OPTIMIZATION OF RECOMBINATION METHODS AND EXPANDING THE UTILITY OF PENICILLIN G ACYLASE

A Dissertation
Presented to
The Academic Faculty

by

Bernard Liat Wen Loo

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemical & Biomolecular Engineering

Georgia Institute of Technology
December 2007

OPTIMIZATION OF RECOMBINATION METHODS AND EXPANDING THE UTILITY OF PENICILLIN G ACYLASE

Approved by:

Dr. Andreas S. Bommarius, Advisor
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Jay H. Lee
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Mark R. Prausnitz
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

Dr. Stefan Lutz
School of Chemistry
Emory University

Dr. Wei-Shou Hu
Chemical Engineering & Materials Science
University of Minnesota, Twin Cities

Date Approved: October 25, 2007

To my family

ACKNOWLEDGEMENTS

There are a number of people instrumental for my success and I would like to thank them with all my heart. First of all, I would like to thank my family whom had been very supportive of my pursuit of my doctoral studies. Without my family's love and support, it would have been very difficult for me to complete my studies.

I am grateful to my advisor, Andreas (Andy) Bommarius who has provided sound advice and support throughout these years. I am also grateful to my thesis committee: Dr. Wei-Shu Hu, Dr. Stefan Lutz, Dr. Mark Prausnitz and Dr. Jay Lee for offering their time and advice.

Many thanks to Karen Polizzi for her help in proof reading my thesis. I would like to acknowledge Janna Blum for help with HPLC data in the PGA experiments and thank our collaborators Dr. Matthew Realff, Dr. Jay Lee and Anshul Dubey for assistance with the analysis of experimental data using Boolean learning and Support Vector Machines learning algorithms.

Last but not least, I am also indebted to other former and current group members: Javier Chaparro-Riggers, Phillip Gibbs, Eduardo Vazquez-Figueroa, Rongrong Jiang, Karl Huetinger, James Broering, Yanto Yanto, Tracey Thaler, Adrian Katona, Thomas Rogers, Andria Deaguero, Victor Yeh, Yue Liang, Matthew Swisher, Desmond Moore and Alicia Powers.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BLSVM	Boolean Learning and Support Vector Machine
BFP	Blue Fluorescent Protein
CASTing	Combinatorial active-site saturation testing
DE	Directed evolution
DEE	Dead-end elimination algorithm
DNA	Deoxyribonucleic acid
DsRed	*Discosoma* Red Fluorescent Protein
EGFP	Enhanced Green Fluorescent Protein
ep-PCR	Error-prone PCR
GFPs	Green Fluorescent Proteins
HcRed	*Heteractis crispa* Red Fluorescent Protein
HPLC	High performance liquid chromatography
mRFP	monomeric Red Fluorescent Protein
NAD	Nicotinamide adenine dinucleotide
NADP	Nicotinamide adenine dinucleotide phosphate
ProSAR	Protein sequence activity relationship algorithm
PCR	Polymerase chain reaction
PDAB	Para-dimethyl-benzaldehyde
PGA	Penicillin G acylase
RBP	Ribose binding protein
RFPs	Red Fluorescent Proteins
TIM	Triose phosphate isomerase

SUMMARY

Protein engineering can be performed by combinatorial techniques (directed evolution) and data-driven methods using machine-learning algorithms. The main characteristic of directed evolution (DE) is the application of an effective and efficient screen or selection on a diverse mutant library. As it is important to have a diverse mutant library for the success of DE, we compared the performance of DNA-shuffling and recombination PCR on fluorescent proteins using sequence information as well as statistical methods. We found that the diversity of the libraries DNA-shuffling and recombination PCR generates were dependent on type of skew primers used and sensitive to nucleotide identity levels between genes. DNA-shuffling and recombination PCR produced libraries with different crossover tendencies, suggesting that the two protocols could be used in combination to produce better libraries. Data-driven protein engineering uses sequence, structure and function data along with analyzed empirical activity information to guide library design. Boolean Learning Support Vector Machines (BLSVM) to identify interacting residues in fluorescent proteins and the gene templates were modified to preserve interactions post recombination. By site-directed mutagenesis, recombination and expression experiments, we validated that BLSVM can be used to identify interacting residues and increase the fraction of active proteins in the library.

As an extension to the above experiments, DE was applied on monomeric Red Fluorescent Proteins to improve its spectral characteristics and structure-guided protein engineering was performed on penicillin G acylase (PGA), an industrially relevant catalyst, to change its substrate specificity.

CHAPTER 1

INTRODUCTION

With advancement in protein engineering, biocatalysts can be more heavily utilized for chemical reactions than previously thought possible. Protein engineering involves improving enzymes to feature superior specificity, faster kinetics, or stability at higher temperature relative to the native enzyme. Protein engineering has been successfully applied to modify to the properties of the biocatalyst so that it can perform satisfactorily in non-native conditions that are optimal for product formation, thereby improving the utility of biocatalysts for industrial reactions. Examples of successful protein engineering projects include fluorescent proteins, hydrolases, lipases, esterases, hydantoinases, lyases and dehydrogenases and halohydrin dehalogenase (1-5).

1.1 Enzymes are optimized for regulation

Enzymes are usually not optimized for catalyzing the reactions chemists find important. Enzymes obtained from organisms have undergone billions of years of Darwinian selection to improve the survival chances of its hosts. Implicitly, selective advantage was applied to the enzymes for regulation of the metabolic pathways but not for maximal catalytic acceleration. Generally, enzymes are also evolved to exist transiently as the living cell is also in a constant state of metabolic flux and non-thermodynamic equilibrium. Thus, enzymes are synthesized when required and are easily degraded when they have served their purpose are not naturally evolved to be stable over the long run which is a desired characteristic for industrial processes.

1.2 Advantages of using enzymes

Enzymes are suited to a variety of industrial applications that synthetic catalysts do not perform as well. Enzymes perform chemo-, regio-, diastereo- and enantio-selective chemistry very well (6). Chemoselectivity refers to the reaction of the enzyme with specific functional groups thereby reducing side reactions, which complicate purification of the products. Regio and diastereoselectivity refer to the ability of the enzymes to react on functional groups located on different part on the same substrate molecule. Finally, enantioselectivity exists in enzymes because they have plasticity to bind to the (R)-enantiomer transition state than (S)-enantiomer transition state or vice-versa. Thus, a prochiral substrate can be reacted to an optically active one. Another two advantages of biocatalysts include catalyzing reactions under milder conditions than chemically synthesized catalysts are able to do and lastly, enzymes are environmentally friendly as they are biodegradable.

1.3 Paradigm shift: Tune enzymes for optimal reaction conditions

There is a paradigm shift in protein engineering for the application of enzyme technology (7). Instead of using a sub-optimal reaction condition due to enzyme limitations for product formation, one can now optimize the reaction conditions for maximal product formation and engineer the enzyme to work well at such conditions. This is made possible by recent advances in molecular biology and protein knowledge. Improved enzyme variants can now be engineered to confer functional advantages to the enzymes by utilizing molecular biology techniques to mutate residues (8). Online protein databases and the plethora of publications on protein research facilitate identification of key residues in proteins. Smart selection of residues for mutations, coupled with a good screening method (9) for the improved variant, can evolve proteins for the desired reaction conditions that maximize the product yield.

1.4 General methods to engineer enzymes

There is no consensus about the best way to engineer novel enzymes. One method to improve the enzymes is to randomly generate a library of variants that are then screened by a high throughput assay. There are a variety of library generation protocols available, it is not clear however, which library protocols is optimal as there is no prior work that systematically compared library protocols to the same set of proteins. Another relatively new method to improve proteins is to make a directed library in which the key residues are chosen for mutations to a limited set of amino acids (10-12). The residues can be determined either by analysis of sequence-activity relationships or sometimes with insight gained from crystal structure analysis, i.e. the crystal structure data can be used to figure out which residues need to be mutated to accommodate the new substrates. The above considerations inspired research into library generation protocols, the use of a machine learning language to analyze library variants for interacting residue. Fluorescent proteins and penicllin G acylase were chosen as the model proteins used for this investigation as they are relatively well-characterized proteins that would serve as a good starting point for protein evolution experiments. One important thing to note to is that fluorescent proteins are not biocatalysts or enzymes but they have a fluorescence phenotype that can be easily measured, similar to enzymes that catalyzes a reaction in which its products can be measured by an assay method.

1.5 Contributions of this work

This dissertation documents the development and application of protein engineering techniques to improve two proteins, the monomeric Red Fluorescent Protein (mRFP), a useful bioimaging protein, and penicillin G acylase (PGA), an industrially useful enzyme for synthesis of β-lactam antibiotics. The background chapter, Chapter 2, will develop a framework for understanding the various types of protein engineering strategies available. Protein engineering techniques such as rational design methods, directed evolution, and data-driven protein engineering methods are described in detail along with additional information on mRFP and PGA proteins that supplements the main content chapters on these proteins.

Chapter 3 describes the synthesis, cloning, expression, characterization and evolution experiments on the mRFP protein. This work reports the use of a DNA synthesis method involving the use of codon-optimized primers and two polymerase chain reactions (PCRs) to generate the mRFP gene. The mRFP gene is then cloned into vectors that is then transformed into *E.coli* cells and induced to overexpress the mRFP protein. Following purification, the mRFP protein is spectrophotometrically characterized. The mRFP gene is then subjected to mutagenesis experiments and a high-through put fluorescence screening method is applied to pick out the improved variants.

The next contribution in Chapter 4 reports on the application of recombination protocols, DNA shuffling and recombination PCR, on mRFP and other fluorescent proteins. This crucial study sheds insight on the robustness of two recombination protocols, compared systematically on fluorescent proteins. Up to 350 sequences were obtained for recombination scenarios with high and low DNA identity levels. Different templates were used for recombination to check for any correlation to number and type of crossover that can be obtained. Two significant revelations were found. First, we found that the use of two-sided skew primers, designed for annealing to either the heads of maternal gene or the tails of paternal genes, for recombination reduces redundant chimera generation and secondly, while recombination PCR can work to a lower identity level than DNA shuffling, the technique has the potential to have similar tendencies like DNA shuffling to produce non-unique redundant sequences. The results from this investigation will be very useful to protein engineers who use recombination techniques as a method to produce their library.

Although recombination library generation protocols are useful to produce protein sequences randomly, a restricted and smarter library in which residues relevant for activity are selected and mutated could improve the chances of success for enzyme engineering efforts using minimal amount of work. A method of identifying interacting residues using a machine learning language, Boolean Learning and Support Vector Machines (BLSVM) on mRFP and DsRed proteins is detailed in Chapter 5. The mRFP genes and DsRed genes were shuffled and their proteins expressed. The activity of fluorescent proteins was recorded and linked to its sequence. BLSVM was then used to analyze the sequence-activity information to identify interacting residues. Site-directed mutants were then created to confirm the presence of interactions. The insight obtained from BLSVM is also used to engineer templates, which improve the number of active variants that is obtained from recombination experiments.

To make use of the molecular biology skills and insight obtained from the initial experiments on an industrially relevant enzyme, the evolution of PGA towards altered substrate specificity was performed. Chapter 6 describes the analysis of 3-D crystal structures to select residues for mutations and the optimization of a high throughput assay for screening variants with substrate specificity change. An overlap extension PCR method was used to generate degenerate libraries, which were screened for using a Schiff-base assay that detects the hydrolysis product of antibiotics. Secondary assays involving the use of high performance liquid chromatography (HPLC) was also performed to confirm activities of screened variants.

Finally, conclusion and recommendations will be discussed in Chapter 7. This chapter will highlight the contributions of this work and suggest experiments for future investigations into data-driven protein engineering projects, point out another method to evolve an mRFP with brighter spectral emissions, propose intermediate substrates for PGA screening and explore other high throughput fluorimetric assays for PGA.

1.6 References

1. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7877-7882.

2. Shaner, N.C., Steinbach, P.A. and Tsien, R.Y. (2005) A guide to choosing fluorescent proteins. *Nature methods*, **2**, 905-909.

3. Lukyanov, K.A., Chudakov, D.M., Lukyanov, S. and Verkhusha, V.V. (2005) Innovation: Photoactivatable fluorescent proteins. *Nat Rev Mol Cell Biol*, **6**, 885-891.

4. Mena, M.A., Treynor, T.P., Mayo, S.L. and Daugherty, P.S. (2006) Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library. *Nat Biotechnol*, **24**, 1569-1571.

5. Bornscheuer, U.T. and Pohl, M. (2001) Improved biocatalysts by directed evolution and rational protein design. *Current opinion in chemical biology*, **5**, 137-143.

6. Faber, K. (2000) *Biotransformations in organic chemistry : a textbook*. 4th, completely rev. & extended ed. Springer, Berlin ; New York.

7. Burton, S.G., Cowan, D.A. and Woodley, J.M. (2002) The search for the ideal biocatalyst. *Nat Biotechnol*, **20**, 37-45.

8. Arnold, F.H. and Georgiou, G. (2003) *Directed evolution library creation : methods and protocols*. Humana Press, Totowa, N.J.

9. Arnold, F.H. and Georgiou, G. (2003) *Directed enzyme evolution : screening and selection methods*. Humana Press, Totowa, N.J.

10. Reetz, M.T. and Carballeira, J.D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature protocols*, **2**, 891-903.

11. Chaparro-Riggers, J.F., Polizzi, K.M. and Bommarius, A.S. (2007) Better library design: data-driven protein engineering. *Biotechnology journal*, **2**, 180-191.

12. Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C. and Minshull, J. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnology*, **7**, 16.

CHAPTER 2

BACKGROUND: PROTEIN ENGINEERING METHODS

2.1 Improving biocatalyst performance

The performance of enzymes as catalysts in reactions can be improved by changing the solvent, pH, substrates, temperature and by engineering the enzyme via alteration of the DNA sequence coding for the protein. The majority of enzymes are stable in aqueous media of mild pH (1). Activated substrates such as esters or amides can be used to achieve reasonable yields for kinetically controlled chemical reactions (2-5). The effects of ions on protein stability can be characterized by the Hofmeister series (6,7). In brief, some anions such as phosphates and sulphates are stabilizing (kosmotropes) while bromide and chloride destabilize (chaotropes) the proteins possibly due to hydration ion effects on the protein. Proteins are sensitive to the temperature as proteins unfold and eventually become inactivated when exposed to temperatures above its melting point (T_m) or below its cold denaturation temperature (T_c). The kinetic stability and thermal stability can be measured using a residual activity assay after heat deactivation (half-life, $t_{1/2}$) or differential scanning calorimetry (melting temperature, T_m), respectively (8). Finally, modifying the amino acid sequences through mutagenesis can alter the properties of the enzyme or biocatalyst and improve its performance in non-native conditions (9). This powerful method can tailor the proteins for the optimal reaction conditions. This section will discuss the various protein engineering methods that have been developed to date.

2.2 Concept of protein sequence landscape

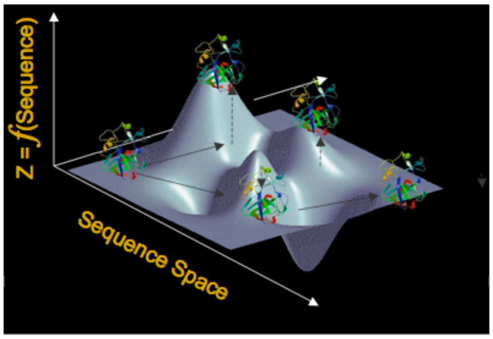


Figure 2.1 Protein fitness landscape. The xy-plane represents the sequence space while the z-axis shows the function of the biocatalyst. Enzymes are shown as colored ribbons and arrows represent mutational trajectories. The peaks and valleys illustrate the complex relationships between sequence and function of the enzyme.

The sequence and the activity of proteins can be visualized as a fitness landscape (10,11). If we plot the sequences of a protein in the two dimensional xy plane, the z-axis can represent the function of the biocatalyst. Figure 2.1 is a hypothetical plot of one such graph. The ribbons represent folded proteins and the arrows show hypothetical protein evolution pathways. The sequence-function or fitness landscape includes valleys that represent unfolded and deactivated proteins while the peaks show proteins with local optima activity. The protein engineering experiments described below traverse the protein landscape in search of a more optimal solution for the reaction at hand. In other words, protein engineers seek to find peaks in the fitness landscape.

2.3 Protein engineering methods

Protein engineering methods can be divided into three groups - rational, combinatorial (directed evolution) and data-driven protein engineering. This section will detail the applications, successes, advantages and pitfalls of various protein engineering methods.

2.3.1 Rational Design

Rational design of proteins usually requires crystal structures and knowledge between the sequence-function as well as mechanism of the catalysis. This method involves knowing what residues to target and replace in the protein to obtained the desired effects. The application of rational design starts with the analysis of the crystal structure. Point mutations are then substituted into the protein via techniques such as site-directed mutagenesis or overlap extension PCR. The mutants are then assayed for improved function. The procedure is low throughput and thus does not require a high through put screen. Rational design has been used successfully on fluorescent proteins to design calcium binding proteins (12), to transform amyloid-like fibrils to monomeric β-sheet proteins (13), to confer triose phosphate isomerase (TIM) activity on ribose binding proteins (RBP) (14), relaxing the nicotinamide cofactor requirement of phosphite dehydrogenase (15) and designing more stable proteins (16).

One good example to highlight the application of rational design method is the work performed on conferring TIM activity on ribose-binding proteins (14,16). TIM catalyzes the interconversion of dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP), constituting an important part of the Embden-Meyerhof pathway (17). Dwyer et al. first tested if RBP can be redesigned to bind to DHAP and GAP. Stereochemical complement ligand-binding surfaces on RBP were identified using a modified dead-end elimination algorithm (DEE) (18,19) that take into account ligand docking and placement of amino side-chain rotamer libraries using van der Waals, hydrogen bonding, solvation and electrostatic energy minimization considerations. The authors found four designs that bind to DHAP and GAP with micromolar affinities that were determined by coupled colorimetric enzyme assays. Subsequently, catalytic activity is introduced by introducing catalytically active residues into the receptor design process. This involves geometrically defining key interactions that contribute to binding (Figure 2.2), identification of positions where placement of catalytic residues and substrate (20) that simultaneously satisfies the geometric constrains and lastly, the complementary surface is generated around the placed substrate. Using the above procedures, Dwyer *et al*. conferred TIM activity to RBP proteins on the order of ~10^5 to 10^6 over background, i.e. k_{cat}/k_{uncat} ~10^5 to 10^6.

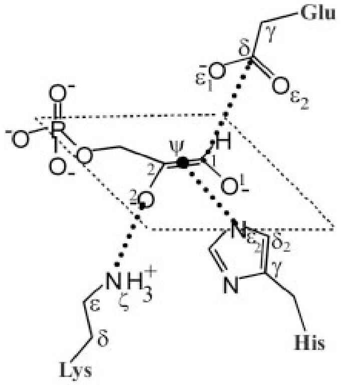
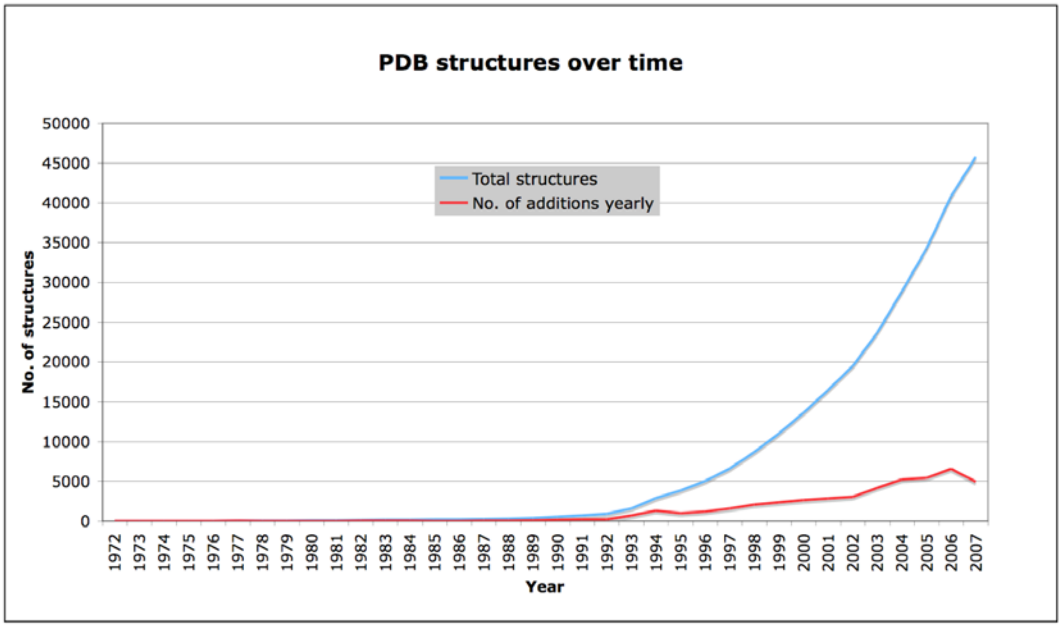


Figure 2.2 Geometric definition used by Dwyer et al for rational design of TIM activity in ribose binding protein. Geometric constraints were obtained in terms of bond lengths, angles and torsions for each residue relative to the enediolate. Figure adapted from (14).

Rational design efforts are limited by the availability of crystal structures and mechanistic knowledge. In 2007, there are currently 46,051 crystal structures with about 10,000 – 20,000 unique sequences (based on BLASTclust, 30% - 95% sequence identity levels) available from the Protein Data Bank (21) that can be used (Figure 2.3). Although there are large numbers of crystal structures in absolute terms, however, relatively speaking, it is small when compared to the diversity of species and number of genes present in each species. Thus, crystal structure data is not always available for the protein studied and not all proteins can also be easily crystallized. One alternative to using crystal structure is homology modeling whereby homologous can be used to make crystal structure for analysis purposes (22). Lastly, mechanistic knowledge of the protein of interest is not always accessible. In the above TIM example, mechanistic knowledge of the TIM isomerization was available. Implicitly, this helped in the selection of residues for substitutions and determination of the type of mutations desired for the rational design of TIM activity in RBP proteins.



*data obtained from Protein Data Bank (21)

Figure 2.3 Growth and total number of PDB structures over time.

2.3.2 Directed evolution

Directed evolution also known as molecular evolution involves the generation of library diversity randomly (> 10^6 variants), which is then followed by a high throughput screening or selection method (23). One key advantage of directed evolution is the lack of requirement for crystal structure data and mechanistic insight. To start with, accessibility of the DNA information is required and a robust high throughput screen. Directed evolution has been used to improve activities and evolve new functionalities for proteins such as esterase, aspartate aminotransferase, biphenyl-dioxygenase, cytochrome P450, penicillin G amidases and monomeric red fluorescent proteins to name a few (9,24-26). Such proteins can have complex sequence-structure relationships that cause rational methods to fail to work. This section discusses the library generation methods and the considerations of library screening.

Random library generation methods can be broadly divided into two types – random mutagenesis and recombination based methods. In short, one conventional random mutagenesis method known as error-prone PCR (ep-PCR) involves the use of manganese ions with skewed deoxyribonucleotides concentrations in the PCR reaction (27). The products resulting from the ep-PCR reaction are prone to erroneous substitution, thus, introducing point mutations into the gene. Another library generation method, recombination, combines the genes to produce chimera variants. In particular, DNA shuffling (28,29) consists of two steps. The genes are first fragmentized using enzymes such as DNAse I and an assembly PCR is done in which fragments of genes form the full length chimeras genes. An additional PCR is performed to amplify more chimeras for cloning.

Library quality is important for success of directed evolution experiments (29). The quality of a library depends on its parental background, sequence and bias in sequence diversity. A library that has high percentage of parental templates and repetitions of non-unique sequences wastes screening efforts and resources. A biased library would limit the sequence-space search for improved variants. All random library generation protocols have inherent bia that can be controlled to some extent (recombination protocols are discussed in Chapter 4).

High throughput selection or screening assays are a pre-requisite for directed evolution experiments (30-32). In general, the selection or screens need to be robust so that it can be easily reproducible and give consistent results. Selection refers to genetic screening methods in which the cells expressing the protein variants either survives or dies depending on the ability of the protein to confer survival advantage to its host cells. Screening methods, on the other hand, involve assays that include color or fluorescence change upon substrate reaction or product formation (Table 2.1).

Table 2.1 Screening methods.

Method	Description	References
Colorimetric	Format: 96-well plates, agar plates, filter-lift Uses: Activity Throughput: High (10^4)	(30,33)
Fluorimetric	Format: 96-well plates, agar plates, fluorescence activated cell sorter Uses: Activity, folding assays, binding Variations: Anchored periplasmic expression (Apex) Throughput: High (10^4 to 10^6)	(34-36)
Phage display	Format: In vitro Uses: Binding Throughput: High (10^4 to 10^6)	(37)
Ribosome display	Format: Agar plates Uses: Stability, binding Throughput: High (10^4 to 10^6)	(38,39)
Gas chromatography (GC)	Format: 96-well plate possible Uses: Assays without known colorimetric or fluorimetric substrates Throughput: Low to medium ($<10^4$)	(40)

One prior work by Zhao and Arnold serves to illustrate the application of directed evolution to evolve thermal stability (41). Subtilisin E genes was subjected to ep-PCR and a recombination method called staggered extension process (StEP) (42). The screen consisted of subjecting the variants to a temperature challenge, which is then followed by a colorimetric assay utilizing succinyl-Ala-Ala-Pro-Phe-p-nitroanilide, improved variants with better thermo-stability were found. The best variant with eight mutations, had a half

–life that was 200 times more than the wild-type $T_{optimal}$ and it is also more active than the wild-type enzyme.

Directed evolution is a useful method but is resource intensive. Tens of thousands of colonies need to be screened before having a good chance to converge at an acceptable solution. It is also not clear if directed evolution would work in all cases of enzyme evolution as enzymes could potentially have evolutionary limits. For example, Miller *et al.* applied directed evolution on β-isopropylmalate dehydrogenase to switch its coenzyme specificity from nicotinamide adenine dinucleotide (NAD) to nicotinamide adenine dinucleotide phosphate (NADP) (43). They found evolutionary constrains on the variants they obtained as the mutants always had lower NADP affinity levels when compared to NAD.

2.3.3 Data-driven protein engineering

Data-driven protein engineering combines bioinformatics, experimental data and computational tools together to evolve proteins in a smarter way, which minimizes experimental work to maximize the probability of obtaining improved variants. There are a couple of variations of data-driven protein engineering to date, but their central theme remains the same – all of them involves creating a targeted library that optimizes the chances of success. One recent review on data-driven protein engineering by Chaparro-Riggers *et al.* (44) classifies the various types of data-driven protein engineering into the following groups listed as follows: structure-based, homology-based, computation-based and combination-based methods. To be consistent with literature, similar terminology and groupings were used in the following paragraphs to illustrate the different types of data-driven protein engineering developed.

2.3.3.1 Structure- and homology- based engineering

Structure-based engineering methods require three-dimensional crystal structures from which intelligent guesses are made to choose the residues for mutation. Morley and Kazlauskas showed that mutations near the active sites would be more useful than random mutations to generate variants with improved enantioselectivity and substrate specificity (45). Figure 2.4 shows the change in activation free energy as a result of the mutations. Evidently, most of the improved variants are centered close to the active sites. These findings suggest that for experiments to change enantioselectivity and substrate specificity, the use of a crystal three dimensional structure coupled with a targeted random mutagenesis method or site-saturation mutagenesis method is a better approach to evolve the protein of interest.

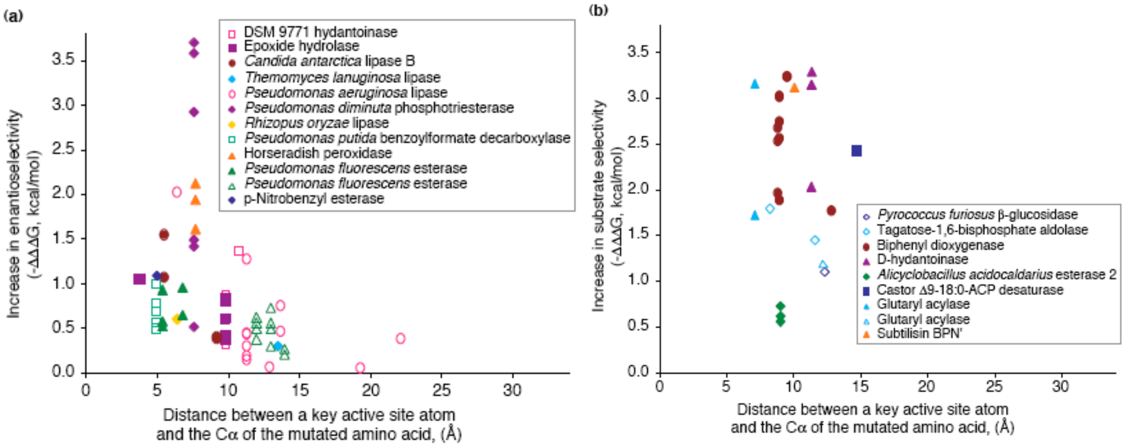


Figure 2.4 (a) Change in free energy for enantioselectivity and (b) Change in free energy for substrate specificity. Figure was adapted from (45).

Structure-based engineering can be combined with homology alignment data to choose residues for mutation. In addition to using the crystal data information, the amino-acids sequence alignment can be utilized to determine residues for mutations. One such method is the consensus approach applied on pencillin G acylase. Polizzi *et al* (46) sought to improve the thermo-stability of penicillin G acylase. They used sequence alignment data on seven variants of penicillin G acylases and reduced the number of mutations from 109 to 21 positions by analyzing the crystal structure data and eliminating mutations close to the active site. They obtained 10 variants with improved thermo-stability without affecting activity significantly.

In some cases, structure-based techniques can be coupled with computational methods and homology data to generate better libraries (47-49). The computational methods for these approaches to identify interacting and important residues include considerations listed as follows: Volumetric, polarity, electrostatics, distances from other amino acids. The libraries can then be designed to minimize incompatibilities. For example, Mena *et al.* evolved a brighter blue fluorescent protein with improved photo-stability and relative quantum yield by first analyzing the crystal structure data and then used a computational algorithm that maximizes preservation of activity to limit size of the library generated which were subsequently screened fluorimetrically.

2.3.3.2 Computational-, Algorithm- and scouting-based engineering

Scouting-based methods can be used to design smart libraries to accelerate convergence towards improved proteins. These methods consist of using sequence-activity data obtained from either parallel experiments or designed single-point mutation experiments – the goal of which is to obtain an initial sequence-activity data on the proteins. Following that, computational analysis involving machine learning algorithms and or linear regression can identify important activity residues and non-linear interactions. Site-saturated libraries can then be designed around these residues, thus, speeding up protein engineering efforts.

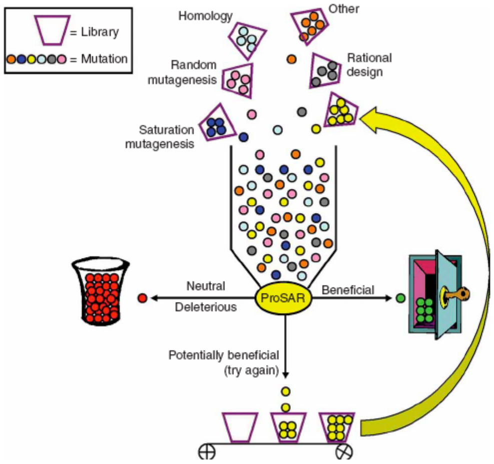


Figure 2.5 Application of ProSAR algorithm. Sequence-activity data is obtained which is then analyzed using ProSAR. The cycle can be iterated to obtained more data until satisfactorily improved variants are obtained. Figure adapted from (50).

One good example to highlight the application of such an approach is the work done by Fox *et. al* on evolving halohydrin dehydrogenase for the production of ethyl (R)-4-cyano-3-hydroxybutyrate, a starting material for producing the cholesterol lowering drug atorvastatin (Lipitor). Figure 2.5 illustrates the methodology of applying the linear regression algorithm, protein sequence activity relationship algorithm (ProSAR). DNA shuffling, random mutagenesis and rational design methods were first used to generate the libraries that were assayed then analyzed using a ProSAR. ProSAR is a statistically based learning method that can infer mutational effects contribution to protein function. Figure 2.6 shows an example of the ProSAR output. The combined method yielded an improved halohydrin dehydrogenase with 4000-fold activity relative to wild-type.

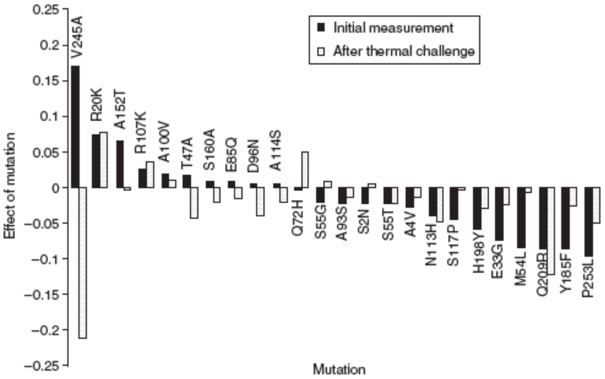


Figure 2.6 Output from ProSAR analysis. The y-axis represents the effects of mutation that is plotted against the type of mutation on the x-axis. The black colored bar represents results taken initially and the clear bars are results taken with thermal challenge. Data obtained from (50).

Although there are a number of different protein engineering approaches, it remains contentious if any particular method is superior. As mentioned above, rational design, directed evolution and data-driven methods all have successful examples of creating improved protein variants. All of them need a library generation method and a good robust screen. The choice of the protein engineering approach or approaches will likely be dependent on the type of protein, the amount of literature on the protein, the type of assays that can be used and the resources available to the protein engineers. To expand understanding of library generation utility on proteins with different homology levels and to validate the applicability of a machine learning language, Boolean Learning Support Vector Machines (BLSVM) for use in guiding directed evolution, this dissertation will report on the performance of recombination and the results of using BLSVM on fluorescent proteins. As an extension, we also looked into evolving monomeric Red Fluorescent Proteins using directed evolution and a structure-guided approach to evolving a variant of penicillin G acylase with an altered substrate specificity.

2.4 References

1. Voet, D. and Voet, J.G. (2004) *Biochemistry*. 3rd ed. John Wiley & Sons, Inc.

2. Faber, K. (2000) *Biotransformations in Organic Chemistry : a Textbook*. 4th, completely rev. & extended ed. Springer, Berlin ; New York.

3. Schellenberger, V., Schellenberger, U., Jakubke, H.D., Zapevalova, N.P. and Mitin, Y.V. (1991) Protease-Catalyzed Peptide-Synthesis Using Inverse Substrates - the Synthesis of Pro-Xaa-Bonds by Trypsin. *Biotechnology and Bioengineering*, **38**, 319-321.

4. Jakubke, H.D., Kuhl, P. and Konnecke, A. (1985) Basic Principles of Protease-Catalyzed Peptide-Bond Formation. *Angewandte Chemie-International Edition in English*, **24**, 85-93.

5. Bommarius, A.S. and Riebel, B.R. (2004) *Biocatalysis - Fundamentals and Applications*. 1st ed. Wiley-VCH Verlag GmbH & Co., Weinheim.

6. Broering, J.M. Thesis (2006).

7. Hofmeister, F. (1888) Zur Lehre von der Wirkung der Salze. Zweite Mitteilung. *Arch. Exp. Pathol. Pharmakol.*, **24**, 247-260.

8. Büchner, J. and Kiefhaber, T. (2004) *Protein folding handbook*. Wiley-VCH, Weinheim.

9. Arnold, F.H. and Volkov, A.A. (1999) Directed evolution of biocatalysts. *Current opinion in chemical biology*, **3**, 54-59.

10. Fox, R., Roy, A., Govindarajan, S., Minshull, J., Gustafsson, C., Jones, J.T. and Emig, R. (2003) Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng.*, **16**, 589-597.

11. Arnold, F.H. (1998) Design by directed evolution. *Accounts of Chemical*

Research, **31**, 125-131.

12. Yang, W., Jones, L.M., Isley, L., Ye, Y., Lee, H.W., Wilkins, A., Liu, Z.R., Hellinga, H.W., Malchow, R., Ghazi, M. *et al.* (2003) Rational design of a calcium-binding protein. *Journal of the American Chemical Society*, **125**, 6165-6171.

13. Wang, W. and Hecht, M.H. (2002) Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. *Proc Natl Acad Sci U S A*, **99**, 2760-2765.

14. Dwyer, M.A., Looger, L.L. and Hellinga, H.W. (2004) Computational design of a biologically active enzyme. *Science (New York, N.Y*, **304**, 1967-1971.

15. Woodyer, R., van der Donk, W.A. and Zhao, H. (2003) Relaxing the nicotinamide cofactor specificity of phosphite dehydrogenase by rational design. *Biochemistry*, **42**, 11604-11614.

16. Sharma, D., Perisic, O., Peng, Q., Cao, Y., Lam, C., Lu, H. and Li, H. (2007) Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proc Natl Acad Sci U S A*, **104**, 9278-9283.

17. Fraenkel, D.G. (1986) Mutants in glucose metabolism. *Annual review of biochemistry*, **55**, 317-337.

18. Looger, L.L., Dwyer, M.A., Smith, J.J. and Hellinga, H.W. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185-190.

19. Looger, L.L. and Hellinga, H.W. (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *Journal of molecular biology*, **307**, 429-445.

20. Hellinga, H.W. and Richards, F.M. (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. *Journal of Molecular Biology*, **222**, 763-785.

21. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

22. Kazlauskas, R.J. (2000) Molecular modeling and biocatalysis: explanations, predictions, limitations, and opportunities. *Current opinion in chemical biology*, **4**, 81-88.

23. Bornscheuer, U.T. and Pohl, M. (2001) Improved biocatalysts by directed evolution and rational protein design. *Current opinion in chemical biology*, **5**, 137-143.

24. Zhou, Z., Zhang, A.-H., Wang, J.-R., Chen, M.-L., Li, R.-B., Yang, S. and Yuan, Z.-Y. (2003) Improving the Specific Synthetic Activity of a Pencillin G Acylase Using DNA Family Shuffling. *ACTA BIOCHIMICA et BIOPHYSICA SINICA*, **35**, 573-579.

25. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7877-7882.

26. Joo, H., Lin, Z. and Arnold, F.H. (1999) Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation. *Nature*, **399**, 670-673.

27. Huimin Zhao, J.C.M., Alex A. Volkov, Frances H. Arnold. (1999) In Arnold L. Demain, J. E. D. (ed.), *Industrial Microbiology and Biotechnology*. 2nd ed. American Society for Microbiology Press.

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

29. Arnold, F.H. and Georgiou, G. (2003) *Directed evolution library creation : methods and protocols*. Humana Press, Totowa, N.J.

30. Arnold, F.H. and Georgiou, G. (2003) *Directed enzyme evolution : screening and selection methods*. Humana Press, Totowa, N.J.

31. Zhao, H. and Arnold, F.H. (1997) Combinatorial protein design: strategies for

screening protein libraries. *Current Opinion in Structural Biology*, **7**, 480-485.

32. Olsen, M., Iverson, B. and Georgiou, G. (2000) High-throughput screening of enzyme libraries. *Current Opinion in Biotechnology*, **11**, 331-337.

33. del Rio, G., Rodriguez, M.-E., Munguia, M.-E., Lopez-Munguia, A. and Soberon, X. (1995) Mutant Escherichia coli penicillin acylase with enhanced stability at alkaline pH. *Biotechnology and Bioengineering*, **48**, 141-148.

34. Schuster, S., Enzelberger, M., Trauthwein, H., Schmid, R.D. and Urlacher, V.B. (2005) pHluorin-based in vivo assay for hydrolase screening. *Analytical chemistry*, **77**, 2727-2732.

35. Griswold, K.E., Kawarasaki, Y., Ghoneim, N., Benkovic, S.J., Iverson, B.L. and Georgiou, G. (2005) Evolution of highly active enzymes by homology-independent recombination. *Proc Natl Acad Sci U S A*, **102**, 10082-10087.

36. Wang, L., Jackson, W.C., Steinbach, P.A. and Tsien, R.Y. (2004) Evolution of new nonantibody proteins via iterative somatic hypermutation. *Proc Natl Acad Sci U S A*, **101**, 16745-16749.

37. Kay, B.K., Winter, J. and McCafferty, J. (1996) *Phage display of peptides and proteins : a laboratory manual*. Academic Press, San Diego.

38. Mattheakis, L.C., Bhatt, R.R. and Dower, W.J. (1994) An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci U S A*, **91**, 9022-9026.

39. Hanes, J. and Pluckthun, A. (1997) In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A*, **94**, 4937-4942.

40. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S. *et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol*, **25**, 338-344.

41. Zhao, H. and Arnold, F.H. (1999) Directed evolution converts subtilisin E into a functional equivalent of thermitase. *Protein engineering*, **12**, 47-53.

42. Zhao, H., Giver, L., Shao, Z., Affholter, J.A. and Arnold, F.H. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat Biotechnol*, **16**, 258-261.

43. Miller, S.P., Lunzer, M. and Dean, A.M. (2006) Direct demonstration of an adaptive constraint. *Science (New York, N.Y*, **314**, 458-461.

44. Chaparro-Riggers, J.F., Polizzi, K.M. and Bommarius, A.S. (2007) Better library design: data-driven protein engineering. *Biotechnol. J.*, **2**, 180-191.

45. Morley, K.L. and Kazlauskas, R.J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol*, **23**, 231-237.

46. Polizzi, K.M., Chaparro-Riggers, J.F., Vazquez-Figueroa, E. and Bommarius, A.S. (2006) Structure-guided consensus approach to create a more thermostable penicillin G acylase. *Biotechnol. J.*, **1**, 531-536.

47. Saraf, M.C. and Maranas, C.D. (2003) Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Eng.*, **16**, 1025-1034.

48. Meyer, M.M., Silberg, J.J., Voigt, C.A., Endelman, J.B., Mayo, S.L., Wang, Z.G. and Arnold, F.H. (2003) Library analysis of SCHEMA-guided protein recombination. *Protein. Sci.*, **12**, 1686-1693.

49. Treynor, T.P., Vizcarra, C.L., Nedelcu, D. and Mayo, S.L. (2007) Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl. Acad. Sci. U S A*, **104**, 48-53.

50. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S. *et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nature biotechnology*, **25**, 338-344.

CHAPTER 3

CLONING AND EVOLVING MRFP

3.1 Introduction

Fluorescent proteins are an important class of proteins for bio-imaging studies, expression monitoring, protein-interaction studies and other functions such as calcium ion biosensors and photo-sensitizers (1-8). Most fluorescent proteins consist of a β-can structure that protects the chromophore from quenching by oxygen and keeps the angle of the chromophore moiety at slightly less than 180° (9,10). As a consequence, fluorescent proteins can also serve as folding reporters since folding is essential for fluorescence to occur. The utility of the fluorescent proteins depend on their photophysical characteristics. Spectral considerations suggest that red fluorescent proteins have better signal to noise ratios than green fluorescent proteins (8). In particular, the monomeric Red Fluorescent Protein (mRFP) is a useful tag-fusion fluorescent protein as it is a monomeric protein with higher signal to noise ratio (Figure 3.1) than Green Fluorescent Proteins (GFPs). However, it suffers from low relative quantum yield of 0.25 and a more red-shifted mRFP from a current maximum at 610 nm is desirable to reduce background noise in the presence of red blood cells. This chapter describes protein engineering effort aimed at preparing an improved variant of mRFP with better quantum yield and more red-shifted wavelength.

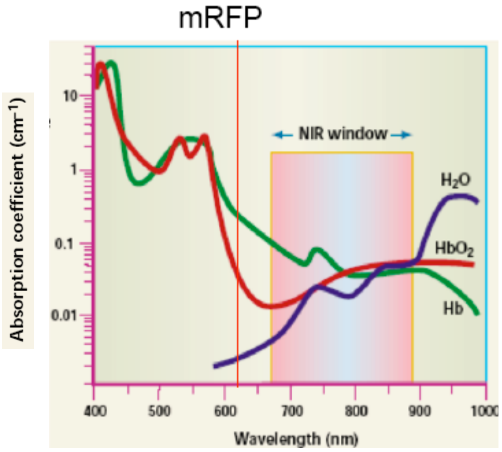


Figure 3.1 Absorption spectral of water, oxygenated haemoglobin and haemoglobin. Adapted from reference (8).

3.1.1 Fluorescent proteins origin

Most of the known fluorescent proteins are derived from organisms living in the ocean. To date, fluorescent proteins have been found and cloned from sea-dwelling organisms such as *Aequorea victoria* jellyfish (11) and corals such as *Anthozoa* (12), *Pectinidae* (13) and *Discosoma species* (14). The GFP protein from *Aequorea victoria* jellyfish was the first fluorescent protein isolated. In 1979, Shimomura deduced the structure of the GFP (15,16), which eventually led to a landmark paper on expression of GFP by Chalfie *et al* in 1994 (17). Since then, other fluorescent protein variants were discovered and engineered, they became, along with GFP, commonly used tools in molecular biology, medicine, and cell biology (9). However, despite many years of research, the function of these fluorescent proteins in these sea-dwelling organisms remains undetermined (4).

3.1.2 Properties of fluorescent proteins

Table 3.1 Lists of selected fluorescent proteins and their applications.

Name	Use(s)	References
DsRed and GFPs, mGFP, mRFP, mCherry, mPlum, mStrawberry, mBanana	Bioimaging, protein-protein interactions, reporter, screening for folded proteins	(19)
CFP and YFP, CFP and YFP	FRET, detection of conformational changes	(18)
Calmodulin linked GFP	Calcium detections	(21-23)
DsRed	Cu^{+} and Cu^{2+} detections	(24)
EGFP, pHluourin	Ratiomeric detection of pH	(25,26)
HyPer	Hydrogen Peroxide detection	(27)
KillerRed	Photosensitizer*	(1)

*Photosensitizer refers to a compound that produces reactive oxygen species upon light excitation

The utility of fluorescent proteins for various applications is dependent on its aggregation tendency, quantum yield, extinction coefficient, photostability, excitation and emission spectral range as well as sensitivity to environmental effects such as pH (18,19). Some fluorescent proteins have the tendency to associate, forming dimers and/or tetramers and dimers, which can prove fatal to cells if it induces aggregation of proteins. The emission intensity of the fluorescent protein is a function of the aggregative state of the proteins, the quantum yield and the extinction coefficient (20). The photostability of the fluorescent protein is a measure of the ability of the fluorescent protein to withstand photobleaching due to destruction of the fluorophore after repeated rounds of excitation and emission (19). The excitation and emissions spectral range of the fluorescent species will determine its application for use in combination with other fluorescent probes for protein-protein interactions investigations and simultaneous tracking of movement of different proteins and cells simultaneously. Some fluorescent proteins are sensitive to pH and thus can be used for pH detection. Examples include pHluorine and EGFP which have been used for measurement of pH *in vivo* and detection of hydrolytic activity (see Table 3.1).

3.1.3 mRFP protein: a useful monomeric red fluorescent protein

The monomeric red fluorescent protein (mRFP) protein is a more useful fluorescent protein than GFPs. The mRFP protein experiences less signal to background noise than GFPs which area frequently used in labs, which are frequently used in labs. The use of GFPs in mammalian systems is disadvantaged as its emissions range from 540 nm to 590 nm which is absorbed by red blood cells (8). Another advantage of mRFP is its monomericity that confers reduced tendency to aggregate when expressed as a tag-fusion protein (28).

The mRFP was engineered from DsRed through directed evolution combined with rational design. Campbell *et al.* engineered the mRFP from the tetrameric DsRed protein by using crystal structures of DsRed protein and introducing arginine mutations into interacting residues I125, T147, H162 and F224 to disrupt the sub-units (28). The fluorescence of the mRFP was then rescued by directed evolution, a process that involves the generation of a large number of mutants using random mutagenesis which is then screened using a cooled charged coupled camera with 560 nm excitation and 610 nm bandpass filters. In total, 33 mutations were introduced into DsRed protein. Of these mutations, 16 mutations were reducing aggregation while 17 mutations were beneficial for brightness and maturation.

3.1.4 Improved mRFP with red-shifted emissions and better relative quantum yield desired

The mRFP protein was one of the few rapidly maturing (< 1hr), red-shifted monomeric fluorescent that can be useful for bioimaging and labeling experiments. However, the mRFP was not red-shifted enough (607 nm emissions) and suffers from a low relative quantum yield (0.26) in comparison to GFPs with relative quantum yields of 0.79 (4). Weissleder (8) suggested that fluorescent probes should ideally have emissions in the near-infrared red region, from ~650 nm to ~850 nm range (Figure 3.1). Most fluorescent probes currently have emissions that suffer from high signal to background noise as a result of absorbance by red blood cells. Thus, to make a more useful mRFP, we decided to engineer mRFP through directed evolution.

3.1.5 Scope of work

The engineering of mRFP towards a more red-shifted wavelength using directed evolution will be described in this chapter. The mRFP library was generated using error-prone PCR (ep-PCR) and the library was screened using fluorescence activated cell sorter. The variants obtained were purified using metal-chelation chromatography and the relative quantum yields were measured.

3.2 Materials & Methods

3.2.1 Materials

Most of the enzymes were bought from New England Biolabs (Beverly, MA). The *Pfu* and *Taq* polymerase were purchased from Stratagene (La, Jolla, California). XL1-Red *E. coli* cells were bought from Stratagene(La, Jolla, California). Tetracycline and chloroamphenicol were purchased from Sigma (St. Louis, Missouri). All oligonucleotides were ordered from MWG-Biotech (High Point, North Carolina). The inducer, anhydrotetracycline (ATC) was made by dissolving 250 mg/l of tetracycline in water and adjusted to pH 3 and autoclaving for 45 mins. Mass spectrometry confirmed that anhyrotetracyline was obtained (See supplement).

3.2.2 Synthesizing codon-optimized mRFP protein

The amino acid sequence of mRFP was obtained from NCBI and *E. coli*-codon optimized primers (supplement 3.7.1) were designed using DNAworks (29) and synthesized. The mRFP gene was obtained via two PCR reactions, one to assemble the codon optimized primers, and the second to amplify the full length product. The gene was amplified using primers with *Esp*3I (italicized) restriction sites (5' -TA*C GTC TCG* TCG ACA TGG CGT CTT CTG AAG ACG TTA TCA AAG AAT TCA TGC GT – 3' and 5' – TA*C GTC TCT* GGC CTA TTA CGC ACC GGT AGA GTG ACG ACC TTC - 3') and digested with *Esp*3I enzyme and ligated using T4 DNA ligase into *Sal*I and *Not*I digested pPROTet vector. Sequencing, expression and characterization consistent with the literature confirmed that the *E. coli* expression optimized mRFP gene was successfully assembled (28).

