

# **IMPROVING PROTEIN SOLUBILITY VIA DIRECTED EVOLUTION**

by

Meagan Z. Perry

A thesis submitted to the Department of Chemistry

In conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(October, 2009)

Copyright ©Meagan Z. Perry, 2009

## Abstract

A major hurdle facing *in vitro* protein characterization is obtaining soluble protein from targets that tend to aggregate and form insoluble inclusion bodies. Soluble protein is essential for any biophysical data collection and new methods are needed to approach this significant problem. Directed evolution can be used to discover mutations which lead to improved solubility using an appropriate screening method. Green fluorescent protein (GFP) has been shown to be an effective solubility reporter which can be used to screen for soluble protein variants. We have chosen three diverse enzymes as targets for improving protein solubility using this technique: arachidonate 5-lipoxygenase—an enzyme which converts fatty acids into leukotrienes, PhnG—an enzyme belonging to the bacterial carbon-phosphorus lyase pathway, and RebG—a glycosyltransferase. Error-prone PCR and DNA shuffling were used to generate libraries of mutants which were subsequently cloned into a GFP-fusion screening vector. From the evolution of 5LO and RebG, much was learned about the optimization of the protocols involved in this methodology, including valuable information about how to avoid common “false-positive” results in which fluorescent colonies arise while screening but do not represent an improvement of the target. Evolution of these two targets did not result in an improvement of solubility, however truncation strategies may still prove to be effective, and more work needs to be done in this area. Evolution of PhnG successfully produced one variant, named clone B6, which showed both an improvement in expression and folding over wild type PhnG. It was also discovered that GFPuv can act as an effective solubility enhancing fusion tag for PhnG. Prior to the current studies PhnG had not been effectively expressed and purified in *E. coli*, however purification and refolding of resolubilized inclusion bodies of the clone B6 PhnG-GFP fusion construct was shown to yield enough soluble protein for future crystallographic studies.

## Acknowledgements

I would first like to thank, with the deepest sincerity, my supervisor Dr. David Zechel for his kindness, understanding, and tremendous support during these last three years at Queen's. I encountered many trials during my studies, both personally and in my research, and he continually showed professional guidance as well as wisdom and understanding. He always gave me the confidence I needed to believe in myself and continue. As a single mother it is very easy to be brought down by the pressure of it all, especially guilt over how your time is divided between the lab and home. During these guilt-stricken moments David continually reminded me that research is often about quality, not necessarily how many hours are spent in the lab, and I want to thank him for that.

I would also like to thank Dr. Derek Pratt and for his always helpful questions and critiques at group meetings, and his willingness to answer all of my (mostly naïve) organic chemistry questions as he would walk through the lab. A special thank you to Jay Hanthorn, who is also very good at answering organic chemistry questions and equally good at keeping everyone's spirits up. There are many other lab members who have helped me immeasurably during my time here: Ryan Latimer, Di Hu, Dr. Anupam Bhattacharya (who trained me, so I would to thank him for his calmness), Shu-Mei He, Dr. Daria Trofimova, Shelly McArthur, and Polly Ho.

An extremely special thank you to my wonderful neighbours and friends in Portsmouth Village, especially Rudy Candela, Krista Asselstine, Annette Willis and Tom Brennan. I have never in all my life experienced such generosity. Thank you for helping me with Naomi whenever I needed it (which was pretty much every day, all day near the end), and thank you for providing Naomi with an extended family in which she feels tremendously loved. Thank you for giving me your friendship and love. We are so lucky to know all of you.

Lastly, to my family. To my parents for giving me the tools to handle all of it, for being there when I needed you the most, for knowing that there is nothing I could ask that you would not try to give. To my brother Tim and his wife Trisha, for spoiling me and spoiling Naomi when I couldn't. Your generosity truly is astounding. To my beautiful daughter Naomi. You are my constant source of joy. No matter how stressful my day is, your beaming smile always takes the heaviest burdens off of my shoulders. I love you, thank you.

# Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents.....	v
List of Figures.....	x
List of Tables.....	xiv
List of Abbreviations.....	xv
Chapter 1 Introduction.....	1
1.1 Protein misfolding: a common problem of recombinant protein expression.....	1
1.2 Protein folding <i>in vitro</i> .....	2
1.3 Protein folding and misfolding in the cell.....	5
1.4 Methods for enhancing recombinant protein solubility in <i>E. coli</i> .....	8
1.4.1 Extrinsic factors affecting protein solubility.....	8
1.4.2 Intrinsic factors affecting protein solubility.....	15
1.5 Directed evolution as a strategy to improve protein solubility.....	17
1.5.1 Diversification.....	19
1.5.1.1 Truncation and fragmentation methods.....	20
1.5.1.2 Random mutagenesis methods.....	21
1.5.1.3 Recombination methods.....	25
1.5.2 Selection and screening strategies for evolving proteins with improved folding and solubility.....	29

1.5.3 Green fluorescent protein as a folding reporter .....	33
1.6 References .....	42
Chapter 2 .....	52
Application of the Green-Fluorescent Protein Solubility Assay to the Directed Evolution of Human 5-Lipoxygenase and RebG .....	52
2.1 Introduction .....	52
2.2 Experimental Procedures and Methods .....	57
2.2.1 Materials .....	57
2.2.2 Construction of the pProEx_GFPuv screening vector .....	58
2.2.3 Cloning of wild type 5LO and RebG into pProEx_GFPuv1 .....	62
2.2.4 Mutagenic PCR of 5LO and RebG .....	63
2.2.5 DNA shuffling of 5LO and RebG .....	64
2.2.6 Rational truncation of lipoxygenases and tagged random-primer PCR .....	66
2.2.7 Screening of library variants for improved fluorescence .....	68
2.2.8 Expression and determination of excitation and emission maxima of GFPuv ..	70
2.3 Results and Discussion .....	72
2.3.1 Optimization of error-prone PCR .....	72
2.3.2 Optimization of DNase I digestion .....	75
2.3.3 Optimization of recombination and amplification .....	79
2.3.4 Screening vector construction .....	81
2.3.5 Considerations for effective screening .....	82

2.3.6 Truncation strategies to improve solubility of 5LO .....	98
2.4 Conclusions .....	105
2.5 References .....	108
Chapter 3 .....	111
Directed Evolution of <i>Escherichia coli</i> Phosphonate Metabolism Protein PhnG .....	111
3.1 Introduction .....	111
3.2 Experimental Procedures and Methods .....	114
3.2.1 Materials .....	114
3.2.2 Cloning of wild type <i>phnG</i> into pProEx_GFPuv1 .....	115
3.2.3 Error-prone PCR of <i>phnG</i> .....	115
3.2.4 DNA shuffling of <i>phnG</i> .....	116
3.2.5 Selection and purification of library variants with improved fluorescence....	118
3.2.6 Expression of PhnG-GFP wild type and clone B6 .....	118
3.2.7 Purification of PhnG-GFP wild type and clone B6 .....	118
3.2.7.1 Purification in the presence of detergent and DTT .....	119
3.2.7.2 Purification under denaturing conditions .....	119
3.2.8 Resolubilization, purification and refolding of PhnG-GFP wild type and clone B6 inclusion bodies .....	119
3.2.9 Fluorescence measurements of PhnG-GFP wild type and clone B6 .....	120
3.2.10 Size exclusion chromatography .....	121
3.3 Results and Discussion .....	122

