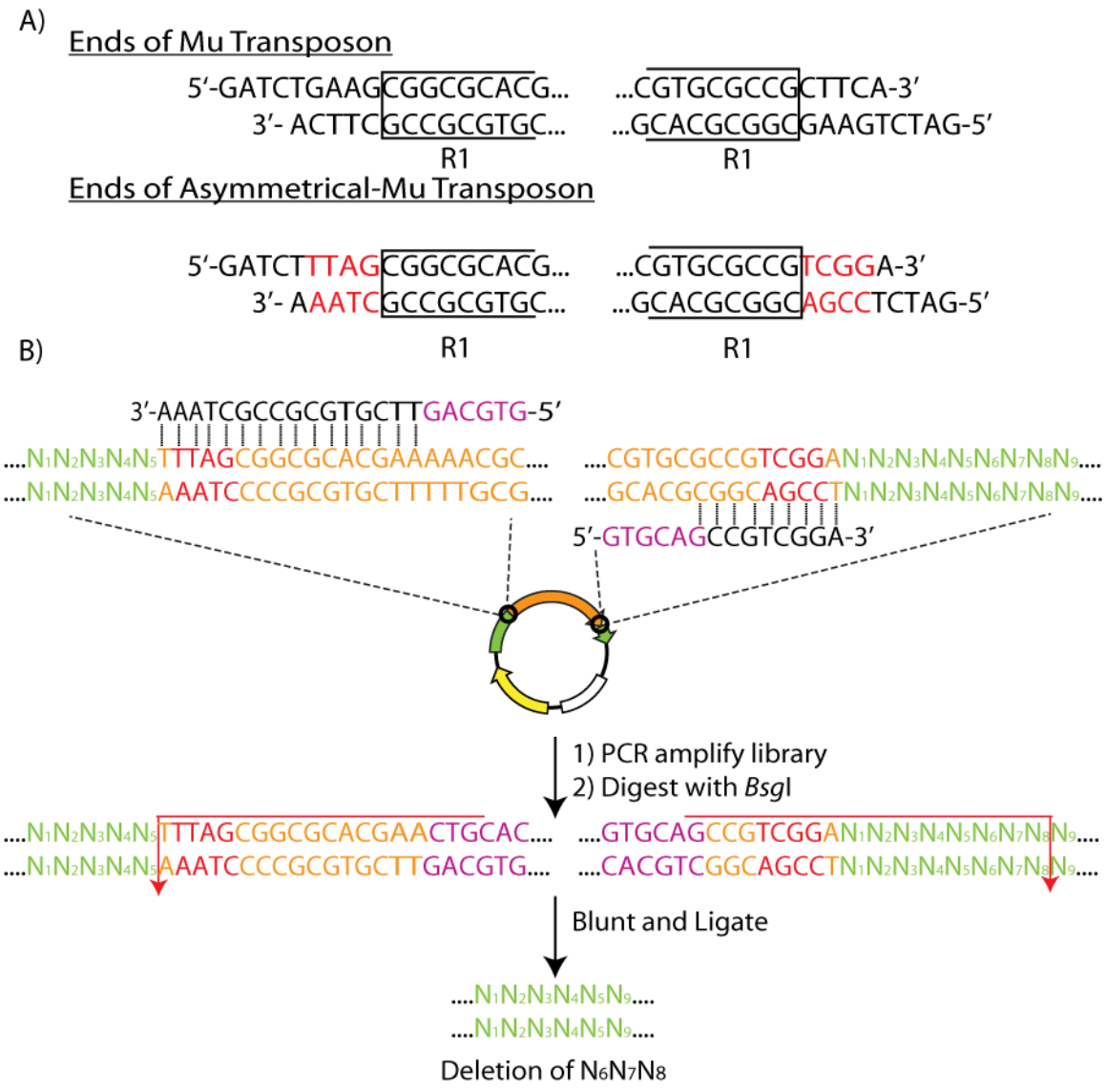


Figure 6.2: Asymmetrical transposon to delete in-frame codons.
A) The difference in the four bases outside of the recognition site are highlighted in red. B) After successful transposition and selection the library is pooled and subjected to amplification. This step is necessary because mutations in the recognition sequence would result in little to no transposition events. The oligonucleotides introduce the *Bsg*I recognition sequence (purple) and when digested there is a net loss of three nucleotides.

A non-linker mediated codon scanning. The ability to use an asymmetric transposon that can select for the correct reading frame can also be adapted to replace a single codon. This modified method would eliminate the need to ligate in a segment of DNA, which can be a limiting step in the process. Work that is being carried out by Jia

Liu is showing to be promising and would greatly enhance the speed at which CSM can be accomplished. Using a non-linker mediated codon scanning method may only require two library transformations and allow for libraries of mutations to be created within one week. The proposed method uses an asymmetrical transposon, one end has the *Mly*I restriction site and the other side the nucleotide sequence from Jia's asymmetrical transposon, see Figure 6.3 A. A transposon reaction would first be performed, transformed and transposon insertions selected for on selectable plates. Since the DNA between the recognition sequences does not effect the transposition reaction, either of the two reading frame selectable systems could be used. Additionally, within the transposon sequence would contain the mutational codon and the restriction sites for *Mly*I and *Bsg*I. After selection, all colonies from the library can be pooled and plasmid DNA isolated. Part of the transposon can be removed by *Mly*I digestion, followed by ligation. This first step would result in deleting $N_2N_3N_4N_5$ from the original gene sequence. Next a PCR reaction can be done on the ligation product, where one oligonucleotide would be an exact complement and the other would introduce an additional *Bsg*I site. The PCR product can then be digested with *Bsg*I to remove $N_1N_2N_3N_4$, blunted and ligated to leave the mutational codon in place of $N_2N_3N_4$. While Figure 6.4 shows the *Mly*I digestion before the PCR amplification step, it is possible to amplify the library first, followed by the *Bsg*I digestion and *Mly*I digestion. It is also shown in Figure 6.4 the mutational codon to be TAG, however just as in the mutagenesis method that was developed in this dissertation any codon of the users choice can be inserted. With this method it is also possible to scan multiple codons, by adjusting the position of the *Bsg*I site.