3.2.3 Creating variants

The error-prone PCR (ep-PCR) protocol from Zhao *et al* was followed. The following solutions were made. 10X Mutagenic buffer consisting of 70 mM $MgCl_2$, 500 mM KCl, 100 mM Tris and 0.1% (wt/vol) gelatin was used. A solution of 5 mM $MnCl_2$ and a dNTP mix consisting of 2 mM dGTP, dATP and 10 mM of dCTP and dTTP. The final PCR mix consists of 10 μl of 10X mutagenic buffer, 30~50 pmol of each primer, 2 fmol of template DNA and an amount of distilled H_2O to bring the final volume to 96 μl. 3 μl of 5 mM $MnCl_2$ is then added to the PCR solution which is mixed either by pipetting or flicking. 1 μl of *Taq* polymerase is added (~5 U/l). The PCR mixture is cycled as follows: 14 cycles of 30 s at 94°C, 30s at 50°C, and 30 s at 72°C. The products were gel purified and digested with the *Sal*I and *Not*I enzymes and cloned into same restriction sites in pPROTeT.E133 vector.

A mutator *E. coli* cell strain, XL1-Red was also used to generate error-prone variants according to the product manual. (Catalog # 200129, Revision #064003, Stratagene). Briefly, the mRFP template in pPROpTeT plasmid was transformed into XL1-Red cells. Approximately 200 colonies were picked and grown in ~5-10 ml of LB broth overnight at 37°C. The colonies were then minipreped and retransformed into XL1-blue cells, which were then picked for screening.

3.2.4 Expression and normalization of cell expression

Freshly transformed cells were picked from LB choramphenicol agar plates and grown in 96-well plates. Each well had 200 μl of LB chloramphenicol in 96-well plates overnight at 37°C and 150 rpm till stationary phase. Each plate had a perimeter of noninoculated wells to mediate evaporation and mRFP positive control cells. Following that, 20 μl of culture was then transferred to another 96-well plate with 180 μl of LB chloramphenicol and inducer, ACT in each well. The final concentrations of ACT and CM in each well were 1 μg/ml and 20 μg/ml respectively. The cultures were then incubated at 30°C at 150 rpm until readings were taken ~30 hrs later using the Fluostar Galaxy fluorimeter with excitation filters and emission filters set at 540 nm and 590 nm respectively.

3.2.5 Batch purification of His-Tagged mRFP

In general, protocols from PROTet^{TM} 6xHN Bacterial Expression System User Manual and BD TALON^{TM} Metal Affinity User Manual were followed. Freshly transformed colonies were picked and grown overnight in 5 ml LB CM (20 μg/ml) at 37°C and 150 rpm. The overnight culture was transferred to fresh LB CM media and diluted 1:50 times the next day. The culture was incubated at 37°C for ~4 hrs, when the culture reached an OD_{600} of 0.5, it was induced by adding ACT to a final concentration of 1 μg/ml. After 24hrs, the culture was harvested by centrifugation at 1000-3000 x g for 15 min at 4°C. The cell pellet was resuspended by vortexing in 5 ml of 1X Extraction/Wash Buffer (4°C, 50mM $NaPO_4$, 300mM NaCl, pH 7) per 500 ml of culture. The cell extract was sonicated on ice for 6 x 30 s. The cell extract was centrifuged at 10,000-12,000 x g for 20 min at 4°C. The supernatant was then transferred over to a clean tube. About 1 to 5ml of the re-equilibrated BD TALON cobalt resin was added to bind the his-tagged proteins. After mixing the resin with the clarified cell extract for 20 min, the suspension was spun down at 700 x g at 4°C for 5 mins. The supernatant was removed and the resin washed twice with 10-20 bed volumes of 1X Extraction/Wash buffer. Each wash consists of 10 min on the platform shaker and centrifugation of the suspension at 700 x g at 4°C for 5 mins. 5 bed volumes of the elution buffer (50 mM $NaPO_4$, 300mM NaCl, 150mM imidazole, pH 7) were then added to the column to obtain the purified proteins. The eluted proteins were dialyzed (30,000 MWCO) in a 1 L solution of 1 x Extraction/Wash buffer overnight at 4°C. Protein gels confirmed that > 95% pure mRFP was obtained (see supplement).

3.2.6 Measuring relative quantum yield

The emission spectral of equally absorbing species of Rhodamine 101 and fluorescent protein were taken using a Photon technology fluorimeter (Model 814, PTI international, Birmingham, New Jersey). The relative areas under the emissions curve were determined for mRFP to Rh101 – this gave the relative quantum yield.

3.2.7 Fluorescence activated cell sorting (FACs) of variants

Freshly transformed cells were grown overnight and induced with ACT (1 μg/ml final concentration) after 18 hrs. The cultures were spun down and resuspended in sterile phosphate buffer saline, pH 7.5 and placed on ice. FACs sorting was done using a BD immunocytometry system (Model: FACSVantage SE DIVA, San Jose, California) in combination with the excitation/emission filters. APC-A: 647 nm/675 nm, FL-6: 488 nm/ 660 nm and PerCp-Cy5-5: 488 nm/ 630 nm. Colonies with the highest intensity were sorted for. 1 ml of sterile PBS were added to the sorted variants. The variants were then plated on LB CM plates and incubated at 30°C overnight.

3.3 Results

3.3.1 Characterization of cloned mRFP

3.3.1.1 Synthesis and sequencing of the mRFP gene

The full-length gene of mRFP protein was obtained via gene synthesis. This involved two PCR steps, of which the first PCR involved gene assembly and the secondary PCR amplified full-length product. The DNA electrophoresis gel pictures for both the PCRs are shown below in Figure 3.2. Gel A shows the DNA gel electrophoresis of the primary PCR in which oligonucleotides were assembled together to form the full-length mRFP gene. It can be observed that lanes four, five and six show a PCR band the size of ~700 bp. Gel B shows the secondary PCR of the primary PCR bands. PCR bands were around the size of 700 bp which is the expected gel band that should be obtained from the mRFP full-length gene.

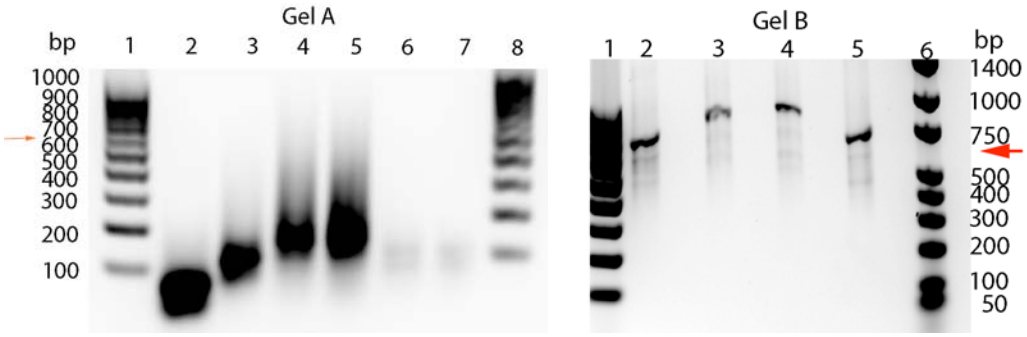
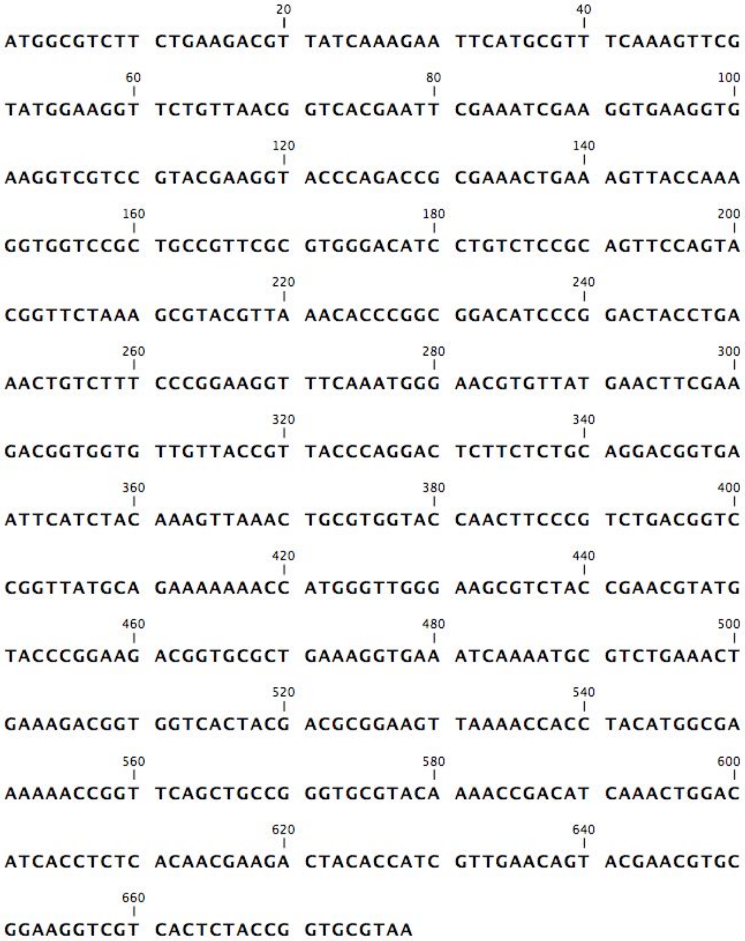


Figure 3.2 DNA gel electrophoresis of primary PCR product and secondary PCR product. Lanes: A1 and B1: 100 bp Perfect DNA Ladder. Lanes B6 and B10: 0.05 – 10 kbp Perfect DNA ladder. Lanes A2 to A7: primary PCR with various primer concentrations. Lanes B2 to B5: secondary PCR from primary PCR as a template. Red arrows mark expected gene size of 678 bps.

The mRFP gene was cloned into cloning vector, pDrive plasmid, and sent for sequencing to confirm that we obtained the mRFP full-length gene. Figure 3.3 shows the obtained DNA sequence of the mRFP gene, which matches the expected sequence, and Table 3.2 shows the expected amino acid sequence of the recombinant mRFP. The sequenced verified full-length mRFP gene is then cloned into pPROTet.E133 plasmid, transformed into DH5αpro *E. coli* cells, induced using anhydrotetracycline to express the mRFP protein. Evidently, we were successful in synthesizing the mRFP gene since a PCR band around 600 to 700 bp was obtained and the expressed protein is fluorescent *in vivo* (see Figure 3.4).



```
                            20                            40
ATGGCGTCTT CTGAAGACGT TATCAAAGAA TTCATGCGTT TCAAAGTTCG
        60                            80                   100
TATGGAAGGT TCTGTTAACG GTCACGAATT CGAAATCGAA GGTGAAGGTG
                           120                           140
AAGGTCGTCC GTACGAAGGT ACCCAGACCG CGAAACTGAA AGTTACCAAA
       160                           180                   200
GGTGGTCCGC TGCCGTTCGC GTGGGACATC CTGTCTCCGC AGTTCCAGTA
                           220                           240
CGGTTCTAAA GCGTACGTTA AACACCCGGC GGACATCCCG GACTACCTGA
       260                           280                   300
AACTGTCTTT CCCGGAAGGT TTCAAATGGG AACGTGTTAT GAACTTCGAA
                           320                           340
GACGGTGGTG TTGTTACCGT TACCCAGGAC TCTTCTCTGC AGGACGGTGA
       360                           380                   400
ATTCATCTAC AAAGTTAAAC TGCGTGGTAC CAACTTCCCG TCTGACGGTC
                           420                           440
CGGTTATGCA GAAAAAAACC ATGGGTTGGG AAGCGTCTAC CGAACGTATG
       460                           480                   500
TACCCGGAAG ACGGTGCGCT GAAAGGTGAA ATCAAAATGC GTCTGAAACT
                           520                           540
GAAAGACGGT GGTCACTACG ACGCGGAAGT TAAAACCACC TACATGGCGA
       560                           580                   600
AAAAACCGGT TCAGCTGCCG GGTGCGTACA AAACCGACAT CAAACTGGAC
                           620                           640
ATCACCTCTC ACAACGAAGA CTACACCATC GTTGAACAGT ACGAACGTGC
       660
GGAAGGTCGT CACTCTACCG GTGCGTAA
```

Figure 3.3 DNA sequence of the synthesized mRFP gene.

Table 3.2 Comparison of amino-acid sequence of mRFP from NCBI and the histidine tagged mRFP protein sequence expressed from the recombinant mRFP pPROTeT.E133 plasmid.

Description	**Amino acid sequence**
Amino-acid of mRFP sequence from NCBI: AAM54544	MASSEDVIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKG GPLPFAWDILSPQFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGG VVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPVMQKKTMGWEASTERMYPE DGALKGE IKMRLKLKDG GHYDAEVKTTYMAKKPVQLPGAYKTDIKLD ITSHNEDYTI VEQYERAEGR HSTGA
mRFP expressed in pPropTeT	HNHNHNHNHNHNGGDDDDKVVDMASSEDVIKEFMRFKVRMEGSVNGHEF EIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDILSPQFQYGSKAYVKHPADI PDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPS DGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGGHYDAEVKTT YMAKKPVQLPGAYKTDIKLDITSHNEDYTIVEQYERAEGRHSTGA

*underlined portion shows the histidine tag along with the linker region to the mRFP protein

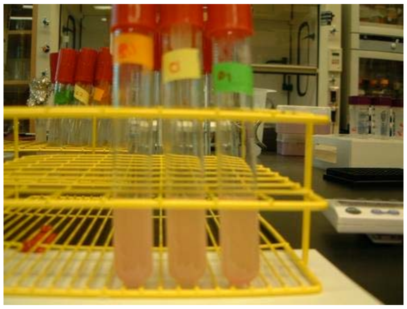


Figure 3.4 Figure showing the expressed recombinant N-terminal His-Tagged mRFP expressed DH5apro cells using pPropTET.E133 plasmid.

The expression of mRFP was monitored over a period of 70 hours. The recombinant DH5αpro cells was first grown overnight and then transferred in 1:200 ratio to fresh LB containing chloramphenicol and induced with ACT when the O.D of cells reach 0.5. Figure 3.5 shows the expression curve for mRFP recombinant cells. It can be determined that the optimal expression time is approximately 24 hours and the optimal concentration of ACT to induce expression is 1 μg/ml.

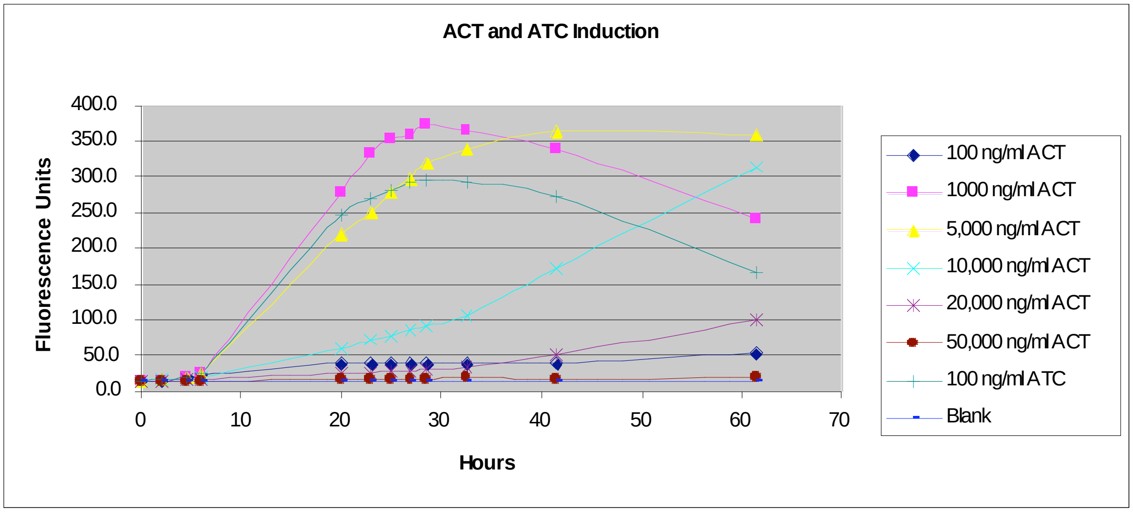


Figure 3.5 Induction with autoclaved tetracycline (ACT) concentrations ranging from 100 to 50,000 ng/ml and 100 ng/ml of anhydrotetracycline (ATC). The optimal inducer concentration was determined to be 1000 ng/ml for ACT.

Purified mRFP was obtained from metal-chelation chromatography of a 1 L batch culture. From a protein gel picture (Figure 3.6), we could determine that up to 20% of overall protein production was mRFP protein. 95% purity of mRFP protein could be obtained from metal-chelation purification. We managed to obtain 30 mg purified mRFP from an estimated 100 mg of mRFP per 1L batch culture.

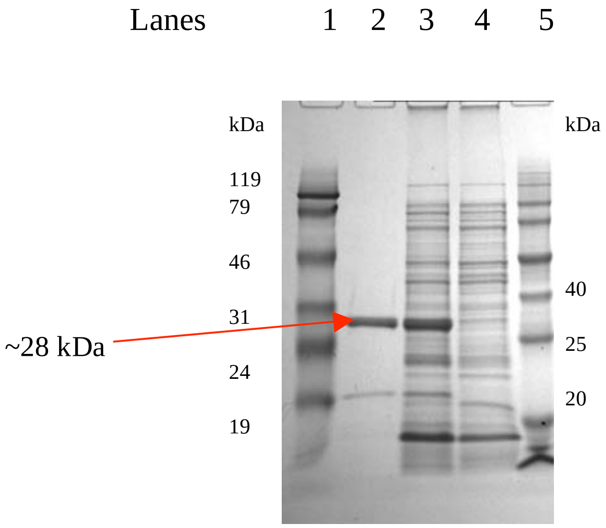
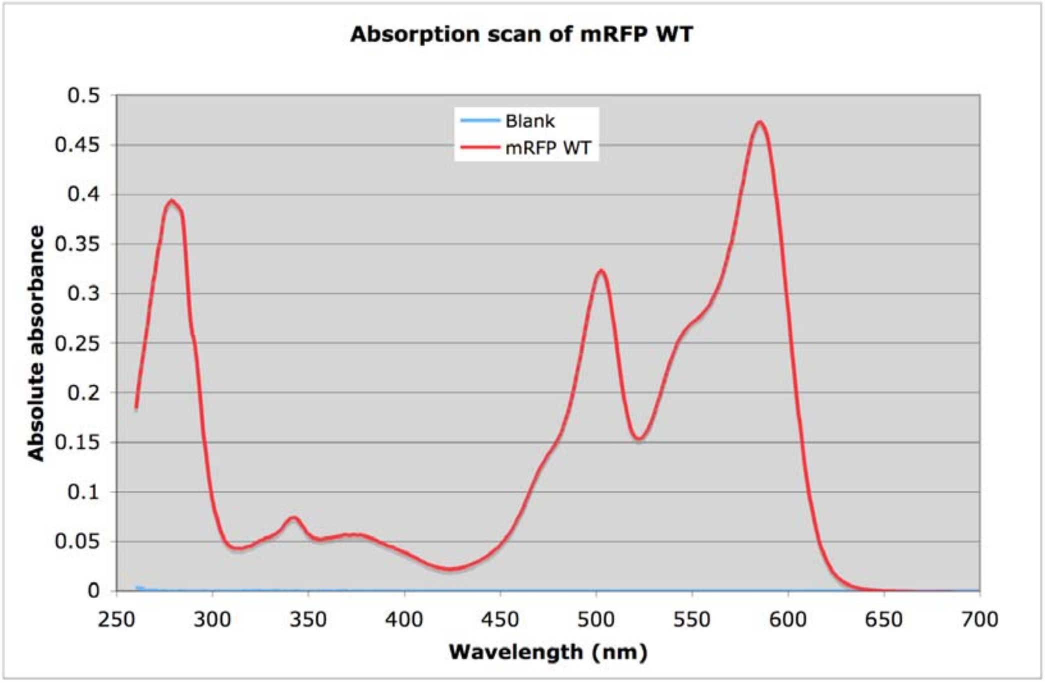


Figure 3.6 Protein gel of mRFP. 1 & 5: Fermentas unstained and Prosieve unstained ladders. 2: Purified mRFP (arrow), 3: mRFP cell lysate induced & 4: Cell lysate uninduced.

3.3.1.2 Spectroscopic scan and measurement of relative quantum yield

The absorption spectrum, emission spectrum, and relative quantum yield of mRFP were found to be consistent with literature (28). The absorption and emissions spectral are shown in Figure 3.7. The mRFP protein has an excitation maximum at 584 nm and an emission maximum at 607 nm. The emission spectral (Figure 3.8) of equally absorbing species of mRFP and rhodoamine 101 were integrated and the ratio of the integration was calculated to obtain the relative quantum yield of mRFP to rhodamine 101. The relative quantum yield of mRFP to rhodamine 101 was found to be 0.25.



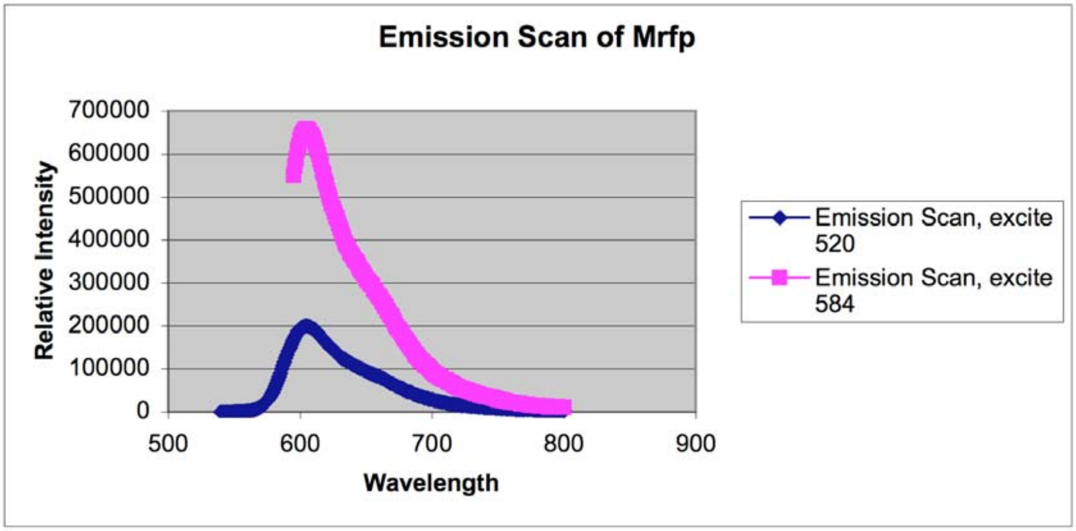


Figure 3.7 Absorption (top) and emissions spectral (bottom) of mRFP protein

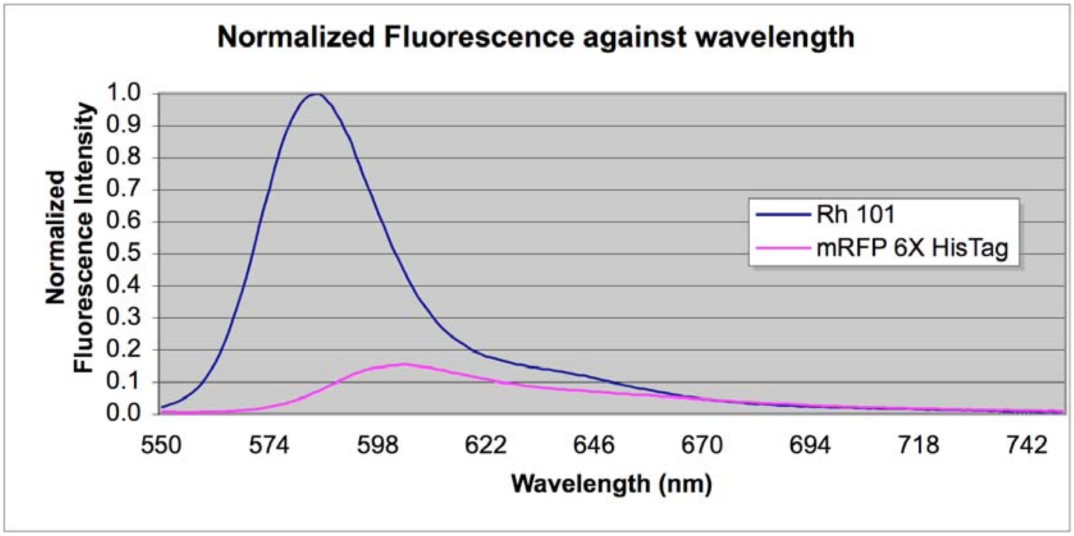


Figure 3.8 Emissions curve of rhodamine 101 (ethanol solution) and 6x His-Tagged mRFP (Na_2HPO_4/NaH_2PO_4 buffer, pH 7.0). The relative quantum yield of mRFP was calculated to be 0.25.

3.3.2 Generation of mutants of mRFP using error-prone PCR and XL1-Red mutator cells

Error-prone PCR was used to create variants of mRFP. In error-prone PCR, one can vary the rate of mutation by adjusting the concentration of $MnCl_2$ or starting template concentration in the PCR mix. The increase of $MnCl_2$ concentration should in theory also increase the rate of mutation. Decreasing the template concentration used for ep-PCR will also increase the rate of mutation as there will be less wild-type template relative to error-prone amplified templates for PCR. The ep-PCR amplified templates serve as new templates for further amplification, thereby increasing the rate of mutation.

To estimate the mutation frequency, we expressed 96 samples of mRFP variants along with the wild-type mRFP and sequenced ep-PCR variants that were generated using various $MnCl_2$ concentrations. A convenient means to estimate the rate of mutation is to use activity screens (30). An increase of the rate of mutation correlates with an increase of deactivation due to the accumulation of deleterious mutations. As it can be observed in Figure 3.9, an increase in the $MnCl_2$ concentration corresponds to a decrease of activity as the number of variants that are active (> 0.2 normalized fluorescence). We also sequenced ep-PCR variants to confirm the effect of varying $MnCl_2$ on rate of mutation (Table 3.3).

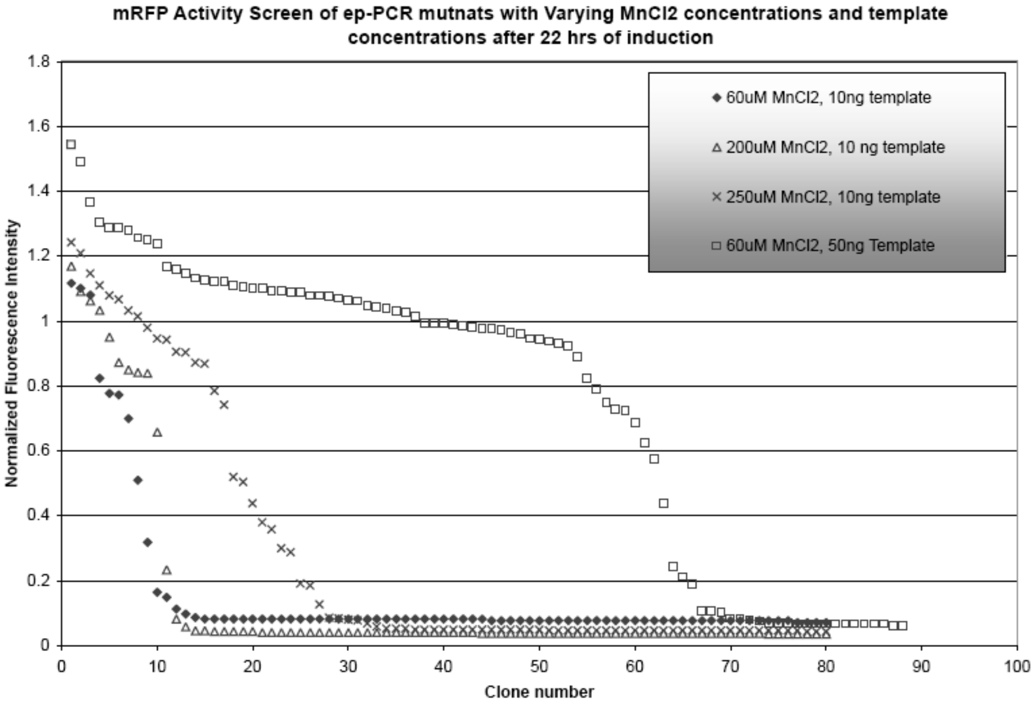


Figure 3.9 Fluorescence of mRFP ep-PCR mutants.

Table 3.3 Sequence-activity data on selected ep-PCR variants. ep-PCR variants produced using 25 ng of template and varied concentrations of MnCl$_2$.

[MnCl2] (uM)	Mass DNA template (ng)	Colony ID	Relative Fluor.	Number of DNA mutations	Number of Amino Acid Mutations	Mutation Frequency (%)	Amino Acid Substitutions
60	25	B03	1.18	1	1	0.147	K184N
	25	C01	1.2	0	0	0.000	-
	25	C04	1.21	1	0	0.147	-
	25	E01	1.24	0	0	0.000	-
	25	H10	0.07	6	5	0.885	S3F, E4G, D5G, I73 (silent), G23R
	25	H01	1.11	3	2	0.442	T201S, N205K
200	25	F01	1.06	0	0	0.000	-
	25	B05	0.06	5	5	0.737	K122 Stop, M149V. D205N, E210Stop, R214S
	25	E05	1.126	1	0	0.147	-
	25	C12	0.98	0	0	0.000	-
	25	E01	0.15	4	4	0.590	D80G, V186D, N205K, D207E
	25	C05	0.99	0	0	0.000	-
	25	G07	1.12	1	1	0.147	-

We also transformed the ep-PCR plasmids and WT mRFP plasmids into XL1-Red mutator cells to generate variants. XL1-Red mutator cells have a deficient DNA replication system that introduces mutations in the plasmids. Thus, we effectively performed two rounds of mutations by using the ep-PCR variants as starting point and one round of ep-PCR by using WT plasmids as the template for XL1-Red cells, respectively. The variants were also screened using FACs.

3.3.3 Screening of mRFP using Fluorescence Activated Cell sorter (FACs)

The mRFP variants with the highest fluorescent intensities at the red-shifted wavelength were screened for using FACs. Figure 3.10 shows the fluorescence intensity (axis plotted in logarithmic scale) of the point mutation variants. We can observe that the point mutations variants have more diverse (more spread out fluorescence population) fluorescence intensity (panels A, B, D in Figure 3.10) than the WT mRFP control cells (panel C). Gates P2 and P3 were used to select the desired colonies that we want. The

FACs sorted variants obtained from P3 are grown overnight and 60 fluorescent variants are randomly picked and expressed. The top variants were his-tagged purified and characterized by fluorimetry (Figure 3.11).

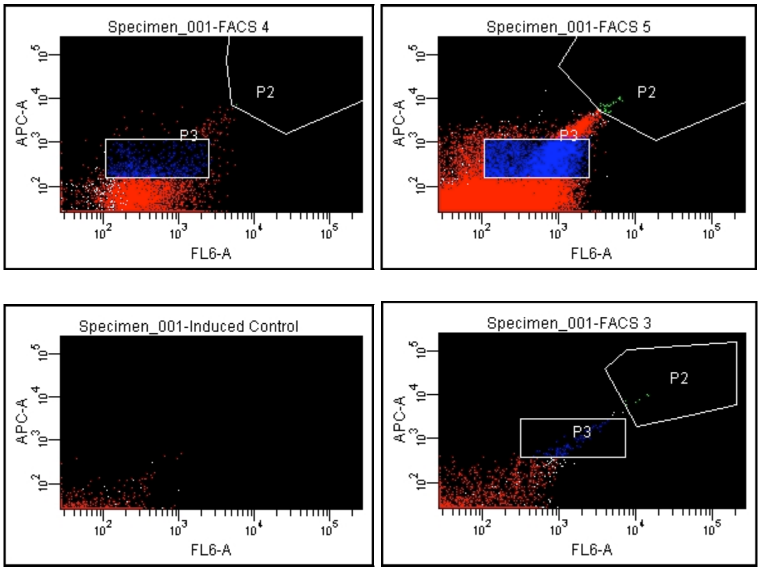
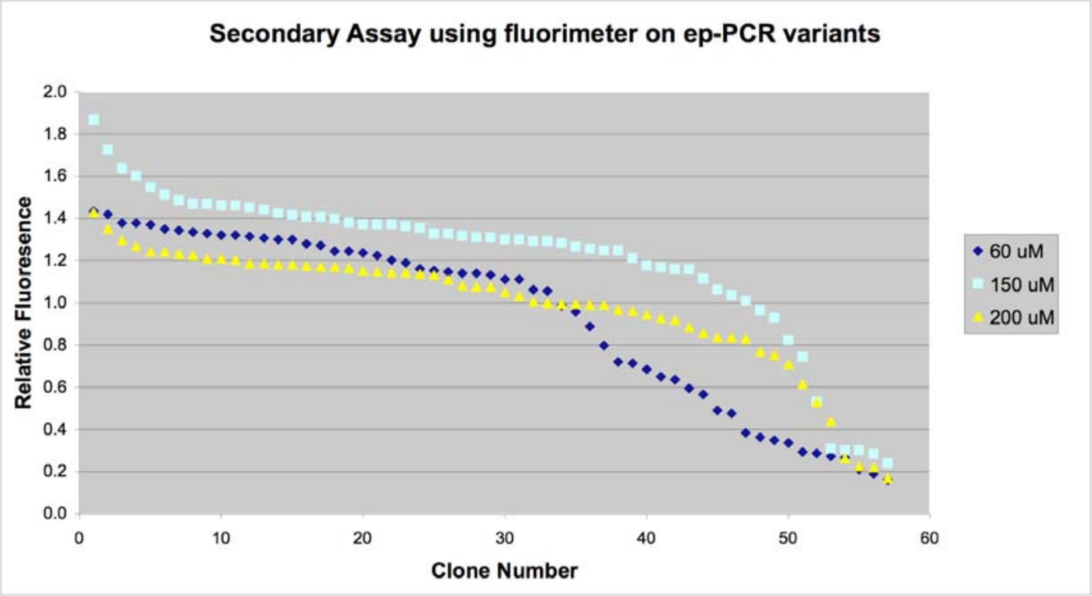
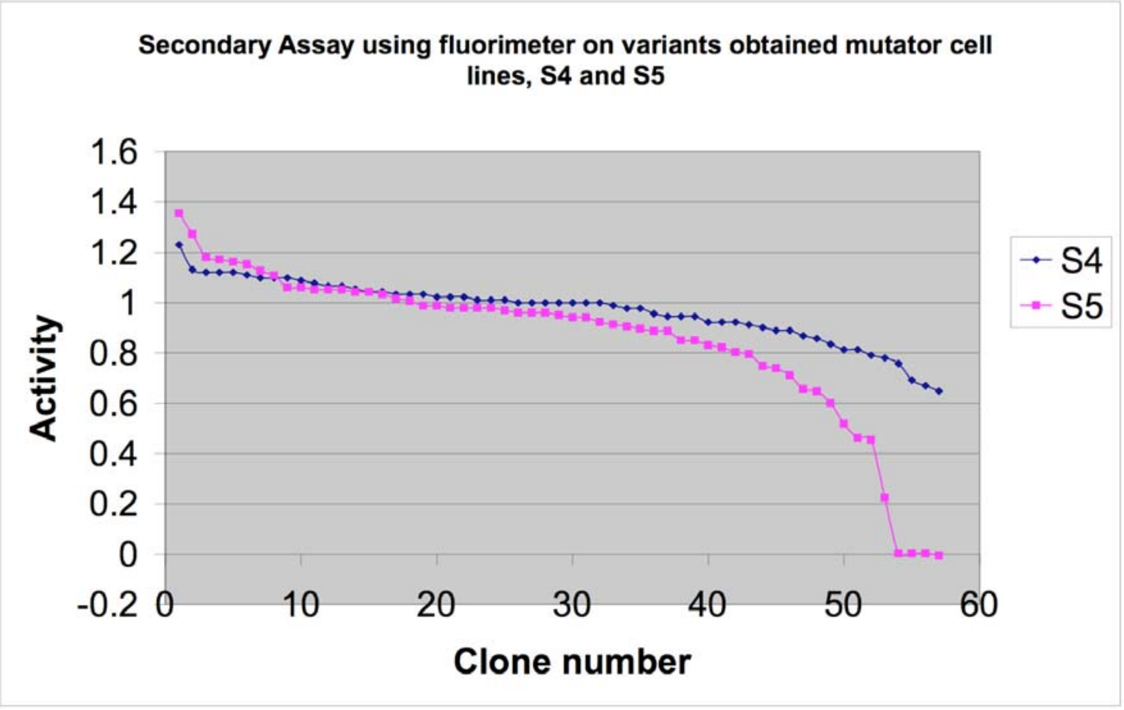


Figure 3.10 FACs selection on WT XL1-red cells (A:top left), round 2 point mutation variants (B: top right), WT mRFP control cells (C: bottom left) and 200 μM ep-PCR cells (D: lower right). APC-A and FL6-A are excitation and emissions filter settings. APC-A: excitation 647 nm, emission 675 nm. FL6-A excitation, 488 nm, emission 660 nm. P2 represents variants screened for while P3 are variants that screened against.



A)



B)

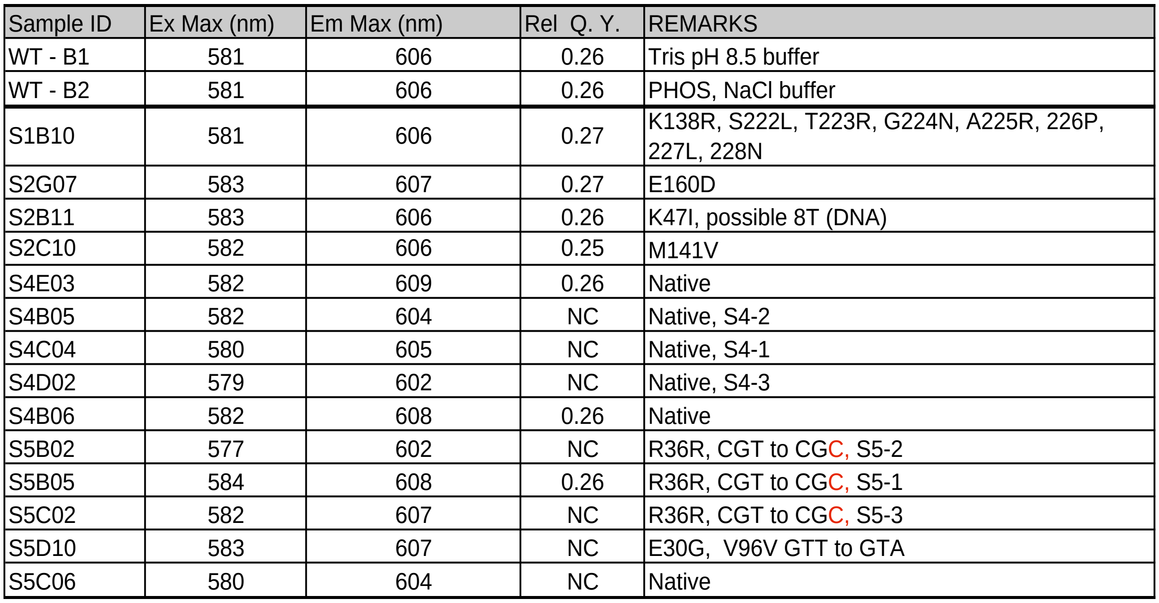
Figure 3.11 Secondary screen on mRFP variants using 96-well plate fluorimeter. A) Normalized fluorescence of FACs sorted mRFP variants generated using different

concentrations of $MnCl_2$. The legend lists various concentrations of $MnCl_2$ used to generate the variants. B) Normalized activity assay on FACs sorted variants generated using XL1-Red mutator cells. S4 used wild-type mRFP gene as a source while S5 experiments used mRFP mutant templates from 150 mM $MnCl_2$ ep-PCR experiments as a source. All readings were normalized to mRFP WT control cells.

3.3.4 Characterization of top variants using Fluorimeter

The top variants obtained from the secondary screen were expressed and purified using metal-chelation chromatography methods (Table 3.4). The relative quantum yields for the purified variants were determined and it ranged from 0.25 to 0.27. The maxima of the emissions wavelength ranged from 602 nm to 609 nm.

Table 3. 4 Characterized top variants from secondary screening.



Sample ID	Ex Max (nm)	Em Max (nm)	Rel Q. Y.	REMARKS
WT - B1	581	606	0.26	Tris pH 8.5 buffer
WT - B2	581	606	0.26	PHOS, NaCl buffer
S1B10	581	606	0.27	K138R, S222L, T223R, G224N, A225R, 226P, 227L, 228N
S2G07	583	607	0.27	E160D
S2B11	583	606	0.26	K47I, possible 8T (DNA)
S2C10	582	606	0.25	M141V
S4E03	582	609	0.26	Native
S4B05	582	604	NC	Native, S4-2
S4C04	580	605	NC	Native, S4-1
S4D02	579	602	NC	Native, S4-3
S4B06	582	608	0.26	Native
S5B02	577	602	NC	R36R, CGT to CGC, S5-2
S5B05	584	608	0.26	R36R, CGT to CGC, S5-1
S5C02	582	607	NC	R36R, CGT to CGC, S5-3
S5D10	583	607	NC	E30G, V96V GTT to GTA
S5C06	580	604	NC	Native

3.4 Discussion

3.4.1 The mRFP gene

We managed to successfully synthesize, clone the mRFP gene, and express and purify mRFP itself. From sequencing information, we were able to verify that the correct genes were synthesized using the procedures described by Hoover *et al.* (29) The procedures were, in general, straightforward to apply involving two PCRs steps. The first step involves gene reassembly and the second amplification of the assembled genes. The expressed mRFP proteins had similar absorption and emission spectral when compared to literature (28). We managed to over-express mRFP up to 20% of protein expression in *E.*coli cells and obtained 30mg of purified mRFP from an estimated 100 mg of mRFP per l L batch culture. We determined that metal-chelation chromatography is an efficient method to purify histidine-tagged mRFP proteins.

3.4.2 Library diversity generated by ep-PCR

The results from sequencing (Table 3.3) indicate that we successfully obtained a diverse set of sequences but parental background was also obtained. Typical ep-PCR efforts using varying concentrations of $MnCl_2$ in the range of 60 μM to 200 μM to change the rate of mutations (31) report 0.11 to 2% mutation frequencies (1 to 20 nucleotides per 1kb). In comparison, we obtained mutation frequencies of 0.15% to 0.89% for mRFP gene whenever variants were generated. We also obtained a 50% background of WT variants – this probably arose from a small gene length (678 bp), which reduces the efficiency of ep-PCR due to a lower probability of mismatching complimentary base-pairing. The presence of background was probably not due to the presence of WT mRFP plasmids that contaminate the PCRs and transformation as the gel purification steps should, in principle, have removed any contaminating plasmids.

3.4.3 Two rounds of evolution and ep-PCR are insufficient to obtain improved variants

The mRFP is a plastic protein as it is found to be very tolerant to single point mutations and two rounds of directed evolution is insufficient to engineer an improved red-shifted variant. In our directed evolution experiments, we randomly mutated the mRFP gene using library generation protocols such error-prone PCR and *E. coli* mutagenic strain XL1-Red. We did not find improved variants after subjecting the library to FACs screening. Two publications (32,33) on the directed evolution of mRFP from Roger Tsien's lab substantiate our view that red shifting the mRFP through directed evolution using ep-PCR to generate libraries is likely to be an unrealistic goal.

We found mRFP to be a protein of high plasticity as it can tolerate a number of amino acids substitutions and still retain its native characteristics. We have applied error-prone-PCR on mRFP, generated a library using XL1-Red *E. coli* mutator strain and screened the variants by fluorescence activated cell-sorting (Figure 3.10). We expressed, purified and characterized the variants showing the highest intensity. Of the screened variants (Table 3.4), we did not find an improved variant suggesting that more mutations were probably required for evolving a red-shifted variant of mRFP.

Recently, Wang *et al*. showed the evolution of mPlum (32), a more red-shifted monomeric red fluorescent protein from mRFP by 23 rounds of somatic hypermutation. Somatic hypermutation (SHM) utilizes activation induced cytidine deaminase and error-prone DNA repair to introduce point mutations into the genes located at the V regions of Ig locus. Wang *et. al*. cloned the mRFP gene into the V regions and performed 29 rounds of SHM with FACs screening. The effort yielded mPlum, a monomeric variant of mRFP that has an emissions maximum at 649 nm, 42 nm more red-shifted than mRFP. The authors pointed out that conventional random mutagenesis method could not generate shifted mRFP variants of more than 623 nm. Seven mutations were introduced into mRFP to change it to mPlum and a much larger sequence space was searched.

Shaner *et al.* (33) reported on the evolution of improved monomeric fluorescent proteins in which they added four amino acids to the N-terminal and seven amino acids of GFP to the C-terminus of mRFP prior to applying directed evolution to the protein. They evolved a fluorescent protein, mCherry, whose emissions is 3 nm more red-shifted than mRFP.

The above mentioned research findings substantiate our view that two rounds of evolution is insufficient to search for an improved red-shifted variant of mRFP and that conventional directed evolution methods utilizing ep-PCR as a method to generate variants is not optimal.

3.5 Conclusion

The mRFP gene was successfully synthesized, cloned and expressed. Subsequently, the analysis of the directed evolution efforts of the mRFP gene combined with recent work from other groups suggest to us that utilization of non-conventional library generation techniques and extensive mutagenesis were required to yield slightly red-shifted mRFP proteins – mPlum and mCherry.

3.6 References

1. Bulina, M.E., Chudakov, D.M., Britanova, O.V., Yanushevich, Y.G., Staroverov, D.B., Chepurnykh, T.V., Merzlyak, E.M., Shkrob, M.A., Lukyanov, S. and Lukyanov, K.A. (2006) A genetically encoded photosensitizer. *Nat Biotechnol*, **24**, 95-99.

2. Chudakov, D.M., Verkhusha, V.V., Staroverov, D.B., Souslova, E.A., Lukyanov, S. and Lukyanov, K.A. (2004) Photoswitchable cyan fluorescent protein for protein tracking. *Nat Biotechnol*, **22**, 1435-1439.

3. Poppenborg, L., Friehs, K. and Flaschel, E. (1997) The green fluorescent protein is a versatile reporter for bioprocess monitoring. *J Biotechnol*, **58**, 79-88.

4. Tsien, R.Y. (1998) The green fluorescent protein. *Annual review of biochemistry*, **67**, 509-544.

5. Tsien, R.Y. (1999) Rosy dawn for fluorescent proteins. *Nature Biotechnology*, **17**, 956-957.

6. Verkhusha, V.V. and Sorkin, A. (2005) Conversion of the monomeric red fluorescent protein into a photoactivatable probe. *Chem Biol*, **12**, 279-285.

7. 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.

8. Weissleder, R. (2001) A clearer vision for in vivo imaging. *Nat Biotechnol*, **19**, 316-317.

9. Zimmer, M. (2002) Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. *Chemical reviews*, **102**, 759-781.

10. Maddalo, S.L. and Zimmer, M. (2006) The role of the protein matrix in green fluorescent protein fluorescence. *Photochemistry and photobiology*, **82**, 367-372.

11. Shimomura, O. (2005) The discovery of aequorin and green fluorescent protein. *Journal of Microscopy*, **217**, 1-15.

12. Matz, M.V., Lukyanov, K.A. and Lukyanov, S.A. (2002) Family of the green fluorescent protein: journey to the end of the rainbow. *Bioessays*, **24**, 953-959.

13. Fron, E., Flors, C., Schweitzer, G., Habuchi, S., Mizuno, H., Ando, R., Schryver, F.C., Miyawaki, A. and Hofkens, J. (2007) Ultrafast excited-state dynamics of the photoswitchable protein Dronpa. *Journal of the American Chemical Society*, **129**, 4870-4871.

14. Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2000) Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proc Natl Acad Sci U S A*, **97**, 11984-11989.

15. Zimmer, M. (2005) *Glowing genes : a revolution in biotechnology*. Prometheus Books, Amherst, N.Y.

16. Shimomura, O. (1979) Structure of the Chromophore of Aequorea Green Fluorescent Protein. *Febs Lett*, **104**, 220-222.

17. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. and Prasher, D.C. (1994) Green fluorescent protein as a marker for gene expression. *Science (New York, N.Y*, **263**, 802-805.

18. Giepmans, B.N., Adams, S.R., Ellisman, M.H. and Tsien, R.Y. (2006) The fluorescent toolbox for assessing protein location and function. *Science (New York, N.Y*, **312**, 217-224.

19. Shaner, N.C., Steinbach, P.A. and Tsien, R.Y. (2005) A guide to choosing fluorescent proteins. *Nature Methods*, **2**, 905-909.

20. Lakowicz, J.R. (2006) *Principles of fluorescence spectroscopy*. 3rd ed. Springer, New York ; Berlin.

21. Miyawaki, A., Llopis, J., Heim, R., McCaffery, J.M., Adams, J.A., Ikura, M. and Tsien, R.Y. (1997) Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature*, **388**, 882-887.

22. Yang, W., Wilkins, A.L., Ye, Y., Liu, Z.R., Li, S.Y., Urbauer, J.L., Hellinga, H.W., Kearney, A., van der Merwe, P.A. and Yang, J.J. (2005) Design of a

calcium-binding protein with desired structure in a cell adhesion molecule. *Journal of the American Chemical Society*, **127**, 2085-2093.

23. Yang, W., Jones, L.M., Isley, L., Ye, Y., Lee, H.W., Wilkins, A., Liu, Z.R., Hellinga, H.W., Malchow, R., Ghazi, M. *et al.* (2003) Rational design of a calcium-binding protein. *Journal of the American Chemical Society*, **125**, 6165-6171.

24. Sumner, J.P., Westerberg, N.M., Stoddard, A.K., Hurst, T.K., Cramer, M., Thompson, R.B., Fierke, C.A. and Kopelman, R. (2006) DsRed as a highly sensitive, selective, and reversible fluorescence-based biosensor for both Cu(+) and Cu(2+) ions. *Biosensors & bioelectronics*, **21**, 1302-1308.

25. Schuster, S., Enzelberger, M., Trauthwein, H., Schmid, R.D. and Urlacher, V.B. (2005) pHluorin-based in vivo assay for hydrolase screening. *Analytical Chemistry*, **77**, 2727-2732.

26. Valetti, F., Sadeghi, S.J., Meharenna, Y.T., Leliveld, S.R. and Gilardi, G. (1998) Engineering multi-domain redox proteins containing flavodoxin as bio-transformer: preparatory studies by rational design. *Biosens Bioelectron*, **13**, 675-685.

27. Belousov, V.V., Fradkov, A.F., Lukyanov, K.A., Staroverov, D.B., Shakhbazov, K.S., Terskikh, A.V. and Lukyanov, S. (2006) Genetically encoded fluorescent indicator for intracellular hydrogen peroxide. *Nature Methods*, **3**, 281-286.

28. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7877-7882.

29. Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research*, **30**, -.

30. Arnold, F.H. and Georgiou, G. (2003) *Directed enzyme evolution : screening and selection methods*. Humana Press, Totowa, N.J.