3.3.1 Mutagenic PCR on <i>phnG</i> —round 1 .....	122
3.3.2 DNA shuffling of <i>phnG</i> —rounds 2 and 3 and back-crossing .....	129
3.3.3 Solubility and fluorescence analysis of clone B6 .....	132
3.3.4 Purification of PhnG wild type and clone B6 fusions in the presence of detergent .....	137
3.3.5 Size exclusion chromatography of PhnG wild type and clone B6 fusions.....	139
3.3.6 Purification under denaturing conditions and refolding of PhnG wild type- and clone B6-GFP fusions .....	142
3.3.7 Expression of PhnG wild type and clone B6 as non-fusions.....	151
3.4 Conclusions .....	153
3.5 References .....	156
Chapter 4 Conclusions .....	159
Appendix A .....	161
Mutagenesis and Kinetic analysis of <i>TmGH1</i> in Preparation for Atomic Force Microscopy.....	161
A.1 Introduction .....	161
A.2 Experimental Procedures and Methods.....	168
A.2.1 Materials .....	168
A.2.2 Mutagenesis of wild type <i>TmGH1</i> .....	169
A.2.3 Expression and purification of wild type <i>TmGH1</i> and <i>TmGH1</i> -Mut .....	170
A.2.4 Kinetic analysis of wild type <i>TmGH1</i> and <i>TmGH1</i> -Mut.....	171
A.3 Results and Discussion.....	172

A.3.1 Expression and purification of wild type <i>TmGH1</i> and <i>TmGH1</i> -Mut .....	172
A.3.2 Kinetic analysis of wild type <i>TmGH1</i> and <i>TmGH1</i> -Mut.....	174
A.4 Conclusions .....	177
A.5 References .....	178

## List of Figures

Figure 1-1	Energy landscapes for protein folding. ....	5
Figure 1-2	Basic components of an <i>E. coli</i> expression vector .....	9
Figure 1-3	Basic steps involved in laboratory evolution of proteins. ....	18
Figure 1-4	Base analogs P and 8-oxoG pairing with A. ....	23
Figure 1-5	A graphical comparison of phenotypic optimization achieved by random mutagenesis methods and recombination methods .....	26
Figure 1-6	Structure of GFP and its chromophore.....	34
Figure 1-7	Post-translational synthesis of GFP chromophore.. ....	35
Figure 2-1	Reaction scheme for 5LO.....	53
Figure 2-2	Active sites of rabbit reticulocyte LO (PDB accession code 2P0M) and human 12LO (PDB accession code 3D3L).....	54
Figure 2-3	Rebeccamycin biosynthetic pathway.. ....	55
Figure 2-4	Insert used for modification of the pProEx multiple cloning region.....	58
Figure 2-5	Insert used for modification of pGFPuv cloning region.....	59
Figure 2-6	Overview of pProEx_GFPuv vector construction.....	62
Figure 2-7	Alignment of human 5LO, human 15LO, and mouse 8LO with truncated human 12LO to determine starting residue of truncated variants. ....	67
Figure 2-8	Effect of Mg <sup>2+</sup> concentration on PCR yield of RebG.....	73
Figure 2-9	Effect of Mn <sup>2+</sup> concentration on the yield of a RebG mutagenic PCR. ....	74
Figure 2-10	Dnase I digestion time trials for 5LO (A) and RebG (B).....	77
Figure 2-11	Fragments from the digestion of 5LO (A) and RebG (B ,C) .....	79
Figure 2-12	Recombination of 5LO with Taq (A), Pfu (B), and Herc (B) polymerase.....	80
Figure 2-13	Amplification of recombined 5LO with Vent, Pfu, and Herculanse II polymerases .....	81