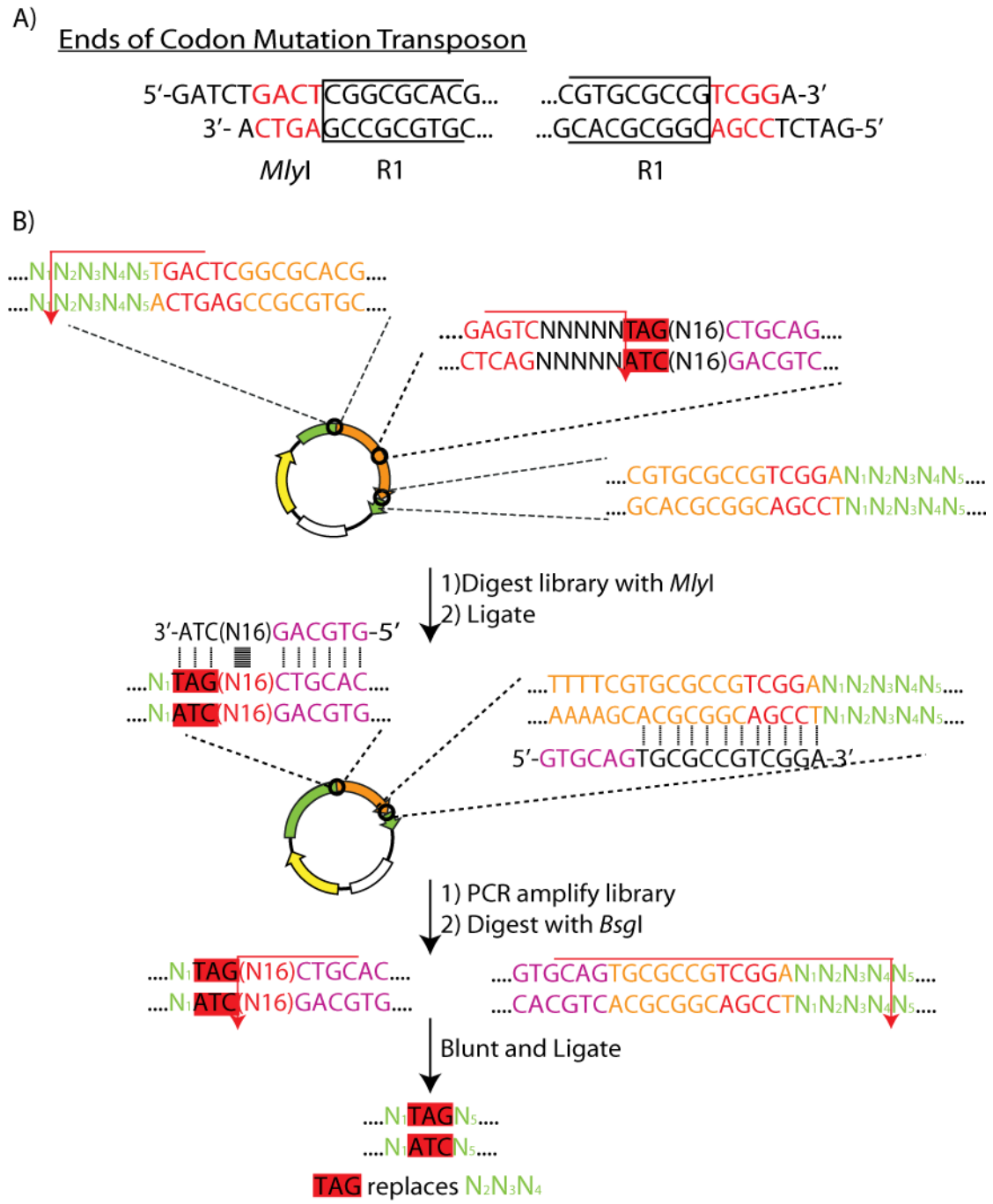


Figure 6.3: CSM without the need for linker ligation

A) Proposed ends of an asymmetrical transposon where one end would have the recognition sequence for *Mly*I. B) After the transposon reaction and selection the library can be pooled. The codon to be mutated is in the transposon itself, with an *Mly*I and *Bsg*I site on either side. Digestion with *Mly*I removes nucleotides up to N_1, ligation would replace $N_2N_3N_4$ with TAG. The library can be amplified to insert a *Bsg*I site, removal of the remainder of the transposon and ligation, results in a replacement of $N_2N_3N_4$.

Change the specificity of the MuA transposase. Engineering the MuA transposase to target A T rich regions of DNA would be a difficult task, since the transposition reaction is done *in vitro*. All mutant transposases would need to be purified separately and reactions done in individual wells with the donor DNA and A T rich target DNA. The five base pair gap would have to be repaired *in vitro* by adding a polymerase and dNTP's after stopping the transposition reaction. After which, the DNA can be amplified by either PCR or rolling circle amplification (RCA).[84] RCA, may produce greater amounts of DNA and the need for specific oligonucleotides is not necessary. If using RCA, an additional digest would be performed using an enzyme that has one restriction site in the target DNA. Either the linear RCA products or the PCR products would be analyzed by gel electrophoresis and in any lane with a band or larger band present would represent a successful transposition event.

Recently, it has been reported that the target site preference of the transposase can be altered by adjusting both the buffer conditions and using a truncated transposase.[85] The authors had hypothesized that by increasing the amount of glycerol or DMSO would alter the target site preference and activity. This was based on previous reports stating that additives affect enzyme activity[86] and can alter DNA binding.[87] The buffer composition that was found to result in the most efficient transposition reaction as well as have a decreased target site preference is as follows; 25 mM Tris-Cl (pH 8.0), 5 mM $MgCl_2$, 0.05% (v/v) Triton X-100, 110 mM NaCl and 10% (v/v) glycerol. The supplied buffers with the transposase do not contain glycerol or DMSO, Epicentre 1X HyperMuA reaction buffer: 150 mM KOAc, 50 mM Tris-OAc (pH 7.5), 10 mM MgOAc, and 4 mM spermidine. It would be interesting to perform simultaneous CSM on the same target,

with the difference being the composition of the buffer used in the transposition reaction. Both libraries can then be sequenced by amplicon sequencing to determine if mutations are more evenly dispersed over the target gene sequence.

6.3 Proposed Future Applications of CSM

The developed CSM method is extremely versatile. The desired amino acid incorporated into the library is user defined. Additionally, the codon that is chosen to create the library can be optimized depending on the organism in which the assay would be carried out. Linkers for all 20 amino acids can easily be produced and only requires that the linker sequence is amplified using an oligonucleotide carrying the codon of choice.

Since, the CSM method allows for the any codon to be randomly distributed throughout any gene. This method should prove to be particularly useful in gaining insight in structure by scanning cysteines and protecting those that are solvent exposed, outlined in Figure 6.4. The CSM method coupled with the technique of labeling cysteines both intra and extra cellular (section 1.1), would be beneficial to understanding membrane bound protein structure. The data obtained from thiol protection assays could then be used to determine where to incorporate an unnatural amino acid, such as *p*-benzoylphenylalanine site-specifically. The mutated protein can then be irradiated with 360 nm light to promote cross-linking, followed by tryptic digests and analyzed by MALDI. This is also an example that could utilize the unique properties of the [D_{11}]-*p*Bpa to analyze data efficiently.

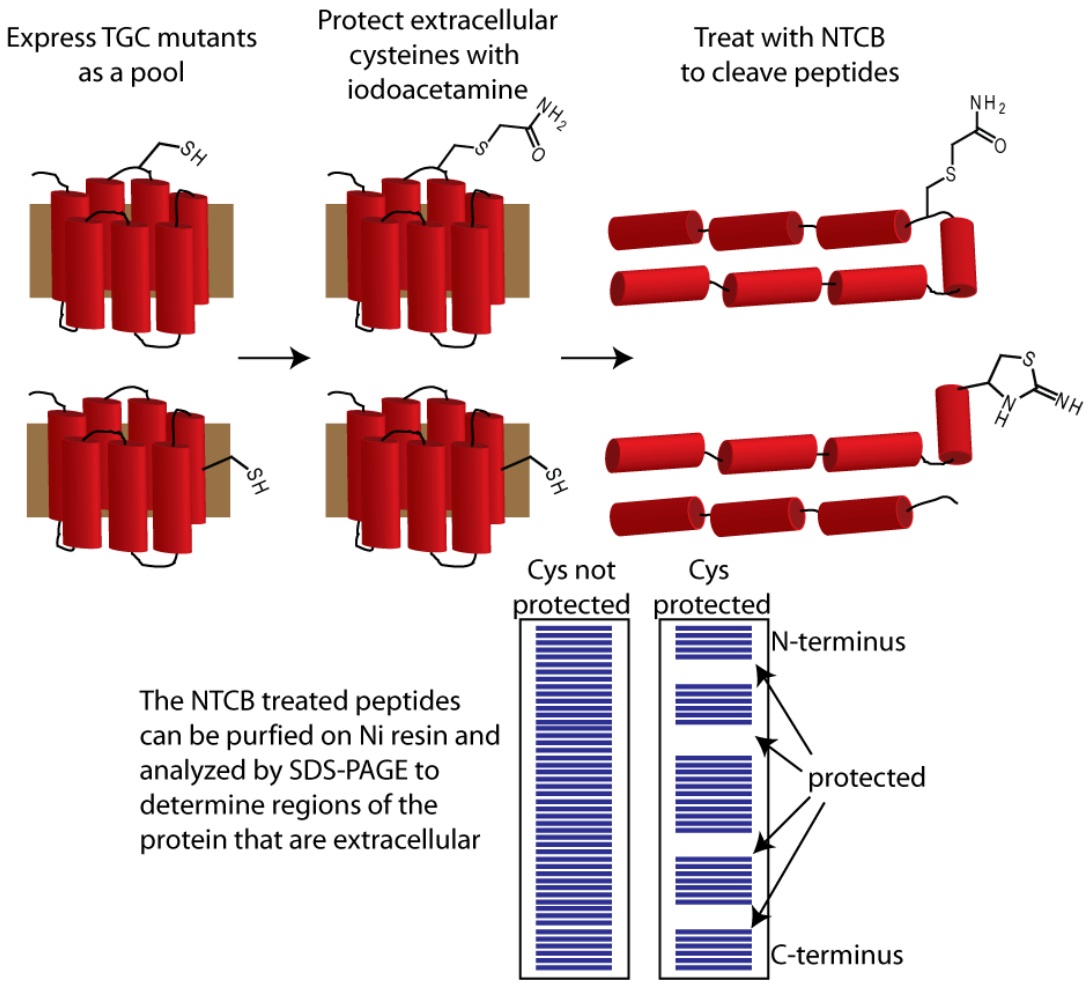


Figure 6.4: Cysteine scan to identify topology of a membrane bound protein.
Using CSM, TGC mutations would be incorporated. The pool of mutations would be expressed with a C-terminal His-tag in one flask. The sample would then be split and one protected with iodoacetamine. Both samples would then be treated with NTCB, purified and analyzed by SDS-PAGE. The non-iodoacetamine treated sample would give rise to multiple bands showing where as the iodoacetamine treated sample would have regions lacking a band, showing that those residues are solvent exposed.

Naturally, if scanning a membrane bound protein, it would be feasible to simply scan the amber codon, TAG and incorporate *p*Bpa in all places. However, as can be seen in with scanning *p*Bpa though GST, there was not incorporation at every position, and this was only on a library of 10 mutants. By first scanning cysteine, non-functional/non-folded members can be eliminated and only a small fraction of the original pool will be assayed with *p*Bpa.