31. Arnold, F.H. and Georgiou, G. (2003) *Directed evolution library creation : methods and protocols*. Humana Press, Totowa, N.J.

32. Wang, L., Jackson, W.C., Steinbach, P.A. and Tsien, R.Y. (2004) Evolution of new nonantibody proteins via iterative somatic hypermutation. *Proc Natl Acad Sci U S A*, **101**, 16745-16749.

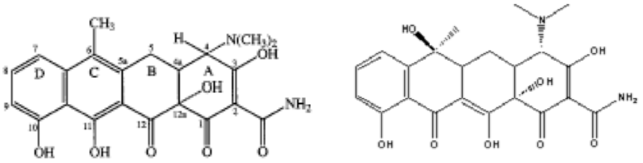
33. Shaner, N.C., Campbell, R.E., Steinbach, P.A., Giepmans, B.N., Palmer, A.E. and Tsien, R.Y. (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat Biotechnol*, **22**, 1567-1572.

3.7 Supplement

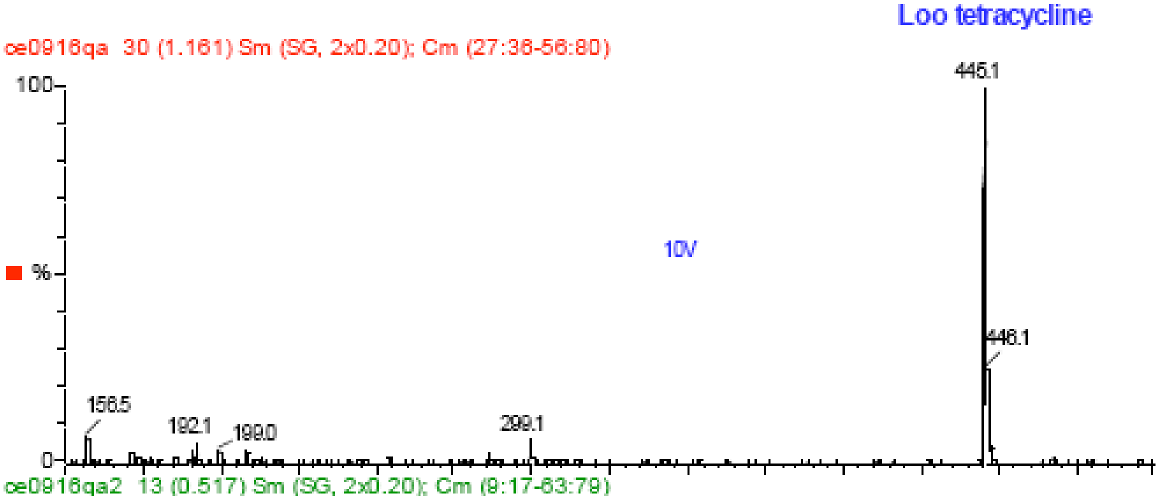
3.7.1 *E. coli* expression optimized primers used for the reassembly of mRFP gene

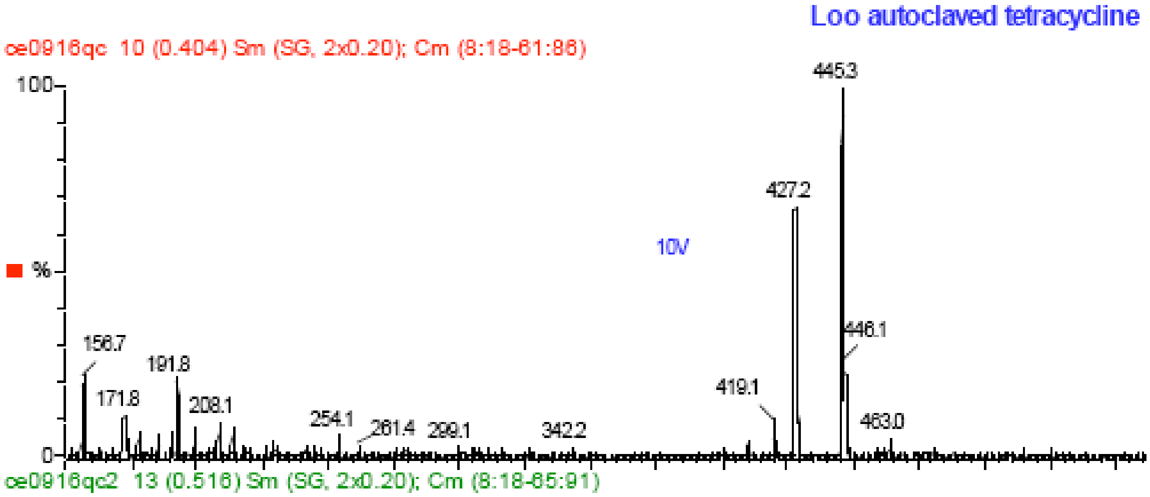
no.	Primer Sequences 5’-3’
1	ATGGCGTCTTCTGAAGACGTTATCAAAGAATTCATGCGTTTCAAAGT
2	TCGTATGGAAGGTTCTGTTAACGGTCACGAATTCGAAATCGAAGGTG
3	AAGGTGAAGGTCGTCCGTACGAAGGTACCCAGACCGCGAAACTG
4	AAAGTTACCAAAGGTGGTCCGCTGCCGTTCGCGTGGGAC
5	ATCCTGTCTCCGCAGTTCCAGTACGGTTCTAAAGCGTACGTTAAACACCCGG
6	CGGACATCCCGGACTACCTGAAACTGTCTTTCCCGGAAGGT
7	TTCAAATGGGAACGTGTTATGAACTTCGAAGACGGTGGTGTTGTTA
8	CCGTTACCCAGGACTCTTCTCTGCAGGACGGTGAATTCATCTACAA
9	AGTTAAACTGCGTGGTACCAACTTCCCGTCTGACGGTCCGG
10	TTATGCAGAAAAAAACCATGGGTTGGGAAGCGTCTACCGAACGTAT
11	GTACCCGGAAGACGGTGCGCTGAAAGGTGAAATCAAAATGCG
12	TCTGAAACTGAAAGACGGTGGTCACTACGACGCGGAAGTTAAAACC
13	ACCTACATGGCGAAAAAACCGGTTCAGCTGCCGGGTGCG
14	TACAAAACCGACATCAAACTGGACATCACCTCTCACAACGAAGACTACACCA
15	TCGTTGAACAGTACGAACGTGCGGAAGGTCGTCACTCTACCGGTGCG
16	TTACGCACCGGTAGAGTGACGACCTTC
17	CGCACGTTCGTACTGTTCAACGATGGTGTAGTCTTCGTTGTGAGAGGTGAT
18	GTCCAGTTTGATGTCGGTTTTGTACGCACCCGGCAGCTGAA
19	CCGGTTTTTTCGCCATGTAGGTGGTTTTAACTTCCGCGTCGTAGT
20	GACCACCGTCTTTCAGTTTCAGACGCATTTTGATTTCACCTTTCAG
21	CGCACCGTCTTCCGGGTACATACGTTCGGTAGACGCTTCCCA
22	ACCCATGGTTTTTTTCTGCATAACCGGACCGTCAGACGGG
23	AAGTTGGTACCACGCAGTTTAACTTTGTAGATGAATTCACCGTCCTG
24	CAGAGAAGAGTCCTGGGTAACGGTAACAACACCACCGTCTTCGAAG
25	TTCATAACACGTTCCCATTTGAAACCTTCCGGGAAAGACAGTTTC
26	AGGTAGTCCGGGATGTCCGCCGGGTGTTTAACGTACGCTTTAGAACCGT
27	ACTGGAACTGCGGAGACAGGATGTCCCACGCGAACGGCA
28	GCGGACCACCTTTGGTAACTTTCAGTTTCGCGGTCTGGGTACC
29	TTCGTACGGACGACCTTCACCTTCACCTTCGATTTCGAATTCGTGA
30	CCGTTAACAGAACCTTCCATACGAACTTTGAAACGCATGAATTCTTT

3.7.2 Structures and Mass spectrometry of anhyrotetracyline and autoclaved tetracycline



Left anhydrotetracycline(MW: 426). Right tetracycline (MW: 444).





The top graph shows the mass spectrometry data on tetracycline while the bottom graph shows the mass spectrometry data on autoclaved tetracycline. ATC expected molecular weight 426. Tetracycline expected molecular weight 444. Conversation rates were estimated to be up to 65% of ATC based on tetracycline.

CHAPTER 4

OPTIMIZING RECOMBINATION PROTOCOLS

4.1 Introduction

A diverse library is essential for success of directed evolution (1-3). Directed evolution involves i) generating a library with large diversity of variants and ii) the application of a screening or selection assay to pick out the improved variants (4-6). This sequence of diversity generation and screening or selection can be repeated iteratively a number of times until the improved variants are obtained. Screening and selection efforts are wasted if the library is not sufficiently diverse (See Figure 4.1). The two main in-vitro library generation methods for molecular diversity can be classified as random mutagenesis and recombination methods (2,3). Random mutagenesis has been reported to be a highly limited and biased method by recent reviews (7,8). On average, it can access 3.14 – 7.40 amino acid substitution per residue due to the redundancy of the genetic code and its organization to minimize mutational errors. In contrast, recombination protocols have not been systematically evaluated on the same set of proteins to date. In this investigation, we applied two protocols, PCR based recombination and DNA shuffling, on fluorescent protein genes to generate libraries and will report on the performance of both recombination protocols.

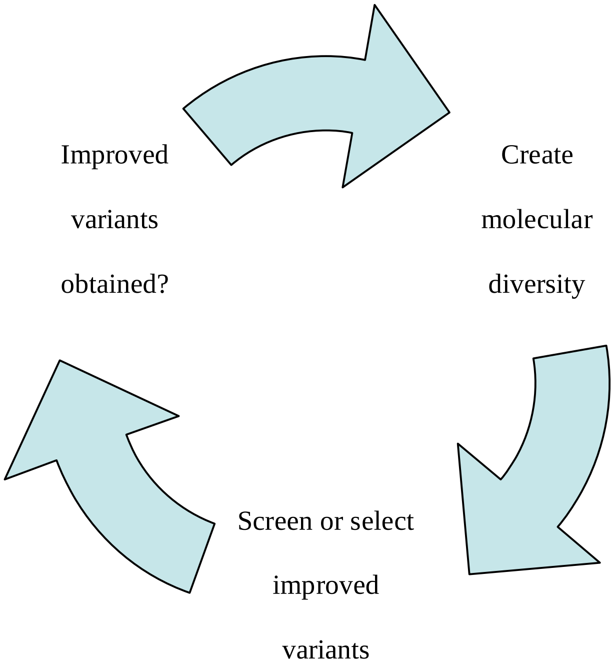


Figure 4.1 Diagram showing the process of directed evolution. Molecular diversity is created and the improved variants can be found after screening and/or selection. The process can be iterated a number of times to obtain better variants.

4.1.1 In-vitro library generation methods

There are a number of library generation methods available to date for recombination. Table 4.1 provides a short description of the key protocols to date. In general, the recombination protocols require at least one parental gene template that can be used as a scaffold for the creation of mutants. The library protocols usually involve a method to create sections of the genes that eventually get assembled together into a full-length chimera. One key disadvantage of a number of libraries generated by some of these protocols such as DNA shuffling and StEP is that they contain a high percentage of parental background or unshuffled clones which increases the amount of effort required to screen for the improved variants.

Table 4.1 The various in-vitro recombination-protocols for library creation (3,23).

Method	**Description of method and characteristics of library**	**References**
DNA shuffling: Single and double stranded	The gold standard for recombination. Involves fragmentation of templates by DNAse I, followed by a reassembly PCR and a secondary PCR to obtain gene fragments. Suffers from parental background. Spiked oligonucleotides can be added during primer synthesis to introduce diversity. CLERGY* uses DNA shuffling to generate initial diversity and homologous recombination during transformation in *S*. cerevisiae creates additional diversity.	(9-13) (3)
StEP: Single and double stranded	StEP: **St**aggered **E**xtension **p**rocess. Short extension steps to promote crossovers. Suffers from parental background. Single stranded step improves performance.	(14)
Recombination PCR/SUUPER	SUUPER: **S**huffling **U**sing **U**npaired **P**rimers. Uses cycling protocols similar to StEP but involves the use of skew primers to force crossovers.	(15)
RACHITT	RACHITT: **Ra**ndom **Chi**meragenesis on **T**ransient **T**emplates Uses fragmented single stranded DNA and parental DNA as templates to make chimera templates. High crossover frequency was reported for this technique. Implementation is technically difficult due to the need to isolate single-stranded DNA in high yield	(16)
Incremental truncation ITCHY and SCRATCHY** SHIPREC: Sequence homolog-independent protein recombination	ITCHY: **I**ncremental **T**runcation for the **C**reation of **H**ybrid Enz**y**mes. SCRATCHY: Combination of ITCHY and DNA shuffling Two genes are incrementally truncated using nucleases and eventually ligated and transformed. The truncation can involve the use of nucleotide analogs Additionally, an in frame selection vector can be used to select in frame chimeras. Single crossover library is only possible. No sequence homology required. SHIPREC: **S**equence **H**omolog-**I**ndependent **P**rotein **Rec**ombination. This is a similar method to ITCHY.	(17-22)

*, ** The full names of these abbreviations were not specified in the literature.

4.1.2 DNA shuffling and Recombination PCR: Background and problems

4.1.2.1 DNA shuffling background

Since its introduction in 1994, DNA shuffling has been widely used for library creation (24-36). Figure 4.2 outlines the steps involved in DNA shuffling. In brief, two or more gene templates are fragmented using DNAse I endonuclease enzyme. After an assembly PCR of the fragments to generate the full-length gene and secondary PCR to amplify the assembled genes, chimeras are obtained. DNA shuffling has become a useful and standard tool for protein evolution evident from the fact that the original DNA shuffling papers (9,24) have been cited more than 500 times each to date. DNA shuffling, however, has inherent disadvantages that limit its utility as a library generation protocol.

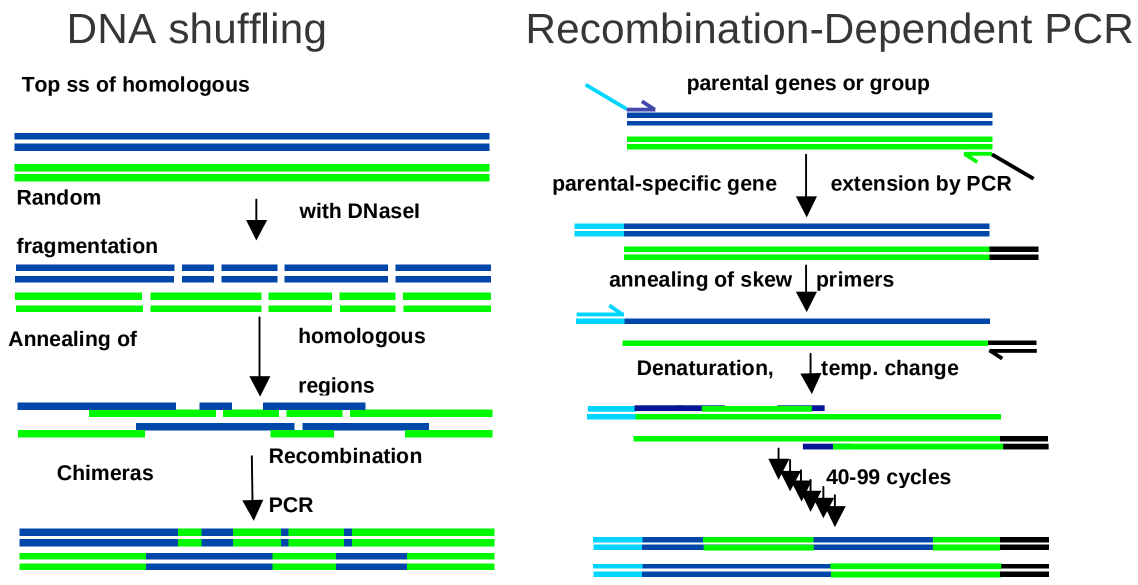


Figure 4.2 DNA shuffling and Recombination PCR (RD-PCR) are depicted on the left and right. DNA-shuffling involves the fragmentation of wild-type parental genes, followed by two PCR steps to recover the full length chimeric genes. RD-PCR involves the use of skew primers and short annealing and extension cycling steps to force crossovers.

4.1.2.2 Recombination PCR: Background

Recombination PCR is an alternate library generation protocol recently developed in 2003 that is reported to eliminate parental background (15). A recent innovation involving the use of templates with skew annealing regions and skew primers by Ikeuchi *et al*. and Milano and Xiao (15,37), generates a library by using PCR cycling with short extension steps and annealing time as illustrated by Figure 4.2. Due to the use of skew primers, only chimeras are amplified and available for cloning with restriction enzymes.

Ikeuchi *et al.* investigated the use of skew primers, non-parental end annealing primers, in combination with short cycles and annealing and extension PCR cycling protocols on enhanced yellow fluorescent protein (EYFP) and green fluorescent protein (GFP) of 72% nucleotide identity (15). They obtained a library with negligible background and high percentage of single crossovers. They also found that recombination PCR required less continuous homologous nucleotide identity than DNA shuffling for crossovers to occur.

Staggered extension protocol (StEP) is a similar protocol to recombination PCR but it does not involve the use of skew primers to reduce parental background. Zhao *et al.* (14) described the application of StEP on two variants of thermostable subtilisin E from *Bacillus subtilis* that differ by ten point mutations. They reported generating libraries with 100% novel chimeras based on sequencing information from ten variants using this method.

4.1.2.3 Problems with DNA shuffling and recombination PCR

One reported key problem with using DNA shuffling to generate a library is the presence of high parental background (12, 15). Parental background is the presence of non-shuffled templates that decreases the diversity of the library and thus, waste screening and/or selection efforts. A diverse library with negligible parental background is highly desired in protein engineering to initiate a more productive search through the sequence space that will increase the chances of success to obtain improved variants.

The lowest limits of nucleotide identity level between genes at which DNA shuffling and recombination PCR can still be applied is not established. With increasing access to homologous genes, the prospect of recombining genes with low levels (< 50%) nucleotide identity is becoming more likely to access a more diverse sequence space. The lowest reported successful DNA shuffling is 56% that led to chimera generation (51).

No head to head comparison of DNA shuffling and recombination PCR has been reported in the literature. DNA shuffling and recombination PCR have not been performed on the same set of proteins and the sequence diversity of the libraries produced by the protocols have not been analyzed extensively. It is crucial that protein engineers have access to information on the inherent characteristics of recombination protocols that may bias their sequence space exploration during directed evolution experiments.

The above issues inspire us to investigate the performance of DNA shuffling and recombination PCR protocols on fluorescent proteins.

4.1.3 Scope of work: Comparison of performance of DNA shuffling and recombination PCR on fluorescent proteins

The objective of this work was to compare the performance of optimized DNA shuffling and recombination PCR protocols on fluorescent proteins with a range of nucleotide identity levels from 45% to 74.5% (Table 4.2). To evaluate performance, the following factors were analyzed: number of crossovers, crossover points, parental background, chimera background, percentage of useful sequences, minimum number of base-pairs required for crossovers and lower limit of nucleotide identity between two genes required for recombination to work. To date, this is the first known head-to-head comparison of the performance of DNA shuffling and recombination PCR on the same set of proteins. It is expected that the insight gained from the characterization of this library will benefit the future users of DNA shuffling and recombination PCR protocols. In particular, the following questions will be answered:

- Which recombination protocol is easier to apply?
- Which protocol can still work at the lowest identity level (45%)?
- How does the sequence diversity of the libraries generated by DNA shuffling and recombination PCR compare?
- Can we optimize recombination PCR?

Table 4.2 Nucleotide identity and amino acid identity of various pair of fluorescent proteins used for recombination experiments. mRFP: monomeric Red Fluorescent Protein, DsRed: *Discosoma* Red Fluorescent Protein, HcRed: *Heteractis crispa* Red Fluroescent Protein, GFP: Green Fluorescent Protein.

Templates	Nucleotide Identity (%)	Amino Acid Identity (%)
mRFP/DsRed	74.5	84.1
DsRed/HcRed	66.6	44.8
mRFP/GFP	45.0	24.2

4.2 Materials & Methods

4.2.1 Construction of the parental fluorescent proteins plasmids

The amino acid sequence of mRFP was obtained from NCBI and *E.* coli-codon optimized primers (supplement 4.7.1) were designed using DNAworks (38) and synthesized. To synthesize the mRFP gene, two PCR reactions were done, one to assemble the codon optimized primers, and the second to amplify the full-length product. The gene was amplified using primers with *Esp*3I (italicized) restriction sites (5’ -TA*C GTC TCG* TCG ACA TGG CGT CTT CTG AAG ACG TTA TCA AAG AAT TCA TGC GT – 3’ and 5’ – TA*C GTC TCT* GGC CTA TTA CGC ACC GGT AGA GTG ACG ACC TTC - 3’) and digested with *Esp*3I enzyme and ligated using T4 DNA ligase into *Sal*I and *Not*I digested pPROTet vector. Sequencing, expression and characterization consistent with the literature confirmed that the *E. coli* expression optimized mRFP gene was successfully assembled (39).

The HcRed gene was cloned from pHcRed1-N1/1 plasmid (BD Biosciences Clontech, Plao Alto, CA) with primers containing *Sal*I and *Not*I restriction sites (CGG GAT TCC ACA TAG TCT CAG GTA *GTC GAC* ATG GTG AGC GGC CTG CTG AAG GAG AGT ATG – 3' and 5`- TTC CGA TAA GTT CAT AGG CCG TG*G CGG CCG C*TC AGT TGG CCT TCT CGG GCA GGT CGC T– 3') and cloned into the pPROTet plasmid.

The GFP gene was amplified from pQBIT7-GFP plasmid (QBIOgene, Carlsbad, CA) primers with *Esp*3I restriction sites (5' - TAC GGT TA*C GTC TCG* TCG ACA TGG CGT CTT CTG AAG ACG TTA TCA - 3' and 5'- TAC GGT TA*C GTC TCG* TCG ACA TGG CTA GCA AAG GAG AAG AAC TCT TCA -3'), digested using *Esp*3I and ligated into *Sal*I and *Not*I digested pPROTet vector.

The DsRed gene was amplified from DsRed2-1 plasmid (BD Biosciences Clontech, Palo Alto, CA) with primers containing *Esp*3I restriction sites (5'-TAC GGT TA*C GTC TCG* TCG ACA TGG CCT CCT CCG AGA ACG TCA -3' and 5'-CAT TAC TA*C GTC TCT* GGC CTA CTA CAG GAA CAG GTG GTG GCG G -3') and cloned in a similar way to mRFP and GFP.

4.2.2 Template preparation for DNA shuffling

The mRFP, DsRed, HcRed and GFP genes were amplified by *Pfu* polymerase using primers that anneal to the pPROTet vector (5`-CTT TCG TCT TCA CCT CGA

GTC C-3`, 5`-CCT ACT CAG GAG AGC GTT CAC C-3`), which added 122 bp to the 5`-terminus and 155 bp to the 3`-terminus. The PCR products were gel purified using 1.2% agarose gel and QIAEX II kit (Qiagen, Valencia, CA).

4.2.3 DNA shuffling

DNA shuffling was performed according to Joern (40), which uses a hybrid method derived from Stemmer *et al.* (9) and Abècassis *et al.* (41). After optimizing the DNaseI concentration and digestion time, 2 μg of an equimolar mixture of the desired parental templates was digested. Fragments of less than 300 bp were isolated by agarose gel purification using QIAEX II (Qiagen, Valencia, CA). 500-750 ng DNA-fragments were mixed with 5 μl *Pfu* buffer (10X reaction buffer: 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 200 mM Tris-HCl (pH 8.8), 20 mM MgSO4, 1% Triton®X-100, 1 mg/ml BSA), 1 μl of *Pfu* polymerase (2.5 U/μl) and water to a final volume of 50 μl and temperature-cycled following the protocol from Abècassis *et al.* (41): 96ºC, 90s; 35 cycles of (94ºC, 30s; 65ºC, 90s; 62ºC, 90s; 59ºC, 90s; 53ºC, 90s; 50ºC, 90s; 47ºC, 90s; 44ºC, 90s, 41ºC, 90s; 72ºC, 4 min); 72ºC, 7 min; 4ºC hold. Following reassembly, 1: 10, 1: 100, 1:1000 dilutions of the reassembled fragments were amplified using nested primers, primers that bind within the products of the first PCR, with *Pfu* polymerase and buffer to determine the optimal dilution ratio. The genes were then amplified using the optimal dilution ratio. The following nested primers were used for the amplification of the fluorescent proteins (5`-ATG GGT CAT AAT CAT AAT CAT AAT CAT AAT C-3` and 5`-GTC TTT CGA CTG AGC CTT TCG T-3`).

4.2.4 Template preparation for the recombination-dependent PCR (RD-PCR)

For the amplification of full-length mRFP, DsRed, HcRed and GFP genes, parent-specific primers were designed, which added a specific overhang (either 5`-CGG GAT TCC ACA TAG TCT CAG GTA-3`, 5’–GCTA CGC ATG AAT GCG TAC T–3’ or 5’-GGA TTC CAC ATA GTC TCA GG-3’) at the 5`-terminus of the one parent and a different overhang (either 5`-TTC CGA TAA GTT CAT AGG CCG TGG-3`, 5’ – GAC GCT TCT GAA GAA GTC CT – 3’ or 5’-TGC CGG ATA CTT GAA TAG CC-3’ ) at the 3`-terminus of the other parent. The amplification of truncated mRFP, DsRed, HcRed and GFP genes required primers that annealed to the interior of the genes. Figure 4.3 shows a pictorial overview of the templates used.

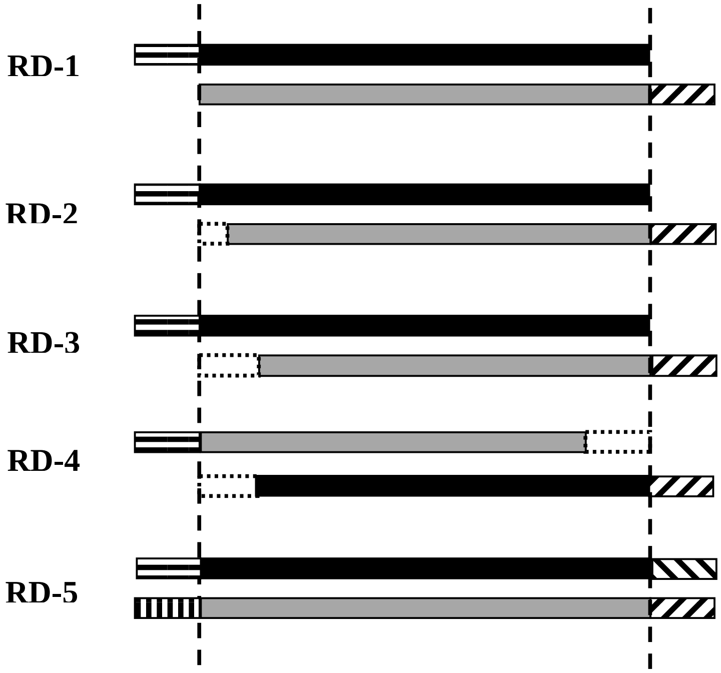


Figure 4.3 Various templates used for the recombination of fluorescent protein genes. RD-1 to RD-5 are labels given to the various combination of templates used. The black bars and grey bars represent double-stranded DNA of the fluorescent proteins. The vertical lines represent the edges of the genes. Boxes outside of the vertical lines represent overhangs attached to the genes. Dotted boxes inside the vertical lines show

truncation regions in the genes.

4.2.5 Recombination PCR

Equimolar mixtures of the extended parental templates were used for the recombination PCR. Up to 50 ng of the templates were used in 25 μl reaction volumes. A small reaction volume ensures that thermal equilibrium is reached faster during the thermal cycling.

Two variations of Ikeuchi's protocol (15) were used: 1) 94°C 2 min, 99 cycles of (94°C 1 min, 63-67°C, 5 sec), 72°C 7 min, hold at 4°C; 2) 98ºC for 5 mins; 40 cycles of (98ºC, 30s; 40-45ºC, 5s; 72°C for 3s); 10 cycles of (94ºC, 30s; 50ºC, 30s; and 72ºC 30s); hold at 4ºC. RD-PCR reactions were performed with *Taq* polymerase and optimized for PCR yield unless otherwise specified.

4.2.6 Sequencing

All sequencing was performed at the FAME Center Sequencing facility located at Emory University, Atlanta, GA by Perry Mars using an ABI Prism® 310 DNA sequencer (Applied Biosystems, Foster City, CA).

4.3 Results

Directed evolution library generation protocols such as DNA shuffling and recombination PCR were applied on fluorescent protein genes from 45% to 75% DNA identity levels (Table 4.2). The results can be categorized into two parts; the recombination of fluorescent proteins with high DNA identity levels (> 70%) which represents the typical recombination experiments and the recombination of fluorescent proteins with low DNA identity levels (< 70%). Also, since DNA sequencing information was obtained for each clone, the parental background, crossover points, number of crossovers and number of mutations per variant were calculated. 390 variants were sequenced and analyzed. Supplement 4.7.2 contains the sequence data that are discussed in this section.

4.3.1 Recombination of fluorescent proteins with high DNA identity levels (74.5%): mRFP and DsRed

The fluorescent protein genes from mRFP and DsRed were recombined using DNA shuffling and recombination PCR. 67 variants were sequenced from the DNA shuffling library and 228 variants were sequenced from the recombination PCR library. The recombination PCR library variants were generated using different templates to test the relation between the quality of the chimeras and the templates used for the recombination protocol and the sequencing results can be sub-divided into RD-1, RD-2, RD-3, RD-4 and RD-5. Figure 4.4 shows a combined plot of RD-1 to RD-4 (One skew

Primer per Parent), RD-5 (Two skew primers) as well as the sequences from DNA shuffling.

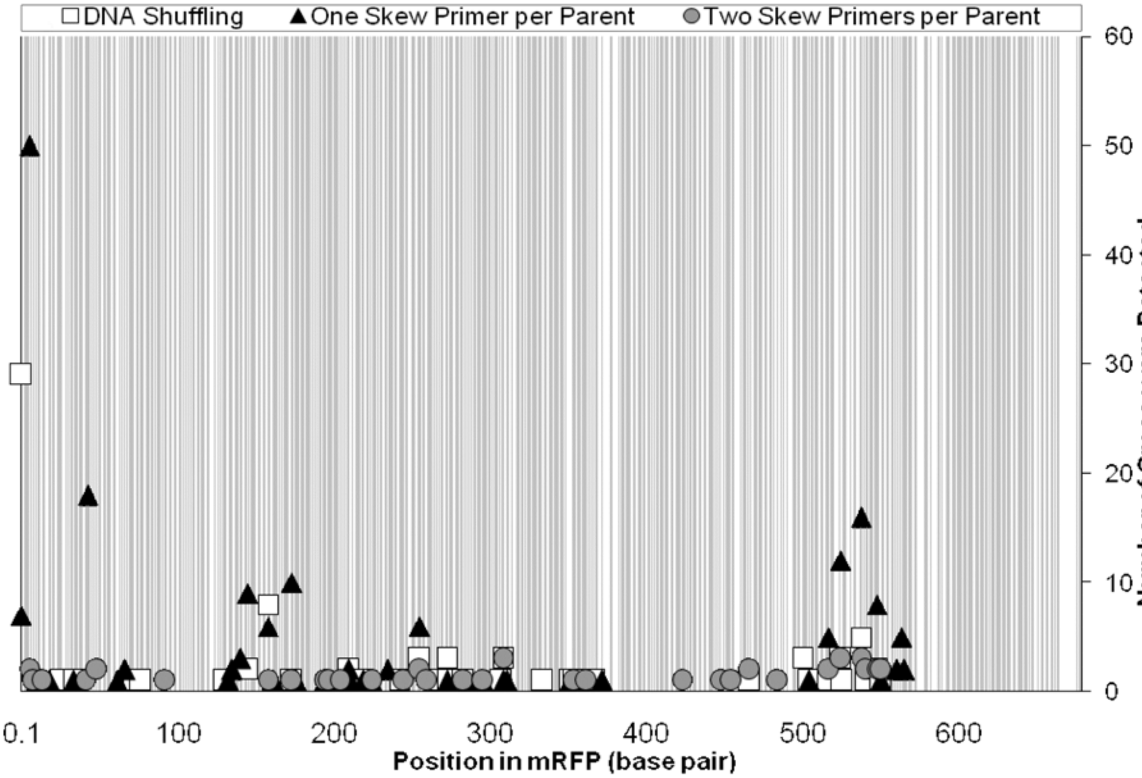


Figure 4.4 The frequency and location of crossovers in DNA shuffled and recombination PCR libraries made from DsRed and mRFP genes. Shaded areas indicate that the bases are identical in DsRed and mRFP while non-shaded areas indicate otherwise. Crossovers are denoted at the position where the first base pair differs between the two sequences. Sequences with multiple crossovers were marked at each crossover position separately. ‘One skew primer per parent’ combines the results from RD-PCR 1-4. A total of 295 sequences are represented in the chart.

The DNA shuffling library had 49% parental sequences (also known as parental background) and this is a documented unwanted side-product as a result of using DNA shuffling protocol. A similar DNA shuffling recombination experiment of green fluorescent protein (GFP) and yellow fluorescent protein (YFP) by Ikeuchi *et al.* (15) also had high parental background as restriction length polymorphism analysis indicated a fragment pattern similar to unshuffled green fluorescent protein. Other recombination experiments involving the use of DNA shuffling reported 20% and 16% parental background when genes of 2100 bp and 1500 bp are shuffled, respectively. These results support the view that parental background has to be expected when DNA shuffling protocols is used (41,42).

To test the correlation between the use of different templates on the recombination PCR protocol and the diversity of the library generated, five sets of recombination PCR libraries were generated using different combinations of primers as illustrated by Figure 4.3.

- The first recombination PCR library (RD-PCR 1) was created using a single skew primer. From the sequencing data of this library, we found zero percent parental background and 76% duplicate sequences, which are also termed as chimera background, at base pair position 6 of mRFP. Therefore, 24% of the library variants were useful sequences for screening, as multiplicity of the same variant sequence does not contribute to diversity of the library. 96% of the library had one crossover and four percent of library had 3 crossovers.

- To reduce the chimera background at base-pair position 6 of mRFP, we performed recombination PCR using truncated DsRed templates (first five bases were removed) with full length mRFP templates. The sequencing data of the RD-

PCR 2 library show that the bias could not be removed by using such a strategy. Of 21 variants analyzed, 43% unique chimeras were obtained. Further truncation of the first 44 base pairs of the DsRed parental gene (RD-PCR 3), lead to a bias towards crossovers at the 3' end of the genes. However, the crossovers were not localized to a single position. Of the 66 variants that were sequenced, 35% were unique variants. When we used mRFP and DsRed truncated templates (mRFP: 40bp 5' end truncated, DsRed 43bp 3'end truncated) to create the RD-PCR 4 library, we found a localization of crossovers at position 50. From 50 variants, 35% useful sequences were found. The statistics for the library are shown in Figure 4.5.

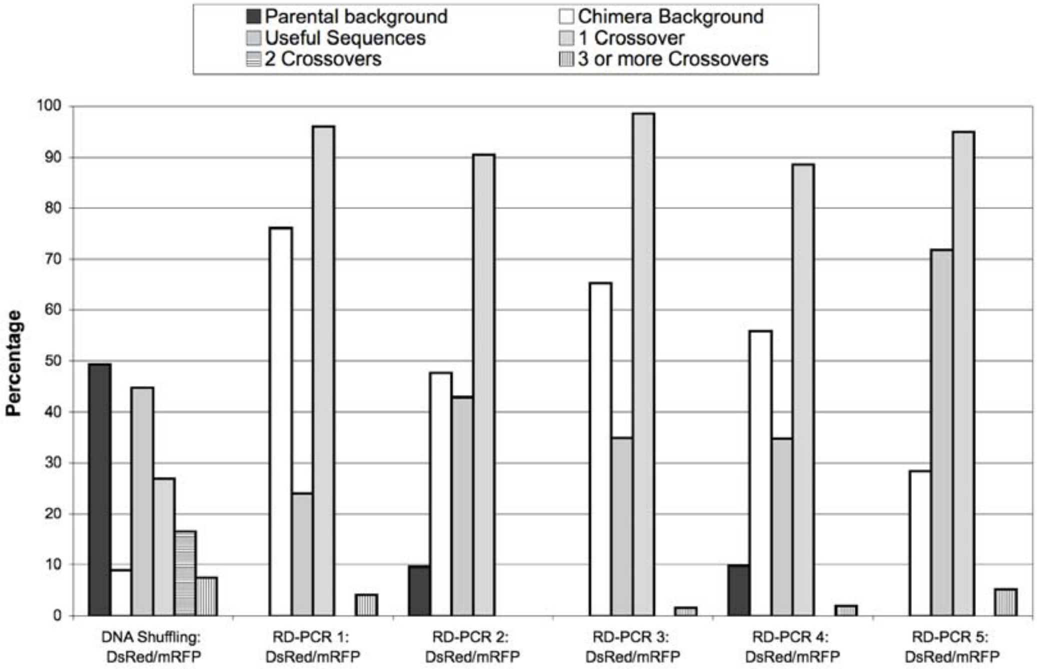


Figure 4.5 Distributions of DNA shuffling and recombination PCR library made from mRFP and DsRed genes. All the sequences are represented as a percentage of the total library for each set of recombination experiments. Chimera background is calculated by dividing the number of non-native sequences that are over-representing a crossover sequence by the total number of sequences for each set of library that were analyzed and

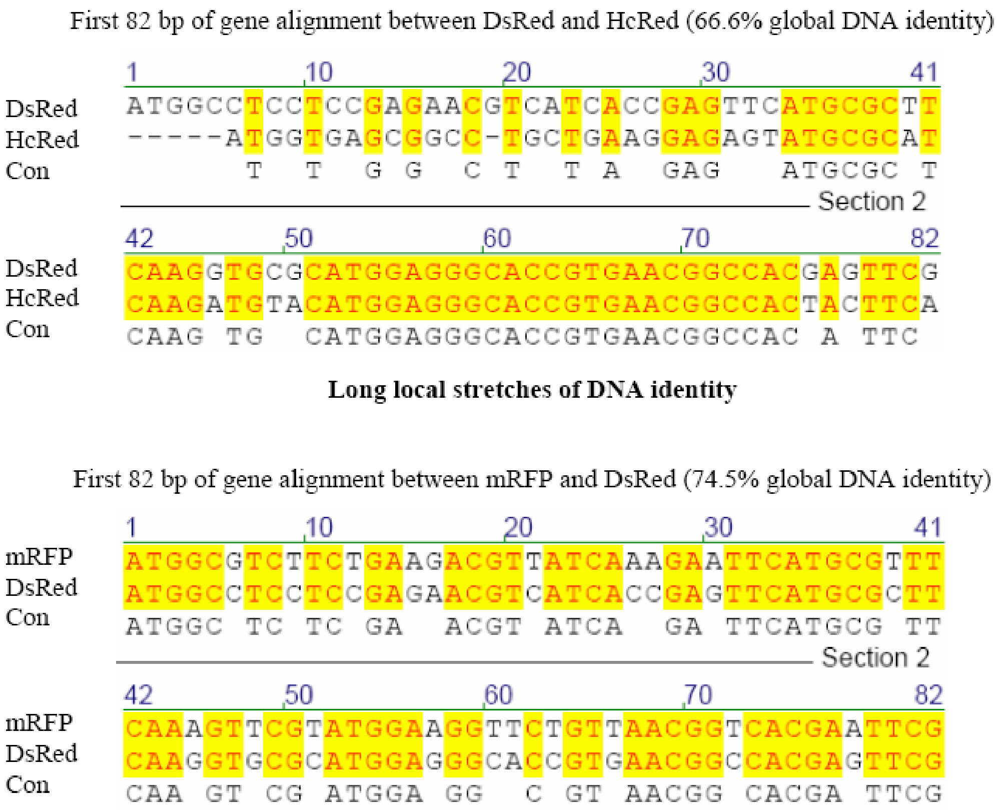
multiplying by a factor of 100. Useful sequences are the number of unique sequences over the total number of sequences for each set of library multiplied by a factor of 100. One, two and three crossovers percentages are calculated by dividing the number of sequences with one, two and three crossovers, respectively by the total number of sequences for each set of library and multiplied by a factor of 100. A total of 295 sequences (67 from DNA shuffling, 228 from RDA-PCR) are represented in the graph.

Surprisingly, we obtained parental background in approximately 10% of sequences for RD-PCR 2 and RD-PCR 4 libraries, despite the expectation that the use of skew primers should eliminate presence of parental background as crossovers must occur between the templates for amplification to proceed.

When templates extended in both directions (RD-PCR 5) were used, parental background was eliminated. Of 39 colonies sequenced, 72% contained unique sequences with predominantly one crossover per gene. One sequence with three crossovers and one with five crossovers were obtained. Crossover points were also more evenly distributed than the other recombinant PCR libraries made using one skew primer, which tended to show significant bias towards the end of the genes (Figure 4.4).

4.3.2 Recombination of fluorescent proteins with moderate DNA identity levels (66.6%): HcRed and DsRed

HcRed and DsRed genes represent a recombination scenario where there are large stretches of identical DNA which bias the crossovers towards these regions (Figure 4.6). For the DNA shuffling library, full-length HcRed and DsRed templates were used for recombination. Truncated gene templates of HcRed and DsRed were used for recombination PCR to reduce chimera background as there were large stretches of identical DNA regions in the front and back end of the HcRed and DsRed genes.



First 82 bp of gene alignment between DsRed and HcRed (66.6% global DNA identity)

```
        1        10        20        30        41
DsRed   ATGGCCTCCTCCGAGAACGTCATCACCGAGTTCATGCGCTT
HcRed   -----ATGGTGAGCGGCC-TGCTGAAGGAGAGTATGCGCAT
Con           T  T  G G  C T  T A  GAG   ATGCGC T
                                            Section 2
        42      50        60        70        82
DsRed   CAAGGTGCGCATGGAGGGCACCGTGAACGGCCACGAGTTCG
HcRed   CAAGATGTACATGGAGGGCACCGTGAACGGCCACTACTTCA
Con     CAAG TG  CATGGAGGGCACCGTGAACGGCCAC A TTC
```

Long local stretches of DNA identity

First 82 bp of gene alignment between mRFP and DsRed (74.5% global DNA identity)

```
        1        10        20        30        41
mRFP    ATGGCGTCTTCTGAAGACGTTATCAAAGAATTCATGCGTTT
DsRed   ATGGCCTCCTCCGAGAACGTCATCACCGAGTTCATGCGCTT
Con     ATGGC TC TC GA  ACGT ATCA  GA TTCATGCG TT
                                            Section 2
        42      50        60        70        82
mRFP    CAAAGTTCGTATGGAAGGTTCTGTTAACGGTCACGAATTCG
DsRed   CAAGGTGCGCATGGAGGGCACCGTGAACGGCCACGAGTTCG
Con     CAA GT CG ATGGA GG  C GT AACGG CACGA TTCG
```

Figure 4.6 DNA alignment of DsRed/HcRed and mRFP/DsRed. Long local stretches of identical DNA regions can be observed in the alignment of DsRed/HcRed.

Twenty variants were sequenced for the DNA shuffling library while 23 variants were sequenced for the recombination PCR library. A comparison of the DNA shuffling library and recombination PCR library revealed that the DNA shuffling library had a better diversity. The DNA shuffling library has 20% parental background, ten percent chimera background and 70% unique sequences. Of the DNA shuffling library unique sequences, 50% had one crossover, 36% had two crossovers and 14% three crossovers. The recombination PCR library has zero percent parental background, 57% chimera background and 43% unique sequences. Of the recombination PCR library unique sequences, 90% had one crossover while 10% of sequences had three crossovers (Figure

4.7).

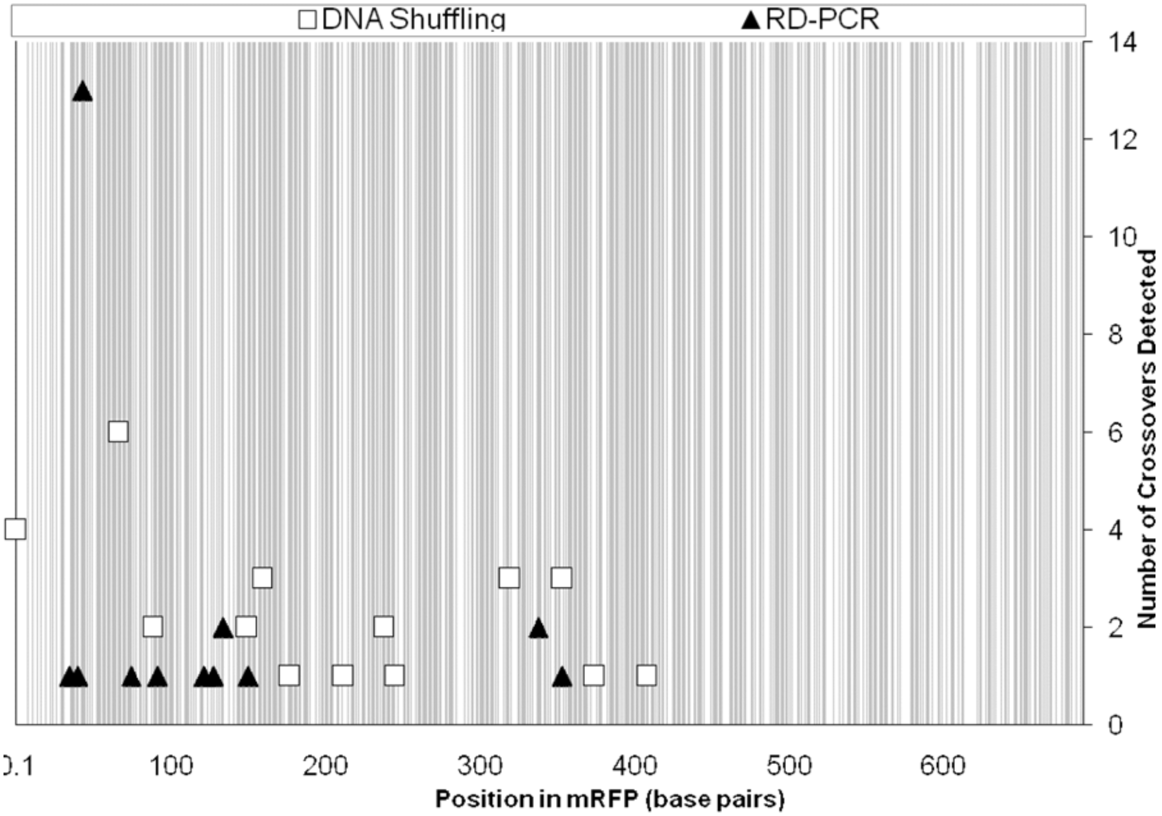


Figure 4.7 The frequency and location of crossovers in DNA shuffled and recombination PCR libraries made from HcRed and DsRed genes. A total of 43 sequences are represented in the chart.

4.3.3 Recombination of fluorescent proteins with low DNA identity levels (45%): mRFP and GFP

To explore the lowest limit of DNA identity level where the recombination protocols can still be performed, the mRFP and GFP gene templates with a 45% DNA identity level were chosen. To date, the lowest reported successful DNA shuffling performed on genes was 56% (42,44, 51). Thus, in this recombination scenario, it is not immediately obvious if either of the protocols could produce useful libraries at all.

From the sequencing data obtained on DNA shuffling library and recombination PCR library, it was evident that DNA shuffling did not work on mRFP and GFP templates as we only found parental sequences, recombination PCR produced a library with zero percent parental background, 82% chimera background and 18% unique sequences (Figure 4.8). Of the unique sequences, all of them (100%) had one crossover.

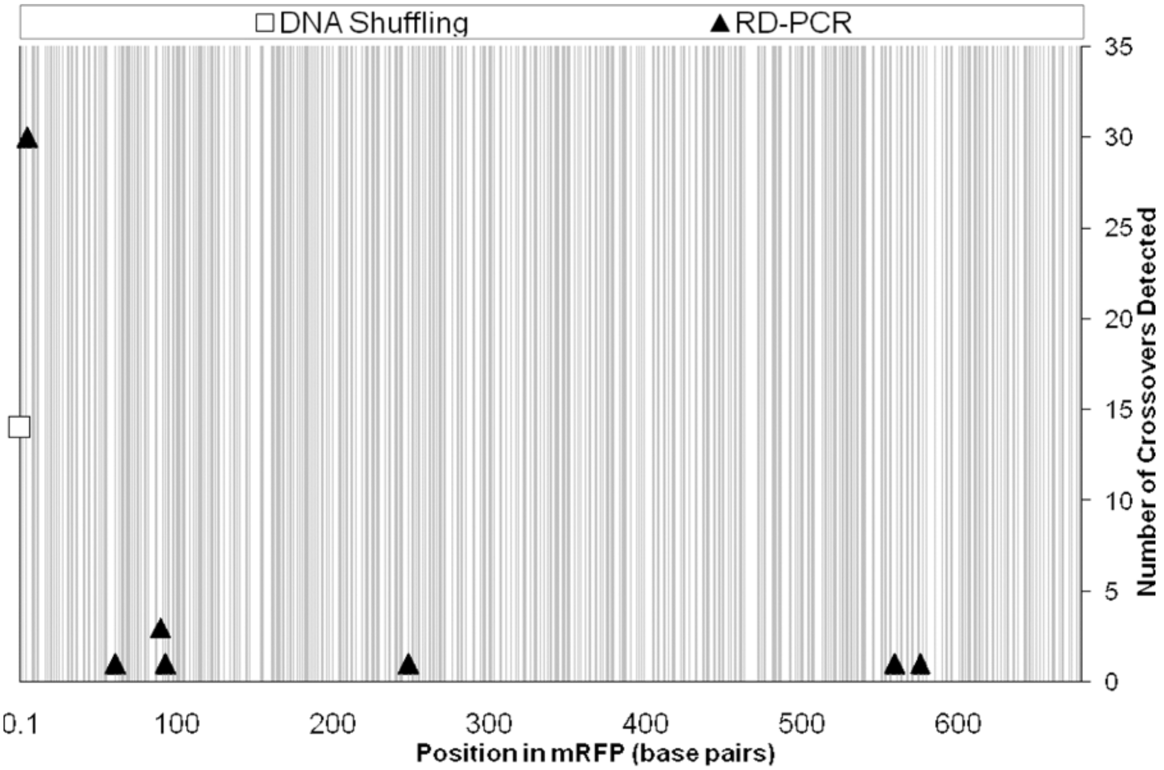


Figure 4.8 The frequency and location of crossovers in DNA shuffled and recombination PCR libraries made from mRFP and GFP genes. A total of 52 sequences are represented in the chart.