Figure 2-14	Screening vector constructs pProEx_GFPuv (A) and pProEx_GFPuv1 (B).....	82
Figure 2-15	Determination of the excitation and emission maxima for the GFP folding reporter. ....	84
Figure 2-16	Streaks and colony PCR of clones selected from round 1 of 5LO and RebG.....	86
Figure 2-17	Examples of the fluorescence obtained from clones picked from round 2 for 5LO (A) and RebG (B).....	87
Figure 2-18	Restriction digest analysis of 5LO round 2 clones showing an insert from colony PCR. ....	88
Figure 2-19	5LO round 2 purified clones (A) and colony PCR of the clones (B).....	89
Figure 2-20.	Sequencing data from RebG (A) and 5LO (B) clones exhibiting very bright fluorescence but no insert.....	91
Figure 2-21	Streaks of fluorescent clones found during the second attempt at round 1 (A) and restriction digest analysis of those clones (B).....	93
Figure 2-22	Sequence comparisons of the normal and mutant screening vectors. ....	95
Figure 2-23	Selected mutants from the first round of evolution on mouse 8LO (A) and the second round of evolution on human 12LO.....	97
Figure 2-24	Fluorescence measurements on the soluble fraction of 12LO mutant cell lysates. ....	98
Figure 2-25	Fluorescence comparison between truncated lipoxygenase fusions and the corresponding full-length wild type fusions.....	99
Figure 2-26	Products from tagged random primer PCR on 5LO.....	100
Figure 2-27	Fluorescence of clones selected after tagged random-primer PCR.....	101
Figure 2-28	Restriction digest analysis of unpurified clones selected after TP-PCR (A), and after purification (B). ....	101
Figure 2-29	(A) DNA sequences giving rise to fluorescent clones after TP-PCR (B) The sequences of the inserts.....	103
Figure 3-1	The CP-lyase pathway acts on organophosphonates to cleave the CP bond and produce a hydrocarbon and orthophosphate.....	111

Figure 3-2	Clones picked from round 1 of evolution of PhnG (A and B) and comparison of mutant 4 and wild type PhnG colonies (C). .....	123
Figure 3-3	PCR amplification of purified mutant genes from round 1 of PhnG evolution .	124
Figure 3-4	Multiple sequence alignment of PhnG homologs .....	127
Figure 3-5.	Alignment of unique <i>phnG</i> mutants from round 1 .....	128
Figure 3-6	DNA shuffling of <i>phnG</i> .....	129
Figure 3-7	Fluorescence of the 6 brightest clones from rounds 2 and 3, and the backcrossing round. ....	130
Figure 3-8.	PhnG clone B6 before and after purification.....	134
Figure 3-9	Test expression of PhnG wild type and clone B6 .....	135
Figure 3-10	Fluorescence of <i>E. coli</i> supernatants containing wild type PhnG-GFP and clone B6.....	136
Figure 3-11	SDS-PAGE of wild type PhnG-GFP (A) and clone B6 PhnG-GFP (B) fractions after IMAC purification under standard conditions. ....	137
Figure 3-12	Structure of n-dodecyl- $\beta$ -D-maltoside. ....	138
Figure 3-13	SDS-PAGE of PhnG wild type (A) and clone B6 (B) after purification in the presence of DDM and DTT.....	139
Figure 3-14	Size exclusion chromatograms of the semi-pure PhnG wild type (A) and clone B6 (B) fusions in DDM and DTT.....	141
Figure 3-15	SDS-PAGE analysis of <i>phnG</i> wild type and clone B6 supernatants after IMAC purification under denaturing conditions. ....	143
Figure 3-16	Absorbance spectra and SDS-PAGE analysis of wild type PhnG-GFP and clone B6 PhnG-GFP denatured in 6 M urea prior to refolding, and refolded PhnG-GFP fusions following dialysis .....	144
Figure 3-17	Precipitates of PhnG wild type (A) and clone B6 (B) fusions after dialysis.....	145
Figure 3-18.	SDS-PAGE gel of fractions obtained from denaturing purification of wild type and clone B6 inclusion bodies.....	146