Protein-protein or protein-ligand interactions can also be addressed. This can be done by scanning natural amino acids and using a prokaryote two-hybrid system[88] to look for disruption of interactions between proteins. Or the [D_{11}]-*p*Bpa[51] can be scanned, followed by cross-linking and analysis using MALDI to determine interactions. As stated in the previous section, introducing cysteine residues followed by thiol protection assays would also aid in determining these interactions.

Additionally this can allow for the creation of a novel protein with new activity. A transcriptional regulator protein can be evolved to be a light sensitive repressor by scanning the unnatural amino acid phenyalanine-4'-azobenzene (AzoPhe). The conformation of the double bond can be controlled by light by switching from cis to trans at 420 nm light and back to cis at 334 nm light, shown in Figure 6.5.

Besides for scanning only one codon throughout a gene, multiple mutations can easily be created by repeating the CSM process. The ability to scan multiple mutations would be more feasible using the β-lactamase/ThyA linker, since selections can be done in the same plasmid. Doing multiple rounds of scanning, would introduce specific codon mutations. Further diversity can be introduced by recombining the mutant genes using DNA shuffling or non-homologus recombination.

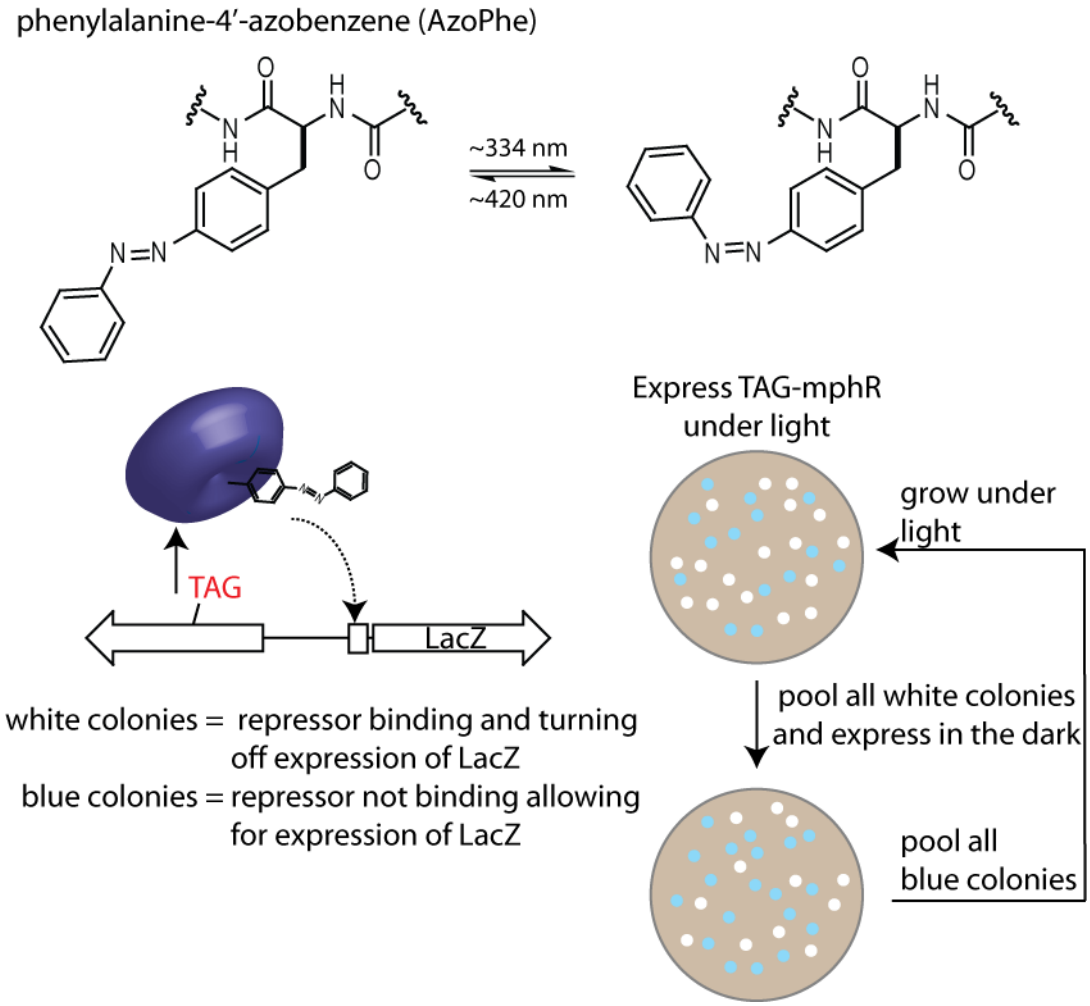


Figure 6.5: Creating a light sensitive genetic switch.
When AzoPhe isomerizes, both the conformation and dipole are changed, this would affect binding of the repressor to DNA. A system where the repressor (blue) controls the expression of a reporter (LacZ) can be constructed to assay the light sensitive genetic switch. The library of TAG-repressor mutants can be expressed under light in the presence of AzoPhe, when assayed in the presence of X-gal, any that are white, have wild-type repressor function. The white colonies can be expressed in the dark, isomerizing the AzoPhe and resulting in blue colonies, showing that there was a change in the binding site. This process can be repeated, each time amplifying the optimal mutation for a light sensitive genetic switch.

6.4 Significance of Codon Scanning Mutagenesis

This is the first example of combining two methodologies typically used in creating random protein diversity for inserting a specific codon mutation at a random place. The fusion of transposon mediated mutagenesis with the ability to select for the correct reading frame has endless possibilities. The only limitations with the method is

that the target DNA does not contain any of the restriction sites necessary to process the mutants, but with the ability to synthesize a gene any silent mutation can be incorporated. While CSM can mutate any position within a protein coding sequence having a high-throughput selection for desired phenotypes of the target protein is desirable. The described method is also not useful in the case where one would want to only create a several site-specific mutants. This method should prove to be the most useful in cases where there is little or no structural data known. As well as, cases where the goal is to create new or enhanced protein function.

The aim of this dissertation was to develop a universal random mutagenesis method that incorporates a user specified mutation, allow for any mutation to be created, regardless of the nucleotide sequence and specifically introduce the mutation in the correct reading frame. This universal method, termed Codon Scanning Mutagenesis utilizes the unique properties of the MuA transposase to insert the type IIs restriction site *Mly*I at multiple non-specified locations within the DNA sequence. Removal of the transposon by digesting with *Mly*I removed a codon from the protein coding sequence. A linker segment that produces a seletable phenotype when inserted in the correct reading frame as well as carries the codon of choice is then inserted. Subsequent removal of the linker will leave the desired mutation without any net addition or deletion of nucleotides.

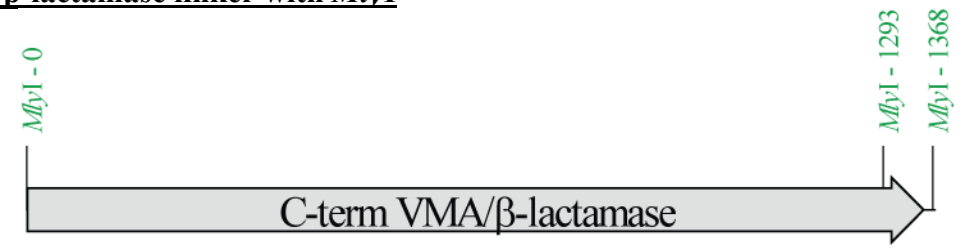
In summary, Codon Scanning Mutagenesis, is the only method to date that allows the user to choose the codon that is to be randomly distributed throughout a protein coding sequence. Further, this is the only protein diversification method that can incorporate a defined number of in-frame mutations, without the use of oligonucleotides.

Appendix: Plasmid Maps and Sequences

***Mly*I-Mu transposon**

GCTTAGATCTGACTCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGCGAAAACGGATCGATCCTA
TCGTCAATTATTACCTCCACGGGGAGAGCCTGAGCAAACTGGCCTCAGGCATTTGAGAAGCACACGGTCAC
ACTGCTTCCGGTAGTCAATAAACCGGTAAACCAGCAATAGACATAAGCGGCTATTTAACGACCCTGCCCTG
AACCGACGACCGGGTCGAATTTGCTTTCGAATTTCTGCCATTCATCCGCTTATTATCACTTATTCAGGCGT
AGCAACCAGGCGTTTAAGGGCACCAATAACTGCCTTAAAAAAATTACGCCCCGCCCTGCCACTCATCGCAG
TACTGTTGTAATTCATTAAGCATTCTGCCGACATGGAAGCCATCACAAACGGCATGATGAACCTGAATCGC
CAGCGGCATCAGCACCTTGTCGCCTTGCGTATAATATTTGCCCATGGTGAAAACGGGGGCGAAGAAGTTGT
CCATATTGGCCACGTTTAAATCAAAACTGGTGAAACTCACCCAGGGATTGGCTGAGACGAAAAACATATTC
TCAATAAACCCTTTAGGGAAATAGGCCAGGTTTTCACCGTAACACGCCACATCTTGCGAATATATGTGTAG
AAACTGCCGGAAATCGTCGTGGTATTCACTCCAGAGCGATGAAAACGTTTCAGTTTGCTCATGGAAAACGG
TGTAACAAGGGTGAACACTATCCCATATCACCAGCTCACCGTCTTTCATTGCCATACGTAATTCCGGATGA
GCATTCATCAGGCGGGCAAGAATCTGAATAAAGGCCGGATAAAACTTGTGCTTATTTTTCTTTACGGTCTT
TAAAAAGGCCGTAATATCCAGCTGAACGGTCTGGTTATAGGTACATTGAGCAACTGACTGAAATGCCTCAA
AATGTTCTTTACGATGCCATTGGGATATATCAACGGTGGTATATCCAGTGATTTTTTTCTCCATTTTAGCT
TCCTTAGCTCCTGAAAATCTCGACAACTCAAAAAATACGCCCGGTAGTGATCTTATTTCATTATGGTGAAA
GTTGGAACCTCTTACGTGCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCCCAGGGCTTCCCGGTATCAA
CAGGGACACCAGGATTTATTTATTCTGCGAAGTGATCTTCCGTCACAGGTATTTATTCGGTCGAAAAGGAT
CGATCCGTTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTCGTGCGCCGAGTCAGATCTAAGC