4.3.4 Discussion of all the recombination results

DNA shuffling and recombination PCR were applied on mRFP/DsRed, HcRed/DsRed and mRFP/GFP genes. There were no clear trends observed on the effect of DNA identity level on the percentage of useful sequences for DNA shuffling while recombination PCR features a monotonic decrease of useful sequences as the DNA identity level decreases (Figure 4.9). Comparison of the sequencing results across different levels of DNA identity levels indicate that when DNA shuffling can be used, it produced unique sequences with larger number of crossovers (Figure 4.10). For DsRed/mRFP and DsRed/HcRed DNA shuffled sequences, 52% of combined useful sequences featured more than one crossover while recombination PCR sequences featured 40% percent of combined useful sequences with more than one crossover.

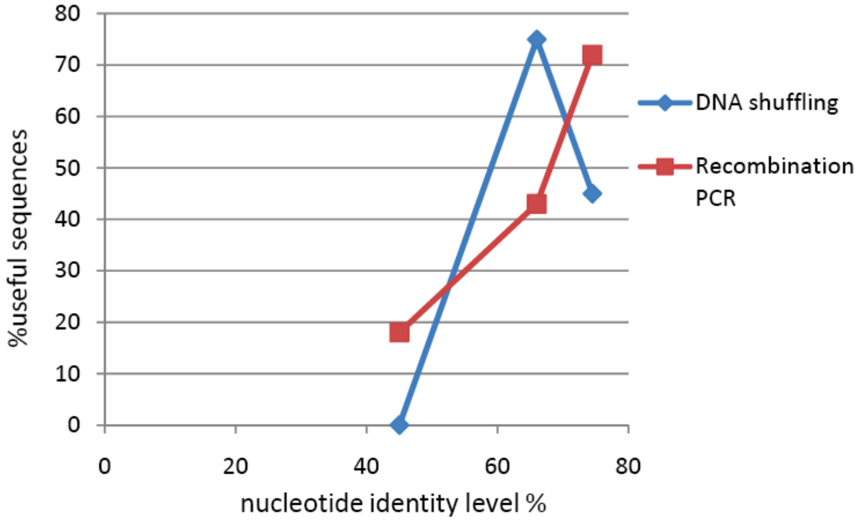


Figure 4.9 Distributions of useful sequences using DNA shuffling and recombination PCR on templates with different nucleotide identity levels.

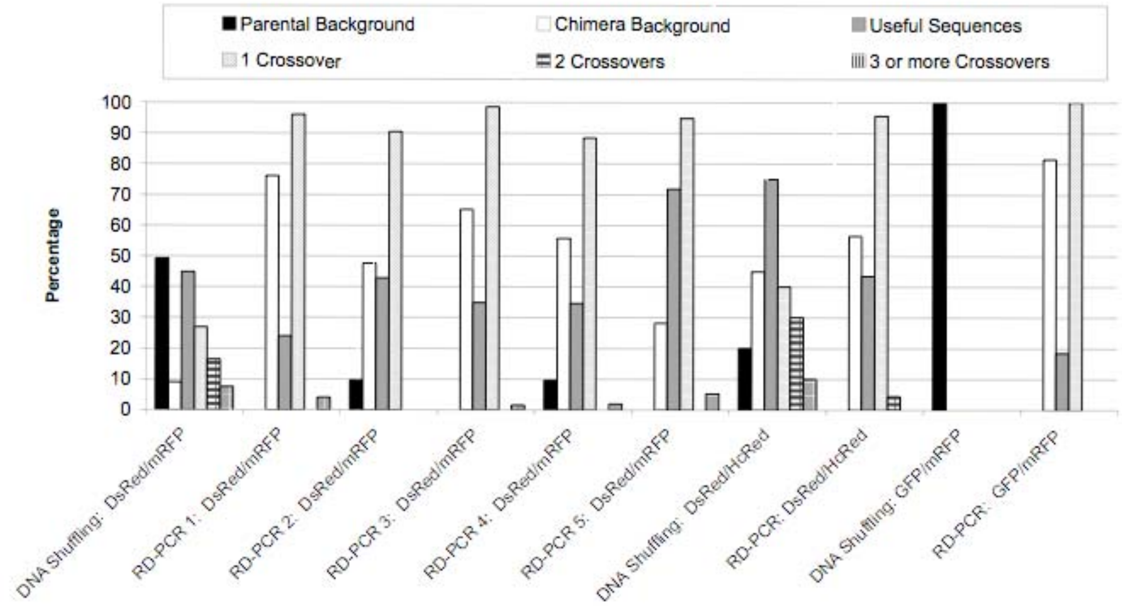


Figure 4.10 The distributions of DNA shuffling and recombination PCR library for all the fluorescent proteins. The various gene template combinations used for recombination are listed as follows: DsRed/mRFP: 75%, DsRed/HcRed: 66%, mRFP/GFP: 45% nucleotide identity level.

Overall, recombination PCR performed similarly to DNA shuffling in generating useful sequences. Of 289 combined sequences, recombination PCR library contained 37% useful sequences while DNA shuffling have 45% useful sequences of 101 sequences. To validate that the two recombination protocols produce equally useful number of sequences, Z-statistical analysis was performed. In this case, the null hypothesis states that the probability of obtaining useful sequences for both libraries is the same. In order to show that the two libraries are similar, it is sufficient to show that we cannot reject the null hypothesis. A p-value of 0.18 is obtained thus there is insufficient evidence that the two libraries are different, thereby validating the claim that the both protocols produced statistically equivalent levels of useful libraries of one or more crossovers.

Table 4.3 Overall mutation rate of DNA shuffling and recombination PCR.

Source	Protocol	overall mutagenic rate %	Total base pairs
mRFP and DsRed	RD-PCR	0.02	33900
mRFP and GFP	RD-PCR	0.04	25200
mRFP and DsRed	DNA-shuffling	0.03	21018
mRFP and GFP	DNA-shuffling	0.02	8136

The analysis of the sequences reveals that the mutagenic rate in recombination PCR is low and comparable to typical DNA shuffling experiments. An overall mutagenic rate of 0.02 to 0.04 % was observed in the sequences obtained from the recombination of mRFP/DsRed and mRFP/GFP genes (Table 4.3). A low mutagenic rate is typically preferred in DNA recombination experiments to limit the search in functional space. Joern *et al.* obtained 0.011% while others have reported values between 0.05 and 0.9% (9, 41, 42, 45).

4.4 Discussion

A systematic comparison on the performance of DNA shuffling and recombination PCR on producing useful libraries of fluorescent protein genes was performed. To date this is the first known reported head-to-head comparison of recombinant library generation protocols, performed on fluorescent proteins. A similar work was done comparing experimental and computationally generated libraries on fluorescent proteins by Treynor *et al.* (46). They compared, through experiments, computationally designed libraries with the libraries of randomly generated sequence diversity using fluorescent proteins. We found from our experiments that there were subtle differences in the quality of the libraries generated by the two protocols.

Depending on the type of gene templates used for recombination, DNA shuffling and recombination PCR library generation protocols perform differently. The next few paragraphs describe in detail, the insights gained from the recombination experiments.

4.4.1 Library diversity: DNA shuffling and recombination PCR

4.4.1.1 DNA shuffling and recombination PCR produced similarly useful libraries of one or more crossovers

Both DNA shuffling and recombination PCR were found to produce libraries of similarly useful diversity. Z-statistical analysis confirmed that both protocols produced equally useful libraries of one or more crossovers. From Figure 4.4, we also observe that recombination PCR libraries created using templates made with two-sided skew primers tended to have a better overall distribution of crossovers than DNA shuffling and one-sided or truncated template recombination PCR library. We expect thus, that the recombination PCR with two-sided skew templates would be a better library than DNA shuffling due to the better distribution of crossovers.

4.4.1.2 DNA shuffling and recombination PCR crossover tendencies

In general, DNA shuffling libraries are biased towards crossovers in regions of high nucleotide identity level (> 11 bp) as shown by Figure 4.11. Statistical analysis using Wilcoxon Rank Sum test on the DNA shuffled and recombination PCR libraries reveal that both libraries require different average lengths of nucleotide identity for crossovers to occur ($p = 0.0004$ that libraries are similar). The recombination PCR protocol requires shorter lengths of nucleotide identity for crossovers to occur. The details of the Wilcoxon Rank Sum calculations can be found in the supplement section 4.7.3.1.

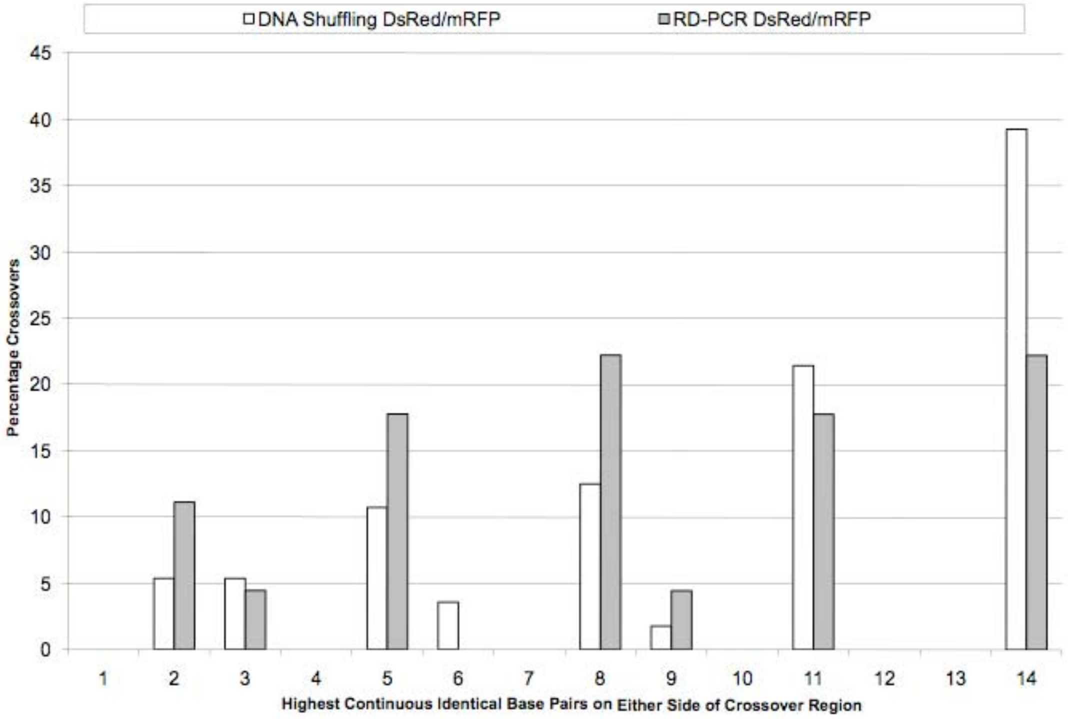


Figure 4.11 The distributions of number of identical nucleotide base-pairs required for crossovers for DNA shuffling and recombination PCR on mRFP/DsRed libraries

4.4.2 Recombinant PCR optimization: Two-sided skew templates libraries had the best diversity

The performance of recombination PCR is highly template-dependent. Figure 4.4 shows that the best recombination PCR library is produced using two-sided skew templates (Figure 4.3). During PCR, heteroduplexing (43), a phenomenon that involves skew extension without recombination via template switching, is more likely to occur when using one-sided skew and truncated templates than library generation using two-sided skew templates, evident from the presence of parental background obtained in the sequences despite the fact that it is theoretically not possible to obtain parental genes in the recombination library (15). In the one-sided skew and truncated libraries, parental background results from the accidental amplification on an unpaired extension containing no crossovers. Another key advantage of using two-sided skew templates for recombination PCR is that different sets of skew primers can be used to change the initial extension point on the gene templates, thereby generating a randomized library with less chimera background and a more spread out crossover points (Figure 4.4).

Other factors that could affect the outcome of recombination PCR are the volume of the PCR reaction, the concentration of templates used, and the cycling protocols used for the PCR. Using a smaller volume ~25 μl instead of 50 μl allows the temperature of the cycling reaction to equilibrate faster. Using a smaller starting concentration of template could theoretically encourage a more diverse library to be generated as there is less overall starting parental templates and more newly generated chimeras templates relative to wild-type templates that can be used for future amplification to generate more sequence diversity. Also, the temperature and number of cycles could in theory be adjusted to obtain a library with more crossovers. However, in the course of our experiments, when we attempted to vary the cycling protocols, we did not get bands that can be cloned. As a result, we settled for optimizing the library for a good PCR band, although possibly, a smeared or fainter gel band could potentially have a library with more chimeras.

4.4.3 Lower identical base pairs requirement required for recombination PCR than DNA shuffling

Recombination PCR has been shown to generate useful library for a nucleotide identity level as low as 45% and requires lower number of identical base-pairs than required for DNA shuffling for crossovers to occur. From Figure 4.8, it is evident that recombination PCR produced chimeras, however, the utility of the libraries produced is questionable as only 18% unique sequences were generated.

The finding that recombination PCR produced chimera library with 45% nucleotide identity level agrees with the observation that recombination PCR requires lesser nucleotide identity levels for crossovers to occur than DNA shuffling (Figure 4.11). Figure 4.8 and Figure 4.10 show that recombination PCR is a more robust protocol and will produce useful libraries when nucleotide identity is as low as 45%.

4.4.4 Inherent characteristics of recombination PCR and DNA shuffling

When recombining two genes, recombination PCR can only produce an odd number of crossovers to get amplification of the templates because of the skew annealing primer regions on the templates. On the other hand, DNA shuffling can produce chimeras with even and odd number of crossovers. One way to overcome the potential reduction in diversity when using recombination PCR is to use three or more genes. With three gene templates, even numbers of crossovers are possible because two different gene templates can contain the same skew primer annealing regions that allow amplification to occur.

DNA shuffling has parental background while recombination PCR, when performed under the optimized conditions described in this chapter, does not. Recombination by DNA shuffling relies on homologous regions annealing. In reassembly, the probability of the parental gene fragments annealing to reform the full length parental gene is energetically favored over the formation of chimeras with imperfect base pairing. However, a possible way to reduce parental background is to perform multiple rounds of DNA shuffling.

DNA shuffling is a difficult protocol to use compared to recombination PCR. The optimization of DNA shuffling is a time-consuming process that involves generating a high quantity of templates and a trial DNA digestion step to determine the ideal digestion time. Following gel purification of the digested fragments, two PCR steps are required. The first involves the reassembly and the second involves amplification of the chimeras. Experienced users of DNA shuffling can use it as a tool for generating chimeras. Less experienced users, however, would stand to benefit more from recombination PCR in which the incorporation of skew-annealing regions into the templates and the use of skew-annealing primers with modified temperature protocols would produce a similarly useful library. It should also be noted, however, depending on the genes recombined, recombination PCR could potentially still require some optimization of template concentration and cycling conditions to produce a clonable band.

4.5 Conclusion

A controlled recombination experiment was performed using two recombination protocols, DNA shuffling and recombination PCR on fluorescent protein genes. Generally, both the protocols produced useful libraries but we found that recombination PCR method is an easier technique to apply. Two-sided skew annealing templates should be used when using the recombination PCR protocol to effectively eliminate parental background and to produce a library with a better crossover distribution. Also, we found that recombination PCR could work with genes of lower nucleotide identity level (45%) than DNA shuffling.

4.5.1 Future trends: Data driven protein evolution and synthetic library

Recent trends in library generation involve computational analysis of the library that direct the generation of combinatorial libraries. Fox *et al.* applied 16 rounds of recombination on halohydrin dehalogenase proteins, obtained information about the important residues from computational analysis of sequence along with phenotype and designed libraries spiked with degenerate codons around the important residues (48). They obtained 4000-fold improvement on the activity of halohydrin dehalogenase proteins. Treynor *et al.* developed a structure-based algorithm that guides library design for fluorescent protein and found better functionality with the designed libraries than the randomly generated library (46). Liao *et al.* developed machine-learning guided approaches that involve analyzing the sequence-activity relationships of 59 test variants to design more productive libraries (47). A 20-fold improved proteinase K variant was obtained as a result of such an effort.

With decreasing costs of synthesis of whole genes, designer synthetic recombination libraries can be created. It is now more economical to order oligonucleotides than ten years ago as the price per base-pair dropped from $4 to approximately $0.30 (49). In fact, there are companies such as DNA 2.0 and Codon Devices that synthesize more 40,000 base pairs genes routinely. Thus, individuals could focus on designing libraries and be less involved in the library generation process. The challenge in such cases lies in the identification of key residues for mutagenesis.

This chapter and its contents were recently accepted for publication in BMC Biotechnology.

Javier F. Chaparro-Riggers, Bernard L.W. Loo, Karen M. Polizzi, Phillip R. Gibbs, Xiao-Song Tang, Mark J. Nelson and Andreas S. Bommarius. Revealing Biases Inherent in Recombination Protocols (2008). *BMC Biotechnology.*

4.6 References

1. Neylon, C. (2004) Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res,* **32**, 1448-1459.

2. Lutz, S. and Patrick, W.M. (2004) Novel methods for directed evolution of enzymes: quality, not quantity. *Current opinion in biotechnology*, **15**, 291-297.

3. Arnold, F.H. and Georgiou, G. (2003) *Directed evolution library creation : methods and protocols*. Humana Press, Totowa, N.J.

4. Jaeger, K.E. and Eggert, T. (2004) Enantioselective biocatalysis optimized by directed evolution. *Current opinion in biotechnology*, **15**, 305-313.

5. Bornscheuer, U.T. and Pohl, M. (2001) Improved biocatalysts by directed evolution and rational protein design. *Current opinion in chemical biology,* **5**, 137-143.

6. Taylor, S.V., Kast, P. and Hilvert, D. (2001) Investigating and Engineering Enzymes by Genetic Selection. *Angew Chem Int Ed Engl*, **40**, 3310-3335.

7. Wong, T.S., Roccatano, D., Zacharias, M. and Schwaneberg, U. (2006) A statistical analysis of random mutagenesis methods used for directed protein evolution. *Journal of molecular biology*, **355**, 858-871.

8. Wong, T.S., Zhurina, D. and Schwaneberg, U. (2006) The diversity challenge in directed protein evolution. *Combinatorial chemistry & high throughput screening*, **9**, 271-288.

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

10. Kikuchi, M., Ohnishi, K. and Harayama, S. (2000) An effective family shuffling

method using single-stranded DNA. *Gene*, **243**, 133-137.

11. Shao, Z., Zhao, H., Giver, L. and Arnold, F.H. (1998) Random-priming in vitro recombination: an effective tool for directed evolution. *Nucleic Acids Res*, **26**, 681-683.

12. Abecassis, V., Pompon, D. and Truan, G. (2000) High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic acids research*, **28**, E88.

13. Ness, J.E., Kim, S., Gottman, A., Pak, R., Krebber, A., Borchert, T.V., Govindarajan, S., Mundorff, E.C. and Minshull, J. (2002) Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat Biotechnol*, **20**, 1251-1255.

14. Zhao, H., Giver, L., Shao, Z., Affholter, J.A. and Arnold, F.H. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat Biotechnol*, **16**, 258-261.

15. Akinori Ikeuchi, Y.K., Tomoya Shinbata, and Tsueneo Yamane. (2003) Chimeric gene library construction by a simple and highly versatile method using Recombination-Dependent Exponential Amplification. *Biotechnol. Prog.*, **19**, 1460-1467.

16. Coco, W.M., Levinson, W.E., Crist, M.J., Hektor, H.J., Darzins, A., Pienkos, P.T., Squires, C.H. and Monticello, D.J. (2001) DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat Biotechnol*, **19**, 354-359.

17. Ostermeier, M., Nixon, A.E. and Benkovic, S.J. (1999) Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg Med Chem*, **7**, 2139-2144.

18. Ostermeier, M., Nixon, A.E., Shim, J.H. and Benkovic, S.J. (1999) Combinatorial protein engineering by incremental truncation. *Proc Natl Acad Sci U S A*, **96**, 3562-3567.

19. Ostermeier, M., Shim, J.H. and Benkovic, S.J. (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat Biotechnol*, **17**, 1205-1209.

20. Ostermeier, M.a.B., S.J. (2001) Construction of hybrid gene libraries involving the circular permutation of DNA. *Biotechnology Letters*, **23**, 303-310.

21. Lutz, S., Ostermeier, M., Moore, G.L., Maranas, C.D. and Benkovic, S.J. (2001) Creating multiple-crossover DNA libraries independent of sequence identity. *Proc Natl Acad Sci U S A*, **98**, 11248-11253.

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

23. Bommarius, A.S. and Riebel, B.R. (2004) *Biocatalysis - Fundamentals and Applications*. 1st ed. Wiley-VCH Verlag GmbH & Co., Weinheim.

24. Crameri, A., Dawes, G., Rodriguez, E., Jr., Silver, S. and Stemmer, W.P. (1997) Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nat Biotechnol*, **15**, 436-438.

25. Crameri, A., E.A., W., E., T. and Stemmer, W.P. (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol*, **14**, 315-319.

26. Reetz, M.T. (2004) Controlling the enantioselectivity of enzymes by directed evolution: practical and theoretical ramificiations. *Proc Natl Acad Sci U S A*, **101**, 5716-5722.

27. Zhao, H. and Arnold, F.H. (1997) Functional and nonfunctional mutations distinguished by random recombination of homologous genes. *Proc Natl Acad Sci U S A*, **94**, 7997-8000.

28. Moore, J.C., Jin, H.M., Kuchner, O. and Arnold, F.H. (1997) Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. *Journal of molecular biology*, **272**, 336-347.

29. Schmidt-Dannert, C., Umeno, D. and Arnold, F.H. (2000) Molecular breeding of carotenoid biosynthetic pathways. *Nat Biotechnol*, **18**, 750-753.

30. Powell, S.K., M.A., K., A., P., R., M., I., B., M., P., E., O., Stemmer, W.P. and N.W., S. (2000) Breeding if retroviruses by DNA shuffling for improved stability and processing yields. *Nat Biotechnol*, **18**, 1279-1282.

31. Ness, J.E., M., W., Giver, L., M., B., J.R., C., Borchert, T.V., Stemmer, W.P. and Minshull, J. (1999) DNA shuffling of subgenomic sequences of subtilisin. *Nat Biotechnol*, **17**, 893-896.

32. Soong, N.W., Nomura, L., Pekrun, K., Reed, M., Sheppard, L., Dawes, G. and Stemmer, W.P. (2000) Molecular breeding of viruses. *Nat Genet.*, **25**, 436-439.

33. Christians, F.C., L., S., Crameri, A., G., F. and Stemmer, W.P. (1999) Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nat Biotechnol*, **17**, 259-264.

34. Raillard, S., Krebber, A., Y., C., J.E., N., E., B., Trinidad, R., Fullem, R., C., D., M., W., Seffernick, J. *et al.* (2001) Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chem Biol*, **8**, 891-898.

35. Crameri, A., Raillard, S., Bermudez, E. and Stemmer, W.P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288-291.

36. Chang, C.C., T, C.T., Cox, B.W., Dawes, G.N., Stemmer, W.P., Punnonen, J. and Patten, P.A. (1999) Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.*, **17**, 793-797.

37. Milano, J. and Tang, X.-S. (2004) US 20040014085.

38. Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Research*, **30**, -.

39. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7877-7882.

40. Joern, J.M. (2003) DNA shuffling. *Methods Mol. Biol.*, **231**, 85-89.

41. Abecassis, V., Pompon, D. and Truan, G. (2000) High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res*, **28**, E88.

42. Joern, J.M., Meinhold, P. and Arnold, F.H. (2002) Analysis of shuffled gene libraries. *Journal of molecular biology*, **316**, 643-656.

43. Odelberg, S.J., Weiss, R.B., Hata, A. and White, R. (1995) Template-switching during DNA-synthesis by *Thermus-aquaticus* DNA-polymerase-I. *Nucleic Acids Research*, **23**, 2049-2057.

44. Zhou, Z., Zhang, A.-H., Wang, J.-R., Chen, M.-L., Li, R.-B., Yang, S. and Yuan, Z.-Y. (2003) Improving the specific synthetic activity of a Pencillin G acylase using DNA family shuffling. *ACTA BIOCHIMICA et BIOPHYSICA SINICA*, **35**, 573-579.

45. Zhao, H. and Arnold, F.H. (1997) Optimization of DNA shuffling for high fidelity recombination. *Nucleic Acids Res*, **25**, 1307-1308.

46. Treynor, T.P., Vizcarra, C.L., Nedelcu, D. and Mayo, S.L. (2007) Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A*, **104**, 48-53.

47. Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C. and Minshull, J. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnology*, **7**, 16..

48. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S. *et al.* (2007) Improving catalytic

function by ProSAR-driven enzyme evolution. *Nature biotechnology*, **25**, 338-344.

49. Voigt, A.C. (2007) In Biology Seminar 030607, I. (ed.), Atlanta.

50. Hayter, A.J. (2002) *Probability and Statistics for Engineers and Scientists*. 2nd ed, Pacific Grove.

51. Kaper, T., Brouns, S.J., Geerling, A.C., De Vos, W.M. and Van der Oost, J. (2002) DNA family shuffling of hyperthermostable beta-glycosidases. *Biochem J.*, **368**, 461-470.

4.7 Supplement

4.7.1 *E*. coli expression optimized primers used for the reassembly of mRFP gene

no.	Primer Sequences
1	5'-ATGGCGTCTTCTGAAGACGTTATCAAAGAATTCATGCGTTTCAAAGT -3'
2	5'-TCGTATGGAAGGTTCTGTTAACGGTCACGAATTCGAAATCGAAGGTG -3'
3	5'-AAGGTGAAGGTCGTCCGTACGAAGGTACCCAGACCGCGAAACTG -3'
4	5'-AAAGTTACCAAAGGTGGTCCGCTGCCGTTCGCGTGGGAC -3'
5	5'-ATCCTGTCTCCGCAGTTCCAGTACGGTTCTAAAGCGTACGTTAAACACCCGG -3'
6	5'-CGGACATCCCGGACTACCTGAAACTGTCTTTCCCGGAAGGT -3'
7	5'-TTCAAATGGGAACGTGTTATGAACTTCGAAGACGGTGGTGTTGTTA - 3'
8	5'-CCGTTACCCAGGACTCTTCTCTGCAGGACGGTGAATTCATCTACAA -3'
9	5'-AGTTAAACTGCGTGGTACCAACTTCCCGTCTGACGGTCCGG -3'
10	5'-TTATGCAGAAAAAAACCATGGGTTGGGAAGCGTCTACCGAACGTAT -3'
11	5'-GTACCCGGAAGACGGTGCGCTGAAAGGTGAAATCAAAATGCG -3'
12	5'-TCTGAAACTGAAAGACGGTGGTCACTACGACGCGGAAGTTAAAACC -3'
13	5'-ACCTACATGGCGAAAAAACCGGTTCAGCTGCCGGGTGCG -3'
14	5'-TACAAAACCGACATCAAACTGGACATCACCTCTCACAACGAAGACTACACCA -3'
15	5'-TCGTTGAACAGTACGAACGTGCGGAAGGTCGTCACTCTACCGGTGCG -3'
16	5'-TTACGCACCGGTAGAGTGACGACCTTC -3'
17	5'-CGCACGTTCGTACTGTTCAACGATGGTGTAGTCTTCGTTGTGAGAGGTGAT -3'
18	5'-GTCCAGTTTGATGTCGGTTTTGTACGCACCCGGCAGCTGAA -3'

4.7.1 continued

19	5'-CCGGTTTTTTCGCCATGTAGGTGGTTTTAACTTCCGCGTCGTAGT -3'
20	5'-GACCACCGTCTTTCAGTTTCAGACGCATTTTGATTTCACCTTTCAG -3'
21	5'-CGCACCGTCTTCCGGGTACATACGTTCGGTAGACGCTTCCCA -3'
22	5'-ACCCATGGTTTTTTTCTGCATAACCGGACCGTCAGACGGG -3'
23	5'-AAGTTGGTACCACGCAGTTTAACTTTGTAGATGAATTCACCGTCCTG -3'
24	5'-CAGAGAAGAGTCCTGGGTAACGGTAACAACACCACCGTCTTCGAAG -3'
25	5'-TTCATAACACGTTCCCATTTGAAACCTTCCGGGAAAGACAGTTTC -3'
26	5'-AGGTAGTCCGGGATGTCCGCCGGGTGTTTAACGTACGCTTTAGAACCGT -3'
27	5'-ACTGGAACTGCGGAGACAGGATGTCCCACGCGAACGGCA -3'
28	5'-GCGGACCACCTTTGGTAACTTTCAGTTTCGCGGTCTGGGTACC -3'
29	5'-TTCGTACGGACGACCTTCACCTTCACCTTCGATTTCGAATTCGTGA -3'
30	5'-CCGTTAACAGAACCTTCCATACGAACTTTGAAACGCATGAATTCTTT - 3’

4.7.2 Sequencing data of recombination of fluorescent proteins

Table 4.4 Sequencing results of recombination-PCR on mRFP and DsRed. mRFP: 678bp, DsRed: 678bp. The crossover position refers to the nucleotide numbering of mRFP.

Method	Genes	Crossover position	No. of crossovers	Highest Continuous bp	Frequency
RD-1	mRFP, DsRed	6	1	5	39
	74.5% id	12	1	2	1
		78	1	5	1
		156	1	2	1
		159	1	5	1
		165	1	5	1
		186	1	14	1
		300	1	11	1
		414	1	8	1
		660	1	3	1
		6, 21, 39	3	5, 4, 8	1
		6, 248, 300	3	5, 6, 11	1
					50 sequences total
RD-2	mRFP, DsRed	NA	0	NA	2
	(5 Ft)	6	1	5	9
	74.5%	7*	1	5	2
		72	1	5	1
		186	1	14	1
		204	1	14	1
		300	1	11	1
		417	1	8	1
		618	1	14	1
		633	1	14	2
					21 sequences total
RD-3	mRFP, DsRed	165	1	5	2
	(44Ft)	171	1	14	6

Table 4.4 continued

	74.5%	186	1	14	3
		189	1	14	1
		204	1	14	6
		207	1	2	1
		247	1	6	2
		258	1	5	1
		300	1	11	1
		321	1	11	1
		348	1	11	1
		363	1	11	1
		366	1	2	1
		609	1	14	2
		618	1	14	9
		633	1	14	10
		636	1	8	2
		645	1	8	6
		648	1	2	1
		660	1	3	1
		664	1	3	5
		666	1	3	2
		228, 240, 264	3	5, 8, 5	1
					66 sequences total
RD-4	DsRed, mRFP (43Bt)	NA	1	NA	5
	(40Ft)	50	1	5	18
		159	1	5	1
		171	1	14	3
		186	1	14	1
		190	1	14	1
		204	1	2	3
		231	1	8	1
		276	1	5	2
		300	1	11	1
		348	1	11	1
		438	1	5	1
		594	1	14	1
		609	1	14	3
		618	1	14	2
		633	1	14	4
		636	1	8	1

Table 4.4 continued

	Not possible crossover position	645	1	8	2
		78, 300, 336	3	5, 11, 11	1
					52 sequences total
RD-5	DsRed, mRFP				
	Head & tail extension	6	1	5	1
		9	1	2	1
		15	1	8	1
		48	1	2	1
		57	1	5	2
		108	1	2	1
		186	1	14	1
		228	1	5	1
		240	1	8	1
		288	1	11	1
		300	1	11	2
		306	1	5	1
		333	1	11	1
		348	1	11	1
		363	1	11	2
		417	1	8	1
		426	1	8	1
		498	1	5	1
		549	1	9	2
		570	1	8	1
		609	1	14	2
		618	1	14	3
		633	1	14	1
		633	1	**14**	2
		636	1	8	2
		645	1	8	2
		648	1	2	1
		528,535,648	3	3,3,2	1
		6,204,231,264,363	5	5,14,8,5,11	1

Table 4.4 continued

39 sequences total

* - frame shift with respect to DsRed

Table 4.5 Sequencing results of DNA-shuffling on mRFP and DsRed. The crossover position refers to the numbering of mRFP.

Method	Genes	Crossover position	No. of crossovers	Highest Continuous bp	Frequency
DNA-shuffling	mRFP, DsRed 74.5% id	NA	0	NA	33
		30	1	8	1
		39	1	8	1
		153	1	2	1
		171	1	14	1
		186	1	14	3
		247	1	6	1
		321	1	11	1
		417	1	8	1
		549	1	9	1
		589	1	3	1
		618	1	14	2
		633	1	14	3
		648	1	2	1
		90, 594	2	5,14	1
		186, 247	2	14, 6	1
		186, 258	2	14, 5	1
		186, 321	2	14, 11	1
		186, 618	2	14, 14	1
		285, 300	2	2, 11	1
		300, 432	2	11, 5	1
		321,333	2	11,11	1
		363, 589	2	11, 3	1
		589, 594	2	3, 14	1
		633, 645	2	14, 8	1
		171, 393, 618	3	14, 5, 14	1
		204, 363, 645	3	14, 11, 8	1
		321, 609, 633	3	11, 14, 14	1

Table 4.5 continued

		84, 186, 363, 414	4	5, 14, 11, 8	1
		300, 315, 363, 636	4	11, 11, 11, 11	1
					67 sequences total

Table 4.6 Sequencing results of DNA shuffling and recombination-PCR on DsRed and HcRed (66% sequence identity). DsRed: 678bp, HcRed: 687 bp. Crossover position with respect to DsRed gene.

Method	Genes	Crossover position	No. of crossovers	Highest Continuous bp	Frequency
RD-PCR	HcRed, DsRed	46	1	5	1
	(93 Bt) (44 Ft)	50	1	25	13
	66.0% id	86	1	15	1
		141	1	6	1
		148	1	6	1
		155	1	17	1
		174	1	17	1
		391	1	17	2
		409	1	17	1
		40, 106, 155	3	6, 5, 17	1
					23 sequences total
DNA-shuffling	HcRed, DsRed	NA	0	NA	4
	66.0% id	76	1	25	2
		102	1	15	1
		173	1	17	1
		184	1	9	1
		205	1	8	1
		275	1	17	1
		409	1	17	1
		102, 370	2	15, 15	1

Table 4.6 continued

		275, 409	2	17, 17	1
		283, 409	2	6, 17	1
		370, 472	2	15, 11	1
		76, 184	2	25, 9	2
		76, 173, 370	3	25, 17, 15	1
		76, 245, 433	3	25, 10, 9	1
					20 sequences total

Table 4.7 Sequencing results of DNA shuffling and recombination-PCR on mRFP and GFP. mRFP: 678bp, GFP: 717 bp. Crossover position with respect to mRFP gene.

Method	Genes	Crossover position	No. of crossovers	Highest Continuous bp	Frequency
RD-PCR	mRFP, GFP	12	1	5	30
	45.0% id	72	1	3	1
		106	1	9	3
		110	1	1	1
		294	1	3	1
		579	1	1	1
		661	1	2	1
					38 sequences total
DNA-shuffling	mRFP, GFP	NA	0	NA	14
	45.0% id				14 sequences total

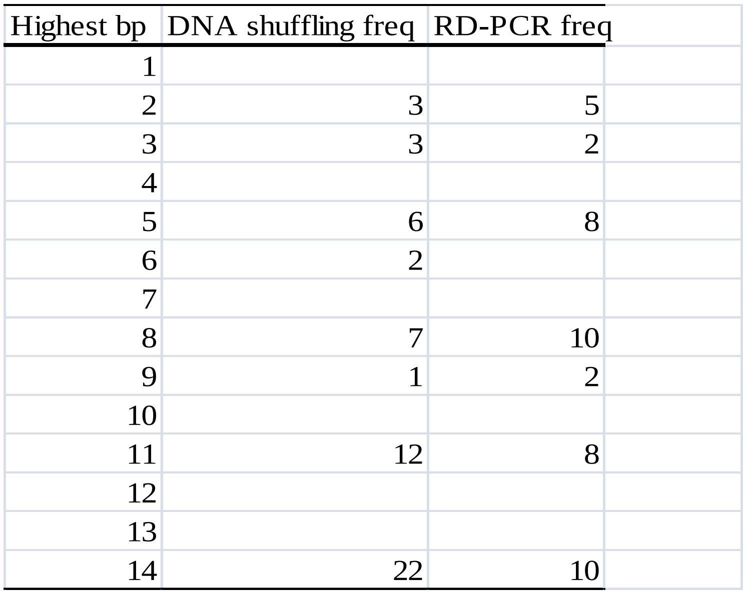
4.7.3 Statistical Calculations

4.7.3.1 Wilcoxon Rank Sum test on DNA shuffled and recombination PCR

The rank sum tests can be used to analyze data sets obtained independently to determine if they have similar distributions. The standard normal distribution can be used to determine the probability of two libraries being similar (50).

To begin with, the null hypothesis is set such that H_o: $L_{dnashuffling} = L_{recombinationPCR}$, where we assume that the distributions of libraries generated by DNA shuffling and recombination PCR are similar.

From Table 4.4 RD-5 and Table 4.5, the following data were obtained.



Highest bp	DNA shuffling freq	RD-PCR freq
1		
2	3	5
3	3	2
4		
5	6	8
6	2	
7		
8	7	10
9	1	2
10		
11	12	8
12		
13		
14	22	10

The highest bp were multiplied by frequency of DNA shuffle or RD-PCR to combine the crossover base pairs into a single number which were then ranked.

Count	Protocol	bp*freq	Protocol	bp*freq
1	DNA shuffling	0	RD-PCR	0
2	DNA shuffling	6	RD-PCR	10
3	DNA shuffling	9	RD-PCR	6
4	DNA shuffling	0	RD-PCR	0
5	DNA shuffling	30	RD-PCR	40
6	DNA shuffling	12	RD-PCR	0
7	DNA shuffling	0	RD-PCR	0
8	DNA shuffling	56	RD-PCR	80
9	DNA shuffling	9	RD-PCR	18
10	DNA shuffling	0	RD-PCR	0
11	DNA shuffling	132	RD-PCR	88
12	DNA shuffling	0	RD-PCR	0
13	DNA shuffling	0	RD-PCR	0
14	DNA shuffling	308	RD-PCR	140

The data were then ranked. For bp*freq values which tied, the two ranks were averaged.

Rank	Protocol	bp*freq	Avg of tied values
1	DNA shuffling	0	7
2	DNA shuffling	0	
3	DNA shuffling	0	
4	DNA shuffling	0	
5	DNA shuffling	0	
6	DNA shuffling	0	
7	RD-PCR	0	
8	RD-PCR	0	
9	RD-PCR	0	
10	RD-PCR	0	
11	RD-PCR	0	
12	RD-PCR	0	
13	RD-PCR	0	
14	DNA shuffling	6	14.5
15	RD-PCR	6	
16	DNA shuffling	9	16.5
17	DNA shuffling	9	
18	RD-PCR	10	
19	DNA shuffling	12	
20	RD-PCR	18	
21	DNA shuffling	30	
22	RD-PCR	40	
23	DNA shuffling	56	
24	RD-PCR	80	
25	RD-PCR	88	
26	DNA shuffling	132	
27	RD-PCR	140	
28	DNA shuffling	308	

M = number of samples for A (DNA shuffling) = 14,

N = number of samples for B (RD-PCR) = 14

S_a = Sum of all ranks = 7+14.5+16.5+17+18+19+20+21+22+23+24+25+26+27+28

U_a = Sa-M*(M+1)/2 = 308-14*(14+1)/2

$$\frac{M \bullet N}{2} = \frac{14 \bullet 14}{2} = 98$$

$$Z = \frac{U_a - \frac{M \bullet N}{2}}{[(M \bullet N)(\frac{M \bullet N + 1}{12})]^{0.5}} = 4.82$$

I approximated 4.82 ~ 3.49 so that I can read the probability from the standard normal distribution table without having to calculate it.

Two sided P-value = 2 *(1- φ(Z)) = 2* (1- φ(3.49)) = 2*(1-(0.9998))=0.0004

The low P-value indicates that it is the data is not statistically significant enough to believe that the two distributions of DNA shuffled library and recombination PCR on mRFP/DsRed are equal.

4.7.3.2 Z-statistics on DNA shuffled and recombination libraries of all fluorescent proteins

The null hypothesis, is set such that H_0: $P_a = P_b$ versus Ha:P_a not equal to P_b. Where Pa = probability of DNA shuffling library having useful sequences with one or more crossovers and Pb = probability of recombination PCR library having useful sequences with one or more crossovers. To show that both libraries are potentially similar, it is sufficient to demonstrate that we cannot reject the null hypothesis.

Protocols	Label	Useful sequences	Label	Total sequences	Label	Probability
DNA shuffle	x	45	45	101	Pa	0.45
Recombination	y	107	107	289	Pb	0.37

Common proportion is calculated by the formula, $p = \frac{(x+y)}{(m+n)} = 0.39$

$$Zvalue = \frac{P_a - P_b}{\left[p \bullet (1-p) \bullet (\frac{1}{n} + \frac{1}{m}) \right]^{0.5}} = 1.34$$

P-value = 2 X φ (-1.34) ~ approx 0.1802 (obtained from the standard normal distribution table)

Thus, it is statistically significant. We cannot reject the null hypothesis. There is evidence present that the two libraries have equally good number of useful sequences with one or more crossovers.

CHAPTER 5

DATA-DRIVEN PROTEIN ENGINEERING USING SUPPORT VECTOR MACHINES AND BOOLEAN LEARNING

5.1 Introduction

There have been recent advances through rational design of proteins for which the 3-D structure and amino acid residues involved in catalysis are well known (1,2). However, there are many proteins where the 3-D structure is not known and structure-to-function map information is not readily accessible. In cases where these pieces of information are available, the engineering of these proteins is still difficult, as sequence-function mapping cannot be easily predicted. In the absence of knowledge about a protein's structure and mechanism, directed evolution can be used as a method to search randomly and iteratively through sequence space to find improved variants. The sequence-activity data obtained from directed evolution experiments can be analyzed to identify interacting residues and parental templates can be engineered to produce a library with more functional variants. A library with more functional variants improves the chances of success for protein engineering experiments. We will report on the application of such an approach on mRFP and DsRed fluorescent proteins.

Directed Evolution (DE) is a process in which mutations are produced, usually at random, in an existing protein sequence in search for a desired property (3,4). Even after considerable advances in library generation and screening methods (5-7), only a small fraction of all the possible sequences (20^L distinct sequences for a protein of length L) can be characterized. Moreover, due to the complex nature of the sequence-to-function map, a majority of the mutations lead to inactive or unfolded proteins. It is known, however, that only a small fraction of the amino acids present in a protein contribute significantly to the protein's properties (8,9). These circumstances provide the motivation to identify such residues, so that experiments can be done more efficiently to increase the chances of success.

It is known that there exist pairs of residues that interact with each other in a protein's three-dimensional structure (10). To create active variants of a protein, it is important to preserve interacting residues. Figure 5.1 shows an example of a recombination that results in non-compatible residues at an interactive distance. Such residues can be identified by applying the concept of feature selection in machine learning to the data that are generated during DE experiments. In theory, this will increase the probability of finding a variant with an improved function since a larger number of the variants generated will be functional. Also, though not yet proven, since a larger number of crossovers can be generated without deactivating the proteins, functional diversity can also be expected to increase.

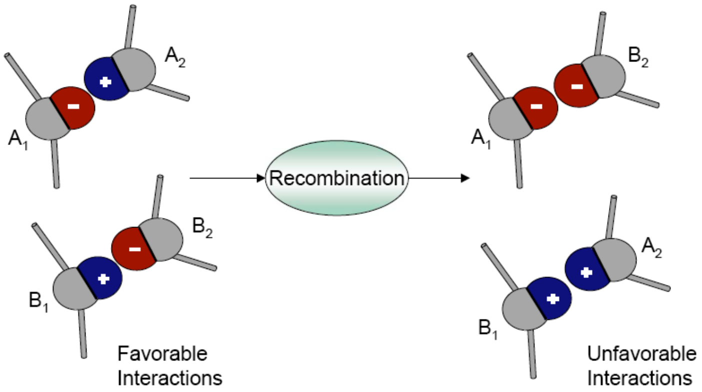


Figure 5.1 Recombination that disrupts favorable interactions.

5.1.1 Previous work performed on the topic

5.1.1.1 Famclash: A procedure that relies on identifying residues positions in the **fam**ily protein sequences to verify conformity to identified conserved properties from which any deviations are denoted as residue-residue **clash**es.

Saraf and Maranas (11) investigated a FamClash based approach that involved using the structure of a protein to find possible interacting residues. These pairs were amino acid residues spatially located close to each other in three-dimensional structures that seemed to have electrostatic, polar, or volumetric compatibility. Saraf and Maranas succeeded in predicting qualitatively the pattern of activity of protein variants of dihydrofolate reductases using this technique. This approach, however, is limited to proteins whose highly accurate three-dimensional structure of less than 1.0 Å rmsd is available. Moreover, experimental data were not used to identify but only to validate the identified interactions.

5.1.1.2 SCHEMA: Clusters of bits that interact favorably.

Meyer *et al.* (10) conducted a similar study on the structure of the protein. They proposed a SCHEMA-guided approach that yields blocks of protein structure, which when swapped between parents result in the minimum amount of disruptions between possible interacting positions. From selection and recombination experiments done on β-lactamases, they found that SCHEMA can be used predict the functionality of variants.

5.1.1.3 Limitations of structure based approaches

Besides requiring a high-resolution three-dimensional structure, another limitation of structure-based approaches is the fact that interaction is assumed to be possible only based on the spatial arrangement of the amino acids. However, it may be possible for two amino acids located far apart from each other in the structure to exhibit an interactive effect (12). This has been observed to be the case for two different pairs of amino acids in *TEM β*-lactamase. In a good example of sign epistasis, whereby the sign of the fitness effect of a mutation is dependent on the alleles present (13), the mutation G238S in the wild-type is known to enhance the cefotaxime hydrolysis but simultaneously increase aggregation and reduce thermodynamic stability. Conversely, another mutation, M182T, reduces hydrolysis while reducing aggregation and increasing thermodynamic stability. Thus, either of these mutations alone reduce the cefotaxime resistance of TEM^{wt}, but together, the double mutant confers increased resistance. Such interactions cannot be identified from three-dimensional structure alone as structure data do not indicate thermodynamic stability and aggregation tendency.

5.1.1.4 Sequence-based methods

Other groups have used statistical tools such as thermodynamic coupling methods and co-variation analysis on sequence-based data to identify interacting and important residues (14-20). Suel *et al.* (20) used a sequence-based thermodynamic method to map the global network of amino acids interactions from which long-range interactions between amino acid can be found. This method involved the considerations of statistical coupling energy (ΔΔG) due to change of amino acid distribution at one position. Lichtarge *et al.* (17) identified functional interfaces by using sequence conservation data and mapping them onto the protein surface to predict functionally important residues. Gaucher *et al.* (15) integrated genomic information and three-dimensional structure with co-varion (variable rates of divergence between trees) based analysis to identify residues that may explain the functional differences between proteins. These techniques are very useful but require large amounts of sequence data on homologous proteins which are not always accessible.

5.1.2 Scope of this work

The above factors inspire a strategy which does not rely on the availability of the three-dimensional structure for any protein, or the availability of large amounts of sequence data from homologous proteins, and is not restricted by the assumption requiring the interacting amino acids to be in close proximity. One solution is to apply a data-driven machine learning approach that can utilize the data generated during directed evolution (21). In this study, we report the experimental validation of a machine learning approach to guide directed evolution on fluorescent proteins. This approach consists of generating a library of variants and assaying for phenotype. Following that, the interacting residues are identified by using machine learning tools (described in the next section) on sequence-activity data. The information can be used to engineer templates to preserve interactions during recombination. This results in a library with a larger fraction of active variants in a library.

5.1.3 Machine learning

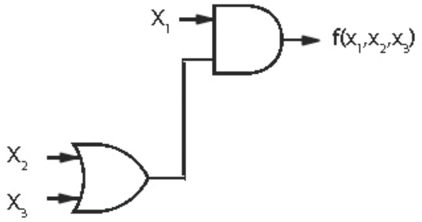
5.1.3.1 Boolean learning

For computer and engineering applications, Boolean functions are a standard representational tool (22). Given a list of examples in 0's and 1's (Boolean form) that are classified as positive or negatives, a Boolean learning algorithm can establish a set of rules or Boolean expressions which classify all the examples correctly. One Clause At a Time (OCAT) is the primary algorithm used to establish Boolean functions. OCAT can be based on either branch-and-bound algorithm or certain heuristics (23,24).

One simple example serves to illustrate the Boolean learning algorithm. Figure 5.2 shows the input (x_1, x_2 and x_3) and output data ($f(x_1,x_2,x_3)$) listed in Boolean form. From the table in Figure 5.2, we can observe that x_1 must always be positive for the function f to be positive. In addition, either x_2 or x_3 must also be positive at the same time for function f to be positive. Thus, the Boolean function, $f(x_1, x_2, x_3) = x_1 \wedge (x_2 \vee x_3)$, classify all the examples correctly (where $\wedge$ represents 'AND' and $\vee$ represents 'OR').

x_1	x_2	x_3	$f(x_1,x_2,x_3)$
0	1	1	0
1	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1

→



$$f(x_1,x_2,x_3) = x_1 \wedge (x_2 \vee x_3)$$

Figure 5.2 Example of inferring Boolean Learning function. Left: Input and output data in Boolean form. Right: Pictorial representation of data using logic gates.

5.1.3.2 Support Vector Machines (SVMs)

Support vector machines (SVMs) are a form of learning algorithm based on the statistical learning theory (25,26). In SVM, the data are mapped from input space to the higher dimensional feature space, using a kernel function, so that a linear function can be fitted (Figure 5.3). Following minimization of the upper bound generalization error by using a linear classification algorithm, the linear function that best separates the data in that space is obtained. Typically, the number of parameters required for this function is equal to the size of the data set used for training, thus avoiding overfitting. SVMs have been widely applied in the field of pattern recognition such as face recognition and for secondary structure prediction of proteins (27-30).

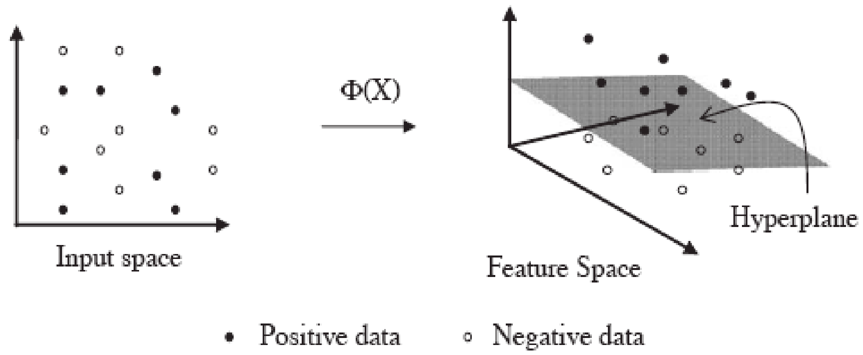
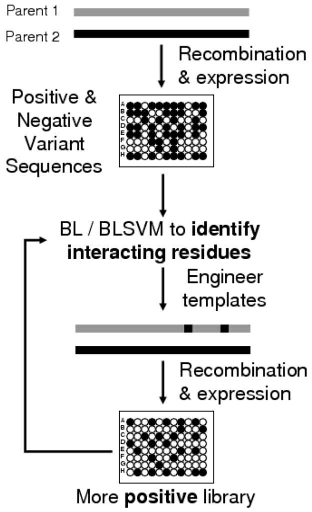


Figure 5.3 An illustration of the transformation from the input space to feature space. Figure adapted from (21).