Figure 3-19	Equalized concentrations of denatured wild type and clone B6 fusions (A) and their relative fluorescence (B).....	147
Figure 3-20.	Absorbance spectra and SDS-PAGE gels of wild type PhnG-GFP and clone B6 PhnG-GFP resolubilized inclusion bodies pre-and post-dialysis. ....	148
Figure 3-21	Size exclusion chromatograms of refolded PhnG wild type (A) and clone B6 (B) fusions.....	150
Figure 3-22	SDS-PAGE and western blot analysis of wild type PhnG and clone B6 as non-fusions. ....	151
Figure 3-23	SDS-PAGE gels of fractions collected from IMAC purification of PhnG wild type and clone B6 as non-fusions.....	152
Figure A-1	Mechanism of <i>TmGH1</i> .....	162
Figure A-2	Structure of $\beta$ -glucosidase inhibitors 1-deoxynojirimycin (A) and isofagomine (B). ....	163
Figure A-3	Slow-onset inhibition of <i>TmGH1</i> by isofagomine .....	163
Figure A-4	Crystal Structure of <i>TmGH1</i> . ....	167
Figure A-5	Purification of wild type <i>TmGH1</i> (A) and <i>TmGH1</i> -Mut (B). ....	173
Figure A-6	Structure of 2-nitrophenyl- $\beta$ -D-glucopyranoside.....	174
Figure A-7	Plots of 2-nitrophenyl- $\beta$ -D-glucopyranoside concentration vs. rate over total enzyme concentration for <i>TmGH1</i> wild type and <i>TmGH1</i> -Mut.....	175
Figure A-8	Reaction of maleimide with the thiol group of a cysteine residue to form a thioether linkage.....	176

## List of Tables

<b>Table 1-1</b>	Vector considerations for protein solubility.....	10
<b>Table 1-2</b>	Promoter sequences in <i>E. coli</i> .....	11
<b>Table 1-3</b>	Rare codon usage in <i>E. coli</i> .....	16
<b>Table 1-4</b>	Literature examples of improvement of solubility via evolution and screening with a GFP folding.....	37
<b>Table 1-5</b>	GFP mutants and their properties.....	38
<b>Table 2-1</b>	Mutagenic PCR reaction mixes for 5LO and RebG.....	64
<b>Table 2-2</b>	Primers for truncation of lipoxygenases.....	67
<b>Table 2-3</b>	Analysis of mutation load after epPCR for 5LO and RebG.....	75
<b>Table 3-1</b>	Mutagenic PCR reaction mix for PhnG.....	116
<b>Table 3-2</b>	Mutation analysis of PhnG after error-prone PCR.....	125
<b>Table 3-3</b>	Amino acid mutations from selected round 1 PhnG mutants.....	125
<b>Table 3-4</b>	Amino acid mutations for selected clones from rounds 2 and 3, and back-crossing of PhnG evolution.....	131
<b>Table A-1</b>	PCR conditions for the mutagenesis of <i>TmGH1</i> .....	170
<b>Table A-2</b>	Kinetic parameters obtained for <i>TmGH1</i> wild type and <i>TmGH1</i> -Mut.....	175

## List of Abbreviations

3-D	three dimensional
5LO	human arachidonate 5-lipoxygenase
AA	arachidonic acid
AFM	atomic force microscopy
AFS	average fragment size
Amp	ampicillin
BSA	bovine serum albumin
CAT	chloramphenicol acetyltransferase
cDNA	complementary DNA
CIAP	calf intestinal alkaline phosphatase
CP	carbon-phosphorus
d8-oxoGTP	8-oxo-2'-deoxyguanosine
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
DDM	n-dodecyl- $\beta$ -D-maltoside
dGTP	2'-deoxyguanosine 5'-triphosphate

DMSO	dimethyl sulfoxide
dPTP	6-(2-deoxy- $\beta$ -D-ribofuranosyl)-3,4-dihydro-8H-pyrimido-[4,5-C][1,2]oxazin-7-one
dTTP	2'-deoxythymidine 5'-triphosphate
dUTP	2'-deoxyuridine 5'-triphosphate
EDTA	ethylene diamine tetraacetic acid
eGFP	enhanced green fluorescent protein
epPCR	error-prone polymerase chain reaction
FACS	fluorescence activated cell sorting
FPLC	fast protein liquid chromatography
GFP	green fluorescent protein
GST	glutathione-S-transferase
GT	glycosyltransferase
HGH	human growth hormone
5-HPETE	5(S)-hydroperoxy-6- <i>trans</i> -8,11,14- <i>cis</i> -eicosatetraenoic acid
HSP	heat shock protein
IMAC	immobilized metal affinity chromatography
IPTG	isopropyl $\beta$ -D-1-thiogalactopyranoside