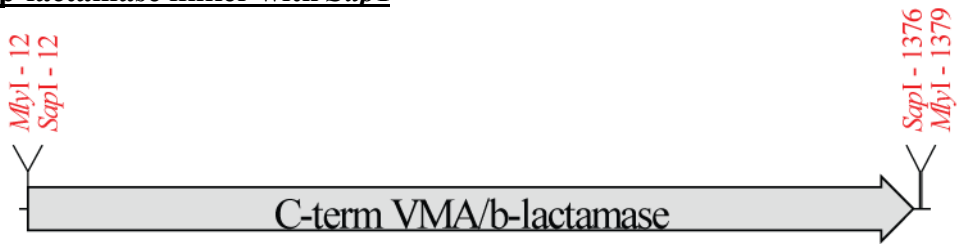
Intein-β-lactamase linker with *Mly*I



GGATCGACTCTCCTGGGTATTCGCAATAATCTTAATACTGAGAATCCATTATGGGACGCTATTGTTGGCTT
AGGATTCTTGAAGGACGGTGTCAAAAATATTCCTTCTTTCTTGTCTACGGACAATATCGGTACTCGTGAAA
CATTTCTTGCTGGTCTAATTGATTCTGATGGCTATGTTACTGATGAGCATGGTATTAAAGCAACAATAAAG
ACAATTCATACTTCTGTCAGAGATGGTTTGGTTTCCCTTGCTCGTTCTTTAGGCTTAGTAGTCTCGGTTAA
CGCAGAACCTGCTAAGGTTGACATGAATGGCACCAAACATAAAATTAGTTATGCTATTTATATGTCTGGTG
GAGATGTTTTGCTTAACGTTCTTTCGAAGTGTGCCGGCTCTAAAAAATTCAGGCCTGCTCCCGCCGCTGCT
TTTGCACGTGAGTGCCCCGGATTTTATTTCGAGTTACAAGAATTGAAGGAAGACGATTATTATGGGATTAC
TTTATCTGATGATTCTGATCATCAGTTTTTGCTTGCCAACCAGGTTGTCGTCCATAATTGCGCTAGTCACC
CAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGAT
CTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGT
TCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATT
CTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAA
TTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACC

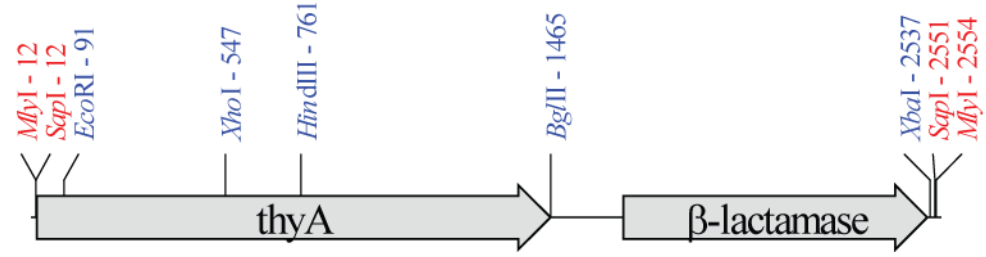
GAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGC
TGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAA
CTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGT
TGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGC
GTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACG
ACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCA
TTGGTAATTGAGTCAGATCGCG

Intein-β-lactamase linker with *Sap*I



AAGAGTCGAACAGCGAGAAGAGCTGAGGGTATTCGCAATAATCTTAATACTGAGAATCCATTATGGGACGC
TATTGTTGGCTTAGGATTCTTGAAGGACGGTGTCAAAAATATTCCTTCTTTCTTGTCTACGGACAATATCG
GTACTCGTGAAACATTTCTTGCTGGTCTAATTGATTCTGATGGCTATGTTACTGATGAGCATGGTATTAAA
GCAACAATAAAGACAATTCATACTTCTGTCAGAGATGGTTTGGTTTCCCTTGCTCGTTCTTTAGGCTTAGT
AGTCTCGGTTAACGCAGAACCTGCTAAGGTTGACATGAATGGCACCAAACATAAAATTAGTTATGCTATTT
ATATGTCTGGTGGAGATGTTTTGCTTAACGTTCTTTCGAAGTGTGCCGGCTCTAAAAAATTCAGGCCTGCT
CCCGCCGCTGCTTTTGCACGTGAGTGCCCCGGATTTTATTTCGAGTTACAAGAATTGAAGGAAGACGATTA
TTATGGGATTACTTTATCTGATGATTCTGATCATCAGTTTTTGCTTGCCAACCAGGTTGTCGTCCATAATT
GCGCTAGTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTAC
ATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAG
CACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCC
GCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAAC
GATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTT
GGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACA
ACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGA
GGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTG
GAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTA
GTTATCTACACGACGGGGAGCCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTC
ACTGATTAAGCATTGGTAAGCTCTTCAGCGTTGCAGACTCTT

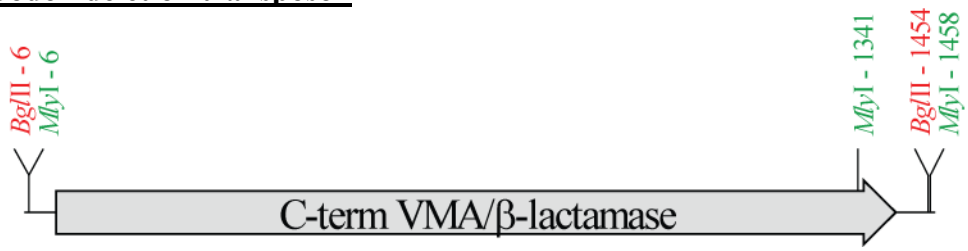
Intein thyA linker



AAGAGTCGAACAGCGAGAAGAGCAAATTCGAGCTCGAACAACAACAACAATAACAATAACAACAACCTCGG
GATCGAGGGAAGGATTTCAGAATTCGCCCTCGCAGAGGGCACTCGGATCTTCGATCCGGTCACCGGTACAA
CGCATCGCATCGAGGATGTTGTCGGTGGGCGCAAGCCTATTCATGTCGTGGCTGCTGCCAAGGACGGAACG
CTGCATGCGCGGCCCGTGGTGTCCTGGTTCGACCAGGGAACGCGGGATGTGATCGGGTTGCGGATCGCCGG
TGGCGCCATCCTGTGGGCGACACCCGATCACAAGGTGCTGACAGAGTACGGCTGGCGTGCCGCCGGGGAAC