5.1.4 Machine learning tools for Combinatorial Protein Engineering

Machine learning by Dubey *et al.* (31) was used to analyze the amino acid sequences and their phenotype. Dubey *et al.* (31) proposed a Support Vector Machines (SVMs) (25,26) based algorithm to identify the individual residues of a protein, which, when mutated, can lead to loss of its function. The data required for the algorithm were the sequences of positive and negative variants of a protein, usually obtained during directed evolution using mutagenesis or recombination. The variants were divided into positive and negative phenotype based on a certain threshold on their activity. This work was extended to show that interactions between different amino acids can also be identified by using Boolean learning and SVMs (21) on sequences obtained from recombination. These sequences were once again divided into positive and negative phenotype based on their activity. This activity can be measured based on the screening procedure used in the Directed Evolution of any protein. The use of both positive and negative examples is a key distinction in light of other methods to determine interacting residues and this approach can be very helpful in identifying the significance of the amino acids. Such an information can be obtained only through Directed Evolution and not from naturally existing proteins.



Parent 1
Parent 2
Recombination & expression
Positive & Negative Variant Sequences
BL / BLSVM to **identify interacting residues**
Engineer templates
Recombination & expression
More **positive** library

Figure 5.4 The steps of applying Boolean learning to identify the interacting residues and engineering the parents to improve the recombinant library and to do a more productive search for improve variants through Directed Evolution.

The procedure (illustrated in Figure 5.4) begins with the recombination of two of more parent sequences. The generated variants are screened to establish their activity, stability, or other dimensions of merit. Based on a threshold value of this measure, the variant sequences above the threshold are considered as positive while the ones below it are considered as negative. It should be noted that the threshold, which can be arbitrarily chosen, dictates how the results are inferred. For example, if the threshold is such that variant with any measurable activity is positive while all the inactive variants are negative, then the interactions of amino acids, which are responsible to retain any activity at all, will be identified. These variant sequences are used as an input for the algorithm developed by Dubey *et al.* (21). This algorithm can then identify the significant interactions. It is assumed that all the important interactions have to be preserved for any variant sequence to show positive. phenotype An interaction is assumed to be preserved if both the amino acids in any variant come from the same parent.

Dubey *et al.* (21) showed through simulations that any pair of residues in the sequence which show interaction or epistasis can be effectively identified, independent of the specific location of these residues in the sequence. It was also suggested that, following this result, the parents can be altered to increase the fraction of positive variants in the recombinant library (Figure 5.4). This was achieved by mutating the amino acids involved in an interaction for any one parent sequence to create a double mutant, such that they are the same as those in the sequence of the other parent. Simulations showed that a recombination between this double mutant of one parent and the native sequence of the other can increase the fraction of positive variants since the interaction cannot be broken through crossovers.

5.1.5 Fluorescent proteins as a model for experimental validation

To apply the approach described above to variant sequences obtained for fluorescent proteins, two fluorescent proteins were chosen, monomeric Red Fluorescent Protein (mRFP) and *Discosoma* Red Fluorescent Protein (DsRed). Fluorescent proteins are a class of proteins important for bio-imaging applications such as gene expression, protein-protein interaction detection, cell and protein tracking (32). They consist of β-can structures that encase a chromophore. The encased chromophore undergoes a series of maturation steps that extend the conjugated double bonds, producing fluorescence. The functionality of fluorescent protein can easily be detected by measuring fluorescence. This assay provides an effective screen for establishing whether a variant sequence is folded or not.

Both mRFP and DsRed are well characterized proteins and have distinct spectral emissions (33,34). mRFP emits with a maximum fluorescence at 607 nm while DsRed emits with a maximum peak at 583 nm. mRFP is derived from DsRed protein (33) by the introduction of 33 mutations which include aggregation-disrupting changes such as R2A, K5E, N6D, I125R, V127T, R153E, H162K, A164R, L174D, I180T, Y192A, Y194K, H222S, L223T, F224G and L225A. Thus, both proteins share a high level of amino acid identity of 84.1%. The mRFP we used was codon-optimized for expression in *E. coli* while the DsRed gene was optimized for mammalian expression. This gave us a gene set with 74.5% nucleotide identity level.

5.2 Materials & Methods

5.2.1 Materials

Most of the enzymes were bought from New England Biolabs (Beverly, MA), with the exception of *Pfu* and *Taq* polymerase, which were purchased from Stratagene (La, Jolla, California). Tetracycline, ampicillin and chloramphenicol were purchased from Sigma (St. Louis, Missouri). The inducer, anhydrotetracycline (ACT) was made by dissolving 250 mg/l of tetracycline in water and adjusted to pH 3 and autoclaving for 45 mins (35).

5.2.2 Parental plasmids construction

The mRFP amino acid sequence was obtained from NCBI and *E* . coli-codon optimized primers were designed using DNAworks (36) and synthesized. After two PCR reactions, the mRFP gene was obtained and reamplified with primers with *Esp*3I restriction sites (5' - TA**C GTC TCG** TCG ACA TGG CGT CTT CTG AAG ACG TTA TCA AAG AAT TCA TGC GT - 3' and 5' - TA**C GTC TCT G**GC CTA TTA CGC ACC GGT AGA GTG ACG ACC TTC - 3'), digested with *Esp*3I enzymes and cloned into *Sal*I and *Not*I restriction sites of digested pPROTet vector using T4 DNA ligase. DH5αpro *E*. coli strains were used for transformation. The mRFP gene is successfully obtained, evident from sequencing data. We can obtain up to 20% of total protein expression of mRFP in cells. The excitation and emission spectral of the fluorescent protein were identical with the literature (33).

The DsRed gene was amplified with primers that contain *Esp*3I restriction sites (5' - TAC GGT TA**C GTC TCG** TCG ACA TGG CCT CCT CCG AGA ACG TCA - 3' and 5' CAT TAC TA**C GTC TCT G**GC CTA CTA CAG GAA CAG GTG GTG GCG G - 3) from DsRed2-1 plasmid (BD Biosciences Clontech, San Jose, California). The DsRed insert was digested with *Esp*3I enzyme and cloned into *Sal*I and *Not*I site of digested pPROTet vector using T4 DNA ligase.

5.2.3 Templates for DNA-shuffling

The mRFP and DsRed genes were amplified by *Pfu* polymerase using pPROTeT vector specific primers (5`-CTT TCG TCT TCA CCT CGA GTC C-3`, 5`-CCT ACT CAG GAG AGC GTT CAC C-3`) that add 122 bp to the 5'-terminus and 155 bp to the 3'-terminus. The PCR products were gel purified thereafter.

5.2.4 DNA-shuffling procedure

A hybrid method for DNA-shuffling from (37) was used. We used up to 4 μg of templates for DNase I digestion. The digestion time was optimized to produce fragments of around 50 bp. The resultant fragments were agarose gel-purified using QIAEX II® Gel Extraction Kit (Qiagen, Valencia, California). The purified fragments were mixed with 5μl of 10X *Pfu* buffer and water to a final volume of 50 μl and cycled using the protocol from (38); 96°C, 90s; 35 cycles of (94 °C, 30 s; 65 °C, 90 s; 62 °C, 90 s; 59 °C, 90s; 53 °C, 90 s; 50 °C, 90~s; 47 °C, 90s; 44 °C, 90s, 41 °C, 90 s; 72 °C, 4 min); 72 °C, 7 min; 4 °C hold. Nested primers (5'-ATG GGT CAT AAT CAT AAT CAT AAT CAT AAT C-3', 5'-GTC TTT CGA CTG AGC CTT TCG T-3') were used to amplify the genes from the diluted reassembled products.

5.2.5. Templates for RDA-PCR

Parent-specific skew primers were designed which added overhangs to the 5'-terminus (5'- CGG GAT TCC ACA TAG TCT CAG GTA-3') and 3'-terminus (5'-TTC CGA TAA GTT CAT AGG CCG TGG-3') to mRFP and DsRed genes, respectively.

5.2.6. Recombination dependent-PCR procedure

An adaptation of protocol by Ikeuchi et al.(39) for recombination was followed.

Taq DNA polymerase was used and the PCR was optimized for a higher yield of crossovers. The cycling protocols used are as follows: 98°C for 5 mins; 40 cycles of (94°C, 30 s; 40-45°C, 5 s; 72°C for 3s); 10 cycles of (94°C, 30s; 50°C, 30s; and 72°C for 30 s); 4°C hold. 2) 95°C, 2 mins; 20 cycles of (95°C, 30 s; 60-67°C, 1+1 s); 60 cycles of (95°C, 30s, 60-67°C, 20 s); 72°C, 10 mins. 3) 95°C, 2 mins; 99 cycles of (95 °C, 30 s; 65°C, 20 s); 72°C for 10 mins.

5.2.7. Creation of point mutation variants of mRFP and DsRed

The Stratagene Quickchange® site-directed mutagenesis protocol was used to create single and double mutants. Sense and anti-sense primers of the point mutations to be introduced were designed and synthesized. The mutations introduced into mRFP are I197A (5' - TGC GTA CAA AAC CGA C**GC C**AA ACT GGA CAT CAC CTC TC - 3', 5' - GAG AGG TGA TGT CCA GTT T**GG C**GT CGG TTT TGT ACG CA - 3'), A217T (5' - CGT TGA ACA GTA CGA ACG T**AC C**GA AGG TCG TCA CTC TAC CG - 3', 5' – CGG TAG AGT GAC GAC CTT C**GG T**AC GTT CGT ACT GTT CAA CG - 3'), L83K (5'- ACC CGG CGG ACA TCC CGG ACT AC**A AG**A AAC TGT CTT TCC CGG AAG GTT TCA - 3', 5' - TGA AAC CTT CCG GGA AAG ACA GTT T**CT T**GT AGT CCG GGA TGT CCG CCG GGT - 3') for single mutants, I197A/A217T (double mutants), I197A/L83K, A217T/L83K and I197A/L83K/A217T (triple mutants). The mutations introduced into DsRed are A197I (5' - CGG CTA CTA CTA CGT GGA C**AT C**AA GCT GGA CAT CAC CTC C - 3',5' - GGA GGT GAT GTC CAG CTT **GAT** GTC CAC GTA GTA GTA GCC G - 3'),T217A (5' - GGA GCA GTA CGA GCG C**GC C**GA GGG CCG CCA CCA C - 3', 5' - GTG GTG GCG GCC CTC **GGC** GCG CTC GTA

CTG CTC C - 3'), K83L (5' - ACC CCG CCG ACA TCC CCG ACT ACC TGA AG**C** **TG**T CCT TCC CCG AGG GCT T, 5' – AAG CCC TCG GGG AAG GA**C** **AG**C TTC AGG TAG TCG GGG ATG TCG GCG GGG T - 3'), A197I/T217A, A197I/K83L, A217A/K83L and A197I/K83L/A217A.

5.2.8 Expression normalization and determination of phenotype on Fluorescent Proteins

The PCR products from DNA-shuffling or recombination dependent-PCR were cloned into the *Sal*I and *Not*I or *Ase*I restriction sites of digested pPROTet using the respective restriction endonucleases and T4 DNA ligase. DH5αpro *E.coli* cells were transformed with recombinant vectors using standard chemical protocol (40) and plated on 20 μg/mL chloramphenicol LB plates. The generation of a diverse library is confirmed by sequencing at least 10 randomly picked colonies. To ensure that the expression levels of the fluorescent proteins in the cell cultures are normalized, we followed the protocols described in (41). Briefly, random colonies of mRFP and DsRed chimeras were either hand-picked or robot-picked using QPix2 colony picking machine (Genetix, Boston, Massachusetts) from chloramphenicol plates and incubated for 24 hrs at 37^{o}C and 150 rpm until growth reached the stationary phase of *E.* coli. Twenty μl of culture were transferred to another 96-well plate containing 180 μl of LB supplemented with choloramphenicol and ACT to induce protein expression. The transferred cultures were incubated 37 ^{o}C and shaken at 150 rpm. After 30 hours, the emission-spectrum scan (excitation 540 nm, emission 500-700 nm range) using SPECTRAmax GEMINI (Molecular Devices Corp, Sunnyvale, California) was obtained. Positive controls of mRFP, DsRed and GFP recombinants and negative controls using LB blanks were also measured. Since the peak fluorescence for mRFP (607 nm) and DsRed (583 nm) differ, to classify a variant as active or inactive, its measured fluorescence intensity was first matched with the appropriate parent and then the peak intensity was compared. Several variants were found which did not share the peak wavelength with any of the parent. For such cases, it was decided that the peak intensity will be compared to the parent with the lower intensity. This decision, however, was not critical in classifying the variants into positive and negative since the difference between the active and inactive variants was very pronounced (see Results and Discussion)

5.2.9 Applying Boolean learning and BLSVM

For Boolean learning these variant sequences were transformed into Boolean vectors as described in Appendix 5.8.1 and the OCAT algorithm was used on the data set of sequences. Along with Boolean learning, the BLSVM algorithm, which was described in Appendix A.1, was also used on the same data set. The Boolean function given in Equation A.1 was identified by these two algorithms from 83 variant sequences. The weighing parameter for BLSVM in Equation (A.5) was set equal to 2 based on the simulation results of Dubey *et al* (21). The Lagrangian variables, α, were obtained by using SVMs on the data set of variant sequences. Since, as discussed in Appendix A.2, the input data needs to be in the form of vectors, **x**, the sequences were represented as a vector of amino acids. The naturally occurring amino acid residues were represented by numbers from 1 to 20 in the input variant sequences (Dubey *et al* (21)). A polynomial kernel (Appendix A.2), $K(x_i, x_j) = (1 + x_i \cdot x_j)^2$, was used since it can account for interactions between different attributes of the positions of the sequence (21). The optimization problem of Equation (A.4) was solved by using Sequential Minimal Optimization (42) in MATLAB®.

5.2.10. Random Sampling of Variant Libraries

In random sampling, multiple samples of the two initial sets are created by first combining them, and then randomly splitting them in two different sets. This process is performed without replacement, so, any point can belong to either of the set but not to both of them. This procedure is repeated for a large number of times and for each split, the difference in the means, *d*, is recorded. From the resulting distribution of *d*, it is possible to calculate the probability of obtaining the difference greater than or equal to the one between the original sets. If this probability is below a certain threshold, β, the null hypothesis can be rejected.

5.3 Results

5.3.1 Amino acid residues 197 and 217 are identified as interacting

The sequence-activity relationships of 83 variants were obtained from recombination and expression experiments. DNA shuffling and recombination-PCR were applied to mRFP and DsRed fluorescent proteins to obtain chimeras which were then expressed in 96 well plates. Figure 5.5 shows the distribution of the crossovers obtained for the 83 variants with an average of 1.73 crossover for the set. Most of the variants (> 60%) of the variants had one crossover. The detailed crossover positions can be found in the Supplement Table 5.7.1. The relative fluorescence intensities of the variants were plotted in Figure 5.6. Since fluorescence intensity is a property that depends on experimental conditions, we plotted the fluorescence intensities relative to the parental wild-type fluorescence to reduce experimental noise and to normalize the fluorescence intensity to a standard reference. If the peak wavelength of the fluorescence intensity of the variant did not match either parent, then it was normalized relative to the parent with the lower wavelength. The solid lines indicate the wildtype fluorescence of mRFP or DsRed protein and the cutoff for deciding activity which was set at ten percent of WT fluorescence. It can be observed from Figure 5.6 that there are no variants very close to the cutoff value. The classifications are thus resistant to fluctuations of fluorescence intensities and the decision to compare the intensity of the variants which did not match the peak wavelength of either parent with the parent with the lower intensity was not critical.

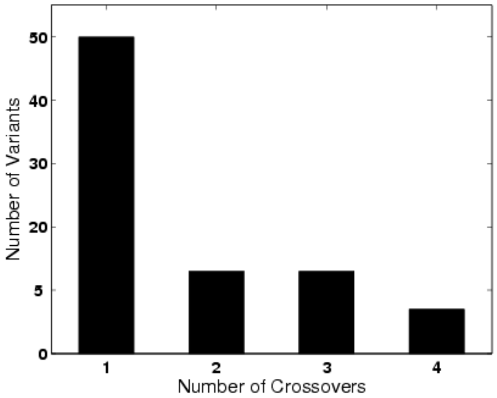


Figure 5.5 The distributions of crossovers in the variants.

The sequence-activity relationships data was analyzed using machine learning algorithms and residues 197 and 217 were identified as interacting. Dubey *et al.* (21) determined through simulations that for sequences comparable to mRFP and DsRed proteins, which consist of 225 amino acids, on the order of 100- 200 sequences were required for reasonably accurate identification of the Boolean function. We used 83 sequences for the computations, which is similar to the number required. In the mRFP protein, isoleucine and alanine were in positions 197 and 217, respectively. For the DsRed protein, alanine and threonine were in positions 197 and 217, respectively.

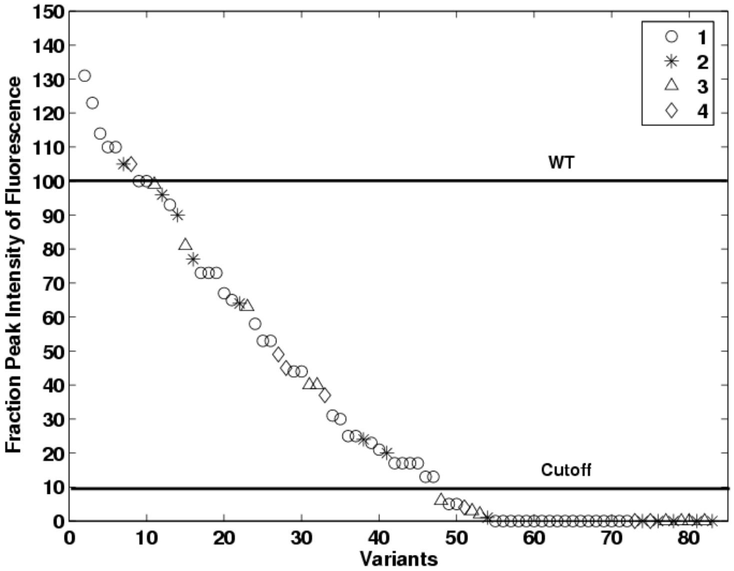


Figure 5.6 The Fraction Peak Intensity of variants. The different geometric symbols indicate the number of crossovers in each variant.

5.3.2 Validating the interactions through point mutations

Evidence of interactions between the residues was checked by introducing single and double point mutations at the residue positions 197 and 217. If there are interactions between these residues, then single point mutation variants should be inactive as the interactions are disrupted. However, if double point mutants are created such that both the residues are swapped for the residues in the other parents, we should expect to see active proteins as the interactions were preserved. Site-directed mutagenesis was used to create point mutants which were fluorimetrically assayed for function. Five different colonies containing the same recombinant protein were scanned for fluorescence. The results were averaged to account for experimental noise. Table 5.1 summarizes the results obtained.

Table 5.1 Relative fluorescence of the various single and double point mutants of mRFP and DsRed proteins.

Variant	% WT intensity	Active?	Predicted
mRFP-I197A	0	No	No
mRFP-A217T	0	No	No
DsRed-A197I	0	No	No
DsRed-T217A	30	Yes	No
mRFP-I197A/A217T	0	No	Yes
DsRed-A197I/T217A	12	Yes	Yes

As expected, mRFP single mutants, I197A, A217T and DsRed single mutant A197I were inactive. Surprisingly, DsRed T217A single mutation seemed to be tolerated although the fluorescence intensity relative to DsRed wild-type seemed to be reduced. The double mutation in mRFP unexpectedly was also not active. However, the double mutation in DsRed was indeed active which demonstrates interaction between the two residues. If there were no interaction between A197 and T217, then the combined effort of mutating them should be linear and the double mutant should be inactive since A197I renders DsRed non-fluorescent. The fluorescence of the double mutant can be viewed as a clear sign of interaction between the two amino acid residues since both the individual mutations have a negative effect on the fluorescence.

5.3.3 Using three-dimensional structures to find interactions to check our predictions

We analyzed the 3D crystal structures of DsRed protein to determine if it was necessary for the residues to be geometrically located next to each other for interaction to occur. To date, the mRFP crystal structure has not been published; an analysis of the residues involved in interactions cannot be performed by simply mutating the DsRed crystal 3D structures. We used comparative modeling, threading and *Ab initio* structure prediction to obtain the mRFP structure (43-47). These programs required the crystal structure or amino acid sequence information along the desired mutations for input. All these methods yielded low resolution β-can protein structures that did not provide the accuracy for probing interacting sites.

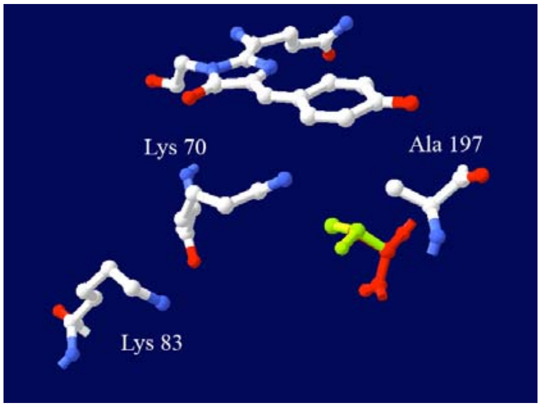


Figure 5.7 Native DsRed structure with T217A mutation illustrated using Rasmol. The threonine residue is colored green while the new alanine reisdue is colored red. Steric conflict is not observed. Variant is active as a result.

The next best alternative to find possible interacting sites is to use the DsRed fluorescent protein structure. The mRFP and DsRed fluorescent proteins differ by 33 amino acids, thus it is likely they will only have a slightly different fold. Nonetheless, we noticed that the key residue 83 that has been known to affect the emission wavelength is different in mRFP (leucine) and DsRed (lysine). The side chain of residue 83 is very close to the chromophore as well as to the residues 70, 197 and 217 (Figure 5.7 – 5.8). The smaller L83 residue in mRFP reduces hydrophobic core packing that could potentially affect the interaction with residue 70 and eventually disrupt chromophore maturation. Thus, the chromophore in mRFP is possibly less tolerant of disruptions in both the residues 197 and 217. From the three-dimensional structure alone, this finding may also suggest that residue 83 can possibly interact with either one of the residues 197 and 217. Note that this residue was not identified by the algorithm to be interacting.

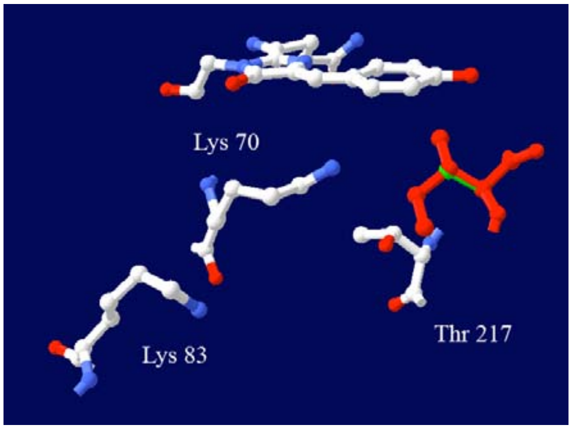


Figure 5.8 Native S197A modified DsRed structure with A197I point mutation. This mutation effectively mutates the mutant back to native residue isoleucine (A-green, I-red). Possible steric hindrance with isoleucine observed.

Experiments were performed to check whether there could be interactions involving residue 83, additional point mutants and double mutants were created and tested for activity. The results for these mutations are listed in Table 5.2. Evidently, L83 is individually important for mRFP since a mutation in that position deactivates the protein. There is no evidence of its interaction with residue I197 since the double and the triple mutants are also inactive. Likewise, in DsRed, the K83 residue also affects the activity because a mutation reduced the intensity (but does not deactivate it). Mutating the active double variant DsRed-A197I/T217A (Table 5.2) at position 83 yielded a triple variant of DsRed with <10% intensity, thus rendering the triple variant inactive according to our threshold. Thus, there is no indication that residue 83 is interacting with residue 217.

Table 5.2 Single, double and triple point mutants in positions 83, 197 and 217 of mRFP and DsRed.

Variant	% WT intensity	Active?
mRFP-L83K	0	No
mRFP-L83K/I197A	0	No
mRFP-L83K/A217T	0	No
DsRed-K83L	~20	Yes
DsRed-K83L/A197I	0	No
DsRed-K83L/T217A	0	No
mRFP-L83K/I197A/A217T	0	No
DsRed-K83L/A197I/T217A	<5	No

5.3.4 Increased fraction of active variants in engineered library

Identification of interacting residues is suggested to aid in directed evolution efforts. Dubey *et. al* (21) suggested a multi-round strategy for using the identification of interacting residues to improve the library of the variants in each round. If point mutations were introduced in the identified pair such that both the parents have the same amino acids, one should expect an increase of the fraction of active variants after recombination of the engineered templates when compared to the library created through the recombination of wild-type templates. Efforts are thus better utilized to search for sequences in the active regions of the sequence space. The improvement in productive search through active space should improve the odds of obtaining a desired variant.

Recombination was applied to WT mRFP and DsRed templates and WT mRFP and engineered DsRed templates to verify that libraries with preserved interactions were more active. DNA shuffling was used on mRFP and active double mutant DsRed-A197I/T217 and mRFP and DsRed wild-type to generate two sets of recombinant libraries. For each set of libraries, 480 randomly picked variants were screened for fluorescence. To ensure that crossovers occur between residues 197 and 217, we sequenced randomly picked variants from the native and engineered library. Table 5.3 shows the results of sequencing. We found 33% crossovers occurring between these amino acids which suggests that recombination is breaking the interactions between residues 197 and 217. The fraction of active variants (at least 10% of wild-type) was 49% for the wild-type recombinant library while the fraction was 67.5% for the engineered library.

To determine whether the results that show a library with higher fraction of active variants was obtained with the engineered templates were statistically significant, testing of the null hypothesis by random sampling was performed (48). The null hypothesis was defined as the assumption that the two different set of results obtained on the fraction of active variants for the WT library and engineered library are the same and any point in one of those sets can belong to either of them. To claim that the two sets are indeed different, it is sufficient to disprove the null hypothesis. Note that since there are a limited number of points in each set, simply comparing the means without further sampling is not statistically sufficient.

The two sets in this study are the active and inactive variants obtained by the recombination of mRFP/DsRed and mRFP/DsRed-A197I/T217A. An active variant was represented by 1 and an inactive variant by 0, so that each set had 480 binary numbers. Even when 10^5 different random samples were created, the probability of getting sets more different than the obtained sets was found to be equal to zero. Thus, the probability, p is less than 10^{-5}. From random sampling, the probability of the null hypothesis was found to be less than 10^{-5}. Conventionally, the null hypothesis is rejected if the probability is less than or equal to 0.05. Thus, it can be concluded that the difference in the results of recombination from the two different pairs of parents is statistically significant.

Table 5.3 The sequences of recombinant libraries of WT library and engineered library generated by DNA-shuffling

Library	Sample no.	Starts with	Crossover positions	No. of Crossovers	Crossover between 197 and 217
WT	1	mRFP	NA	0	
WT	2	mRFP	NA	0	
WT	3	mRFP	NA	0	
WT	4	mRFP	NA	0	
WT	5	mRFP	39	1	
WT	6	mRFP	321	1	
WT	7	mRFP	549	1	
WT	8	mRFP	618	1	X
WT	9	mRFP	186, 247	2	
WT	10	mRFP	285, 300	2	
WT	11	DsRed	300,432	2	
WT	12	mRFP	204, 363, 645	3	X
WT	13	DsRed	321, 609, 633	3	X
				Variants total	9
	Variants with crossover between residue 197 and 217				3
	%Variants with crossover between residue 197 and 217				33.3

Table 5.3 continued

Library	Sample no.	Starts with	Crossover positions	No. of Crossovers	Crossover between 197 and 217
EN	1	mRFP	153	1	
EN	2	mRFP	171	1	
EN	3	mRFP	186	1	
EN	4	mRFP	186	1	
EN	5	mRFP	186	1	
EN	6	mRFP	247	1	
EN	7	mRFP	633	1	X
EN	8	mRFP	186, 322	2	
EN	9	mRFP	186, 618	2	X
EN	10	mRFP	363, 589	2	
EN	11	DsRed	589, 594	2	X
EN	12	mRFP	633, 645	2	
EN	13	DsRed	171, 393, 618	3	X
EN	14	mRFP	84, 186, 363, 414	4	
EN	15	mRFP	300, 315, 363, 636	3	X
				Variants total	15
	Variants with crossover between residue 197 and 217				5
	%Variants with crossover between residue 197 and 217				33.3

5.4 Discussion

5.4.1 Importance of residues 197 and 217 for mRFP and DsRed

The three-dimensional crystal structure of DsRed fluorescent protein (1G7K) obtained from the Protein Data Bank (PDB) can be used to study the positioning of the interacting pair of amino acids. It should be noted that no crystal structures were used initially to identify the interacting amino acids. Figure 5.9(a) shows the two residues highlighted in white within the core of DsRed fluorescent protein. The two residues are very close spatially. The distance between the side chains is 3.79Å (O_β of T197 to C_β of A217). The chromophore is shown in blue while the residue K70 is colored red. Figure 5.9(b) shows an enlarged view of the residues 197 and 217 along with K70. It is interesting to note that both of these residues are very close (O_β of T217 to N_γ of K70 – 3.74 Å and C_β of A197 to N_γ of K70-4.22 Å) to K70, which has been known to exert a significant influence on the maturation of the chromophore.

Residues 197 and 217 have been reported to be involved in stabilizing the anion form of the chromophore (33,34,49,50) which corroborates the prediction that residues 197 and 217 could be interacting. Campbell *et al.* found in the course of evolving mRFP from DsRed, that residues K70, S197 and T217 were important residues that affect the fluorescent properties of DsRed mutants (33).Verkhusha and Sorkin analyzed the residues spatially close to the chormophore and determined the 146, 161 and 197 residues (relative to DsRed) were the main molecular photoconversion determinants for tetrameric fluorescent proteins (50) and created a photoactivatable probe by mutating the mRFP protein in these three positions 146, 161 and 197. Yarbrough *et al.* crystallized DsRed fluorescent protein and mentioned that S197 residue of DsRed interacts indirectly through a water molecule with K70 residue. Replacing S197T affected the final maturation of DsRed fluorescent protein (49). Lastly, Baird *et al.* characterized DsRed fluorescent protein mutants and found S197T mutation increased maturation time of DsRed fluorescent protein (34).

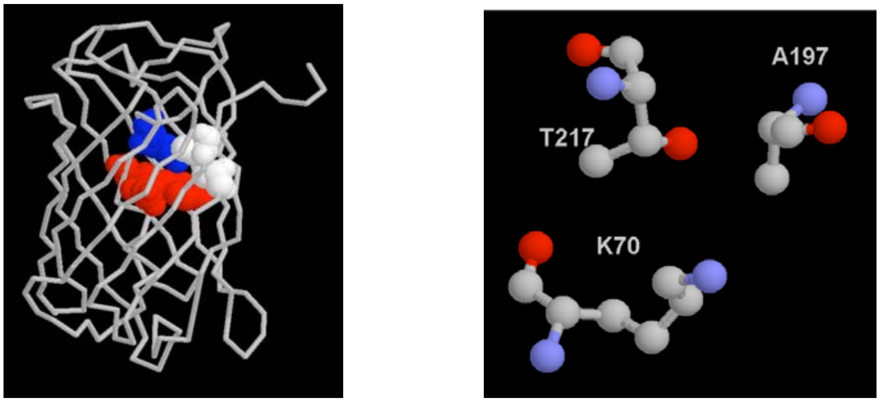


Figure 5.9 (a) Left: The crystal structure of DsRed with the residues in positions 197 and 217 highlighted. The residues 197 and 217 are shown in white. K70 and the chromophore are colored in blue and red respectively. (b) Right: Enlarged view of interacting residues 197A and 217T with K70 highlighted.

5.4.2 Nature of interactions

The exact nature of the interaction could not be determined, but they could be interacting indirectly through the K70 residue (49). The K70 network has been known to affect to affect the maturation of DsRed. Since K70 was conserved between the two sequences, it is possible that the difference in the residues in position 197 and 217 compensate for each other. Such compensating effects, where the loss of fitness through one point mutation can be counteracted by another, have been demonstrated before. Jucovic and Poteete (51) deactivated lysozyme by a single mutation and recovered its function through two additional point mutations. In nature, compensating mutations have been found in elongation factors and in orthologs of related species (52,53). Weinreich *et al*. discovered compensatory mutations in TEM-1*β*-lactamase where the additive effects of the combination of G238S and M182T point mutations increased activity towards cefotaxime.

To explain the unexpected fluorescence of DsRed T217A, the DsRed crystal structure was analyzed to gain insight. The mutation function of Swiss-PDBViewer (54) was used to produce the point variants T217A and A197I from the native structure of DsRed. Figure 5.7 show that the T217A mutation does not seem to cause steric hindrance with either K70 or the chromophore (66-68). The lack of steric hindrance, thus, may still allow for full red chromophore maturation. The reduced intensity could result from the loss of hydrophobic packing that results in the chromophore not optimally positioned for the extension of the conjugated bonds that confers red fluorescence (32,34,55). The

DsRed A197I mutation on the other hand, seems to result in a structure which displays steric hindrance by I197 on the chromophore as shown in Figure 5.8 and thus results in an inactive variant.

It is also possible that interaction between residue 197 and 217 exists in DsRed but not mRFP, even though the two fluorescent proteins feature 84.1% identity. The sequence-activity relationship could be different for mRFP and DsRed protein despite the high level of identity. Since the library which was used as input for the algorithm to determine potentially interacting residues consisted of DsRed and mRFP chimeras, pairs that are important to one parent would still be identified by the algorithm since there would be loss of function in the recombined variant. Table 5.1 shows that the double point mutant of DsRed is active while the double point mutant of mRFP was inactive. This demonstrates that interaction was occurring between residues 197 and 217 for DsRed fluorescent protein while the evidence for interaction for mRFP is not as strong.

5.4.3 Usefulness of BLSVM for directed evolution

While a library with improved fraction of active variants could be obtained by engineering the templates to preserve the interactions, it is not necessary evident that such a technique would be useful for evolving proteins with higher activities. In the current study, we screened for a library with a larger fraction of positives instead of a library with higher activity level. The objective of obtaining a library with a higher fraction of active variants was achieved. Though not yet demonstrated in this study, such a library could be potentially useful for evolving protein towards different substrate specificities as the evolution of different substrate specificity could decrease native activity. Conceivably, an improved variant of a protein would strengthen the case for using machine learning algorithm to identify important and interacting residues and guide directed evolution efforts.

5.4.4 Examples of data-driven protein engineering

Two recent publications highlight the usefulness of using computational algorithm on experimentally obtained sequence-activity relationships to find residues crucial for protein evolution. These proprietary algorithms were developed in protein engineering firms. Fox *et al.* (56) demonstrated the use of regression analysis in parallel to recombination experiments to identify residues that tend to increase the overall activity of the protein. They evolved an improved halohydrin dehalogenase that is 4000-fold more active than wild-type protein. Liao *et. al.* (57) used machine learning language along with a synthetic library to engineer the proteinase K protein. They showed that using only two cycles of protein evolution with machine learning guided approach and only testing just 95 variants, they could obtain proteinase K variants with 20 fold improvement relative to the wild-type protein with respect to specific activity on the hydrolysis of a tetrapeptide substrate (N-Succinyl-Ala-Ala-Pro-Leu p-nitroanilide).

5.5 Conclusion

A method involving the use of Boolean learning (BL) and Support Vector Machines (SVMs) (= BLSVM) to identify potentially interacting residues using sequence-activity data on mRFP and DsRed fluorescent proteins was demonstrated in this study. Single, double and triple point mutants of the fluorescent proteins in positions 83, 197, and 217 were created to test the predictions. All the single mutants except the T217A DsRed variant were inactivated as expected. This might be explained by the lack of steric hindrance on the chromophore by the mutation. The DsRed variant, A197I/T217A was active as expected but surprisingly, the double point mRFP variant I197A/A217T is inactive possibly due to less tolerant hydrophobic interactions. The evidence of non-linear effects of A197I/T217A DsRed variant suggests interaction do occur between residues 197 and 217. We show that using geometric proximity in three-dimensional crystal structure is not a good way to find interacting residues. The results were also used to improve the fraction of active variants by preserving the interactions of crucial residues through template engineering prior to recombination experiments.

This chapter and its contents were recently accepted for publication in Biotechnology Journal.

Bernard L.W. Loo, Anshul Dubey, Matthew J. Realff, Jay H. Lee and Andreas S. Bommarius. Identifying Interacting Residues using Boolean Learning and Support Vector Machines: Case study on mRFP and DsRed proteins (2008). *Biotechnology Journal.*

5.6 References

1. Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364-1368.

2. Park, H.-S., Nam, S.-H., Lee, J.K., Yoon, C.N., Mannervik, B., Benkovic, S.J. and Kim, H.-S. (2006) Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, **311**, 535-538.

3. Chen, K. and Arnold, F.H. (1993) Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U S A*, **90**, 5618-5622.

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

5. Lin, H. and Cornish, V.W. (2001) In Vivo Protein-Protein Interaction Assays: Beyond Proteins. *Angew. Chem. Int. Ed. Engl.*, **40**, 871-875.

6. Daugherty, P.S., Chen, G., Iverson, B.L. and Georgiou, G. (2000) Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl. Acad. Sci. U S A*, **97**, 2029-2034.

7. Neylon, C. (2004) Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res.*, **32**, 1448-1459.

8. Arnold, F.H. (1998) Design by directed evolution. *Accounts of Chemical Research*, **31**, 125-131.

9. Huang, W., Petrosino, J., Hirsch, M., Shenkin, P.S. and Palzkill, T. (1996) Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.*, **258**, 688-703.

10. Meyer, M.M., Silberg, J.J., Voigt, C.A., Endelman, J.B., Mayo, S.L., Wang, Z.G. and Arnold, F.H. (2003) Library analysis of SCHEMA-guided protein

recombination. *Protein. Sci.*, **12**, 1686-1693.

11. Saraf, M.C. and Maranas, C.D. (2003) Using a residue clash map to functionally characterize protein recombination hybrids. *Protein Eng.*, **16**, 1025-1034.

12. Weinreich, D.M., Delaney, N.F., Depristo, M.A. and Hartl, D.L. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, **312**, 111-114.

13. Weinreich, D.M., Watson, R.A. and Chao, L. (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution Int. J. Org. Evolution*, **59**, 1165-1174.

14. Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W. and Dress, A.W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164-178.

15. Gaucher, E.A., Miyamoto, M.M. and Benner, S.A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl. Acad. Sci. U S A*, **98**, 548-552.

16. Larson, S.M., Di Nardo, A.A. and Davidson, A.R. (2000) Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.*, **303**, 433-446.

17. Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342-358.

18. Lockless, S.W. and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295-299.

19. Pollock, D.D., Taylor, W.R. and Goldman, N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, **287**, 187-198.

20. Suel, G.M., Lockless, S.W., Wall, M.A. and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication

in proteins. *Nat. Struct. Biol.*, **10**, 59-69.

21. Dubey, A., Realff, M.J., Lee, J.H. and Bommarius, A.S. (2006) Identifying the interacting positions of a protein using boolean learning and support vector machines. *Computational Biology and Chemistry*, **30**, 268-279.

22. Schneeweiss, W.G. (1989) *Boolean Functions with Engineering Applications and Computer Programs.* Springer-Verlag, Heidelberg.

23. Deshpande, A.S. and Triantaphyyllou, E. (1998) A greedy randomized adaptive search procedure (GRASP) for inferring logical clauses from examples in polynomial time and some extensions. *Math. Comput. Model.*, **27**, 75-99.

24. Triantaphyyllou, E. (1994) Inference of a minimum size boolean function from examples by using a new efficient branch-and-bound approach. *Journal of Global Optimization*, **5**, 69-94.

25. Vapnik, V. (1995). Springer, Berlin.

26. Christianini, N. and Shawe-Taylor, J. (2000) *An introduction to Support Vector Machines and other Kernel-based learning methods.* Cambridge University Press.

27. Deniz, O., Castrillion, M. and Hernandez, M. (2003) Face recognition using independent component analysis and support vector machines. *Face Recognition Letters*, **24**, 2153-2157.

28. Ward, J.J., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2003) Secondary structure prediction with support vector machines. *Bioinformatics*, **19**, 165-1655.

29. Kim, H. and Park, H. (2003) Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng.*, **16**, 553-560.

30. Sanchez, A. and David, V. (2003) Advanced support vector machines and kernel methods. *Neurocomputing*, **55**, 5-20.

31. Dubey, A., Realff, M.J., Lee, J.H. and Bommarius, A.S. (2005) Support vector machines for learning to identify the critical positions of a protein. *J. Theor. Biol.*,

234, 351-361.

32. Verkhusha, V.V. and Lukyanov, K.A. (2004) The molecular properties and applications of Anthozoa fluorescent proteins and chromoproteins. *Nature biotechnology*, **22**, 289-296.

33. Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2002) A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U S A* **99**, 7877-7882.

34. Baird, G.S., Zacharias, D.A. and Tsien, R.Y. (2000) Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral. *Proc. Natl. Acad. Sci. U S A*, **97**, 11984-11989.

35. Barton, D.H. and Magnus, P.D. (1971) Experiments on the synthesis of tetracycline. I. Introduction to the series. *J. Chem. Soc. [Perkin 1]*, **12**, 2164-2166.

36. Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.*, **30**, e43.

37. Joern, J.M., Meinhold, P. and Arnold, F.H. (2002) Analysis of shuffled gene libraries. *J. Mol. Biol.*, **316**, 643-656.

38. Abecassis, V., Pompon, D. and Truan, G. (2000) High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2. *Nucleic Acids Res.*, **28**, E88.

39. Akinori Ikeuchi, Y.K., Tomoya Shinbata, and Tsueneo Yamane. (2003) Chimeric gene library construction by a simple and highly versatile method using Recombination-Dependent Exponential Amplification. *Biotechnol. Prog.*, **19**, 1460-1467.

40. Sambrook, J. and Russell, D.W. (2001) *Molecular cloning : a laboratory manual*. 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

41. Arnold, F.H. and Georgiou, G. (2003) *Directed enzyme evolution : screening and*

selection methods. Humana Press, Totowa, N.J.

42. Platt, J.C. (1998), *Technical Report MSR-TR-98-14*. Microsoft Corporation.

43. Bates, P.A., Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins,* **Suppl 5**, 39-46.

44. Bates, P.A. and Sternberg, M.J. (1999) Model building by comparison at CASP3: using expert knowledge and computer automation. *Proteins*, **Suppl 3**, 47-54.

45. Contreras-Moreira, B. and Bates, P.A. (2002) Domain Fishing: a first step in protein comparative modelling. *Bioinformatics*, **18**, 1141-1142.

46. Kim, D.E., Chivian, D. and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526-W531.

47. Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318-326.

48. Fisher, R. (1966) *The Design of Experiments*. Oliver & Boyd, Edinburgh.

49. Yarbrough, D., Wachter, R.M., Kallio, K., Matz, M.V. and Remington, S.J. (2001) Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-A resolution. *Proc. Natl. Acad. Sci. U S A*, **98**, 462-467.

50. Verkhusha, V.V. and Sorkin, A. (2005) Conversion of the monomeric red fluorescent protein into a photoactivatable probe. *Chem. Biol.*, **12**, 279-285.

51. Jucovic, M. and Poteete, A.R. (1998) Delineation of an evolutionary salvage pathway by compensatory mutations of a defective lysozyme. *Protein. Sci.*, **7**, 2200-2209.

52. Kulathinal, R.J., Bettencourt, B.R. and Hartl, D.L. (2004) Compensated deleterious mutations in insect genomes. *Science,* **306**, 1553-1554.

53. Besire, S., Ludwig, A., Brade, V. and Wichelhaus, T.A. (2005) Compensatory adaptation to the loss of biological fitness associated with acquisition of fusidic acid resistance in staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, **49**, 1426-1431.

54. N., G. and Peitsch, M. (1997) Swiss-model and the swiss-pdbviewer: An environment for comparative protein modeling. *Electrophoresis*, **18**, 2714-2723.

55. Zaveer, M.S. and Zimmer, M. (2003) Structural analysis of the immature form of the GFP homologue DsRed. *Bioorg. Med. Chem. Lett.*, **13**, 3919-3922.

56. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S. *et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nature biotechnology*, **25**, 338-344.

57. Liao, J., Warmuth, M.K., Govindarajan, S., Ness, J.E., Wang, R.P., Gustafsson, C. and Minshull, J. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnology*, **7**, 16.

58. Scholkopf, B. and Smola, A.J. (2002) *Learning with Kernels.* MIT Press, Cambridge.

5.7 Supplement

5.7.1 The input data for BLSVM analysis - Sequence-activity data used as input to BLSVM to identify interacting residues. Interactions are preserved when residues 197 (589-591bp) and 217 (649-651bp) are from the same parents.

S No.	Starts with	Crossover position DNA	No. of Crossovers	Function	Interactions preserved?
1	mRFP	6	1	Y	Y
2	dsRed	6	1	Y	Y
3	mRFP	12	1	Y	Y
4	dsRed	31	1	Y	Y
5	mRFP	72	1	Y	Y
6	mRFP	78	1	Y	Y
7	dsRed	141	1	Y	Y
8	dsRed	153	1	Y	Y
9	mRFP	156	1	Y	Y
10	mRFP	159	1	Y	Y
11	dsRed	159	1	Y	Y
12	mRFP	165	1	Y	Y
13	mRFP	171	1	Y	Y
14	dsRed	171	1	Y	Y
15	mRFP	186	1	Y	Y
16	dsRed	186	1	Y	Y
17	dsRed	190	1	Y	Y
18	mRFP	204	1	Y	Y
19	dsRed	204	1	Y	Y
20	mRFP	207	1	Y	Y

5.7.1 continued

21	dsRed	229	1	N	Y
22	dsRed	231	1	Y	Y
23	mRFP	276	1	Y	Y
24	dsRed	276	1	Y	Y
25	mRFP	282	1	Y	Y
26	mRFP	300	1	N	Y
27	dsRed	300	1	N	Y
28	dsRed	348	1	N	Y
29	mRFP	414	1	Y	Y
30	mRFP	418	1	Y	Y
31	mRFP	435	1	Y	Y
32	dsRed	439	1	N	Y
33	mRFP	549	1	N	Y
34	mRFP	589	1	**N**	**N**
35	dsRed	595	1	**N**	**N**
36	mRFP	609	1	**N**	**N**
37	dsRred	609	1	**N**	**N**
38	mRFP	618	1	**N**	**N**
39	mRFP	618	1	**N**	**N**
40	dsRed	618	1	**N**	**N**
41	dsRed	619	1	**N**	**N**
42	mRFP	633	1	**N**	**N**
43	dsRed	633	1	**N**	**N**
44	mRFP	636	1	**N**	**N**
45	dsRed	636	1	**N**	**N**
46	mRFP	645	1	**N**	**N**
47	dsRed	645	1	**N**	**N**
48	mRFP	660	1	Y	Y

5.7.1 continued

49	mRFP	664	1	Y	Y
50	mRFP	666	1	Y	Y
51	mRFP	171, 186, 247, 282	4	N	Y
52	mRFP	171, 306	2	N	Y
53	mRFP	186, 204	2	Y	Y
54	mRFP	186, 204, 300, 348	4	Y	Y
55	mRFP	186, 204, 456	3	N	Y
56	mRFP	186, 204, 589, 633	4	**N**	**N**
57	mRFP	186, 219, 618	3	**N**	**N**
58	mRFP	186, 259	2	N	Y
59	mRFP	186, 428, 447	3	Y	Y
60	mRFP	189, 223, 618	3	**N**	**N**
61	mRFP	204, 219	2	Y	Y
62	mRFP	204, 223, 252, 259	4	Y	Y
63	mRFP	204, 247	2	Y	Y
64	mRFP	204, 300	2	N	Y
65	mRFP	204, 414, 456	3	Y	Y
66	mRFP	228, 277, 348	3	Y	Y
67	mRFP	300, 337, 609, 637	4	Y	Y
68	mRFP	306, 321	2	Y	Y
69	mRFP	306, 321, 456	3	N	Y
70	dsRed	321, 334	2	Y	Y
71	mRFP	333, 363, 589, 618	4	**N**	**N**
72	dsRed	39, 186, 321	3	N	Y
73	dsRed	40, 348, 468	3	Y	Y
74	mRFP	444, 609	2	**N**	**N**
75	mRFP	45, 186, 609, 633	4	Y	Y
76	mRFP	589, 633	2	**N**	**N**
77	mRFP	594, 618	2	Y	Y
78	mRFP	6, 21, 39	3	Y	Y

5.7.1 continued

79	mRFP	6, 248, 300	3	Y	Y
80	mRFP	609, 633	2	Y	Y
81	dsRed	79, 300, 337	3	N	Y
82	dsRed	84, 240, 258	3	N	Y
83	dsRed	91, 594	2	**N**	**N**

5.8 Appendix

5.8.1 Boolean Learning

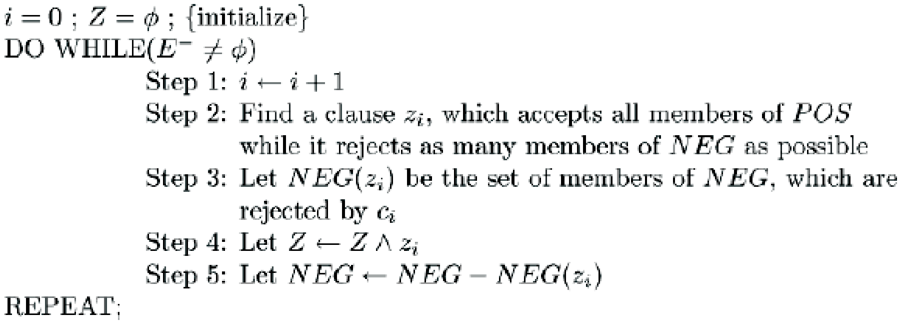
Boolean functions (22) are a standard representational tool for computer and engineering applications. Given a set of examples in Boolean form (0's and 1's), which are classified as either positive or negative, a Boolean learning algorithm can establish a set of rules or Boolean expressions, which classify all the examples correctly. The approach to finding interacting positions is to train a classification algorithm to distinguish between sequences that are active and inactive. The resulting classification scheme is then examined to reverse engineer the interacting positions. Dubey *et al.* (21) showed that interacting residues can be identified from the sequences of both active and inactive variants obtained by recombining two or more parents by expressing the problem in the form of Boolean learning. Boolean learning is a classification problem where a Boolean function is identified from a given set of positive and negative data (22). The data is also in the form Boolean vectors where each dimension of the vector can have a value of 0 or 1. The identified Boolean function is selected because it can classify the given examples correctly into positive or negative. A Boolean expression or function is combination of binary variables joined by logical operators like `AND' or `OR'. These expressions can also be viewed as logical rules that can classify examples into two classes -- positive and negative, based on whether the rules are satisfied or not. If the boolean function is chosen such that it reflects the interaction between different amino acid residues then those amino acids can be identified.