ITCHY	incremental truncation for the creation of hybrid enzymes
LB	Luria Bertani
LT	leukotriene
LTA <sub>4</sub>	leukotriene A <sub>4</sub>
MAP	mutagenesis assistant program
MBP	maltose binding protein
MIC	minimum inhibitory concentration
Ni-NTA resin	nickel nitriloacetic acid resin
NMR	nuclear magnetic resonance
NusA	N-utilizing substance A
PDB	Protein Data Bank
PCR	polymerase chain reaction
Pi	orthophosphate
PMMA	poly (methylmethacrylate)
PRM complex	protein-ribosome-mRNA complex
ProSIDE	protein stability increased by directed evolution
RT-PCR	reverse-transcriptase polymerase chain reaction

SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SMFS	single molecule force spectroscopy
StEP	Staggered extension process
sulfo-EMCS	[N-e-Maleimidocaproyloxy]sulfo succinimide ester
SUMO	small ubiquitin-related modifier
TF	trigger factor
TP-PCR	tagged-random primer polymerase chain reaction
Tris	tris(hydroxymethyl) aminomethane
WT	wild type

# Chapter 1

## Introduction

### 1.1 Protein misfolding: a common problem of recombinant protein expression

Structural biology and bioinformatics have become two of the most powerful ways to learn about life at the molecular level. Information elucidated from the three-dimensional (3-D) structure of proteins, as well as mechanistic analysis through kinetics and mutagenesis, can provide insights into novel catalytic mechanisms, substrate specificity and binding, and evolution. Clearly the benefit of this knowledge is profound and wide-spread, from helping elucidate the causes of certain diseases to improved drug design to the more efficient degradation of toxins in the environment. X-ray crystallography and NMR spectroscopy are two main ways structural biologists determine protein structures and both of these methods, as well as activity assays, require the protein being studied to be soluble, properly folded, and highly pure in solution [1]. Recombinant protein expression in bacterial hosts, especially *Escherichia coli*, is currently the most common method for acquiring large yields of protein for use in structure determination. In 2003, approximately 80% of the 3-D structures submitted into the Protein Data Bank (PDB) were prepared using an *E. coli* expression system [2]. Recombinant protein expression in *E. coli* boasts the advantages of low cost, simple purification techniques, ease of genetic manipulation, and the availability of vast amounts of literature which thoroughly characterizes its capabilities and limits [3]. Because of these advantages it has been of great interest to find general ways of improving the folding and solubility of proteins that misfold when overexpressed in *E. coli*. As it remains, one-half to one-third of prokaryotic proteins and likely a higher fraction of eukaryotic proteins cannot be recombinantly expressed in a well-folded and soluble form in *E. coli* [1, 4].

Recombinant protein misfolding and insolubility arises for many reasons and a great deal of research effort has gone into elucidating the mechanisms behind it. This introduction will include a brief review on what is currently known about protein folding *in vitro*, how it relates to folding and misfolding *in vivo*, and some general techniques to improve folding of recombinant protein upon expression in *E. coli*, with particular emphasis on directed evolution as a strategy.

## **1.2 Protein folding *in vitro***

Proteins have the potential to reach their native state whether they are in a cell or a test tube. Therefore, it is the amino acid sequence of a protein that dictates the final tertiary structure, although chaperones can assist a protein in reaching this final structure in a cell [5, 6]. A denatured protein has limitless conformations which it can search, and if it were to randomly search every possible configuration the folding process would take billions of years as opposed to milliseconds [7]. Thus, instead of random searching the folding process is guided by physiochemical forces. In the final native structure it is the additive strength of numerous hydrogen bonds, electrostatic interactions, van der Waals interactions (that arise from tight packing), and the hydrophobic effect which provide structural stability, with the latter providing the dominant driving force for folding [8-10]. Evidence to support this theory arises from the observation that residues with hydrophobic side chains are always found in the core of proteins, implying that their non-polar nature drives them away from any aqueous environment (due to the favourable entropy associated with releasing water molecules forming clathrate structures around these non-polar surfaces). This observation is also supported by the fact that proteins are easily denatured when dissolved in non polar solvents or aqueous solutions of chaotropes (urea, guanidine hydrochloride) [11]. The essential role these buried hydrophobic residues play in stability is also highlighted by their conservation throughout evolution [12].