TCCGCAAGGGAGACAGGGTGGCGCAACCGCGACGCTTCGATGGATTCGGTGACAGTGCGCCGATTCCGGCG
CGCGTGCAGGCGCTCGCGGATGCCCTGGATGACAAATTCCTGCACGACATGCTGGCGGAAGAACTCCGCTA
TTCCGTGATCCGAGAAGTGCTGCCAACGCGGCGGGCACGAACGTTCGGCCTCGAGGTCGAGGAACTGCACA
CCCTCGTCGCCGAAGGGGTTGTTGTACACAACTGTATGAAACAATACCAAGATTTAATTAAAGACATTTTT
GAAAATGGTTATGAAACCGATGATCGTACAGGCACAGGAACAATTGCTCTGTTCGGATCTAAATTACGCTG
GGATTTAACTAAAGGTTTTCCTGCGGTAACAACTAAGAAGCTCGCCTGGAAAGCTTGCATTGCTGAGCTAA
TATGGTTTTTATCAGGAAGCACAAATGTCAATGATTTACGATTAATTCAACACGATTCGTTAATCCAAGGC
AAAACAGTCTGGGATGAAAATTACGAAAATCAAGCAAAAGATTTAGGATACCATAGCGGTGAACTTGGTCC
AATTTATGGAAAACAGTGGCGTGATTTTGGTGGTGTAGACCAAATTATAGAAGTTATTGATCGTATTAAAA
AACTGCCAAATGATAGGCGTCAAATTGTTTCTGCATGGAATCCAGCTGAACTTAAATATATGGCATTACCG
CCTTGTCATATGTTCTATCAGTTTAATGTGCGTAATGGCTATTTGGATTTGCAGTGGTATCAACGCTCAGT
AGATGTTTTCTTGGGTCTACCGTTTAATATTGCGTCATATGCTACGTTAGTTCATATTGTAGCTAAGATGT
GTAATCTTATTCCAGGGGATTTGATATTTTCTGGTGGTAATACTCATATCTATATGAATCACGTAGAACAA
TGTAAAGAAATTTTGAGGCGTGAACCTAAAGAGCTTTGTGAGCTGGTAATAAGTGGTCTACCTTATAAATT
CCGATATCTTTCTACTAAAGAACAATTAAAATATGTTCTTAAACTTAGGCCTAAAGATTTCGTTCTTAACA
ACTATGTATCACACCCTCCTATTAAAGGAAAGATGGCGGTGTAAAGATCTGACGAAAGGGCCTCGTGATAC
GCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAAT
GTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACC
CTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTC
CCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAA
GATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCG
CCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTG
ACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTC
ACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAA
CACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGAC
ACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTC
CCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGG
CTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGG
CCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGCCAGGCAACTATGGATGAACGAAA
TAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCTCTAGAAGCTCTTCAGCGTT
GCAGACTCTT

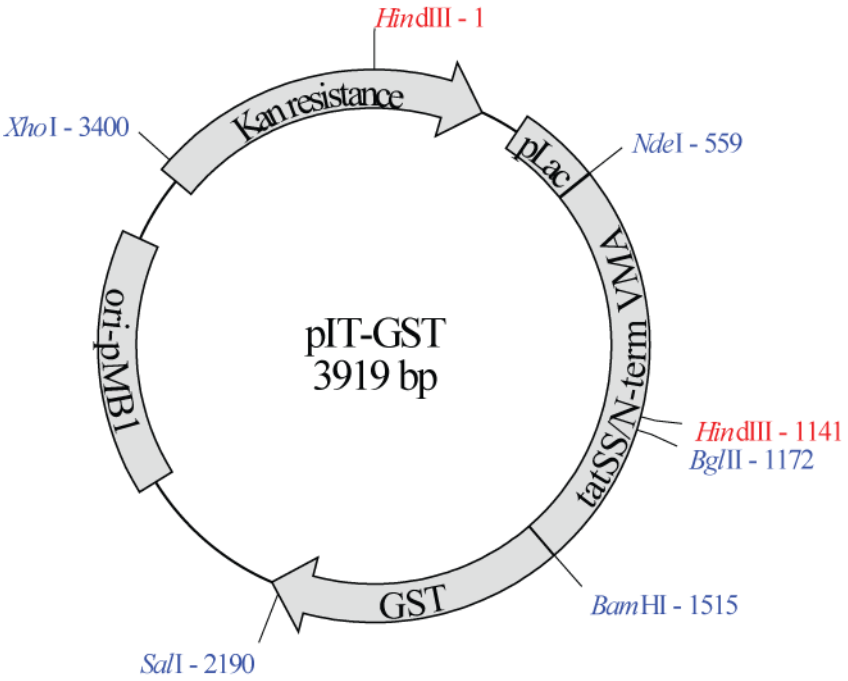
***Mly*I-Codon deletion transposon**



TAATAAGATCTGACTCGGCGCACGAAAAACGCGAAAGCGTTTCACGATAAATGCGAAAAACTGGGTATTCG
CAATAATCTTAATACTGAGAATCCATTATGGGACGCTATTGTTGGCTTAGGATTCTTGAAGGACGGTGTCA
AAAATATTCCTTCTTTCTTGTCTACGGACAATATCGGTACTCGTGAAACATTTCTTGCTGGTCTAATTGAT
TCTGATGGCTATGTTACTGATGAGCATGGTATTAAAGCAACAATAAAGACAATTCATACTTCTGTCAGAGA
TGGTTTGGTTTCCCTTGCTCGTTCTTTAGGCTTAGTAGTCTCGGTTAACGCAGAACCTGCTAAGGTTGACA
TGAATGGCACCAAACATAAAATTAGTTATGCTATTTATATGTCTGGTGGAGATGTTTTGCTTAACGTTCTT
TCGAAGTGTGCCGGCTCTAAAAAATTCAGGCCTGCTCCCGCCGCTGCTTTTGCACGTGAGTGCCCCGGATT
TTATTTCGAGTTACAAGAATTGAAGGAAGACGATTATTATGGGATTACTTTATCTGATGATTCTGATCATC
AGTTTTTGCTTGCCAACCAGGTTGTCGTCCATAATTGCGCTAGTCACCCAGAAACGCTGGTGAAAGTAAAA
GATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGA
GAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTAT
CCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTAC
TCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCAT

GAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGC
ACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGAC
GAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTAC
TCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGG
CCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCA
GCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGA
TGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAATTTTTCGCATTTATCG
TGAAACGCTTTCGCGTTTTTCGTGCGCCGAGTCAGATCTTATTA

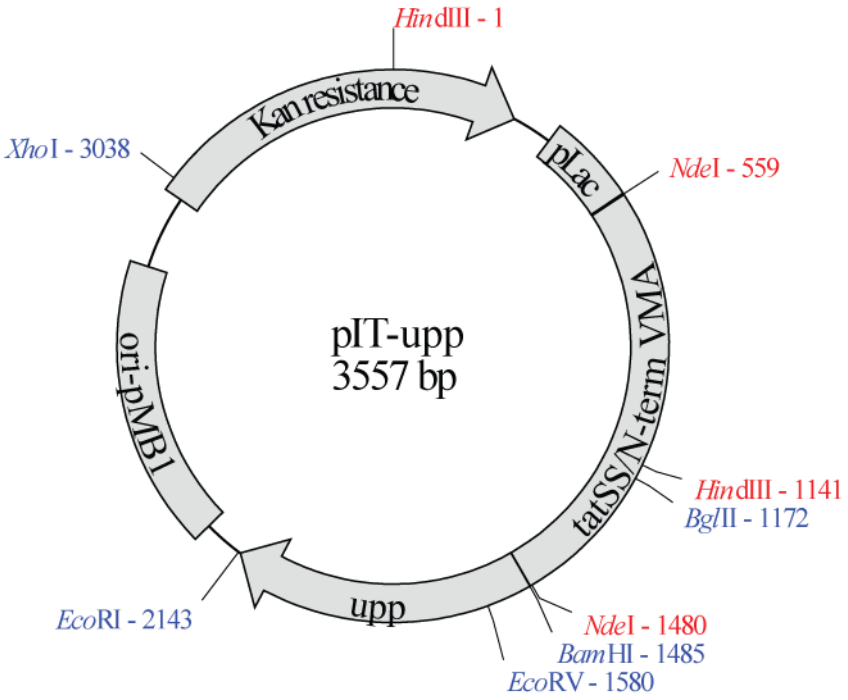
pIT-GST



AAGCTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGA
CGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTAGGAATCGCAGACCGATACCAGGATCTTGCCA
TCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGAT
AATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTG
GTTGTAACACTGGCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAATCGAACTTTTGC
TGAGTTGAAGGATCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGA
AAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTT
ATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAAGGAGATATACATATGAACAA
TAACGATCTCTTTCAGGCATCACGTCGGCGTTTTCTGGCACAACTCGGCGGCTTAACCGTCGCCGGGATGC
TGGGGCCGTCATTGTTAACGCCGCGACGTGCGACTGCGCATAATGGGTGCTTTGCCAAGGGTACCAATGTT
TTAATGGCGGATGGGTCTATTGAATGTATTGAAAACATTGAGGTTGGTAATAAGGTCATGGGTAAAGATGG
CAGACCTCGTGAGGTAATTAAATTGCCCAGAGGAAGAGAAACTATGTACAGCGTCGTTCAGAAAAGTCAGC
ACAGAGCCCACAAAAGTGATTCAAGTCGTGAAGTGCCAGAATTACTCAAGTTTACGTGTAATGCGACCCAT
GAGTTGGTTGTTAGAACACCTCGTAGTGTCCGCCGTTTGTCTCGTACCATTAAGGGTGTCGAATATTTTGA
AGTTATTACTTTTGAGATGGGCCAAAAGAAAGCCCCCGACGGTAGAATTGTTGAGCTTGTCAAGGAAGTTT
CAAAGAGCTACCCAATATCTGAGGGGCCTGAGAGAGCCAACGAATTAGTAGAATCCTATAGAAAGGCTTCA
AATAAAGCTTATTTTGAGTGGACTATTGAGGCCAGAGATCTTTCTCTGTTGGGTTCCCATGTTCGTAAAGC
TACCTACCAGACTTACGCGCCAATTCTTTATGAGAATGACCACTTTTTCGACTACATGCAAAAAAGTAAGT
TTCATCTCACCATTGAAGGTCCAAAAGTACTTGCTTATTTACTTGGTTTATGGATTGGTGATGGATTGTCT
GACAGGGCAACTTTTTCGGTTGATTCCAGAGATACTTCTTTGATGGAACGTGTTACTGAATATGCTGAAAA
GTTGAATTTGTGCGCCGAGTATAAGGACAGAAAAGAACCACAAGTTGCCAAAACTGTTAATTTGTACTCTA
AAGTTGTCAGAGGTACCATGGCCGGATCCCCTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCC
ACTCGACTTCTTTTGGAATATCTTGAAGAAAAATATGAAGAGCATTTGTATGAGCGCGATGAAGGTGATAA
ATGGCGAAACAAAAAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATTATATTGATGGTGATGTTA

AATTAACACAGTCTATGGCCATCATACGTTATATAGCTGACAAGCACAACATGTTGGGTGGTTGTCCAAAA
GAGCGTGCAGAGATTTCAATGCTTGAAGGAGCGGTTTTGGATATTAGATACGGTGTTTCGAGAATTGCATA
TAGTAAAGACTTTGAAACTCTCAAAGTTGATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATGTTCGAAG
ATCGTTTATGTCATAAAACATATTTAAATGGTGATCATGTAACCCATCCTGACTTCATGTTGTATGACGCT
CTTGATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTTTTAAAAAACG
TATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATAGCATGGCCTTTGCAGGGCT
GGCAAGCCACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGATCTGGTTTCTAGAGGCGTCGACCATCAT
CATCATCATCATTGAGTTTAAACGGTCTCCAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTG
ATACAGATTAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTAGCGCGGTGGT
CCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATG
CGAGAGTAGGGAACTGCCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTAT
CTGTTGTTTGTCGGTGAACGATATCTGCTTTTCTTCGCGAATTAATTCCGCTTCGCAACATGTGAGCAAAA
GGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGA
CGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGT
TTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTT
CTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCG
CTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTC
TTGAATCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCG
AGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATT
TGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAA
CCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAA
GATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCAT
GAGTTGTGTCTCAAAATCTCTGATGTTACATTGCACAAGATAAAAATATATCATCATGAACAATAAAACTG
TCTGCTTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCC
GCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAG
GTGCGACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGC
GTTGCCAATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCAT
CAAGCATTTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGGAAAACAGCATTCC
AGGTATTAGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTG
CATTCGATTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTAGCTCAGGCGCAATCACG
AATGAATAACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCT
GGAAAGAAATGCAT

pIT-upp



AAGCTTTTGCCATTCTCACCGGATTCAGTCGTCACTCATGGTGATTTCTCACTTGATAACCTTATTTTTGA
CGAGGGGAAATTAATAGGTTGTATTGATGTTGGACGAGTAGGAATCGCAGACCGATACCAGGATCTTGCCA
TCCTATGGAACTGCCTCGGTGAGTTTTCTCCTTCATTACAGAAACGGCTTTTTCAAAAATATGGTATTGAT
AATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGATGAGTTTTTCTAATCAGAATTGGTTAATTG
GTTGTAACACTGGCAGAGCATTACGCTGACTTGACGGGACGGCGGCTTTGTTGAATAAATCGAACTTTTGC
TGAGTTGAAGGATCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGA
AAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTT
ATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAAGGAGATATACATATGAACAA
TAACGATCTCTTTCAGGCATCACGTCGGCGTTTTCTGGCACAACTCGGCGGCTTAACCGTCGCCGGGATGC
TGGGGCCGTCATTGTTAACGCCGCGACGTGCGACTGCGCATAATGGGTGCTTTGCCAAGGGTACCAATGTT
TTAATGGCGGATGGGTCTATTGAATGTATTGAAAACATTGAGGTTGGTAATAAGGTCATGGGTAAAGATGG
CAGACCTCGTGAGGTAATTAAATTGCCCAGAGGAAGAGAAACTATGTACAGCGTCGTTCAGAAAAGTCAGC
ACAGAGCCCACAAAAGTGATTCAAGTCGTGAAGTGCCAGAATTACTCAAGTTTACGTGTAATGCGACCCAT
GAGTTGGTTGTTAGAACACCTCGTAGTGTCCGCCGTTTGTCTCGTACCATTAAGGGTGTCGAATATTTTGA
AGTTATTACTTTTGAGATGGGCCAAAAGAAAGCCCCCGACGGTAGAATTGTTGAGCTTGTCAAGGAAGTTT
CAAAGAGCTACCCAATATCTGAGGGGCCTGAGAGAGCCAACGAATTAGTAGAATCCTATAGAAAGGCTTCA
AATAAAGCTTATTTTGAGTGGACTATTGAGGCCAGAGATCTTTCTCTGTTGGGTTCCCATGTTCGTAAAGC
TACCTACCAGACTTACGCGCCAATTCTTTATGAGAATGACCACTTTTTCGACTACATGCAAAAAAGTAAGT
TTCATCTCACCATTGAAGGTCCAAAAGTACTTGCTTATTTACTTGGTTTATGGATTGGTGATGGATTGTCT
GACAGGGCAACTTTTTCGGTTGATTCCAGAGATACTTCTTTGATGGAACGTGTTACTGAATATGCTGAAAA
GTTGAATTTGTGCGCCGAGTATAAGGACAGAAAAGAACCACAAGTTGCCAAAACTGTTCATATGGGATCCG
CTGCCCCTCAACGCAGAAAGAGTATGAAGATCGTGGAAGTCAAACACCCACTCGTCAAACACAAGCTGGGA
CTGATGCGTGAGCAAGATATCAGCACCAAGCGCTTTCGCGAACTCGCTTCCGAAGTGGGTAGCCTGCTGAC
TTACGAAGCGACCGCCGACCTCGAAACGGAAAAAGTAACTATCGAAGGCTGGAACGGCCCGGTAGAAATCG
ACCAGATCAAAGGTAAGAAAATTACCGTTGTGCCAATTCTGCGTGCGGGTCTTGGTATGATGGACGGTGTG
CTGGAAAACGTTCCGAGCGCGCGCATCAGCGTTGTCGGTATGTACCGTAATGAAGAAACGCTGGAGCCGGT
ACCGTACTTCCAGAAACTGGTTTCTAACATCGATGAGCGTATGGCGCTGATCGTTGACCCAATGCTGGCAA
CCGGTGGTTCCGTTATCGCGACCATCGACCTGCTGAAAAAAGCGGGCTGCAGCAGCATCAAAGTTCTGGTG
CTGGTAGCTGCGCCAGAAGGTATCGCTGCGCTGGAAAAAGCGCACCCGGACGTCGAACTGTATACCGCATC
GATTGATCAGGGACTGAACGAGCACGGATACATTATTCCGGGCCTCGGCGATGCCGGTGACAAAATCTTTG
GTACGAAATAAAGAATTCTGCTTTTCTTCGCGAATTAATTCCGCTTCGCAACATGTGAGCAAAAGGCCAGC

AAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCAT
CACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCC
TGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTT
CGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAG
CTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAATC
CAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATG
TAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATC
TGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGC
TGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTT
TGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGTTGT
GTCTCAAAATCTCTGATGTTACATTGCACAAGATAAAAATATATCATCATGAACAATAAAACTGTCTGCTT
ACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGATTA
AATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTGCGAC
AATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGAAACATGGCAAAGGTAGCGTTGCCA
ATGATGTTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTCCGACCATCAAGCAT
TTTATCCGTACTCCTGATGATGCATGGTTACTCACCACTGCGATCCCCGGGAAAACAGCATTCCAGGTATT
AGAAGAATATCCTGATTCAGGTGAAAATATTGTTGATGCGCTGGCAGTGTTCCTGCGCCGGTTGCATTCGA
TTCCTGTTTGTAATTGTCCTTTTAACAGCGATCGCGTATTTCGTCTAGCTCAGGCGCAATCACGAATGAAT
AACGGTTTGGTTGATGCGAGTGATTTTGATGACGAGCGTAATGGCTGGCCTGTTGAACAAGTCTGGAAAGA
AATGCAT

Bibliography

(1) Cunningham, B. C., and Wells, J. A. (1989) High-resolution epitope mapping of Hgh-receptor interactions by alanine-scanning mutagenesis. *Science 244*, 1081-1085.

(2) Lundegaard, C., and Jensen, K. F. (1999) Kinetic mechanism of uracil phosphoribosyltransferase from *Escherichia coli* and catalytic importance of the conserved proline in the PRPP binding site. *Biochemistry 38*, 3327-3334.

(3) Fuchs, S. M., and Raines, R. T. (2007) Arginine grafting to endow cell permeability. *ACS Chem. Biol. 2*, 167-170.