The variant sequences used by Dubey *et al.* (21) were obtained by the recombination (4,39) of two or more parent proteins. Thus, any amino acid in a variant sequence can be traced back to one of the parents. This limits the number of amino acids, which can occur in any position of the variant sequences to the number of parents. To formulate a Boolean learning problem, the sequence of each variant, denoted by **x**, was transformed to yield a Boolean vector **X**. For *k* parents, since each position can have up to *k* amino acid residues, a vector $\mathbf{X} = \left[A_1, A_2 ..., A_k\right]$ can be formed where each A_i is a Boolean vector $\left\lfloor a_{i1,} a_{i2}, ...a_{in} \right\rfloor$ for an enzyme of length *n*. If $a_{ij} = 1$ (or a_{ij} is true), the amino acid in position *j* of the sequence alignment comes from the parent *i* and if it does not, a_{ij}= 0 (or $\overline{a}_{ij}$ is true). There is a small, but finite, probability that the amino acid in a position does not come from either of the parent (if the crossover is made between the codon that is not conserved); in that case, $\overline{a}_{ij}$ is true for all *i*'s. Note that if insertions or deletions exist in any parent, they can be treated as an amino acid belonging to that particular parent. For example, if a variant shows an insertion in position *i*, which is also present in parent *j*, then $\overline{a}_{ij}$ is true.

The Boolean function, which represents binary interactions between different amino acids in the sequence, was formulated by Dubey et al. (21) in the form of a Conjunctive Normal Form (CNF) (22). For two parents, this function can be written as:

$$f(\mathrm{X}) = \underset{l,m \in \Theta}{\Lambda} \left(\left(a_{1l} \vee a_{2m} \right) \left(a_{1m} \vee a_{2l} \right) \right) \tag{A.1}$$

where $\wedge$ stands for `AND' and $\vee$ stands for `OR'. *l* and *m* are a pair of interacting positions and Θ is the set of all interacting pairs. Each $a_1 \vee a_2$ is referred to as a clause and any $\wedge_{i=1}^{N} (\beta_i)$ stands for $\beta_1 \wedge \beta_2 ... \wedge \beta_N$ for clauses β_1 through β_N . If *f*(**X**) is true (*f*(**X**) =1), the sequence **X** is functional. Note that due to the 'AND' sign between the different clauses, all of them need to be true for *f*(**X**) to be equal to 1. A clause being true implies that the condition given by that clause is satisfied for any sequence. For example, in Equation (A.1) if a_{1l} is true a_{1l} =1, then the clause is true since, only one of a_{1l} or a_{2m} need to be true, due to the 'OR' sign between them. However, since a_{2l} has to be false (only one of a_{1l} or a_{2l} can be true because the amino acid in position *l* can come from either parent 1 or parent 2), the term a_{1m} needs to be true for the second clause and thus, the function to be true. This implies that both the positions *l* and *m* in the alignment need to have the amino acids corresponding to parent 1. Similarly, the two clauses will be true if both the amino acids correspond to parent 2. Thus, these clauses represent the interaction between the two positions. This function can also be extended to more than two parents(21). Boolean learning was used to find the set Θ , which consists of the interacting positions. The *One Clause At a Time* (OCAT) algorithm (24) for a CNF function, shown in Figure 5.10, was used for Boolean learning to identify the function in equation (A.1) from the given sequences. The data set is divided into positive sequences, *POS*, and negative sequences, *NEG*. Step 2 of the algorithm finds a clause, which accepts, or is true, for all the positive examples and is not true for the maximum number of the negative examples. This is in agreement with any clause in a CNF function as described earlier, since it should be true for every positive example. By selecting the clause, which is rejects the largest number of negative examples, OCAT selects the most important clause first. It should be noted that any clause in a CNF function can be true for some negative examples, but together, all the clauses in a CNF function must reject all the negative examples. Dubey *et al*. (21) modified the algorithm to search only for clauses of the form shown in Equation (A.1). The found clauses indicate an interaction between the amino acid positions involved in.



```
$i = 0$ ; $Z = \phi$ ; {initialize}
DO WHILE($E^- \neq \phi$)
            Step 1: $i \leftarrow i + 1$
            Step 2: Find a clause $z_i$, which accepts all members of $POS$
                    while it rejects as many members of $NEG$ as possible
            Step 3: Let $NEG(z_i)$ be the set of members of $NEG$, which are
                    rejected by $c_i$
            Step 4: Let $Z \leftarrow Z \wedge z_i$
            Step 5: Let $NEG \leftarrow NEG - NEG(z_i)$
REPEAT;
```

Figure 5.10 OCAT algorithm.

5.8.2 The BLSVM algorithm

Dubey *et al.* (21) also introduced the BLSVM algorithm, which combines Boolean learning with SVMs, a class of non-linear classification algorithms. Thus, given a data set consisting of points in the form of vectors, divided into two classes, they can find a classifying function. This function can then be used to classify any data points, for which the class is unknown. SVMs are based on *Statistical Learning Theory* and as such, they also try to minimize the s*tructural risk* of the solution (25,58). Data, divided into two different classes, are mapped into a higher dimensional feature space, using a kernel. The feature space is formulated in such a way that the data are linearly separable and linear regression can be used to find a classifying function. This function can be used to predict the class for unclassified data. SVMs have been widely applied in the field of pattern recognition, for example, facial recognition (27,30). They have also been used for the secondary structure prediction of proteins (28,29)

The input data for SVMs are in the form of vectors **x** with outputs $y \in [1,-1]$ depending on whether the datum belongs to the positive or negative class. As mentioned earlier, the input data is transformed into a feature space, which is usually a higher-dimensional space than the input space. Each vector is transformed into the feature space to yield a feature vector, ϕ (**x**), for a transformation ϕ. SVMs bypass the intricate task of explicitly defining a feature space for a problem by using a kernel function as the inner product of feature vectors:

$$K(x_1, x_2) = \langle \phi(x_1) \cdot \phi(x_2) \rangle \qquad \text{(A.2)}$$

If $S = ((x_1, y_1,), \ldots, (x_l, y_l))$ is the training data set for an SVM, the learning algorithm constructs a hyperplane that optimally separates the data in the feature space (26,58):

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right) \qquad \text{(A.3)}$$

If *f(x)* is positive, the point **x** belong to the class with *y* = 1 and vice versa. The variables $\alpha = (\alpha_1, \ldots, \alpha_l)$ are called the Lagrangian multipliers and *b* is called the bias. These variables are obtained by minimizing the risk in classifying the data. Using the duality theorem, minimum risk translates into the following optimization problem:

$$\text{maximize } W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_1, x_j) \qquad \text{(A.4)}$$

$$\text{subject to } \sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1,\ldots, l$$

This formulation corresponds to the 1-Norm soft margin classification (58). The bias, *b* is chosen so that $y_i f(x_i)$ =1for any *i* with $0 \le \alpha_i \le C$.

Since the feature vector is never explicitly used in this formulation, a function that can be defined as the kernel is sufficient to define the feature space. For a function to qualify as a kernel, it should satisfy the properties of an inner product in any space, symmetry and the Cauchy-Schwarz inequality $\left((K(x,z))^2 \le K(x,x)K(z,z)\right)$. In addition, according to Mercer's theorem, the kernel matrix, $\mathrm{K} = \left(K\left(x_i, x_j\right)\right)_{i,j=1}^{l}$, should also be positive semi-definite (have non-negative eigenvalues) (26). Different forms of functions, which satisfy these properties, have been used for various problems. The polynomial kernel, $K(x,z) = (1 + x \cdot z)^m$ and the radial basis kernel, $\exp\left(-\|x - z\|^2 / \sigma^2\right)$ are frequently applied.

Since SVMs are based on structural risk, the classification function found by them has is supposed to have a better chance of performing well for data points, which were not included in the given data set, or in other words, have better generalization. This also enables them to tolerate errors in the data set. On the contrary, the OCAT algorithm for Boolean learning, mentioned earlier, is based only on the *empirical risk*, since it finds a function, which classifies all the given data points correctly, or in other words, minimizes the empirical error. Therefore, Dubey *et al.*(21) formulated the BLSVM algorithm by combining Boolean learning and SVMs. This was done using the Lagrangian variables, denoted by α in Equation (A.3). These variables also can be inferred as the significance of the corresponding data point in the classification function. In BLSVM, this concept of assigning significance or weights to each data point was borrowed from SVMs and incorporated in Boolean learning. Thus, the same data set was used by both SVMs as well as for the OCAT algorithm as described earlier. However, the value of α's obtained from SVMs were used to evaluate the weights, *W*, for Boolean learning as given by Dubey *et al.* (21):

$$W(x_i) = 1 + \beta \times \alpha_i / \max(\alpha) \qquad \text{(A.5)}$$

Where β is a weighting parameter, which can be used to vary the importance of α. The OCAT algorithm was modified to accept these weights for each data point (21). Instead of using the number of examples as the criteria for selecting a clause in Figure 8, the sum of the weights for each example was taken into account. The weighing parameter, β, varies the amount of significance given to the result from SVMs.

CHAPTER 6

EXPANDING THE SUBSTRATE SPECIFICITY OF PENICILLIN G ACYLASE

6.1 Introduction

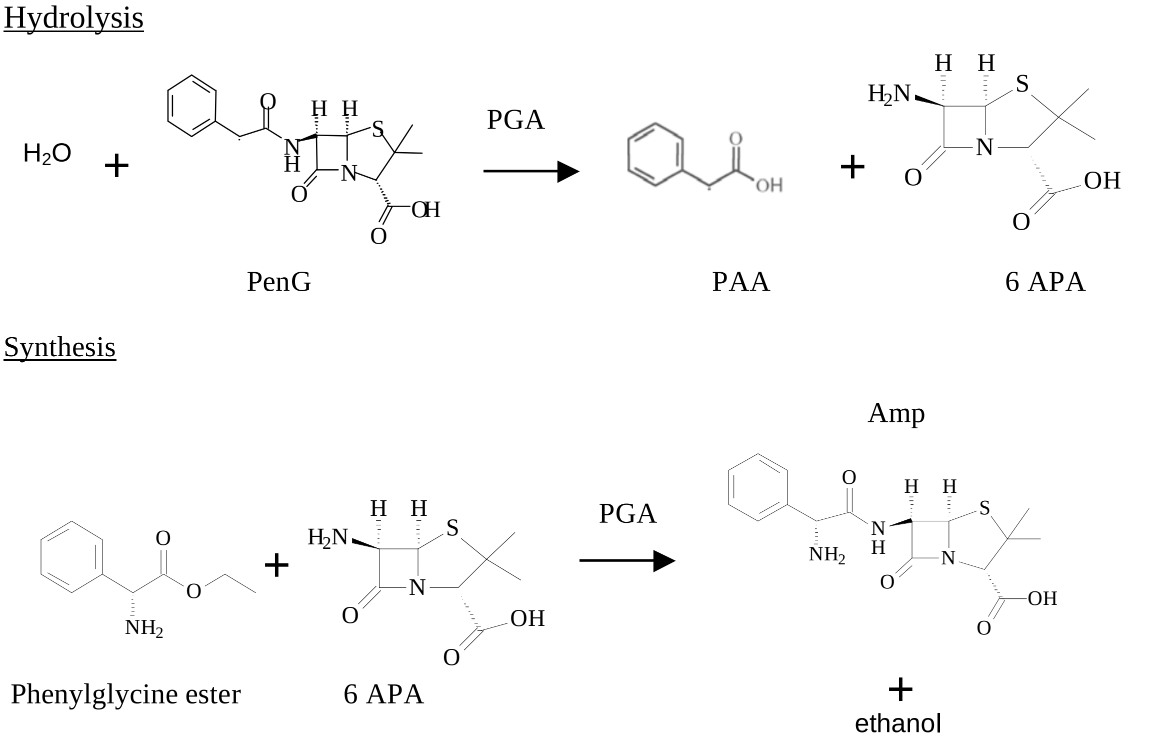
6.1.1 Increasing resistance threat and the need for development of better biocatalyst for antibiotics synthesis

The development of resistance of bacteria towards antibiotics is a major economic and health threat. Antibiotic resistance has a huge economic impact worldwide. In the US alone, the economic impact of antibiotic resistance annually was five billion dollars. Two million cases of infectious diseases occur annually in the US which result in 90, 000 deaths. Of these, 70% of infections that result in deaths are resistant to at least one antibiotic (1). Bacterial resistance towards conventional antibiotics has also been developing rapidly as more organisms such as *S. pneumoniae*, *N. gonorrhoeae* and *Salmonella sp.* have developed resistance towards conventional antibiotics. In addition to those problems, the time span for development of resistance is also short, it typically varies from three to fifteen years (2) which means that antibiotics quickly become useless. The above factors inspire the need to rapidly develop new antibiotics and new enzymes that produce them.

6.1.2 Penicillin G Acylase (PGA): An important biocatalyst

The penicillin G acylase (PGA) protein is an industrially important biocatalyst used for the production of semi-synthetic β-lactam antibiotics. PGA is used industrially for the hydrolysis of penicillin G to produce an important intermediate, 6-amino-penicillanic (6-APA), and phenylacetic acid. PGA also is used to catalyze the synthesis of semisynthetic β-lactam derivatives ampicillin and amoxicillin using phenylglycine or hydroxylphenylglycine ethyl ester with 6-APA, respectively (3,4). Currently, PGA is only active towards penicillin G, ampicillin and amoxicillin

Figure 6.1 and Figure 6.2). A PGA with altered substrate specificity towards β-lactam resistant antibiotics such as oxacillin, cloxacillin and nafcillin would be highly desirable as no biocatalysts are available for these substrates currently.



Abbreviations

PAA:Phenyl acetic acid,6-APA: 6 amino-penicillanic acid, PenG: Penicillin G, Amp: Ampicillin

Figure 6.1 PGA catalyzed reactions.

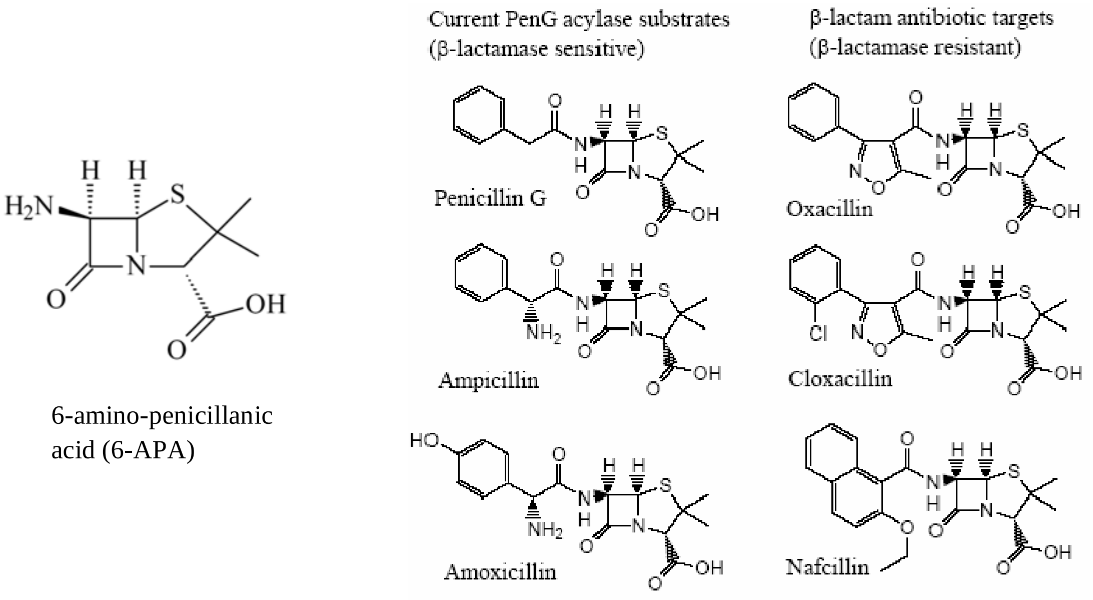

6-amino-penicillanic acid (6-APA)

Figure 6.2 β-lactam ring and various antibiotics.

6.1.3 Characteristics of PGA

PGA (EC 3.5.1.11) from *E. coli* (ATCC 11105) is a 86 kDa heterodimeric amidase that undergoes complex maturation pathways. After the PGA messenger ribonucleic acid is translated to a polypeptide, it exists as a precursor form called PreProPGA that consists of a signal, connecting peptide, α-subunit and β-subunit (Figure 6.3). The precursor is exported to the periplasm where it finally undergoes intramolecular proteolysis to form the mature PGA enzyme (5,6). Numerous studies have been performed on optimizing the expression of PGA and suggest that the complex maturation pathway necessitates a low-temperature (16°C) incubation to produce functional PGA as high-temperature overexpression (>30°C) frequently produces inclusion bodies in the periplasm of overproducing PGA *E.coli* (6-10).

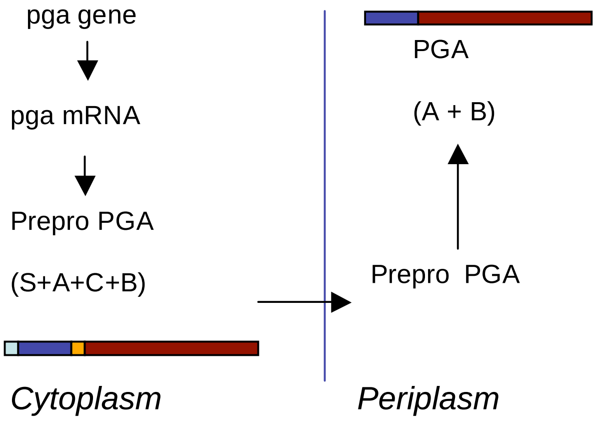


Figure 6.3 Maturation pathway of PGA enzyme. S: Signal peptide (light blue), A: α-chain (blue), C: Linker (orange) and B: β-chain (red).

PGA comes from the family of **N-t**erminal-**n**ucleophile hydrolases that consist of an N-terminal nucleophile (in this case, β-Serine1) acting as the catalytic residue within a four-layer α + β structure with two anti-parallel β sheets that is known as the Ntn fold (Figure 6.4). The catalysis involves the stabilization of the tetrahedral intermediate by Gln B23, Asn B241 and Arg B69 (11) which is then followed by the release of 6-APA. Subsequently, the nucleophilic substrate attacks the alkyl carbonyl group attached to βS1 and releases it from the PGA protein (Figure 6.5).

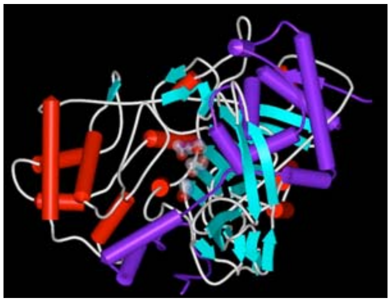


Figure 6.4 Crystal structure of PGA enzyme. Purple chain represent chain A, chain B is represented by other colors. The catalytic residues are shown with their VDW radius as semi-transparent ball and sticks.

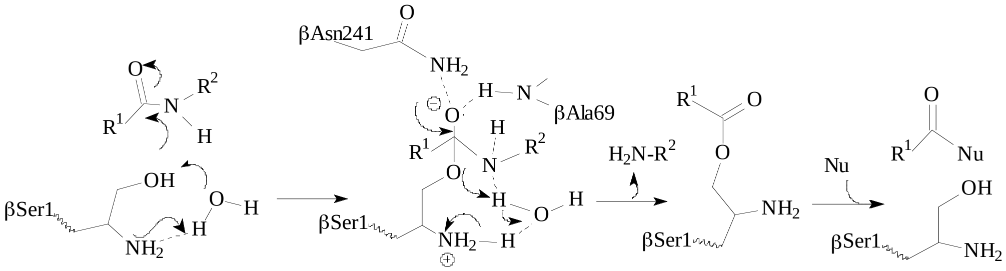


Figure 6.5 Proposed mechanism of PGA catalyzed reaction. The nucleophilic serine attacks the carbonyl carbon of the peptide bond. The oxyanion is stabilized by βN241 and βA69. The products are then released following collapse of the tetrahedral intermediate. R^1 is a phenylacetic acid derivative whereas R^2 may vary.

6.1.4 Prior protein engineering work on PGA

6.1.4.1 Sequence-activity relationships of PGA

Significant protein engineering work has been done on PGA by Alkema, McVey and Duggleby. Alkema and Janssen have crystallized PGA and investigated the functions of various residues which are tabulated in Appendix 6A (12-14). McVey crystallized mutant and wild-type PGA with penicillin G substrate and penicillin G sulfoxide inhibitor to investigate functions of various PGA residues (11). Duggleby postulated a possible mechanism of how PGA catalyzes the amidase reaction which is described Figure 6.5

(15). All these pioneers provided mechanistic insight, crystal structure and residue information on PGA. Efforts by research groups have also been focused on rational design and the optimization of the production of PGA or of the products of the reaction (16,17). Alkema and Janssen have mutated the residues listed in Appendix 6A because they believed that these residues should affect either the binding, catalysis or substrate specificity based on mechanistic insight obtained from three-dimensional crystal structure data (11-13, 15).

6.1.4.2 Improving synthesis-to-hydrolysis ratio of PGA

PGA genes were shuffled to improve its synthesis-to-hydrolysis properties (Figure 6.6). Zhou *et al.* evolved PGA via DNA shuffling on PGA genes with an identity of 62.5% to 96.9% from *Escherichia coli, Kluyvera cryocrescens* and *Providencia rettgeri* (18). Zhou *et al.* selected for higher synthesis-to-hydrolysis activity and managed to obtain a mutant with 40% higher synthetic activity than the wild-type PGA gene. Jager *et al.* shuffled PGA from *E. coli, Kluyvera cryocrescens* and *Providencia rettgeri* and obtained 40-90% increase in the relative rate of acylation of the β-lactam nucleus during ampicillin synthesis (19). Jager *et al.* used a selection prescreen on leucine auxotrophic HB101 cells plated on minimal media supplemented with N-phenyl-acetyl-L-leucine. The active PGAs liberate sufficient leucine required for growth. The selected variants were then screened using HPLC.

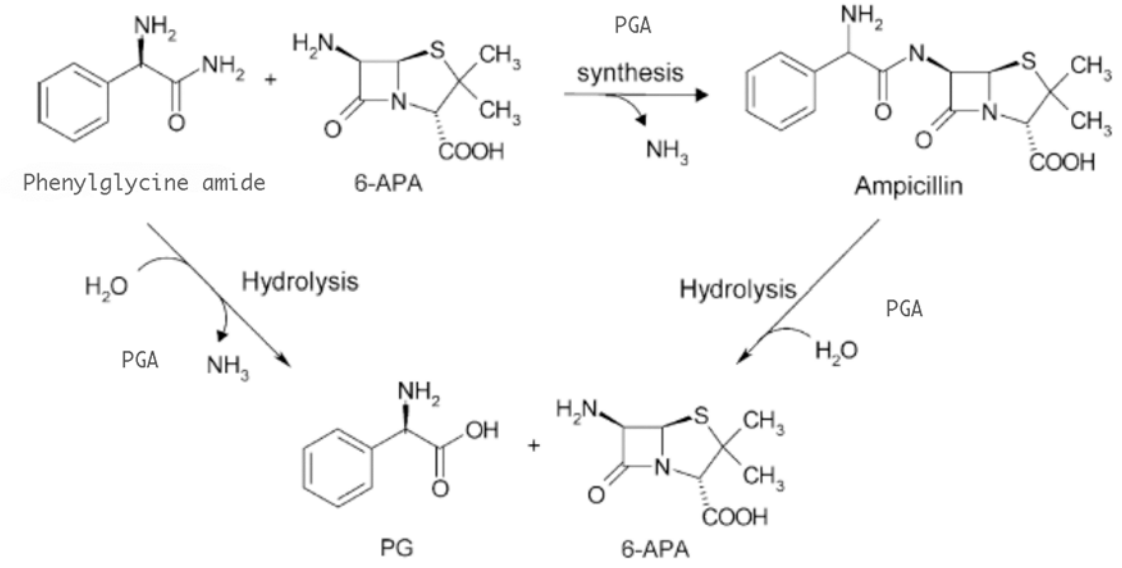


Figure 6.6 Kinetically controlled synthesis of ampicillin. Penicillin G acylase (PGA) can couple the activated side-chain phenylglycine amide to 6-amino-penicillanic acid (6-APA). Both the activated side chain and the formed product, ampicillin can also hydrolyzed to phenylglycine (PG) and 6-APA using PGA as a catalyst. The synthesis-to-hydrolysis ratio can be calculated by dividing the initial rate of formation ampicillin by the rate of formation of hydrolyzed products (PG in this case). Figure adapted from ref (19).

6.1.4.3 Thermo-stability improved PGA variants

Polizzi *et al.* improved the kinetic thermostability of PGA (20). They applied a data-driven, structure-guided consensus approach to select residues potentially important for stability of PGA and avoided mutating the active site of PGA to decrease the risk of affecting its activity. Of 21 point-mutants created, 46% of PGA variants were found to be more thermostable than wild-type PGA. Two PGA variants had an almost threefold longer half-life at 50°C.

6.1.5 Residues important for substrate specificity of PGA

PGA enzyme hydrolyzes a variety of substrates. Documented substrates include 6-nitro-phenylacetamido-benzoic acid, *R*-phenylglycine amide, *R*-phenylglycine methyl ester, *p*-hydroxy-R-phenylglycne amide and 6-nitro-3-*R*-phenylglycine methyl ester Kasche *et al.* investigated the substrate specificity on PGA from *E. coli*, *A. faecalis*, *K. cryocrescens*, *A. viscosus* and found that the substrate specificity (k_{cat}/K_M) for these groups is broad, ranging anywhere from 10^3 to 10^6 (21).

A number of publications on PGA point out that the residues that are responsible for substrate specificity are in the active sites of PGA (22). Residues that bind to the β–lactam ring moiety include βF146 and βF71 through van der Waals interactions of the side chains and by hydrogen bonds between the guanidium group of αR145. In the acyl binding site, βA69 and βS1 interact with the phenylacetyl group. The entrance to this site is formed by βF71 and βF24.

Analysis of the crystal structure of PGA (PDB ID: 1gm7) indicates a relatively small binding pocket (Figure 6.7). This suggests that there could be a potential pocket binding problem if we want to evolve PGA towards a larger substrate such as oxacillin, cloxacillin of nafcillin using random mutagenesis. Morley and Kazlaukas (23) analyzed the effect of mutating residues near the catalytic sites on activity for β-glucosidase, dioxygenase, desaturase, glutaryl acylase and subtilisin. They concluded that site-directed mutagenesis on regions near the active sites (5 – 10 Å) would probably be more useful than random mutagenesis to change the substrate specificities of enzymes as change in activation free energy for enantioselectivity was most significant in that region of proteins (See Figure 2.4). In light of the above, one can conclude that structure-guided approach involving site-saturated mutagenesis close to the active site to evolve PGA towards altered substrate specificity would have a better chance of success than random mutagenesis-based methods.

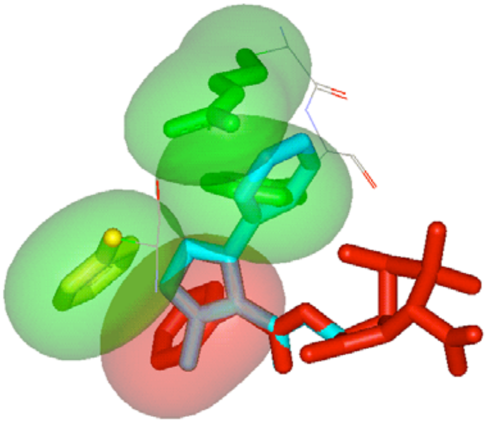
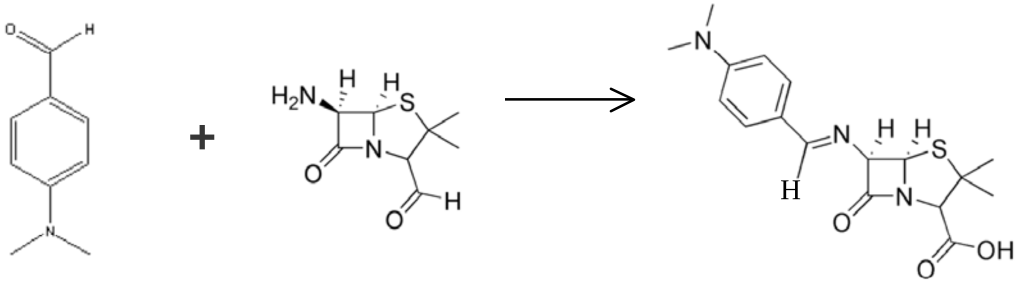


Figure 6.7 Crystal structure of PGA complexed with Pencillin G Sulfoxide. Red: Penicillin G Sulfoxide, Blue: Oxacillin. The amino acids a145, b 146 and b24 (Green), must be modified to accommodate the larger side chain.

6.1.6 PGA assays

6.1.6.1 PDAB Schiff base chromogenic assay

The PDAB assay has been used by a number of groups to screen and characterize PGA related activities. For example, Gabriel *et al.* used a PDAB filter lift assay to screen for PGA variants with improved activity at alkaline pH (24). Two key advantages of the PDAB assay are that it can be used in a high-throughput fashion and it can detect hydrolysis of any β-lactam antibiotics as it detects presence of 6-APA. The disadvantage of the assay is that it can be only used as an end-point assay.



PDAB 6-APA Schiff base chromogenic compound

Figure 6.8 PDAB (p-dimethylamino-benzaldehyde) reacts with 6-APA to form a Schiff base that can be monitored at 415 nm.

6.1.6.2 NIPAB chromogenic assay

NIPAB (2-nitro-5- [(phenylacetyl)amino]benzoic acid, (Figure 6.9) frequently serves as an assay for PGA. PGA catalyzes the hydrolysis of NIPAB and this results in the cleavage of the amide bond, releasing a chromogenic product which results in an increase in absorbance at 405 nm (12,25). However, NIPAB cannot be used for screening oxacillin variant since an oxacillin analogue version of NIPAB is not available.

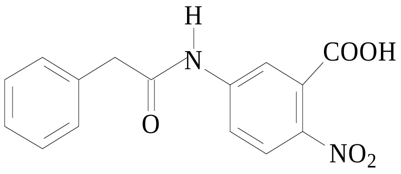


Figure 6.9 NIPAB chemical structure

6.1.6.3 Cresol Red pH based colorimetric assay

The cresol red and phenol red assay is a convenient colorimetric assay that can be used to measure penicillin G hydrolysis. Simons *et al.* successfully used phenol red assay on PGA catalyzed reactions (26). The pH change is induced when phenyl-acetic acid is liberated due to the hydrolysis of penicillin G. This results in a colorimetric change of the pH indicator. This assay is not applicable to assay all β-lactam antibiotics as the side-chains of the group linked to 6-APA have to be sufficiently acidic or alkaline to induce a colorimetric change. This assay could be used in a high-throughput fashion.

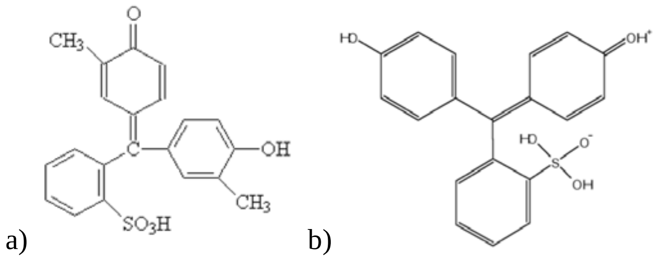


Figure 6.10 (a) Cresol red structure, (b) Phenol red structure

6.1.6.3 HPLC assay

The products of PGA catalyzed reactions can be characterized by HPLC as long as there are appropriate standards available. HPLC has been used to analyze active PGA variants. Usually, C18 columns are used in connection with a pump and a UV detector set at 214 nm (19). The key disadvantage of using HPLC to analyze the reactions is its low throughput.

6.1.7 Scope of work

Protein engineering by structure-guided method on PGA was performed to expand the substrate repertoire. We decided to engineer PGA for activity towards oxacillin first, then check the variant for activity on nafcillin and cloxacillin. To start with, we examined the crystal structure of penicillin G complexed with penicillin G sulfoxide and found residues α145, β146, and β24 of PGA to be potential targets for mutagenesis. We then produced libraries by the overlap extension method and screened the library using p-dimethyl-amino benzaldehyde (PDAB) assays.

6.2 Materials & Methods

6.2.1 Materials

Most molecular biology enzymes were bought from New England Biolabs (Beverly, MA). The *Pfu* and *Taq* polymerase were purchased from Stratagene (La, Jolla, California). Chloroamphenicol was purchased from Sigma (St. Louis, Missouri). The inducer, isopropyl-β-D-thiogalactoside (IPTG) was purchased from EMD Biosciences (San Diego, CA). All oligonucleotides were ordered from MWG-BIOTECH (High Point, North Carolina). Penicillin G Acylase genes from *E*. coli (PEC) and *K*. *cryocrescens* (PKC) were gratefully obtained from Dr. Dick Janssen (University of Groningen, Groningen, Netherlands).

6.2.2 Expression and purification of Penicillin G Acylase

Freshly transformed colonies on choramphenicol LB selection plates containing the PGA gene were picked and grown overnight in LB CM, 37°C using a shaking incubator. 1:200 dilutions of the overnight cultures in 25 ml of LB CM at 37°C follows.

The cultures were grown till the OD_{600} reached 0.8 and IPTG is then added to a final concentration of 0.1mM to induce expression. The cultures were then allowed to express at 24°C at least 16 hours. After 16 hours, the cultures were pelleted by centrifugation and lysed by osmotic shock following the protocol described in ref (12). Briefly, the pellets were resuspended in 1/10 volume of ice-cold osmotic shock buffer (20% sucrose, 100 mM Tris-HCl, pH 8.0, 10 mM EDTA). Follow centrifugation, the pellet was again resuspended in 1/10 ice-cold osmotic shock buffer (1 mM EDTA) and centrifuged again. The final supernatant solution contains the PGA enzymes thereafter.

6.2.3 Library generation using overlap extension

The penicillin G acylase library was generated using overlap extension method described in (27). Degenerate primers containing NNK were ordered from MWG (High point, North Carolina). Primer sequences can be found in the supplement section. All PCR reactions were performed using *Pfu* polymerase and mixed following the vendor's instruction (final concentrations of 0.2 mM dNTP, 1 X Pfu buffer, 0.3 pmol/ μL primers and ~30 ng PEC WT plasmid templates to 50 μl final volume) and cycled using the following program: 95°C for 2 min, hold 95°C add enzyme, 30 X (95°C for 1min, 50°C for 1 min, 72°C for 1 – 3 min depending on length of fragments), 72°C for 10 min, hold 4°C. Fragments containing the NNK codons were then gel purified and used for the secondary reassembly PCR. For the secondary PCR, the full-length degenerate genes were assembled using equimolar concentrations of NNK fragments together with 1 X Pfu buffer and 0.2 mM dNTP. The PCR cycling protocols were: 95°C for 2 min, hold 95°C add enzyme, 30 X (95°C for 1min, 47°C for 1 min, 72°C for 5 min), 72°C for 10 min, hold 4°C. For the tertiary PCR, ten to twenty μl of the assembly PCR mixtures were mixed with the end-primers (0.3 pmol/ ul final concentrations of PECfor68 and PECrev68) containing appropriate restriction sites. The cycling parameters were: 95°C for 2 min, hold 95°C add enzyme, 15 X (95°C for 1 min, 55°C for 1 min, 72°C for 5 min), 72°C for 10 min, hold 4°C. For the tertiary PCR, it is important not to have too many cycles of amplification to reduce the bias of the library.

The insert library was digested using *NdeI* and *HindIII* enzymes and ligated into digested and dephosphorylated PEC vectors. The library was transformed into *E. coli* HB101 strain and plated on CM selective LB agar plates. Similarly digested and dephosphorylated PEC plasmids were used as control to monitor intramolecular ligation background. Typically, we found less than 0.5 % background. At least four of the transformants were sent for sequencing to confirm that we have obtained a diverse library.

6.2.4 Assay of PGA using colormetric means

We followed the procedures described in (26). The assay reagent mixture consisted of 100 mM NaH_2PO_4 pH 8.2, 15 mM penicillin G, 0.002% (w/v) Phenol Red in 1 ml assay volume. Standard curves were made with various concentrations of phenylacetic acid. Typically, 10 ul of purified enzyme mixture was added to 1 ml of assay solution. The Beckmann Coulter DU 800 spectrophotometer (Fullerton, CA) was used to record the readings at absorbance 430 nm.

6.2.5 Assay of PGA using PDAB solution in 96-well plates (End point method)

Para-dimethyl-benzaldehyde solution (PDAB) was used to measure the amount of β-lactam produced by the hydrolysis reaction of β-lactam-based antibiotics such as penicillin G, ampicillin, oxacillin and methicillin. The reaction of PDAB with the β-lactam moiety forms a Schiff base which can be monitored at 415 nm. Typically, a yellow-colored solution can be observed in the presence of β-lactam. PDAB assay is however, an end-point measurement method as the Schiff base reaction requires a low pH to occur.

15mM of antibiotics were mixed with ~ 30 μg of purified enzyme. The reaction is allowed to proceed at 30°C for at least ½ hour. Following that, 20 μl of reaction mixture was transferred to 180 μl PDAB assay solution in transparent Falcon 96-well plates. After 5 mins incubation at room temperature, absorbance readings were taken using the Fluorstar Galaxy spectrophotometer (Durham, NC).

6.2.6 High-through put screening using PDAB filter lift assay

A modified procedure obtained from (24) was used. Freshly transformed HB101 bacterial colonies were lifted using filter papers and the corners were marked with a black ballpoint pen so that we can map the colonies of interest back their original position on the plates. The colonies were lysed using chloroform vapor for ½ hour and then rinsed 3 times with the freshly made antibiotic solution (2wt % in PBS pH 7.8). The filter papers were incubated in a sealed container to prevent excessive evaporation at 37°C for ½ hour. Finally, they were then rinsed three times with PDAB solution (7.5 mL of 3.5 wt% PDAB in methanol, 40 mL 40% glacial acetic acid and 20 mL of 0.05M NaOH, final pH of solution is 2.65, final volume is 67.5 mL). A number of labs have used final concentrations of PDAB ranging from 0.0625 to 0.3% (28-31). Generally, increasing the concentration of PDAB increases the response as well as background. Yellow spots on the filter paper indicate production of β-lactam and implied hydrolysis of target antibiotic.

6.2.7 HPLC assay of reaction mixture

10X dilution of enzymes (~250 μg/ml) was added to a solution of antibiotics (15mM of either penicillin G or oxacillin in PBS buffer pH 7.8) and reacted for at least ½ hr at 30°C. The reaction was then diluted 10X into a solution of 5mM phosphate (pH 3) aqueous mixture of 30% acetonitrile (5mM Phos pH 3) ~10 μL of the solution was ran through a Beckman Coulter System Gold® reverse phase HPLC (Fullerton, CA) with a C18 column and 168 Detector. Elution was done using a solution containing 340 mg/l SDS, 5mM phosphate, 30% acetonitrile, adjusted to pH 3.0 with phosphoric acid. Flowrate was set at 1 ml/min. Retention time for 6-APA, oxacillin and penicillin G are 4.12 min, 15.00 min and 9.58 min, respectively.

6.3 Results

6.3.1 Purification and activity assay on PGA from *E .coli and K. cryocrescens*

PGA from *E. coli* (pEC plasmid) and *K. cryocrescens* (pKC plasmid) were expressed in HB101 cells and purified using the osmotic shock method as described in the Materials & Methods section. Figure 6.11 shows that that we obtained the expected two main protein bands from the osmotic shock of pEC and pKC recombinant cells. The heterodimeric PGA protein for *E. coli* and *K. cryocrescens* consists of an α-subunit ~21 kDa and a β- subunit ~65 kDa. Colorimetric assays revealed that the pEC and pKC recombinant proteins show activity on ampicillin (4.04, 3.55 μmol/(mg•min) respectively), penicillin G (15.1, 19.9 μmol/(mg•min) respectively) but not on oxacillin, nafcillin and cloxacillin (<0.1 μmol/(mg•min)).

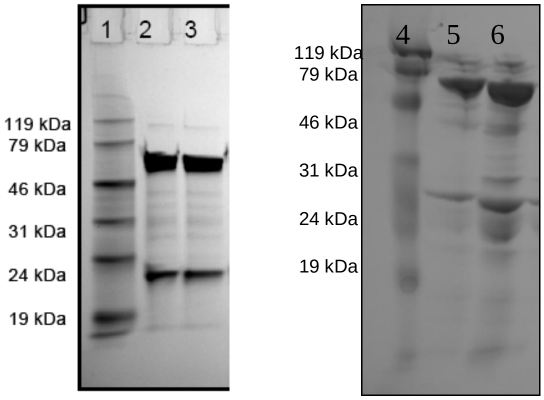


Figure 6.11 Protein gel picture of PGA expressed. (1) Ladder, (2) Periplasmic supernatant for pEC, (3) Dialyzed fraction, (4) Ladder, (5) Perisplasmic supernatant for pEC, (6) Perisplasmic supernatant for pKC. Expected size of pEC consisted of main unit 65kDa and subunit 21kDA. At least 90% purity of PGA were obtained for both pKC and pEC purification. ~2% of total cell protein production was PGA.

6.3.2 Library generation using overlap extension

Following the characterization of the PGA genes, we used generated a library via overlap extension using pEC gene as a template. Figure 6.12 show that the primary, secondary and tertiary PCR steps for the library creation. The library was digested and ligated to a digested and dephosphorylated PEC vector as described in the Materials and Methods section. Table 6.1 lists the results of sequencing of four randomly picked transformants. It can be observed that we did not see any bias of DNA and amino acids for the four sequences. This provided evidence that that a diverse and unbiased library was generated.

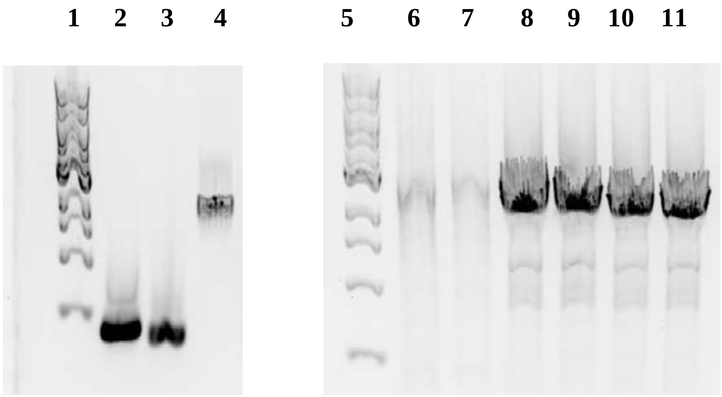


Figure 6. 12 DNA gel of primary, secondary and tertiary PCR for generating library. Expected gene size is approximately 2100 bp. (1) & (5) 0.05 – 10 kbp Ladder, (2) Fragment 1, (3) Fragment 2, (4) Fragment 3, (6) & (7) reassembly overlap PCR, (8) – (11) Secondary amplifications on (6) and (7).

Table 6. 1 Sequencing results of 4 randomly picked pEC variants.

Residues	β24	α145	α146
Library	NNK	NNK	NNK
Original Sequence	TTT (F)	CGG (R)	TTC (F)
Sequence 1	CTG (L)	TTT (F)	TAT (Y)
Sequence 2	CCG (P)	TTG (L)	TAT (Y)
Sequence 3	ACG (T)	CTT (L)	ATG (M)
Sequence 4	AAT (N)	CCT (P)	ACG (T)

6.3.3. Validation/Development of PDAB filter lift assay for screening hydrolysis of oxacillin and penicillin V

6.3.3.1 96-well plates screen

We confirmed the linearity of absorbance at 405nm with increasing concentrations of β-lactam in PDAB reaction solution. Figure 6.13 shows various concentrations of β-lactam dissolved in PBS, pH 7.8 and mixed with PDAB solution. Of the two curves shown, the lower curve shows the titration curve with 100μl of PDAB solution (0.2 wt% final PDAB concentration in a 200 μl final reaction volume) while the upper curve shows the titration curve using 175μl of PDAB solution (0.35wt% final PDAB concentration in 200μl final reaction volume). We observe that by increasing the concentrations of PDAB assay solution, we can obtain a more sensitive response range as the slope varied by a factor of two when we increased the PDAB solution from 100 μl to 175 μl. We also show that the absorbance increases linearly with increasing amounts of β-lactam.

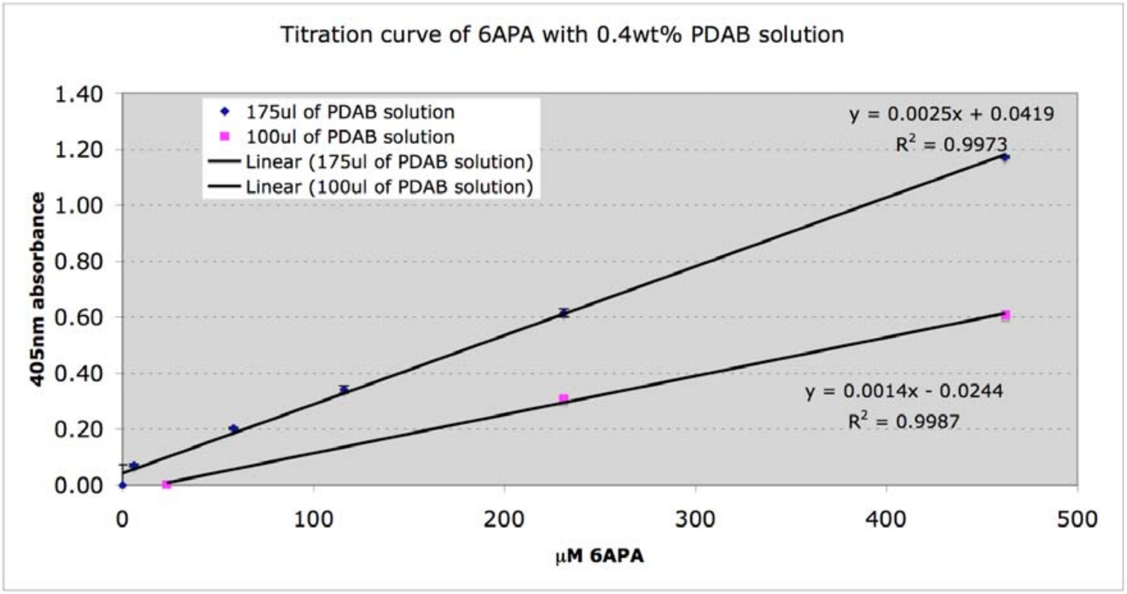


Figure 6.13 Titration curve of 6 amino-penicillanic acid (6 APA) with different final concentrations of PDAB solution. All data points were done in triplicates and final

volume of all reactions were 200ml. Readings were taken using Fluostar Galaxy 96-wellplate spectrophotometer.

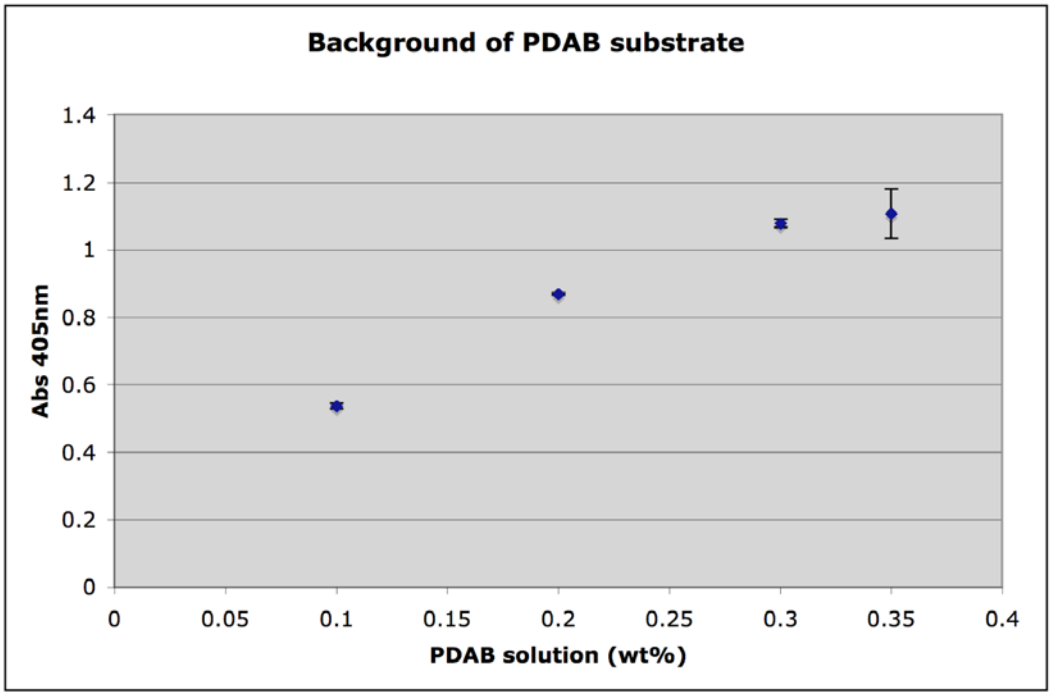


Figure 6.14 The effect of increasing PDAB solution on absorbance at 405nm. All data points were obtained in triplicates using 96-well plate Fluostar Galaxy instrumentation.

To determine the effects of PDAB solution on background absorbance, we titrated PDAB solution and measured its absorbance at 405 nm. Figure 6.14 shows 405 nm absorbance readings measured at various concentrations of PDAB solution. In general, the amount of background absorbance increases as we increase the concentration of PDAB solution.

6.3.3.2 Checking the sensitivity of Filter lift assay

The sensitivity of the filter lift assay was checked by adding various concentrations of β-lactam solution onto filter papers and checked visually for yellow coloration. We tested the extent of our ability to distinguish the yellow coloration of the reaction of PDAB solution with β-lactam on nitrocellulose and VWR filter papers. Table 6.2 shows the results of observing yellow color by using various concentrations of 6-APA from 0 to 800 μM that were spotted on filter papers that were subsequently rinsed with 0.4 wt% PDAB solution. From the experiments, we determined that we could detect down to 470 μM of 6-APA using the filter lift assay with nitrocellulose paper. Figure 6.15 show an example of the filter lift assay on pEC HB101 cells on penicillin G substrates.

Table 6.2 Observable concentrations of yellow stain by the PDAB solution reaction with various concentrations of β-lactam on filter paper. + indicates observation of yellow stain while – indicates absence of yellow stain.