(4) Fuchs, S. M., Rutkoski, T. J., Kung, V. M., Groeschl, R. T., and Raines, R. T. (2007) Increasing the potency of a cytotoxin with an arginine graft. *Protein Eng. Des. Sel. 20*, 505-509.

(5) Hutchings, M. G., Grossel, M. C., Merckel, D. A. S., Chippendale, A. M., Kenworthy, M., and McGeorge, G. (2001) The structure of m-Xylylenediguanidinium sulfate: A putative molecular tweezer ligand for anion chelation. *Cryst Growth Des 1*, 339-342.

(6) Wender, P. A., Mitchell, D. J., Pattabiraman, K., Pelkey, E. T., Steinman, L., and Rothbard, J. B. (2000) The design, synthesis, and evaluation of molecules that enable or enhance cellular uptake: peptoid molecular transporters. *Proc. Natl. Acad. Sci. U.S.A. 97*, 13003-13008.

(7) Umezawa, N., Gelman, M. A., Haigis, M. C., Raines, R. T., and Gellman, S. H. (2001) Translocation of a β-Peptide Across Cell Membranes. *J. Am. Chem. Soc. 124*, 368-369.

(8) Rothbard, J. B., Jessop, T. C., Lewis, R. S., Murray, B. A., and Wender, P. A. (2004) Role of membrane potential and hydrogen bonding in the mechanism of translocation of guanidinium-rich peptides into cells. *J. Am. Chem. Soc. 126*, 9506-9507.

(9) Rothbard, J. B., Jessop, T. C., and Wender, P. A. (2005) Adaptive translocation: the role of hydrogen bonding and membrane potential in the uptake of guanidinium-rich transporters into cells. *Adv. Drug Deliv. Rev. 57*, 495-504.

(10) Leland, P. A., Schultz, L. W., Kim, B. M., and Raines, R. T. (1998) Ribonuclease A variants with potent cytotoxic activity. *Proc. Natl. Acad. Sci. U.S.A. 95*, 10407-10412.

(11) Zhu, Q., and Casey, J. R. (2007) Topology of transmembrane proteins by scanning cysteine accessibility mutagenesis methodology. *Methods 41*, 439-450.

(12) Wu, J., and Watson, J. T. (1998) Optimization of the cleavage reaction for cyanylated cysteinyl proteins for efficient and simplified mass mapping. *Anal. Biochem. 258*, 268-276.

(13) Matthews, C. R. (2003) Pathways of protein folding. *Annu. Rev. Biochem. 62*, 653-683.

(14) Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol. 24*, 79-88.

(15) Peters, M. W., Meinhold, P., Glieder, A., and Arnold, F. H. (2003) Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc. 125*, 13442-13450.

(16) Kubo, T., Peters, M. W., Meinhold, P., and Arnold, F. H. (2006) Enantioselective epoxidation of terminal alkenes to (R)- and (S)-epoxides by engineered cytochromes P450 BM-3. *Chemistry 12*, 1216-1220.

(17) Cropp, T. A., and Schultz, P. G. (2004) An expanding genetic code. *Trends Genet. 20*, 625-630.

(18) Hutchison, C. A., 3rd, Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., and Smith, M. (1978) Mutagenesis at a specific position in a DNA sequence. *J. Biol. Chem. 253*, 6551-6560.

(19) Sanger, F., and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol. 94*, 441-448.

(20) Mullis, K. B., and Faloona, F. A. (1987) Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Meth. Enzymol. 155*, 335-350.

(21) Singer, B., and Kusmierek, J. T. (1982) Chemical mutagenesis. *Annu. Rev. Biochem. 51*, 655-693.

(22) Sinha, R. P., and Hader, D. P. (2002) UV-induced DNA damage and repair: a review. *Photochem. Photobiol. Sci. 1*, 225-236.

(23) Kunkel, T. A., Roberts, J. D., and Zakour, R. A. (1987) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Meth. Enzymol. 154*, 367-382.

(24) Stemmer, W. P., and Morris, S. K. (1992) Enzymatic inverse PCR: a restriction site independent, single-fragment method for high-efficiency, site-directed mutagenesis. *Biotechniques 13*, 214-220.

(25) Cadwell, R. C., and Joyce, G. F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl 2*, 28-33.

(26) Stemmer, W. P. C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature 370*, 389-391.

(27) Bittker, J. A., Le, B. V., and Liu, D. R. (2002) Nucleic acid evolution and minimization by nonhomologous random recombination. *Nat. Biotechnol. 20*, 1024-1029.

(28) Graf, R., and Schachman, H. K. (1996) Random circular permutation of genes and expressed polypeptide chains: application of the method to the catalytic chains of aspartate transcarbamoylase. *Proc Natl Acad Sci U S A 93*, 11591-11596.

(29) Poussu E, V. M., Paulin L, Savilahti H. (2004) Probing the alpha-complementing domain of E. coli beta-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition. *Proteins 54*, 681-692.

(30) Selifonova, O., Valle, F., and Schellenberger, V. (2001) Rapid evolution of novel traits in microorganisms. *Appl. Environ. Microbiol. 67*, 3645-3549.

(31) Wang, L., Zhang, Z. W., Brock, A., and Schultz, P. G. (2003) Addition of the keto functional group to the genetic code of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A. 100*, 56-61.

(32) Deiters, A., Cropp, T. A., Mukherji, M., Chin, J. W., Anderson, J. C., and Schultz, P. G. (2003) Adding amino acids with novel reactivity to the genetic code of *Saccharomyces cerevisiae*. *J. Am. Chem. Soc. 125*, 11782-11783.

(33) Deiters, A., Cropp, T. A., Summerer, D., Mukherji, M., and Schultz, P. G. (2004) Site-specific PEGylation of proteins containing unnatural amino acids. *Bioorg. Med. Chem. Lett. 14*, 5743-5745.

(34) Bose, M., Groff, D., Xie, J., Brustad, E., and Schultz, P. G. (2006) The incorporation of a photoisomerizable amino acid into proteins in *E. coli*. *J. Am. Chem. Soc. 128*, 388-389.

(35) Deiters, A., Groff, D., Ryu, Y., Xie, J., and Schultz, P. G. (2006) A genetically encoded photocaged tyrosine. *Angew. Chem. Int. Ed. Engl. 45*, 2728-2731.

(36) Chin, J. W., Martin, A. B., King, D. S., Wang, L., and Schultz, P. G. (2002) Addition of a photocrosslinking amino acid to the genetic code of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A. 99*, 11020-11024.

(37) Huang, L. Y., Umanah, G., Hauser, M., Son, C., Arshava, B., Naider, F., and Becker, J. M. (2008) Unnatural amino acid replacement in a yeast G protein-coupled receptor in its native environment. *Biochemistry 47*, 5638-5648.

(38) Chin, J. W., and Schultz, P. G. (2002) In vivo photocrosslinking with unnatural amino acid mutagenesis. *Chembiochem 3*, 1135-1137.

(39) Wang, Q., Parrish, A. R., and Wang, L. (2009) Expanding the genetic code for biological studies. *Chem. Biol. 16*, 323-336.

(40) Hohsaka, T., Ashizuka, Y., Murakami, H., and Sisido, M. (1996) Incorporation of nonnatural amino acids into streptavidin through *in vitro* frame-shift suppression. *J. Am. Chem. Soc. 118*, 9778-9779.

(41) Hohsaka, T., Kajihara, D., Ashizuka, Y., Murakami, H., and Sisido, M. (1998) Efficient incorporation of nonnatural amino acids with large aromatic groups into streptavidin in *in vitro* protein synthesizing systems. *J. Am. Chem. Soc. 121*, 34-40.

(42) Hohsaka, T., Ashizuka, Y., Taira, H., Murakami, H., and Sisido, M. (2001) Incorporation of nonnatural amino acids into proteins by uing various four-base codons in an *Escherichia coli in vitro* translation system. *Biochemistry 40*, 11060-11064.

(43) Wang, L., Magliery, T. J., Liu, D. R., and Schultz, P. G. (2000) A new functional suppressor tRNA/aminoacyl-tRNA synthetase pair for the *in vivo* incorporation of unnatural amino acids into proteins. *J. Am. Chem. Soc. 122*, 5010-5011.

(44) Chin, J. W., Cropp, T. A., Anderson, J. C., Mukherji, M., Zhang, Z., and Schultz, P. G. (2003) An expanded eukaryotic genetic code. *Science 301*, 964-967.

(45) Kobayashi, T., Nureki, O., Ishitani, R., Yaremchuk, A., Tukalo, M., Cusack, S., Sakamoto, K., and Yokoyama, S. (2003) Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nat. Struct. Biol. 10*, 425-432.

(46) Wang, L., and Schultz, P. G. (2001) A general approach for the generation of orthogonal tRNAs. *Chem. Biol. 8*, 883-890.

(47) Wang, L., Brock, A., Herberich, B., and Schultz, P. G. (2001) Expanding the genetic code of *Escherichia coli*. *Science 292*, 498-500.

(48) Sinz, A. (2003) Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom. 38*, 1225-1237.

(49) Burdine, L., Gillette, T. G., Lin, H. J., and Kodadek, T. (2004) Periodate-triggered cross-linking of DOPA-containing peptide-protein complexes. *J. Am. Chem. Soc. 126*, 11442-11443.

(50) Friedhoff, P. (2005) Mapping protein-protein interactions by bioinformatics and cross-linking. *Anal. Bioanal. Chem. 381*, 78-80.

(51) Wilkins, B. J., Daggett, K. A., and Cropp, T. A. (2008) Peptide mass fingerprinting using isotopically encoded photo-crosslinking amino acids. *Mol. Biosyst. 4*, 934-936.

(52) Maru, Y., Afar, D. E., Witte, O. N., and Shibuya, M. (1996) The dimerization property of glutathione *S*-transferase partially reactivates Bcr-Abl lacking the oligomerization domain. *J. Biol. Chem. 271*, 15353-15357.

(53) McTigue, M. A., Williams, D. R., and Tainer, J. A. (1995) Crystal structures of a schistosomal drug and vaccine target: glutathione *S*-transferase from S*chistosoma japonica* and its complex with the leading antischistomal drug praziquantel. *J. Mol. Biol. 246*, 21-27.

(54) Dorman, G., and Prestwich, G. D. (1994) Benzophenone photophores in biochemistry. *Biochemistry 33*, 5661-5673.

(55) Ryu, Y. H., and Schultz, P. G. (2006) Efficient incorporation of unnatural amino acids into proteins in *Escherichia coli*. *Nat. Methods 3*, 263-265.

(56) McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A. 36*, 344-355.

(57) Mizuuchi, K. (1983) *In vitro* transposition of bacteriophage Mu: a biochemical approach to a novel replication reaction. *Cell 35*, 785-94.

(58) Chaconas, G., Lavoie, B. D., and Watson, M. A. (1996) DNA transposition: jumping gene machine, some assembly required. *Curr. Biol. 6*, 817-820.

(59) Savilahti, H., Rice, P. A., and Mizuuchi, K. (1995) The phage Mu transpososome core: DNA requirements for assembly and function. *Embo J. 14*, 4893-4903.

(60) Haapa, S., Taira, S., Heikkinen, E., and Savilahti, H. (1999) An efficient and accurate integration of mini-Mu transposons *in vitro*: a general methodology for functional genetic analysis and molecular biology applications. *Nucleic Acids Res. 27*, 2777-2784.

(61) Craig, N. L. (1995) Unity in transposition reactions. *Science 270*, 253-254.

(62) Mizuuchi, K. (2003) Transpositional recombination: Mechanistic insights from studies of Mu and other elements. *Annu. Rev. Biochem. 61*, 1011-1051.

(63) Poussu, E., Jantti, J., and Savilahti, H. (2005) A gene truncation strategy generating N- and C-terminal deletion variants of proteins for functional studies: mapping of the Sec1p binding domain in yeast Mso1p by a Mu *in vitro* transposition-based approach. *Nucleic Acids Res. 33*, e104.

(64) Jones, D. D. (2005) Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. *Nucleic Acids Res. 33*, e80.

(65) Baldwin, A. J., Busse, K., Simm, A. M., and Jones, D. D. (2008) Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). *Nucleic Acids Res. 36*, e77.

(66) Baldwin, A. J., Arpino, J. A., Edwards, W. R., Tippmann, E. M., and Jones, D. D. (2009) Expanded chemical diversity sampling through whole protein evolution. *Mol. Biosyst. 5*, 764-766.

(67) Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol. 17*, 691-695.

(68) Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D., and Davidson, A. R. (1999) A simple *in vivo* assay for increased protein solubility. *Protein Sci. 8*, 1908-1911.

(69) Sieber, V., Martinez, C. A., and Arnold, F. H. (2001) Libraries of hybrid proteins from distantly related sequences. *Nat Biotechnol 19*, 456-460.

(70) Lutz, S., Fast, W., and Benkovic, S. J. (2002) A universal, vector-based system for nucleic acid reading-frame selection. *Protein Eng. 15*, 1025-1030.

(71) Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N., Noren, C. J., Thorner, J., and Belfort, M. (1994) Protein splicing elements - inteins and exteins - a definition of terms and recommended nomenclature. *Nucleic Acids Res. 22*, 1125-1127.

(72) Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002) Inteins: structure, function, and evolution. *Annu. Rev. Microbiol. 56*, 263-287.

(73) Gerth, M. L., Patrick, W. M., and Lutz, S. (2004) A second-generation system for unbiased reading frame selection. *Protein Eng. Des. Sel. 17*, 595-602.

(74) DeLisa, M. P., Tullman, D., and Georgiou, G. (2003) Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. *Proc. Natl. Acad. Sci. U.S.A. 100*, 6115-6120.

(75) Bradley, L. H., Kleiner, R. E., Wang, A. F., Hecht, M. H., and Wood, D. W. (2005) An intein-based genetic selection allows the construction of a high-quality library of binary patterned de novo protein sequences. *Protein Eng. Des. Sel. 18*, 201-207.

(76) Wood, D. W., Wu, W., Belfort, G., Derbyshire, V., and Belfort, M. (1999) A genetic system yields self-cleaving inteins for bioseparations. *Nat. Biotechnol. 17*, 889-892.

(77) Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A. 97*, 6640-6645.

(78) Patrick, W. M., Firth, A. E., and Blackburn, J. M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng. 16*, 451-457.

(79) Wong, Q. N., Ng, V. C., Lin, M. C., Kung, H. F., Chan, D., and Huang, J. D. (2005) Efficient and seamless DNA recombineering using a thymidylate synthase A selection system in *Escherichia coli*. *Nucleic Acids Res. 33*, e59.

(80) Cherepanov, P. P., and Wackernagel, W. (1995) Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene 158*, 9-14.

(81) Daggett, K. A., Layer, M., and Cropp, T. A. (2009) A general method for scanning unnatural amino acid mutagenesis. *ACS Chem. Biol. 4*, 109-113.

(82) Mardis, E. R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet. 9*, 387-402.

(83) Kim, Y. C., Lee, H. S., Yoon, S., and Morrison, S. L. (2009) Transposon-directed base-exchange mutagenesis (TDEM): a novel method for multiple-nucleotide substitutions within a target gene. *Biotechniques 46*, 534-542.

(84) Christ, D., Famm, K., and Winter, G. (2006) Tapping diversity lost in transformations--*in vitro* amplification of ligation reactions. *Nucleic Acids Res. 34*, e108.

(85) Kim, Y. C., and Morrison, S. L. (2009) N-terminal domain-deleted mu transposase exhibits increased transposition activity with low target site preference in modified buffers. *J. Mol. Microbiol. Biotechnol. 17*, 30-40.

(86) Wiggers, H. J., Cheleski, J., Zottis, A., Oliva, G., Andricopulo, A. D., and Montanari, C. A. (2007) Effects of organic solvents on the enzyme activity of Trypanosoma cruzi glyceraldehyde-3-phosphate dehydrogenase in calorimetric assays. *Anal. Biochem. 370*, 107-114.

(87) Conlan, L. H., Jose, T. J., Thornton, K. C., and Dupureur, C. M. (1999) Modulating restriction endonuclease activities and specificities using neutral detergents. *Biotechniques 27*, 955-960.

(88) Clarke, P., Cuiv, P. O., and O'Connell, M. (2005) Novel mobilizable prokaryotic two-hybrid system vectors for high-throughput protein interaction mapping in *Escherichia coli* by bacterial conjugation. *Nucleic Acids Res. 33*, e18.

Curriculum Vitae

Kelly A. Daggett

Education:

- Ph.D. in Chemistry, University of Maryland, August 2004 – December 2009
- B.S. in Biochemistry, Manhattan College, May 2004

Honors and Awards:

- Distinguished Teaching Assistant 2007 - 2008
- GAANN Teaching Fellow, August 2006 – August 2007
- Chemistry Honors Medal, Manhattan College, 2004
- Sigma Xi, Research Honor Society, Manhattan College, 2004
- Howard Hughes Medical Grant, Summer 2003
- Gamma Sigma Epsilon, Chemistry Honor Society, 2003

Research Interests:

Evolution of proteins, incorporation of unnatural amino acids into proteins, bio-organic chemistry.

Publications:

Daggett, K.A., Layer, M. and Cropp, T.A. A General Method for Scanning Unnatural Amino Acid Mutagenesis. *ACS Chem. Biol.* **2009**, *4*, 109.

Wilkins, B.J., Daggett, K.A. and Cropp, T.A., Peptide Mass Fingerprinting Using Isotopically Encoded Photo-crosslinking Amino Acids. *Mol Biosyst.* **2008**, *4*, 934.

McCullagh, J.V. and Daggett, K.A. Synthesis of Triarylmethane and Xanthene Dyes Using Electrophilic Aromatic Substitution Reactions. *J Chem Educ.* **2007**, *84*, 1799.