		Paper Type	
6APA (mg/ml)	6APA (mM)	Nitrocellulose paper	VWR filter paper
800	3.72	+++	+++
400	1.86	++	+++
200	0.93	+	++
100	0.47	+-	-
10	0.05	-	-
Control or 0	0.00	-	-

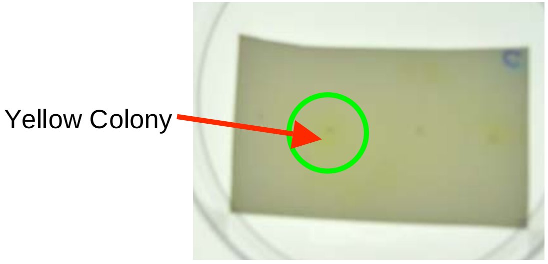


Figure 6.15 Filter lift paper assay on pEC expressing HB101 cells. A yellow smudge indicates active PGA enzyme.

6.3.4 Screening of variants using oxacillin substrates and secondary assay on screened variants

30, 000 variants were generated using overlap PCR using degenerate NNK codons (where N represents DNA codes: A/T/C/G and K represents DNA codes: T/G) at positions, β24, α145 and α146, as mentioned in 6.3.2. The variants were screened using the PDAB filter lift method on oxacillin substrates. Eight colonies around a yellow smudge on the filter paper were picked, grown overnight, expressed and osmotically shocked to obtain purified proteins. The purified proteins were reacted overnight with oxacillin substrate along with negative controls. The samples were then assayed using HPLC. Results from HPLC indicated that no oxacillin activities were detected (See Table 6.3).

Table 6.3 HPLC assay on eight screened oxacillin variants along with positive and negative controls. C1: PBS buffer, PGA, oxacillin, C2: PBS buffer, penicillin G + OS buffer, C3: PBS buffer, oxacillin, OS buffer. P4: PBS buffer, PGA, penicillin G. EP is negative control using purified protein from empty pEC plasmids expression. Oxa1 to Oxa8 are the eight screened variants.

	HPLC Activity	
Sample	Pen G	Oxacillin
C1 - Negative Control	n/a	-
C2 - No Enzyme	-	n/a
C3 - No Enzyme	n/a	-
P4 - Positive Control (WT penGA)	+	n/a
EP- Empty Vector	-	n/a
Oxa1	+	-
Oxa2	+	-
Oxa3	-	-
Oxa4	-	-
Oxa5	+	-
Oxa6	-	-
Oxa7	+	-
Oxa8	-	-

+ indicates 6-APA, a product of hydrolysis reaction of b-lactam antibiotics was detected

- indicates that no 6-APA was detected

6.4 Discussion

6.4.1 The PGA gene

We successfully expressed, purified and characterized PGA. We obtained a specific activity of 15 to 20 μmol/(mg min) for pEC protein on penicillin G within the expected range. The pEC protein does not have detectable activity towards cloxacillin, naficillin and oxacillin in agreement with the absence of reported activity towards these substrates. This is likely due to the presence of a large side-chain group which prevents the antibiotic substrate from binding the PGA enzyme, thereby preventing catalysis altogether. Literature values for the activity of PGA were within one order of magnitude penicillin to our results (26,32).

6.4.2 Library diversity generated by overlap extension PCR

Sequencing of four randomly picked variants revealed that a diverse library had been generated. Only four sequences were obtained to characterize the library due to sequencing quality issues and moreover, it is not immediately obvious, how many sequences were really required to characterize a library with a theoretical size of 8000 variants. We introduced NNK mutations (N represents DNA code: A/T/C/G and K represents DNA code: T/G) at three positions in PGA. From Table 6.1, it is evident that for β24 and α145, all four sequences had a different amino acid. For α146, although two out of four amino acids were the same, at the DNA level, their codons were different. As a combined effect of the above, there were also no similar amino acids combinations for each of the sequences. This implies that a non-biased library was generated.

6.4.3 PDAB filter lift assay

As is evident from Table 6.2, the human eye can only see a yellow coloration formed by the reaction of a minimum of 0.47 μM of 6-APA with PDAB substrate, suggesting that the variant needs at least 1% conversion (47μM 6-APA detected over 56mM antibiotics substrate used for assay) to be screened for. Figure 6.13 indicates that the PDAB assay can detect down to nM ranges of 6-APA using a spectrophotometer. Rio *et al.* used a similar assay to screen for improved variants of PGA that is more stable at pH 8.5. Out of screening 1500 variants, they obtained two variants that were more stable than wild-type PGA at a more basic pH (24). One of the variants had a twofold increase in the half-life of the mutant enzyme at pH 8.5 as compared to wild-type PGA. To the best of our knowledge, no other group has yet evaluated the PDAB filter lift assay sensitivity.

6.4.4 Results of oxacillin screen

No improved variants were obtained after screening 30,000 variants of PGA. We expect that the library was thoroughly screened for improved variants as a three-fold screen was done on a library of 20^3 ~ 8000 variants. HPLC secondary assay on the screened variants showed no activity towards oxacillin. This suggests that we could have obtained false positives. Two reasons that no improved variants were obtained could be that the directed library is not optimal for substrate specificity change and PGA could potentially reach its substrate evolutionary limit or that local rearrangements are not going to create enough room to accommodate the bulky substrates. Unfortunately, though, that would mean that a combinatorial approach would be necessary because substantial repacking to reposition the backbone of the protein to reduce or remove steric hindrance to the large substrates (oxacillin, cloxacillin, naficillin) would be needed. While the latter is impossible to test for stringently, the former could be explored by the creation of another library involving different residues. The key here is selecting the right residues for evolving the desired improvement. Possible techniques to select additional residues for mutagenesis could include protein docking, data-driven techniques such as CASTING (33-35), ProSar (36) and application of techniques such as BLSVM (37,38) that is described in more detail in Chapter five of this thesis.

6.5 Conclusion

Protein engineering work was done on PGA to change its substrate specificity by structure-guided approach. The PGA gene from *E. coli* and *K. cryocrescens* were expressed, purified and characterized. Based on crystal structure analysis, we targeted three residues β24, α145 and α146 with site-saturation mutagenesis using NNK randomization. After confirming the library to be diverse through DNA sequencing, we screened the variants using PDAB assay. No improved variants were obtained for the oxacillin library suggesting that that other approaches involving choosing and selecting different residues for mutagenesis and the use of intermediate substrates for evolution might be a better strategy. Additional work done using protein docking methods and the application of structure-guided methods such as CASTing, possibly in combination with NDT libraries, and/or data-driven methods such as ProSar or BLSVM to select potential residues for mutations could be useful to evolve PGA variants with altered substrate specificity. Alternatively, other enzymes such as α-amino-ester hydrolases (AEH) which catalyzes the hydrolysis and synthesis of esters and amides of a-amino acids, can be used as a biocatalyst for synthesizing oxacillin instead of PGA as the substrate pocket of AEH is bigger than PGA (22, 39-41).

6.6 References

1. McGowan, J.E.J. (2001) Economic Impact of Antimicrobial Resistance. *Emerging Infectious Diseases*, **7**, 286-292.

2. Cohen, M.L. (1992) Epidemiology of drug resistance: implications for a post-antimicrobial era. *Science (New York, N.Y*, **257**, 1050-1055.

3. Ulijn, R.V., De Martin, L., Halling, P.J., Moore, B.D. and Janssen, A.E. (2002) Enzymatic synthesis of beta-lactam antibiotics via direct condensation. *J Biotechnol*, **99**, 215-222.

4. Goncalves, L.R.B., Sousa, R., Fernandez-Lafuente, R., Guisan, J.M., Giordano, R.L.C. and Giordano, R.C. (2002) Enzymatic synthesis of amoxicillin - Avoiding limitations of the mechanistic approach for reaction kinetics. *Biotechnology and Bioengineering*, **80**, 622-631.

5. Chou, C.P., Yu, C.C., Lin, W.J., Kuo, B.Y. and Wang, W.C. (1999) Novel strategy for efficient screening and construction of host/vector systems to overproduce penicillin acylase in Escherichia coli. *Biotechnol Bioeng*, **65**, 219-226.

6. Kasche, V., Lummer, K., Nurk, A., Piotraschke, E., Rieks, A., Stoeva, S. and Voelter, W. (1999) Intramolecular autoproteolysis initiates the maturation of penicillin amidase from Escherichia coli. *Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology*, **1433**, 76-86.

7. Dai, M., Zhu, Y., Yang, Y., Wang, E., Xie, Y., Zhao, G. and Jiang, W. (2001) Expression of penicillin G acylase from the cloned pac gene of Escherichia coli ATCC11105. Effects of pacR and temperature. *Eur J Biochem*, **268**, 1298-1303.

8. Sriubolmas, N., Panbangred, W., Sriurairatana, S. and Meevootisom, V. (1997) Localization and characterization of inclusion bodies in recombinant Escherichia coli cells overproducing penicillin G acylase. *Appl Microbiol Biotechnol*, **47**, 373-378.

9. Lin, Y.H., Hsiao, H.C. and Chou, C.P. (2002) Strain improvement to enhance the production of recombinant penicillin acylase in high-cell-density Escherichia coli cultures. *Biotechnol Prog*, **18**, 1458-1461.

10. Lin, W.J., Huang, S.W. and Chou, C.P. (2001) High-level extracellular production of penicillin acylase by genetic engineering of Escherichia coli. *Journal of Chemical Technology and Biotechnology*, **76**, 1030-1037.

11. McVey, C.E., Walsh, M.A., Dodson, G.G., Wilson, K.S. and Brannigan, J.A. (2001) Crystal structures of penicillin acylase enzyme-substrate complexes: structural insights into the catalytic mechanism. *J Mol Biol*, **313**, 139-150.

12. Alkema, W.B., Hensgens, C.M., Kroezinga, E.H., de Vries, E., Floris, R., van der Laan, J.M., Dijkstra, B.W. and Janssen, D.B. (2000) Characterization of the beta-lactam binding site of penicillin acylase of Escherichia coli by structural and site-directed mutagenesis studies. *Protein Eng*, **13**, 857-863.

13. Alkema, W.B., Dijkhuis, A.J., De Vries, E. and Janssen, D.B. (2002) The role of hydrophobic active-site residues in substrate specificity and acyl transfer activity of penicillin acylase. *Eur J Biochem*, **269**, 2093-2100.

14. Alkema, W.B.L., Prins, A.K., de Vries, E. and Janssen, D.B. (2002) Role of alpha Arg(145) and beta Arg(263) in the active site of penicillin acylase of Escherichia coli. *Biochemical Journal*, **365**, 303-309.

15. Duggleby, H.J., Tolley, S.P., Hill, C.P., Dodson, E.J., Dodson, G. and Moody, P.C. (1995) Penicillin acylase has a single-amino-acid catalytic centre. *Nature*, **373**, 264-268.

16. Arroyo, M., de la Mata, I., Acebal, C. and Castillon, M.P. (2003) Biotechnological applications of penicillin acylases: state-of-the-art. *Appl Microbiol Biotechnol*, **60**, 507-514.

17. Margreth A. Wegman, M.H.A.J., Fred van Rantwijk, Roger A. Sheldon. (2001) Towards Biocatalytic Synthesis of b-lactam Antibiotics. *Adv. Synth. Catal.*, **343**, 6-7.

18. Zhou, Z., Zhang, A.-H., Wang, J.-R., Chen, M.-L., Li, R.-B., Yang, S. and Yuan, Z.-Y. (2003) Improving the Specific Synthetic Activity of a Pencillin G Acylase

Using DNA Family Shuffling. *ACTA BIOCHIMICA et BIOPHYSICA SINICA*, **35**, 573-579.

19. Jager, S.A.W., Jekel, P.A. and Janssen, D.B. (2007) Hybrid penicillin acylases with improved properties for synthesis of beta-lactam antibiotics. *Enzyme and Microbial Technology*, **40**, 1335-1344.

20. Polizzi, K.M., Chaparro-Riggers, J.F., Vazquez-Figueroa, E. and Bommarius, A.S. (2006) Structure-guided consensus approach to create a more thermostable penicillin G acylase. *Biotechnology Journal*, **1**, 531-536.

21. Galunsky, B., Lummer, K. and Kasche, V. (2000) Comparative study of substrate- and stereospecificity of penicillin G amidases from different sources and hybrid isoenzymes. *Monatshefte Fur Chemie*, **131**, 623-632.

22. Bruggink, A. (2001) *Synthesis of beta-lactam antibiotics*. 1 ed. Kluwer Academic Publishers, Netherlands.

23. Morley, K.L. and Kazlauskas, R.J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol*, **23**, 231-237.

24. del Rio, G., Rodriguez, M.-E., Munguia, M.-E., Lopez-Munguia, A. and Soberon, X. (1995) Mutant Escherichia coli penicillin acylase with enhanced stability at alkaline pH. *Biotechnology and Bioengineering*, **48**, 141-148.

25. Alkema, W.B.L., Floris, R. and Janssen, D.B. (1999) The use of chromogenic reference substrates for the kinetic analysis of penicillin acylases. *Analytical Biochemistry*, **275**, 47-53.

26. Simons, H. and Gibson, T.D. (1999) Rapid continuous colorimetric enzyme assay for penicillin G acylase. *Biotechnology Techniques*, **13**, 365-367.

27. Sambrook, J. and Russell, D.W. (2001) *Molecular cloning : a laboratory manual*. 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

28. Shewale, J.G., Kumar, K.K. and Ambekar, G.R. (1987) Evaluation of determination of 6-aminopenicillanic acid by p-dimethyl aminobenzaldedyde. *Biotechnology Techniques*, **1**, 69-72.

29. Arroyo, M., Torres-Guzman, R., De La Mata, I., Castillon, M.P. and Acebal, C. (2002) A kinetic examination of penicillin acylase stability in water-organic solvent systems at different temperatures. *Biocatalysis and Biotransformation*, **20**, 53-56.

30. Oh, B., Kim, K., Park, J., Yoon, J., Han, D. and Kim, Y. (2004) Modifying the substrate specificity of penicillin G acylase to cephalosporin acylase by mutating active-site residues. *Biochemical and Biophysical Research Communications*, **319**, 486-492.

31. Balasingham, K., Warburton, D., Dunnill, P. and Lilly, M.D. (1972) Isolation and kinetics of penicillin amidase from Escherichia coli. *Biochimica et Biophysica Acta, Enzymology*, **276**, 250-256.

32. Alkema, W.B.L., de Vries, E., Floris, R. and Janssen, D.B. (2003) Kinetics of enzyme acylation and deacylation in the penicillin acylase-catalyzed synthesis of beta-lactam antibiotics. *European Journal of Biochemistry*, **270**, 3675-3683.

33. Reetz, M.T., Wang, L.W. and Bocola, M. (2006) Directed evolution of enantioselective enzymes: iterative cycles of CASTing for probing protein-sequence space. *Angew Chem Int Ed Engl*, **45**, 1236-1241.

34. Reetz, M.T., Carballeira, J.D., Peyralans, J., Hobenreich, H., Maichele, A. and Vogel, A. (2006) Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. *Chemistry (Weinheim an der Bergstrasse, Germany)*, **12**, 6031-6038.

35. Clouthier, C.M., Kayser, M.M. and Reetz, M.T. (2006) Designing new Baeyer-Villiger monooxygenases using restricted CASTing. *The Journal of organic chemistry*, **71**, 8431-8437.

36. Fox, R.J., Davis, S.C., Mundorff, E.C., Newman, L.M., Gavrilovic, V., Ma, S.K., Chung, L.M., Ching, C., Tam, S., Muley, S. *et al.* (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol*, **25**, 338-344.

37. Dubey, A., Realff, M.J., Lee, J.H. and Bommarius, A.S. (2006) Identifying the interacting positions of a protein using boolean learning and support vector machines. *Computational Biology and Chemistry*, **30**, 268-279.

38. Dubey, A., Realff, M.J., Lee, J.H. and Bommarius, A.S. (2005) Support vector machines for learning to identify the critical positions of a protein. *J Theor Biol*, **234**, 351-361.

39. Barends, T.R.M., Polderman-Tijmes, J. J., Jekel, P.A., Williams, C., Wybenga, G., Janssen, D.B. and Dijkstra, B. W. (2003). Acetobacterturbidans alpha-amino-acid ester hydrolase: how a single mutation improves an antibiotic-producing enzyme. *J Biol Chem*, **281**, 5804-5810.

40 Polderman-Tijmes, J. J., Jekel, P. A., De Vries, E. J., Van Merode, A. E. J., Floris, R., Van Der Laan, J. M., Sonke, T. and Janssen, D. B. (2002). *Appl Environ Microbiol*, **68**, 211-218.

41 Takahashi, T., Kato, K., Yamazaki, Y., and Kato, K. (1974). *Biochem J*, **137**, 497-503.

6.7 Supplement

6.7.1 Primer sequences of degenerate sequences

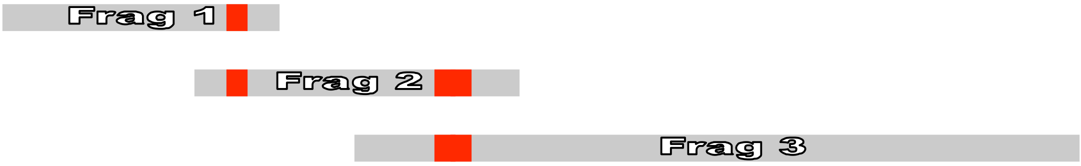


Figure 6. 16 Fragments of PGA used to assemble PGA library. The figures are not drawn to scale. Red boxes represent relative degenerate regions.

<u>Primers and sequences</u>

oxa145146r2_blw

CGATTTCGCTAGTGCTATCAGAMNNMNNGTTTGCCATGGTGCCCACAAATAT

CATCG

oxab24r2_blw

CCATAAGTATACGCAGGCGCATACCAGCCMNNCTGCGGACCATTTACCATGA

TTGC

pga24for GGCTGGTATGCGCCTGCGTATACTTATGGTATTGG

pga145-6for TCTGATAGCACTAGCGAAATCGATAATCTG

PECfor68

AGGAAAAACATATGAAAAATAGAAATCGTATGATCGTGAACTGTGTTACTGC

PECrev68 AGGCCTGCAATTTCTTTTCCAACTGTTC

PCR fragments

Fragment 1: PECfor68 + oxab24r2blw + pEC plasmid ~400bp

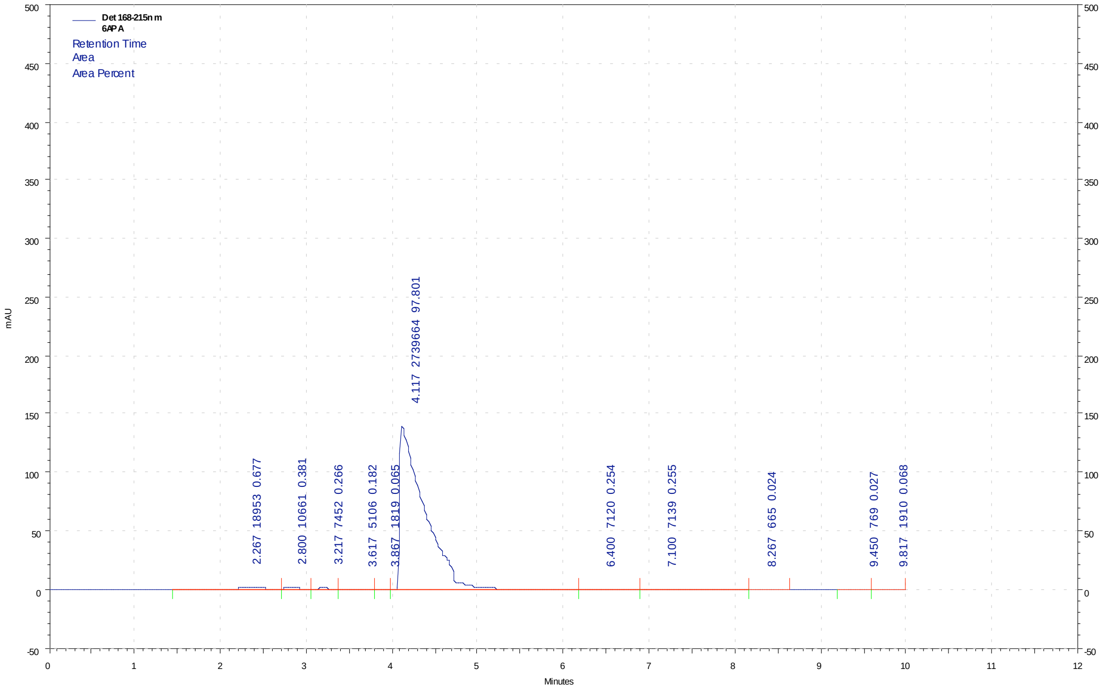
Fragment 2: pga145-6for + oxa145146r2_blw + pEC plasmid ~400bp

Fragment 3: pga24for + PECrev68 + pEC plasmid ~1400bp

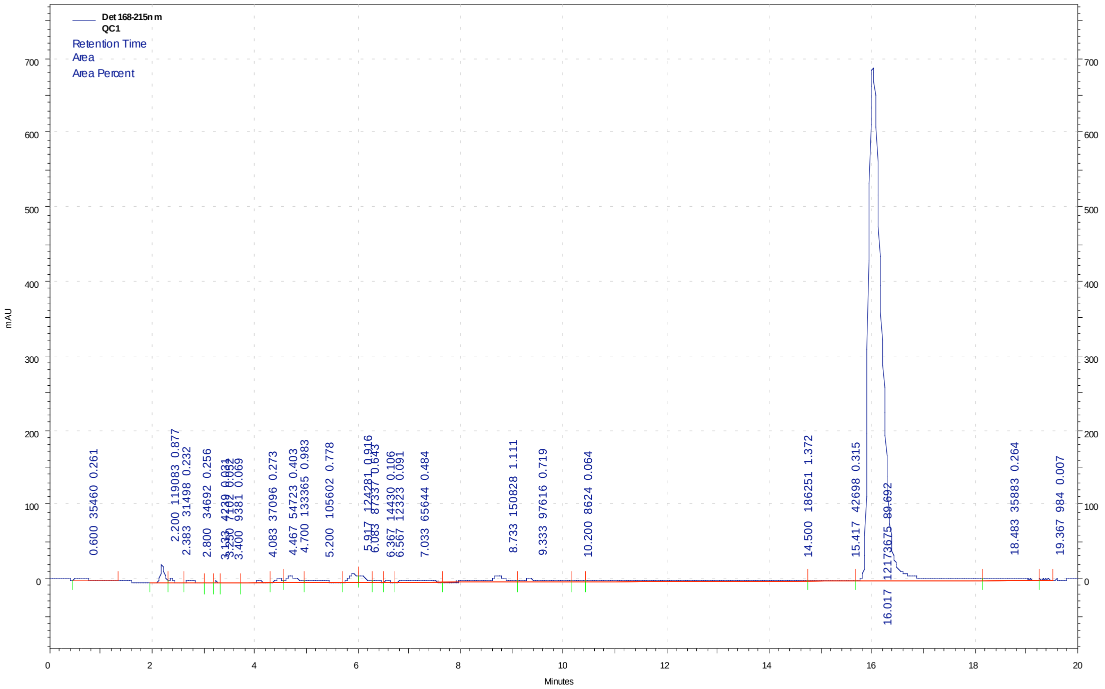
All fragments must be gel purified to remove pEC parental plasmid. To generate full length PEC library insert, use combine Fragment 1 – 3 in molar ratio. PCR to reassemble then secondary PCR using PECfor68 and PECrev68.

6.7.2 HPLC data on screened oxacillin library

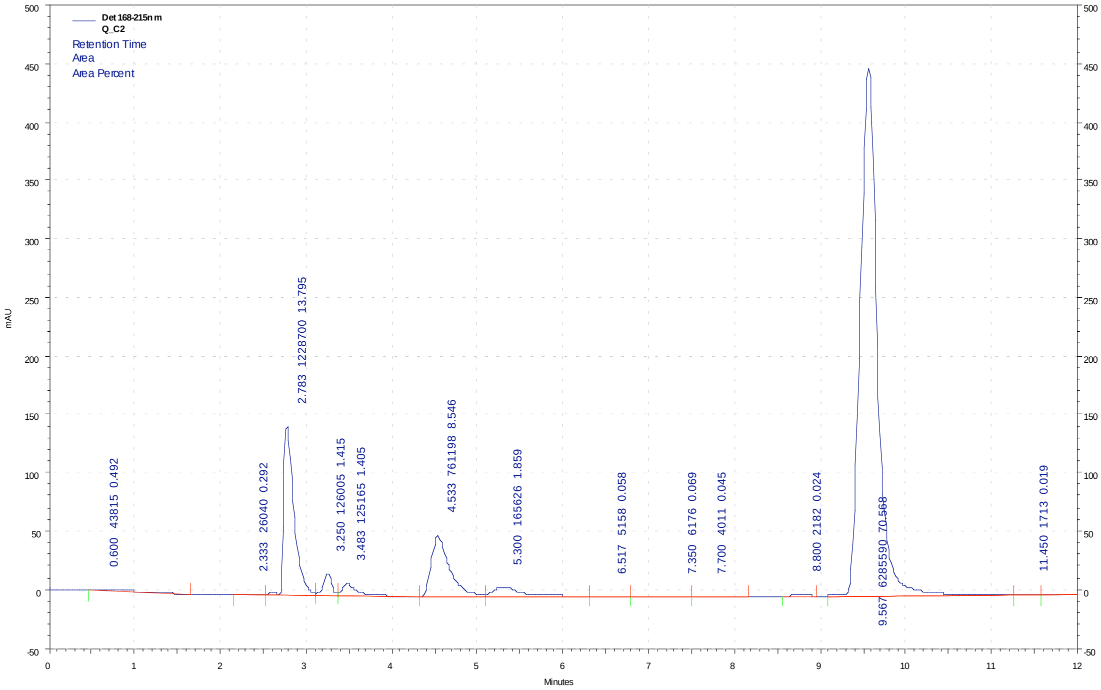
6APA



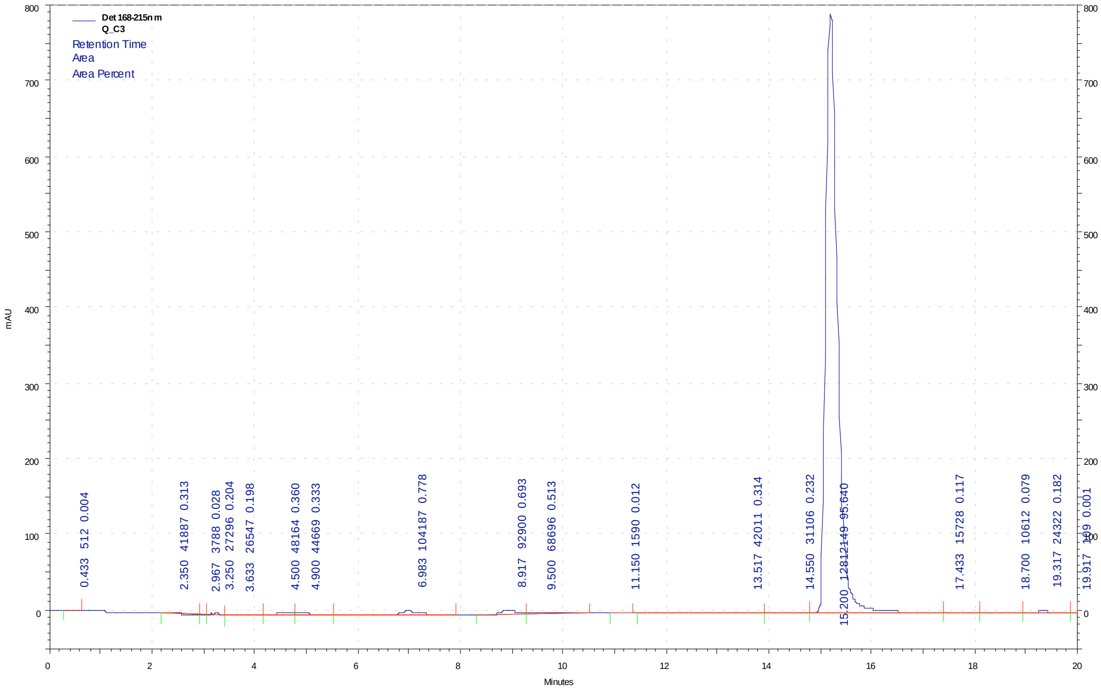
C1: PBS buffer + Oxa + PGA



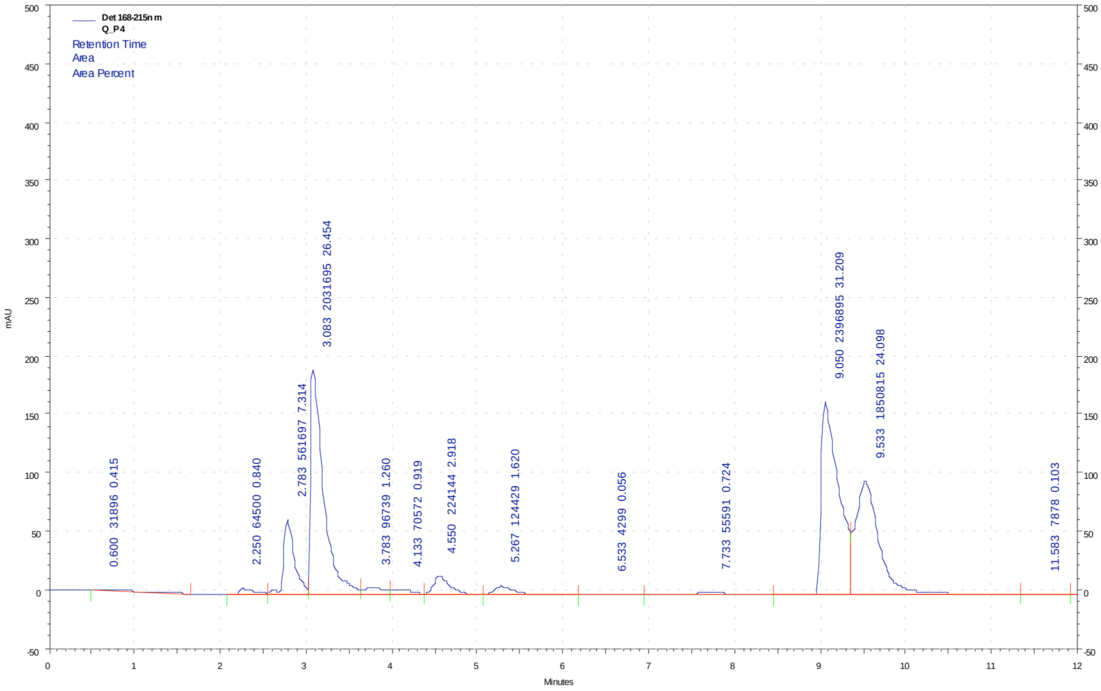
C2: PBS Buffer + PenG+OS Buffer



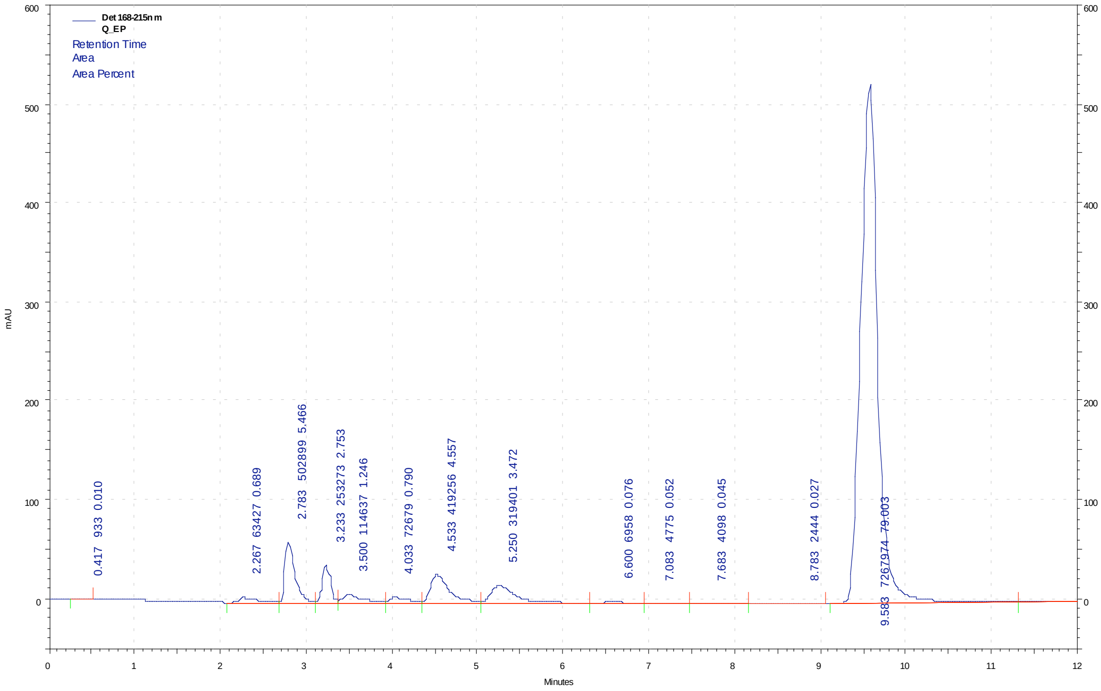
C3: PBS Buffer + Oxa+OS Buffer



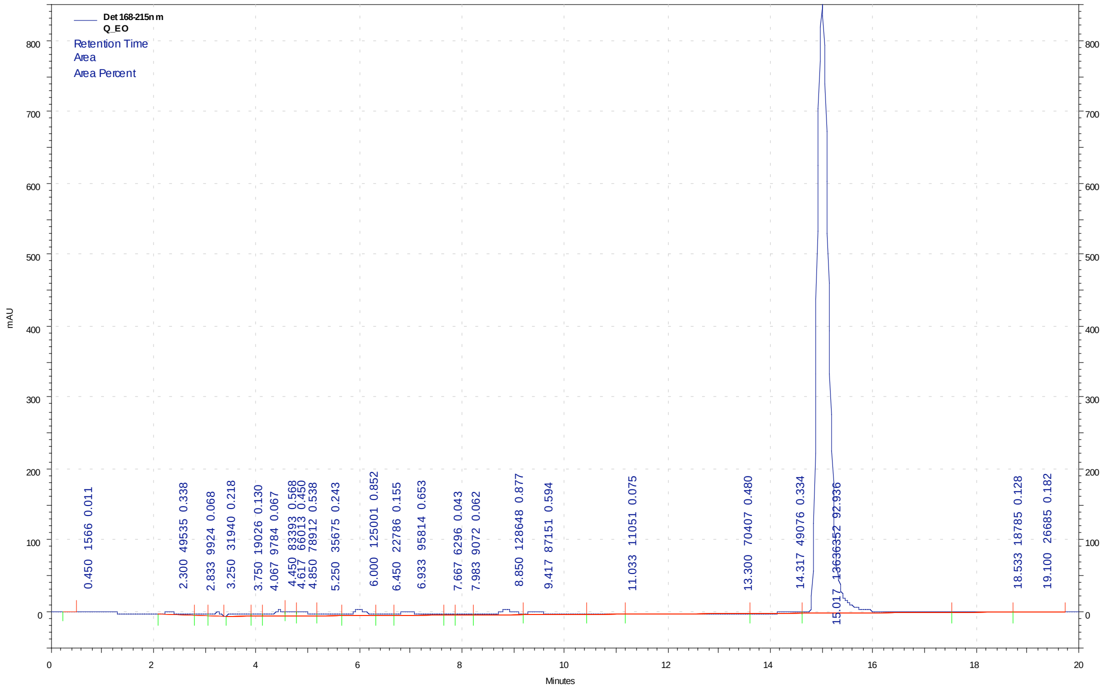
P4: PBS Buffer + PenG+PGA_WT



EP – Empty Vector on Penicillin G

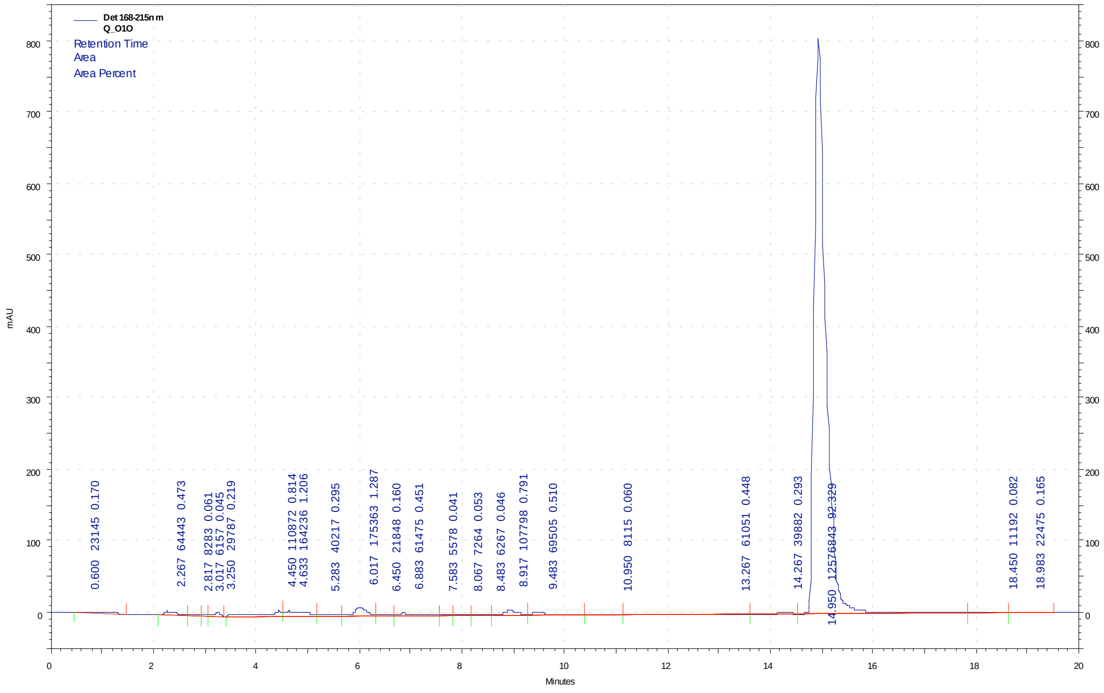


EO_ Empty Vector on Oxacillin

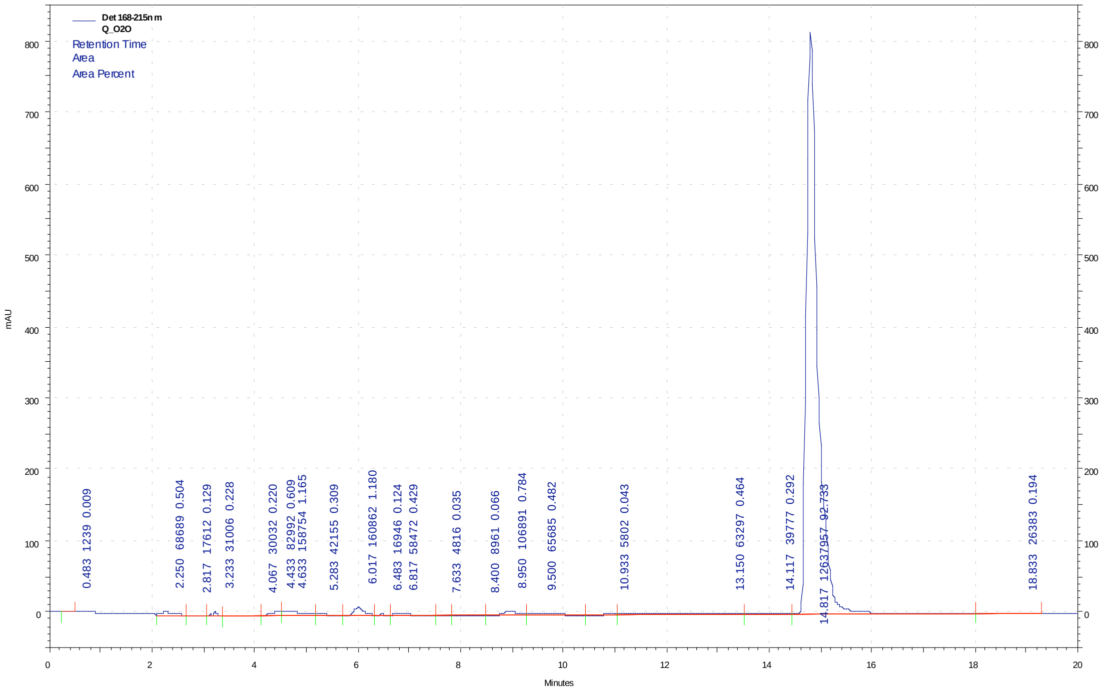


6.7.2.1 Oxacillin variants tested on oxacillin substrates

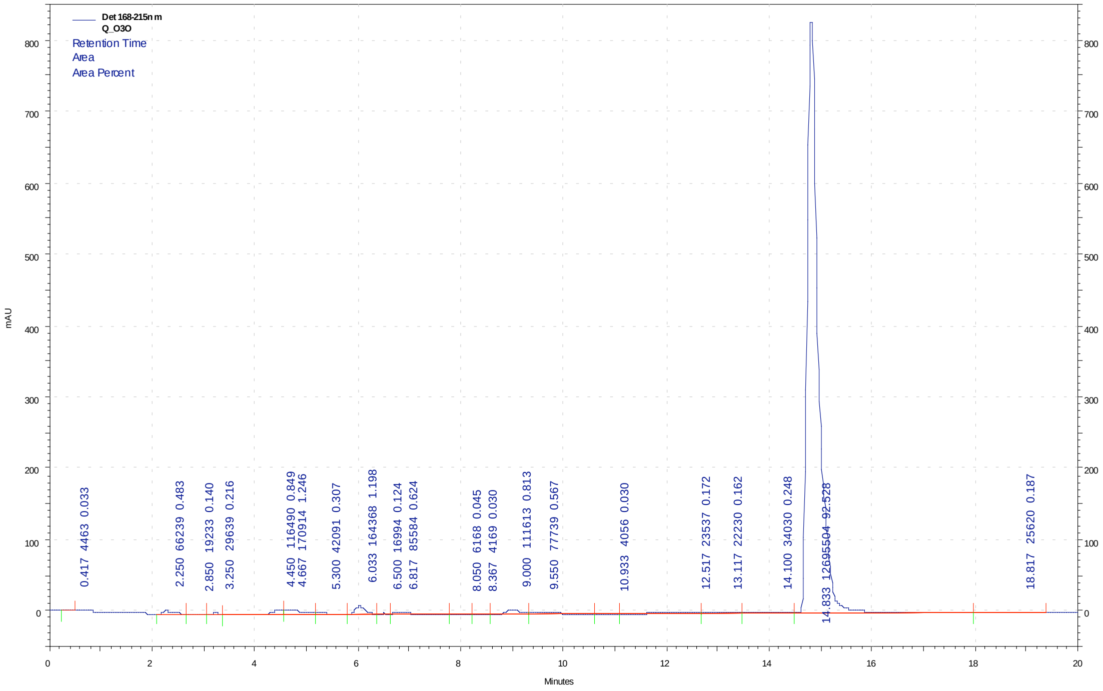
O1O – Oxacillin Mutant 1 on Oxacillin



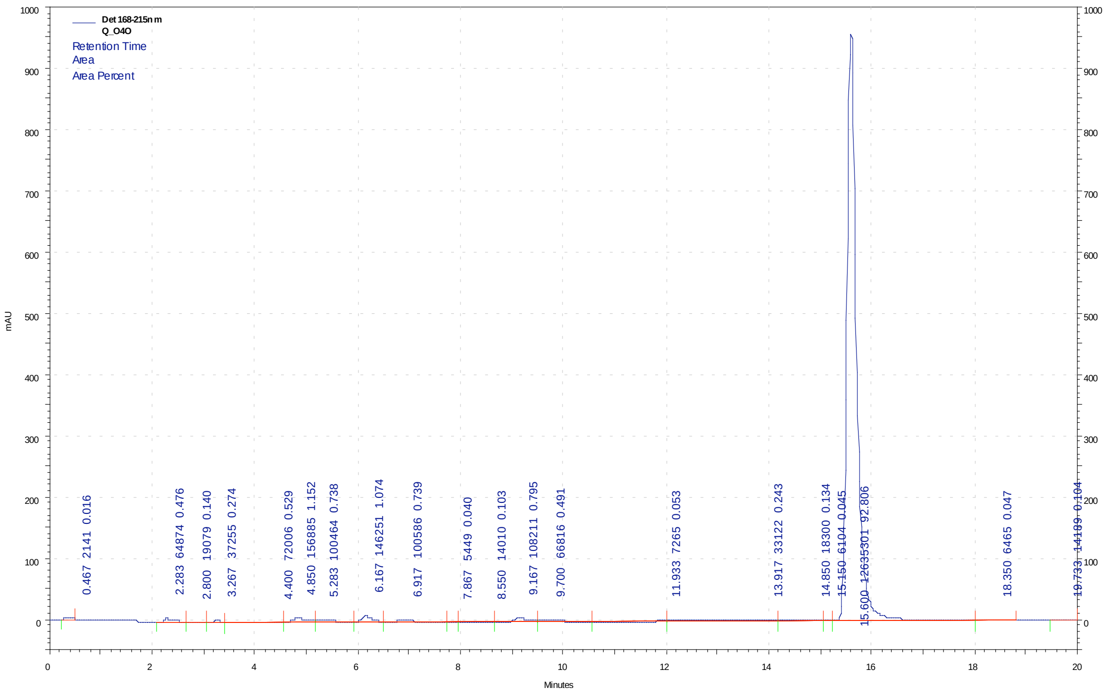
O2O_ Oxacillin Mutant 2 on Oxacillin



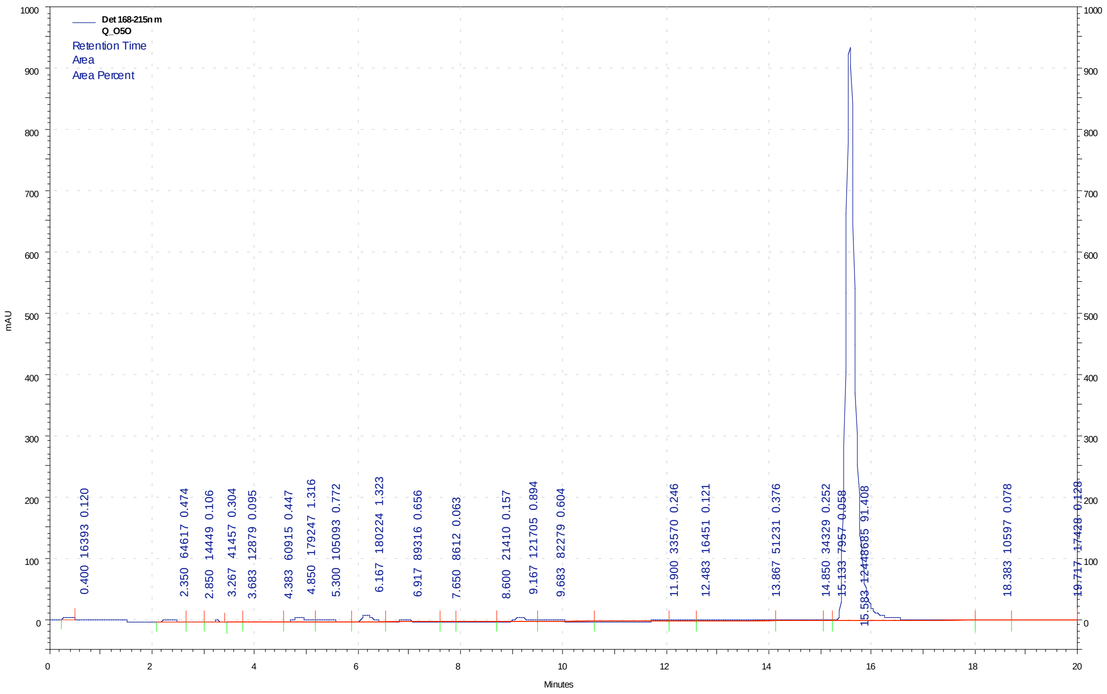
O3O_ Oxacillin Mutant 3 on Oxacillin



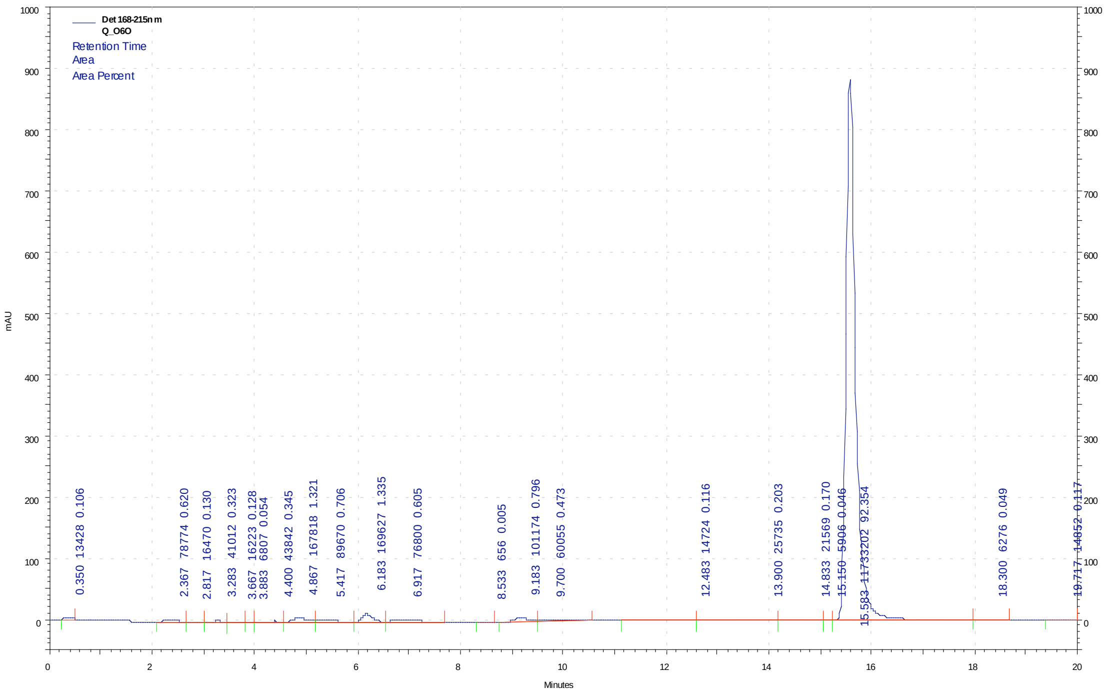
O4O_ Oxacillin Mutant 4 on Oxacillin



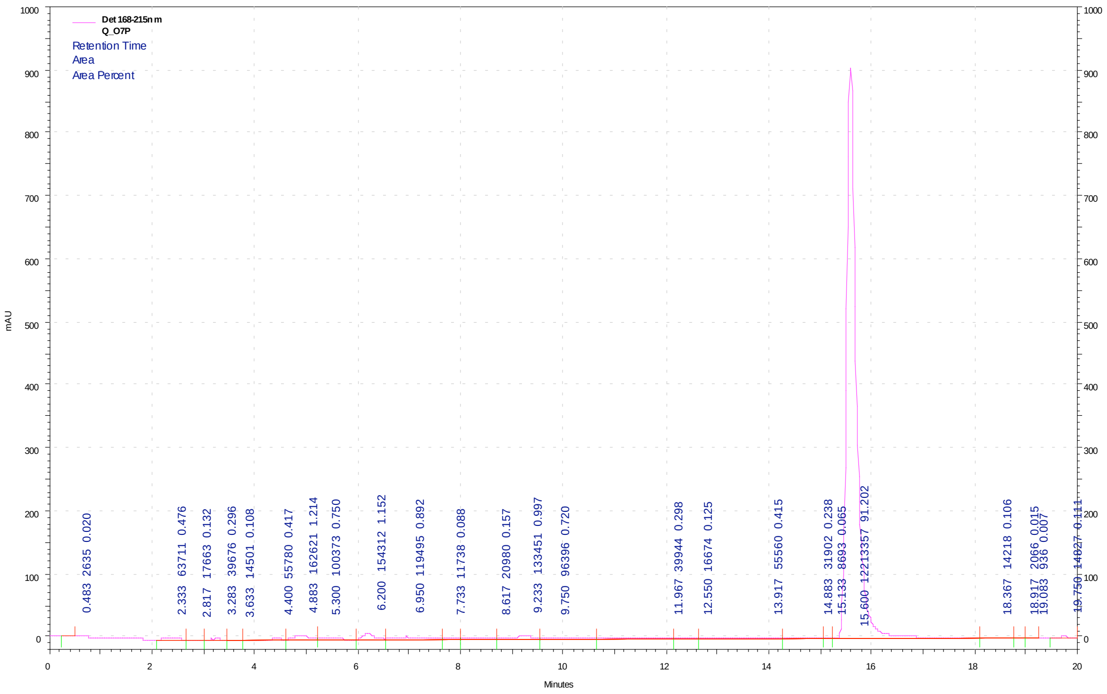
O5O



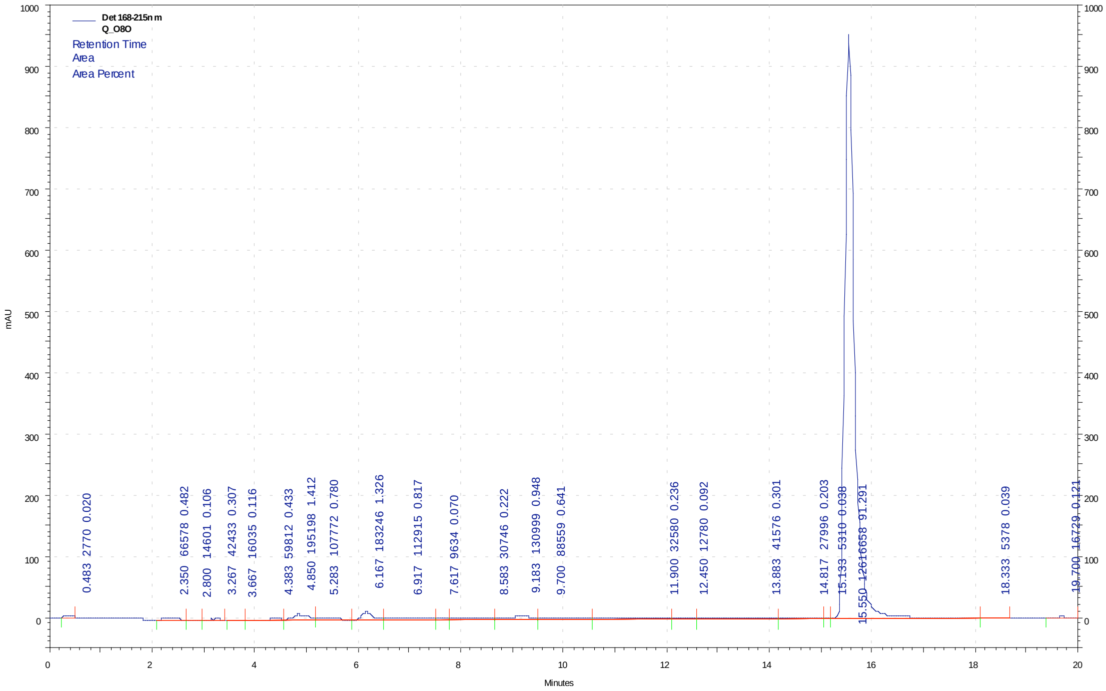
O6O



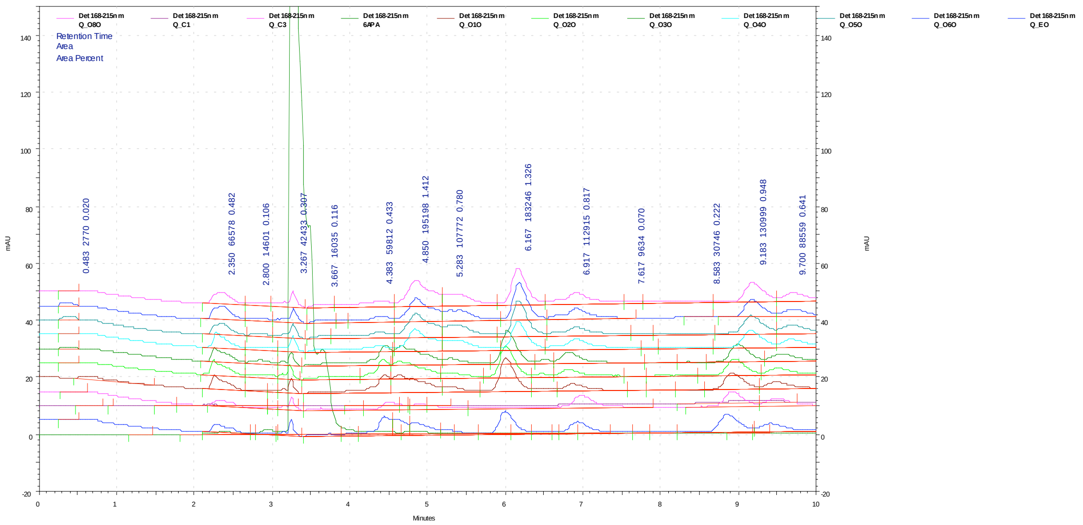
O7O (labeled as O7P)



O8O

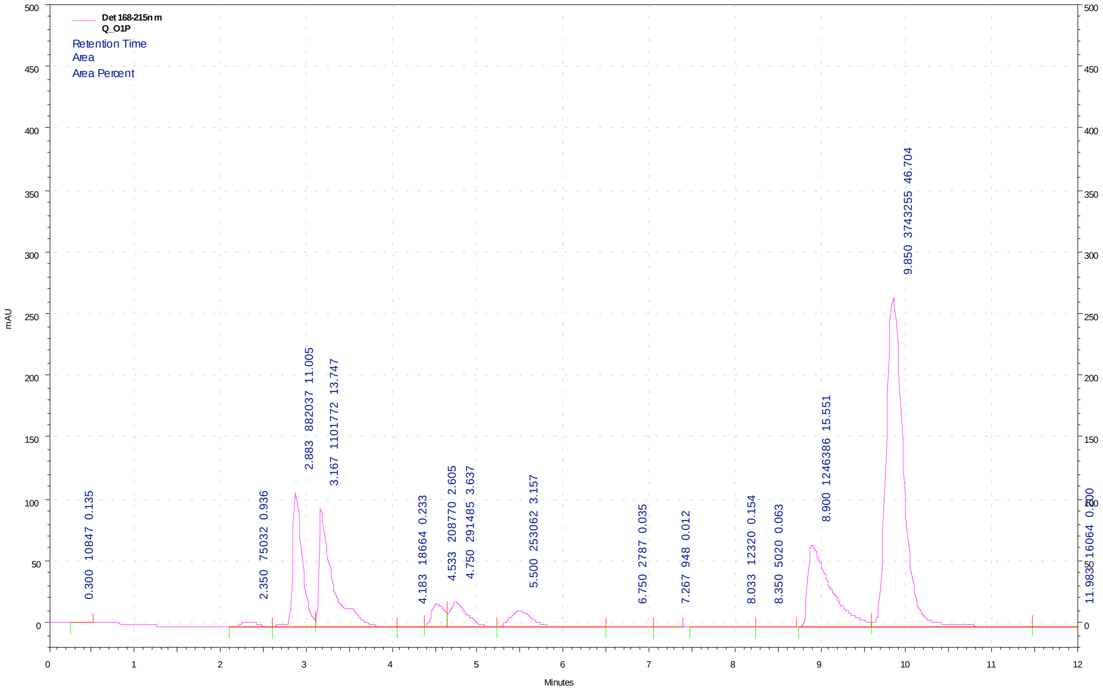

Det 168-215nm
Q_O8O
Retention Time
Area
Area Percent
mAU
Minutes
0
2
4
6
8
10
12
14
16
18
20
0
100
200
300
400
500
600
700
800
900
1000
0.483 2770 0.020
2.350 66578 0.482
2.800 14601 0.106
3.267 42433 0.307
3.667 16035 0.116
4.383 59812 0.433
4.850 195198 1.412
5.283 107772 0.780
6.167 183246 1.326
6.917 112915 0.817
7.617 9634 0.070
8.583 30746 0.222
9.183 130999 0.948
9.700 88559 0.641
11.900 32580 0.236
12.450 12780 0.092
13.883 41576 0.301
14.817 27996 0.203
15.133 5310 0.038
15.550 12616658 91.291
18.333 5378 0.039
19.700 16729 0.121

Overlay of small peaks for Oxa mutants on Oxacillin

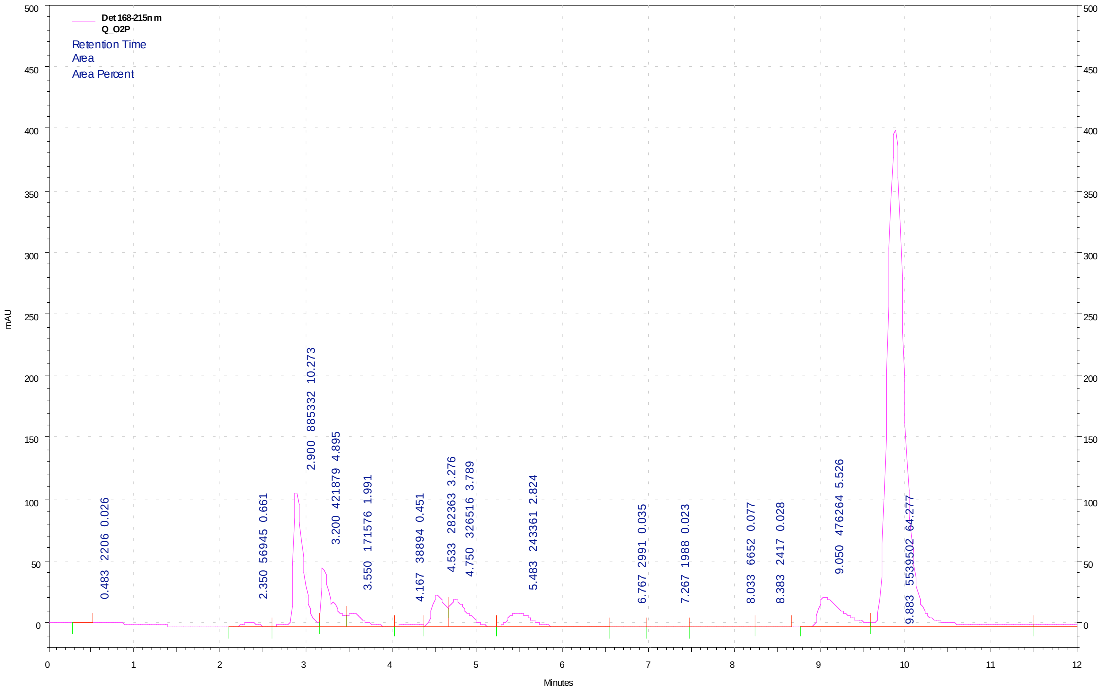


6.7.2.2 Oxacillin library assayed on penicillin G substrates

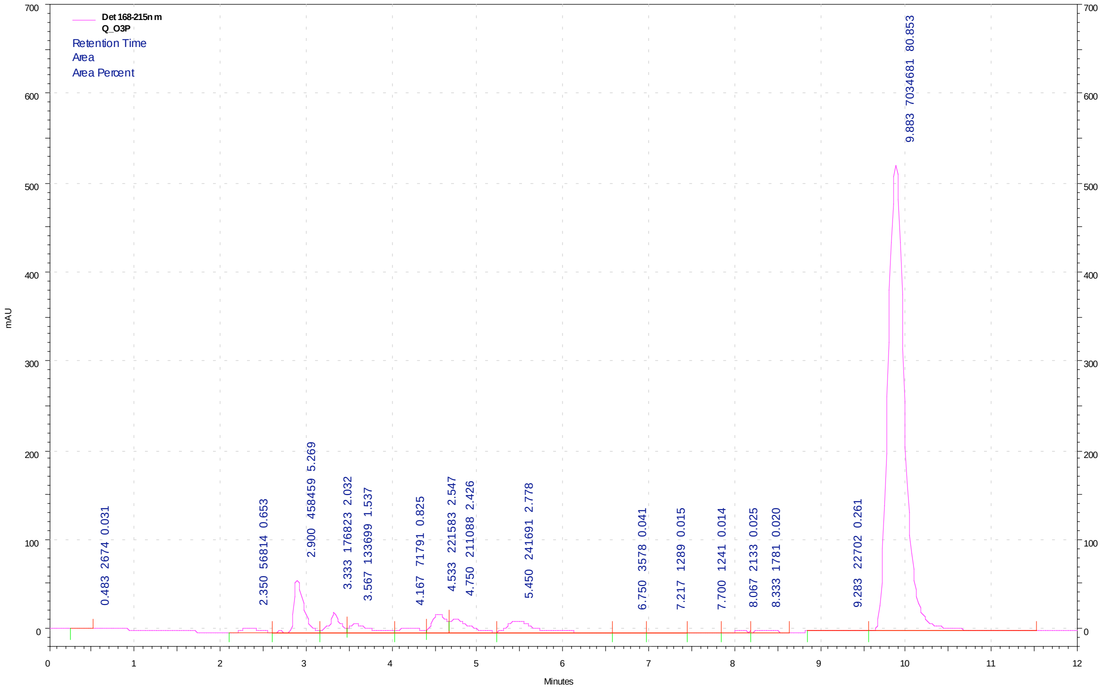
O1P Oxacillin Mutant 1 on Pencillin G



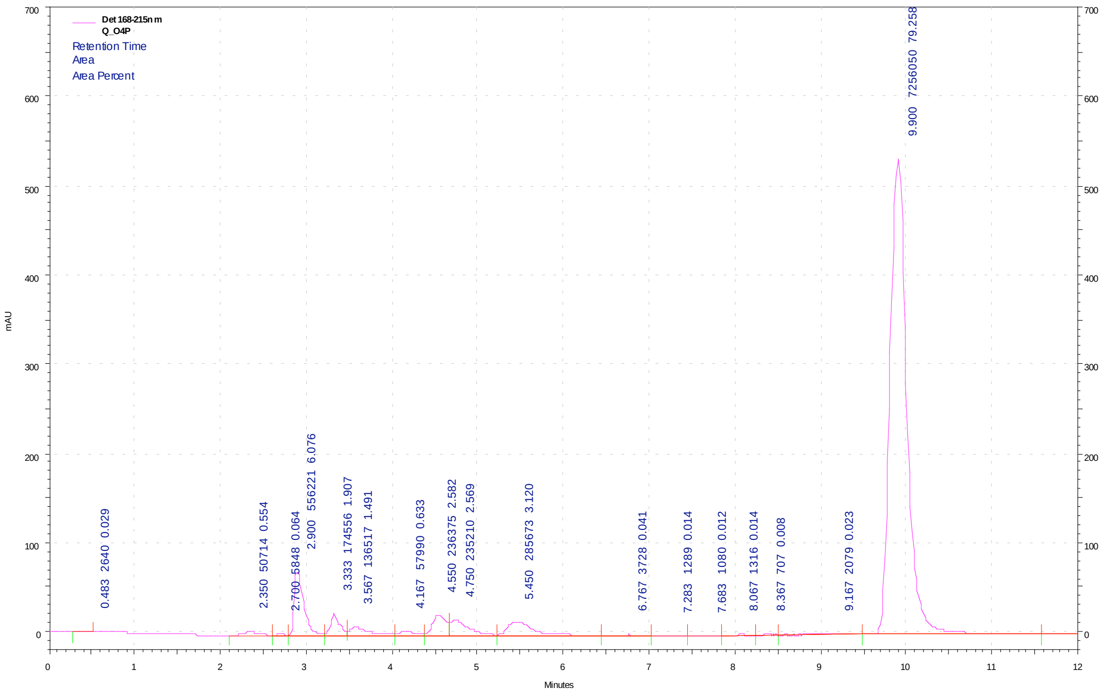
O2P


500
450
400
350
300
250
200
150
100
50
0
mAU
Det 168-215nm
Q_O2P
Retention Time
Area
Area Percent
0.483 2206 0.026
2.350 56945 0.661
2.900 885332 10.273
3.200 421879 4.895
3.550 171576 1.991
4.167 38894 0.451
4.533 282363 3.276
4.750 326516 3.789
5.483 243361 2.824
6.767 2991 0.035
7.267 1988 0.023
8.033 6652 0.077
8.383 2417 0.028
9.050 476264 5.526
9.883 5539502 64.277
0
1
2
3
4
5
6
7
8
9
10
11
12
Minutes

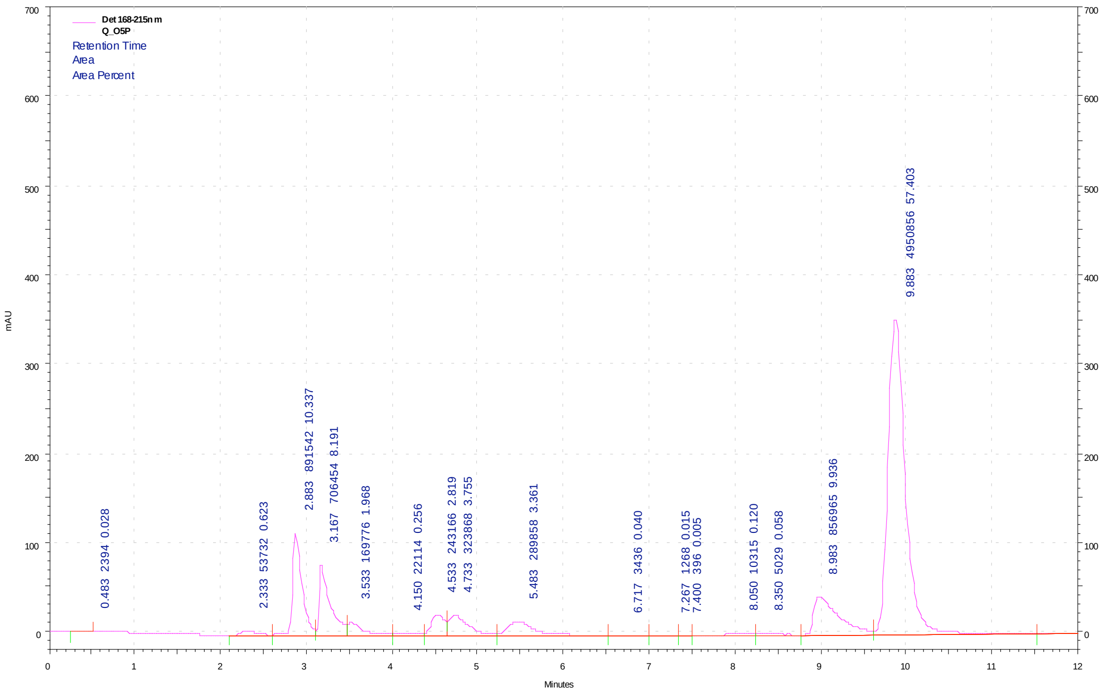
O3P


700
600
500
400
300
200
100
0
mAU
Det 168-215nm
Q_O3P
Retention Time
Area
Area Percent
0.483 2674 0.031
2.350 56814 0.653
2.900 458459 5.269
3.333 176823 2.032
3.567 133699 1.537
4.167 71791 0.825
4.533 221583 2.547
4.750 211088 2.426
5.450 241691 2.778
6.750 3578 0.041
7.217 1289 0.015
7.700 1241 0.014
8.067 2133 0.025
8.333 1781 0.020
9.283 22702 0.261
9.883 7034681 80.853
0
1
2
3
4
5
6
7
8
9
10
11
12
Minutes

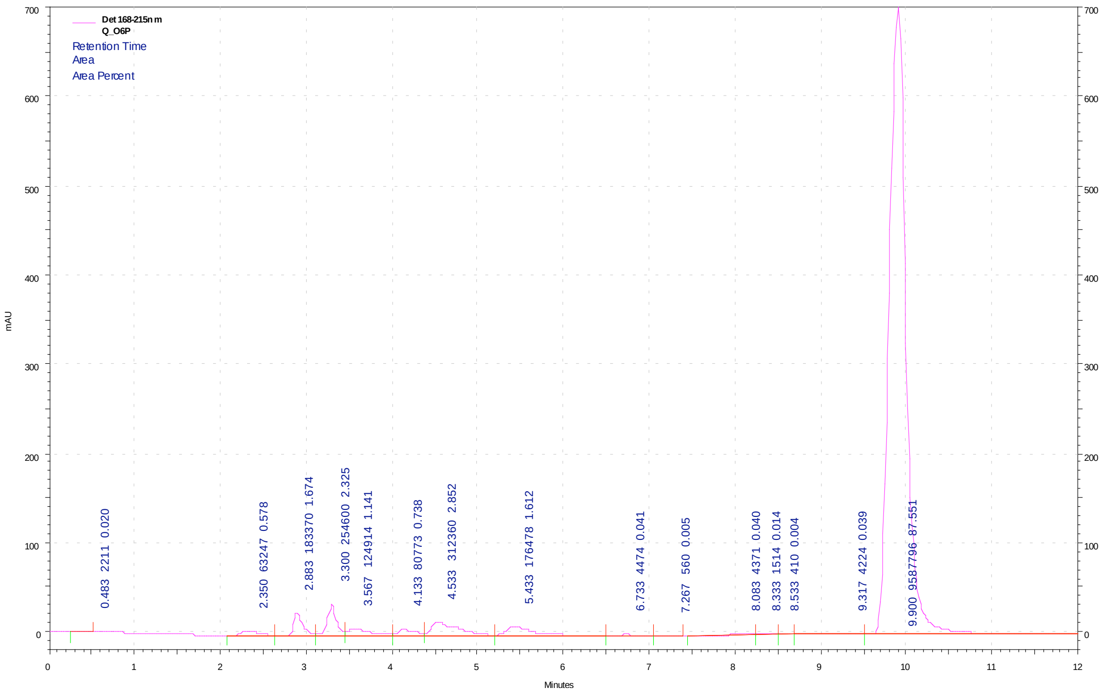
O4P


Det 168-215nm
Q_O4P
Retention Time
Area
Area Percent
mAU
Minutes
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6
7
8
9
10
11
12
0.483 2640 0.029
2.350 50714 0.554
2.700 5848 0.064
2.900 556221 6.076
3.333 174556 1.907
3.567 136517 1.491
4.167 57990 0.633
4.550 236375 2.582
4.750 235210 2.569
5.450 285673 3.120
6.767 3728 0.041
7.283 1289 0.014
7.683 1080 0.012
8.067 1316 0.014
8.367 707 0.008
9.167 2079 0.023
9.900 7256050 79.258

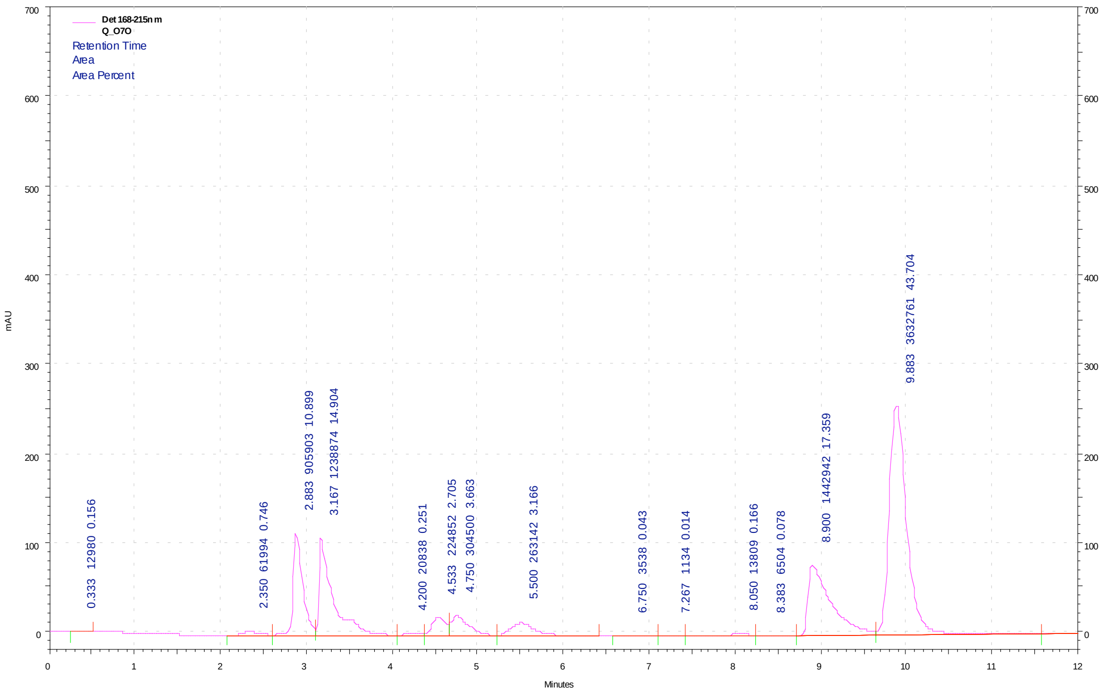
O5P


Det 168-215nm
Q_O5P
Retention Time
Area
Area Percent
mAU
Minutes
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6
7
8
9
10
11
12
0.483 2394 0.028
2.333 53732 0.623
2.883 891542 10.337
3.167 706454 8.191
3.533 169776 1.968
4.150 22114 0.256
4.533 243166 2.819
4.733 323868 3.755
5.483 289858 3.361
6.717 3436 0.040
7.267 1268 0.015
7.400 396 0.005
8.050 10315 0.120
8.350 5029 0.058
8.983 856965 9.936
9.883 4950856 57.403

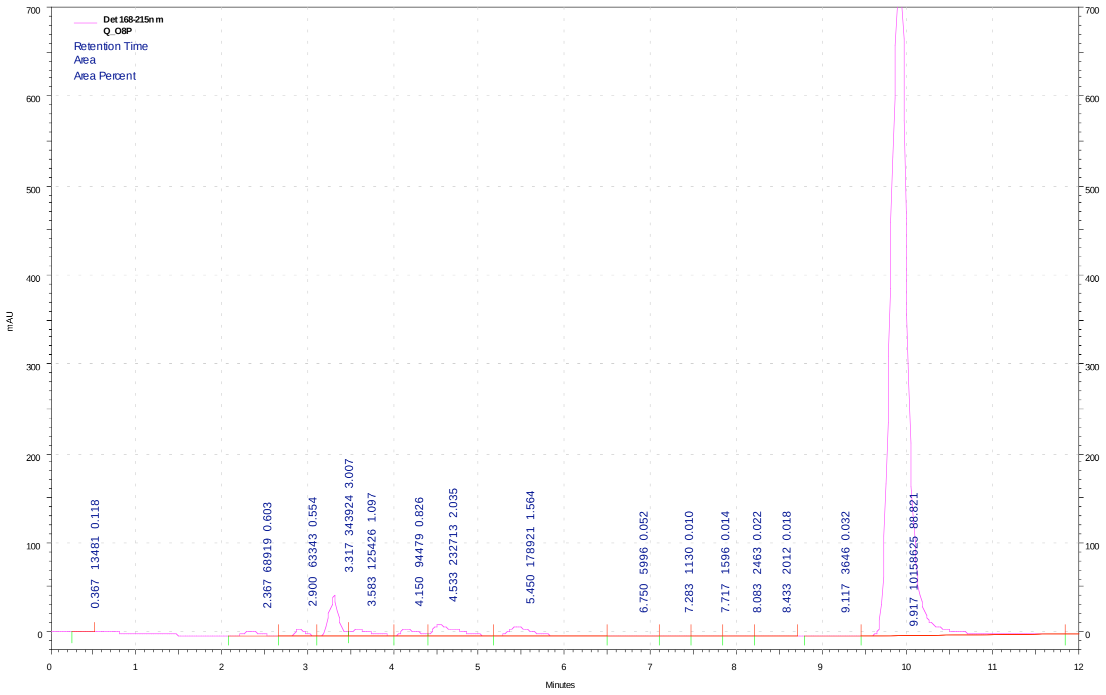
O6P



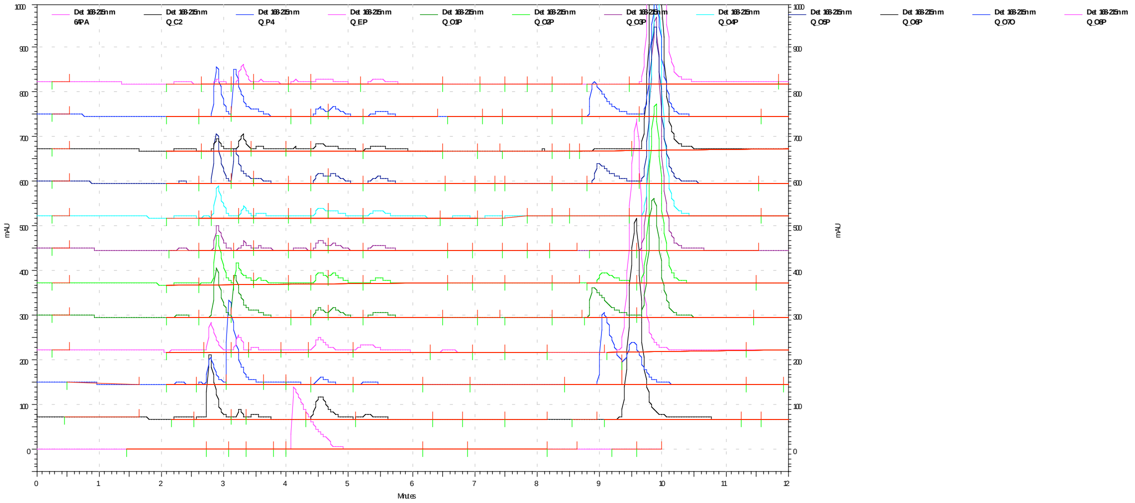
O7P (Labeled O7O)



O8P



Stacked Pen G on Penicillin G



6.8 Appendix

Residue	Function	Source
Arg B263	Invariant residue. Mutation leads to accumulation of a protein precursor form.	(11)
Asn B241	Implicated in the stabilization of tetrahedral intermediate. Mutation inactivates enzyme but does not prevent autocatalytic processing or ability to bind to substrates. Believed to have important role in maintaining the active site geometry and in productive substrate binding. Crucial for catalysis.	(11)
Ser B1	Involved in direct nucleophilic attack on the scissile amide. Cys mutation still allows processing but makes it prone to oxidation. Involved in catalystis. Alkema believes that a water molecule bridges the nucleophilic attack of the substrate.	(11,15)
Gln B23 Ala B69	Involved in stabilization along with Asn B241 of the oxyanion hole similar to serine proteases. Gln B23 probably have Van der Waals interactions with the carbonyl oxygen of b-lactam ring. Gln B23 is usually conserved.	(11,12,15)
Arg A145	Sides chains move significantly when ligands bind. (Coil and helical conformation). Moves away from active site when binded to substrate. Side chain connected to carboxylate group of the ligand through H bond via water molecules.	(11,12)
Phe A146	Sides chains move significantly when ligands bind. (Coil and helical conformation). Moves away from active site when binded to substrate. Plays an important role in substrate binding. Linked via a calcium ion to Phe B71. Side chains bind to B-lactam. Van der Waals interactions with the thiazolidine ring of penicillin G. Residue affects synthesis capability of PenG. Active site phenylalanines.	(11-13)
Met A142 Ala A143	May affect binding due to its position in the pocket. Met 142 aromatic residue interact with PAA side chain.	(11)
Phe B71	Side chains bind to b-lactam. Van der Waals interactions with the thiazolidine ring of penicillin G. Phe B71 forms part of extended loop Asp B73, Val B75 and Asp B76 that bind to calcium ion. The calcium ion help to define the structure of binding pocket.	(11,12)

Appendix 6.8 continued		
Asp B73 Val B75 Asp B76	Binds to calcium ion along with Phe B71	(11)
Glu A152	End of loop with Phe A146 through which A chain makes its contribution to calcium coordination.	(11)
Ile 177	Non polar residue whose aromatic ring interacts with PAA side chain.	(11)
Phe B24	Aromatic ring interacts with PAA side chain. Active site pheylalanine.	(11,13)
Tyr B31	Residue which Phe A146 moves towards to open active site for binding PenG	(12)
Phe B57	Active site phenylalanines	(13)

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

The success of protein engineering efforts is heavily dependent on the quality of library or the methods to select residues for targeted mutagenesis, assuming that good robust screening assays are available. The quality of the library generated using gene recombination technologies can potentially bias the protein sequence space being probed by the screening assays, thereby, leading to a suboptimal variant. The methods used for selection of residues for targeting mutagenesis efforts can focus the search for improved variants in functional protein sequence space. Hence, it can accelerate and increase productivity of protein engineering efforts. This work mainly i) contributes to data-driven protein engineering research using Boolean Learning and Support Vector Machines (BLSVM) to identify and preserve interacting residues and engineering a library to preserve interactions and ii) adds to the knowledge of how to generate more useful libraries with better diversity. We also investigated the importance of using a good strategy for protein engineering by directed evolution experiments on monomeric Red Fluorescent Protein (mRFP) and structure-guided protein engineering efforts on penicillin G acylase (PGA).

7.1 Comparison of recombination protocols

There are a large number of library generation protocols that protein engineers can use to mutate the proteins. However, the performance of these library generation protocols have not been compared experimentally on the same sets of proteins of different homology levels. A head-to-head comparison of the sequence diversity of library generation protocols serves to illuminate potential issues that can arise from generation of libraries. Considerations such as the amount of parental templates present (parental background), duplication of chimeras (chimera background), crossover distributions and number of unique sequences are paramount for productive and efficient probe of protein sequence space during screening assays. By comparing the sequence diversity of libraries generated from DNA shuffling and recombination PCR on fluorescent proteins with different homology levels, we addressed deficiencies in library generation protocol found in the literature.

One significant finding from our work on comparing the performance of recombination protocols is that the templates used for recombination of high homology proteins (high level of identity) influence the quality of the library heavily. The experiments in Chapter 4 showed that two-sided skew templates produced the optimal mRFP and DsRed recombination PCR chimera library that has more evenly distributed crossovers, low parental background and the highest percentage of unique chimeras.

Another important finding from the recombination experiments from Chapter 4 is that DNA shuffling and recombination PCR produces libraries that require different lengths of nucleotide identity level for crossovers to occur, thus producing libraries with different diversities. The Wilcoxon rank sum statistical test and Z-statistical test applied on the sequences obtained from DNA shuffling and recombination PCR suggests that the distributions of both libraries are different and that both protocols produced similarly useful libraries with one or more crossovers. Hence, to reduce bias caused by dependence on nucleotide identity and increase diversity, one can envisage applying DNA shuffling and recombination PCR together to produce a more diversified library that accesses more protein sequence space.

Recombination PCR can shuffle genes with lower DNA identity level than DNA shuffling but it is not clear if such an application would result in an active protein. We managed to obtained chimeric sequences using recombination PCR on mRFP and GFP genes (nucleotide identity level of 45%). However, it remains to been validated, whether recombination of genes with low homology yields functional proteins.

7.2 Data-driven protein engineering

Data-driven protein engineering experiments using machine learning language such as BLSVM to identify interacting residues applied on red fluorescent proteins demonstrate the utility of the algorithm for identifying interacting residues and potentially for improving library generation efforts. In Chapter 5, we used BLSVM to analyze the sequence-activity information obtained from recombination experiments on mRFP and DsRed proteins. We proceeded to engineer the templates prior to recombination and compared the fraction of active variants obtained from recombined engineered as well as wild-type library. We found a statistically significant improved percentage of the number of active variants in the engineered templates, suggesting that the preservation of the interactions can improve the quality of a library.

One interesting observation from point mutagenesis experiments in Chapter 5 is that interactions between residues can exist more strongly in one protein than another. It can be argued that the same set of interactions that exist in one protein may not exist in another homologous protein. We found conclusive evidence of non-linear interactions in DsRed proteins based on the empirical observation that the fluorescence of double point mutant of DsRed is reactivated non-linearly. For mRFP however, the double point mutant was inactive unexpectedly.

The next logical extension for the data-driven protein engineering experiments is to look at improving industrially relevant enzymes by identifying interacting residues and important residues (Figure 7.1). In brief, amino acid alignments of the protein of interest can be performed to identify potentially important residues to be mutated, shuffled and obtain its sequence-activity relationship via screening assays. Following that, BLSVM can be used to identify interacting and important residues. A directed library consisting of a reduced set of amino acids can then be created and screened. Following iterations of the above-mentioned procedure, one should have a good chance to evolve an improved protein while validating the method for protein engineering use.

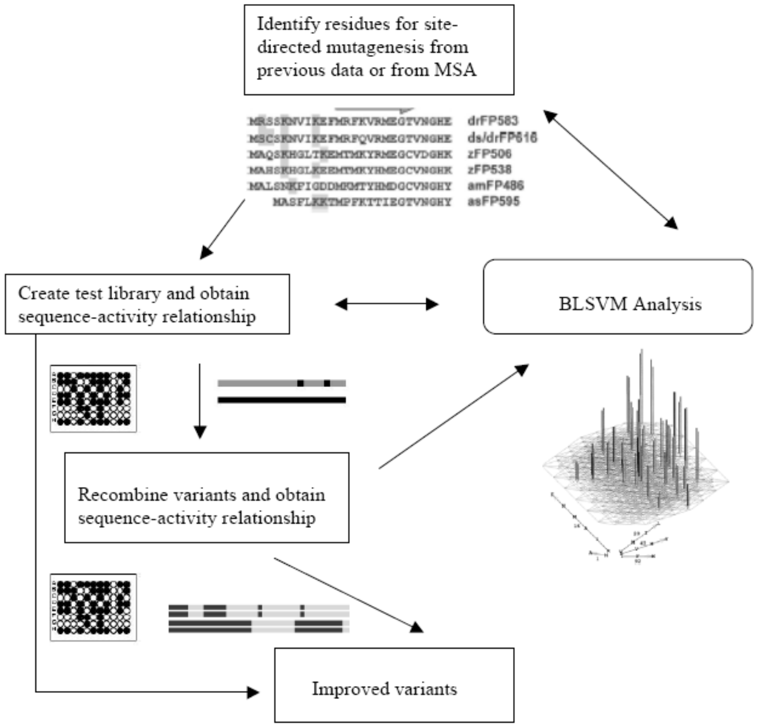

Identify residues for site-directed mutagenesis from previous data or from MSA
MRSSKNVIKEFMRFKVRMEGTVNGHE drFP583
MSCSKNVIKEFMRFQVRMEGTVNGHE ds/drFP616
MAQSKHGLTKEMTMKYRMEGCVDGHK zFP506
MAHSKHGLKEEMTMKYHMEGCVNGHK zFP538
MALSNKFIGDDMKMTYHMDGCVNGHY amFP486
MASFLKKTMPFKTTIEGTVNGHY asFP595
Create test library and obtain sequence-activity relationship
BLSVM Analysis
Recombine variants and obtain sequence-activity relationship
Improved variants

Figure 7.1 Schematic of modified BLSVM procedure for evolving proteins. Sequence-activity landscape picture was adapted from (1).

7.3 Data-driven protein engineering, 3rd wave: Emerging picture of the near future of Protein Engineering

The principle problem of using combinatorial methods for protein engineering or directed evolution (DE), the 2nd wave of protein engineering, is the impossibly huge size of the sequence space and generating huge libraries to partially counteract this problem. Recent efforts in protein engineering today strive to limit library size while increasing the low hit rate of directed evolution. Two elements emerged to be crucial to limit library size limitation: computational data handling and interpretation (such as BL/SVM) and superior molecular biology (NDT vs NNK randomization) combined with structure-guided limitations of mutated amino acid residues.

Table 7.1 illustrates the 3rd wave of protein engineering due to recent advances in tools and methods development for protein. In general, the steps involved for protein engineering can be divided into i) Finding important residues, ii) Finding interacting residues, iii) Improving non-targeted trait, iv) Restricting nucleotide range and v) Restricting amino acid sites for randomization.

Table 7.1 Generic recipe for data-driven protein engineering.*,#

Step	**Technique(s)**	**Tools**	**Used by**
1. find important residues	- min. least squares - DNA2.0** - SVM	- ProSAR	- *Codexis* - *DNA2.0*
2. find interactive residues	- DNA2.0** - computational thermodynamic stability - rotamer G minimization. - Boolean Learning and Support Vector Machine	- SCHEMA - FamClash	- *DNA2.0* - Arnold - Maranas

Table 7.1 continued

3. improve non-target-ed trait& (i.e. stability if target is specificity)	- exhaustive mutagenesis - structure-guided consensus - structure guided consensus	- MSA, overlay - MSA, overlay	- *Genencor* - Arnold
4. restrict nucleotide range	- gene synthesis	- GENEy	- Daugherty
5. restrict amino acid sites for randomization	- degenerate oligos - CASTing	- CASTER	- *Codexis* - Reetz

GENEy Gene degeneracy program developed within the Bommarius group for limiting the nucleotide size of the library

MSA: Multiple sequence alignment, Daugherty: Patrick Daugherty, Maranas: Costas Maranas,

* The above table was consolidated based on input from Dr. Andreas Bommarius who previously had a discussion about the above topic, among others, with Dr. Frances Arnold and Dr. Manfred Reetz.

** DNA2.0 applies a range of techniques that are not published: SVMR, Lasso, Ridge, PLS, LSVM, LPBoostR.

# Bommarius lab has used i) – iii) so far.

7.4 Protein engineering of mRFP and PGA

Protein engineering efforts on mRFP and PGA did not result in improved variants but the insight gained from these experiments can serve to design future experiments that would improve the probability of success. From Chapter 3, we mentioned that a more red-shifted version of mRFP, mPlum, was already obtained and that two rounds of directed evolution was insufficient to find improved variants. However, a version of mRFP with a higher relative quantum yield still remains to be found. A recent successful work done on improving the brightness of blue fluorescent protein (BFP) point to potential methods to improve the relative quantum yield of mRFP. Mena *et al.* created a targeted synthetic library of BFPs based on considerations of distance from chromophore, the direction the side-chains were pointing and its involvement in catalysis of chromophore formation (2). The variants can be subjected to fluorescence activated cell sorting (FACs) to screen for the brightest fluorescent protein. Alternatively, the variants can be expressed on plates and screened using an imaging machine from which fluorescence intensity can be determined.

7.4.1 PGA screen using penicillin V

One of the future directions for PGA experiments should include screening the library for activity towards penicillin V. From the crystal structure data presented in Chapter 6, it is conceivable that the current goal of evolving substrate specificity of PGA towards oxacillin in one step may be too demanding on the evolutionary pathway. Instead, a bridging or intermediate substrate, which displays less substrate hindrance than oxacillin, could be used as a target for evolution first. Following that, one can attempt to evolve PGA to be more active towards more difficult substrates such as oxacillin, cloxacillin and nafcillin.

7.4.2 PGA fluorimetric screen using pH sensitive fluorescent proteins

The screens for PGA experiments could also be improved by using fluorescent proteins co-expressed with PGA enzyme in combination with FACs for screening. Chapter 6 showed that possible background signal issues plague the filter lift assays for the current PGA experiments. A high-throughput fluorimetric screen that has better signal-to-noise ratio for PGA assays would be extremely helpful to find improved variants. One possible high-throughput assay involves the use of pH sensitive fluorescent protein such as pHluorin or enhanced green fluorescent protein (EGFP) that changes spectral characteristics when it detects a change of pH, due to PGA activity on antibiotic substrates that liberates an acidic moiety upon hydrolysis. Schuster *et al.* demonstrated the use pHluorin *in vivo* assay to screen for local hydrolytic induced pH shift in *E.coli* cells (3).

7.4.3 Creation of a novel fluorimetry based screening system using penicillin receptors and fluorescent proteins

Another plausible assay involves the use of penicillin binding receptors such as BlaR1 that changes its conformation upon binding to β-lactam (4). Fluorescent proteins positioned in the relevant loops (Figure 7.2) would move apart during the binding event, thus, inducing a change in fluorescent resonance energy transfer (FRET). Alternatively, cells that have BlaR1 signal transduction machinery can be genetically modified to incorporate GFP gene, either in replacement, or fused to the β-lactamase gene. The cells, in theory, will express GFP in the presence of β-lactam, thus allowing one to identify cells that have β-lactam present fluorimetrically. Thereafter, the genetically engineered cell strains can be used for expressing the PGA variants and detect for activity towards β-lactamase resistant antibiotics.

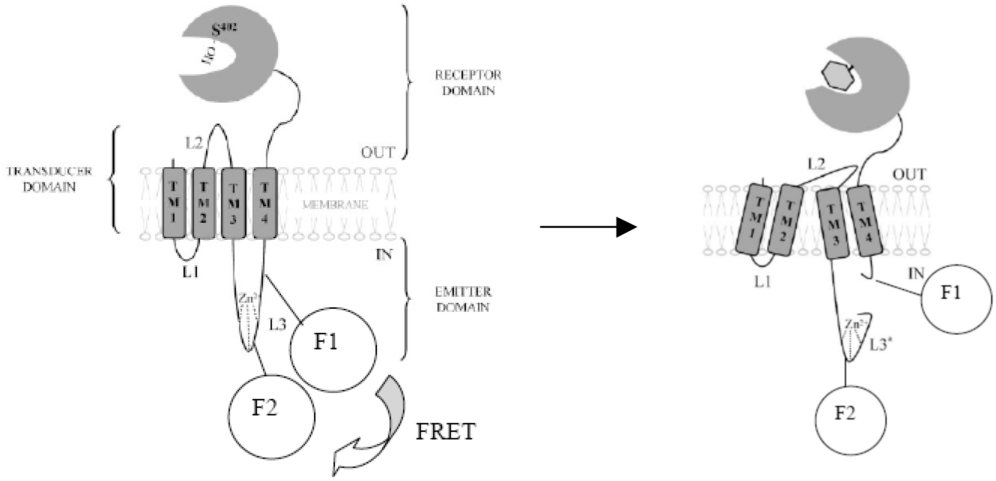


Figure 7.2 – A representation of the BlaR1 binding receptor from *Bacillus lichiniformis*. β-lactam binds to serine residue (S^{402}) and induces a conformational change that results in metallo-proteases cleaving loop L3. Fluorescent proteins (F1, F2) which are placed in loop L3, are separated. This results in FRET decreasing. Figure adapted from (4).

7.4.4 Selection assays to decrease screening efforts

Selection assays utilizing an *E. coli* auxotroph combined with the relevant antibiotic side-chain analog can also be used for assaying improved variants. Recent work by Jaeger *et al.* showed that the use of a prescreen selection assay when evolving hybrid PGA is a viable way to eliminate inactive variants (5). They used an auxotroph such as HB101 *E. coli* strain that requires leucine in combination with minimal LB plates and N-phenyl-acyl-L-leucine to prescreen active shuffled chimeras of PGA from *E.coli*, *Kluyvera cryocrescens* and *Providencia rettgeri.* Active PGA will liberate leucine that is required for growth of *E.coli*. Following that, secondary assays on the selected variants, were done using high performance reverse chromatography to verify synthesis of ampicillin and to obtain activity. Assuming stable and suitable analogues with the targeted side-chains are accessible, we can use selection assays to reduce the number of variants for screening.

7.4.5 New libraries for PGA experiments

Lastly, an important aspect for future PGA work is the design of new libraries using techniques such as protein sequence activity relationship (ProSar), BLSVM and/or combinatorial active-site saturation testing (CASTing) (6). Assuming that a good screen is accessible, one can create a targeted restricted library for PGA for evolution towards different antibiotic substrates. ProSar and BLSVM can identify key residues that affect the activity of the PGA while CASTing limits mutagenesis to residues that would likely be important for substrate specificity change. These smart directed libraries should in theory accelerate convergence towards the desired optimized variants, as the protein sequence search is limited to functionally relevant space.

7.5 References

1. Aita, T., Iwakura, M. and Husimi, Y. (2001) A cross-section of the fitness landscape of dihydrofolate reductase. *Protein Engineering*, **14**, 633-638.

2. Mena, M.A., Treynor, T.P., Mayo, S.L. and Daugherty, P.S. (2006) Blue fluorescent proteins with enhanced brightness and photostability from a structurally targeted library. *Nat Biotechnol*, **24**, 1569-1571.

3. Schuster, S., Enzelberger, M., Trauthwein, H., Schmid, R.D. and Urlacher, V.B. (2005) pHluorin-based in vivo assay for hydrolase screening. *Analytical Chemistry*, **77**, 2727-2732.

4. Hanique, S., Colombo, M.L., Goormaghtigh, E., Soumillion, P., Frère, J.M. and Joris, B. (2004) Evidence of an intramolecular interaction between the two domains of the BlaR1 penicillin receptor during the signal transduction. *The Journal of biological chemistry*, **279**, 14264-14272.

5. Jaeger, K.E. and Eggert, T. (2004) Enantioselective biocatalysis optimized by directed evolution. *Current Opinion in Biotechnology*, **15**, 305-313.

6. Reetz, M.T., Bocola, M., Carballeira, J.D., Zha, D. and Vogel, A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew Chem Int Ed Engl*, **44**, 4192-4196.

VITA

Bernard Loo Liat Wen

Bernard was born in Singapore. He attended public schools in Singapore. After serving about two years in the army for national service as an infantry officer, he came to the US in 1999 to read chemical engineering at the University of Minnesota, Twin Cities. In 2002, he started his chemical engineering doctoral program at Georgia Institute of Technology. In 2007, he was awarded his Ph.D degree. While not doing research, he enjoys a variety of activities – cycling, tennis, golf, hiking, rollerblading and drinking coffee.