The exact mechanism of how forces guide an amino acid sequence to form proper secondary and tertiary structure is still under debate however, it is agreed that a sequential stabilization process occurs in which local contacts are formed first, and that differing mechanisms are likely for different proteins. It is also agreed that the denatured state is quite varied depending on environment, and can have a range of residual structure from random coils to structured intermediates [13, 14]. The degree to which the denatured state is structured is a major determinant of the rate at which the protein will fold [14].

In the proposed nucleation-condensation mechanism, portions of the denatured chain form a nucleus in the folding transition state which contains specific native contacts [13, 15, 16]. The structure of the nucleus may be stable enough to be observed experimentally as an intermediate, however for proteins which display two-state folding this is not the case. Upon formation of the nucleus the unstructured portions of the chain will rapidly “condense” around it (through cooperativity of multiple weak interactions) and the native structure is finally reached. Proteins which have a high degree of residual structure in the denatured state will thus fold rapidly as the nucleus will be reached quickly and will have a high degree of native structure [13].

Another proposed mechanism, known as hydrophobic zipping, describes folding as the initial formation of local contacts which tightens the structure enough so that contacts further away in the chain can be made, and so on until the native state is reached [11, 17]. A single nucleus would not be observed, but rather small, metastable structures will begin to form simultaneously along various portions of the chain. Structures which are primarily stabilized by local contacts such as helices and turns would be the first to form, followed by a progression of stability as these small metastable structures “zip” into larger, increasingly stable native-like structures, until finally the native state is reached when these substructures combine and are locked into place by non-local interactions. Like the nucleation-condensation mechanism

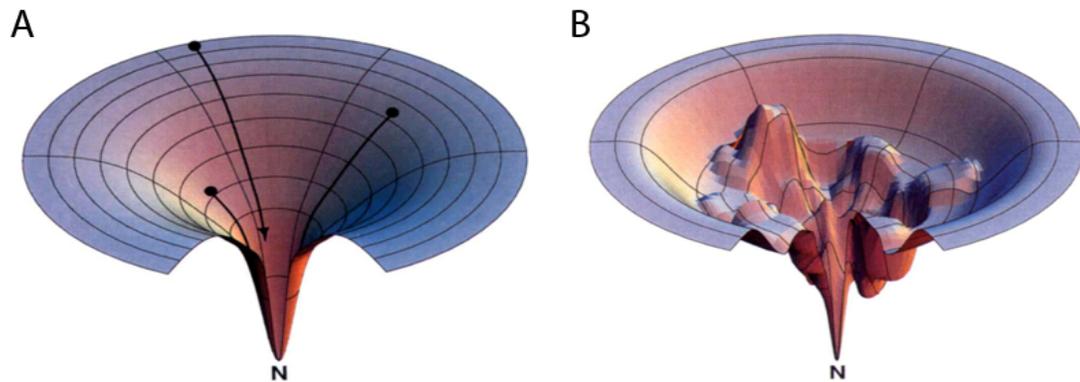
cooperativity between multiple weak interactions is a major determinant of folding. Again, denatured ensembles high in residual structure would lead to rapid folding as many local contacts are already made once folding commences.

It has been shown that there is a correlation between the proximity of native contacts in the amino acid chain and the rate at which the protein folds [18]. Structures where the native contacts are mostly local such as highly helical structures with tight turns generally fold faster than structures with  $\beta$ -sheets, which are stabilized more by non-local contacts [18].

Even though general mechanisms can be envisaged for folding, proteins are not thought to fold along a singular pathway with distinct intermediates. On the contrary, a protein is thought to have multiple that paths to the native state, and the denatured and intermediate states are not singular structures but heterogeneous ensembles of molecules [17]. The environment surrounding the protein affects the path it will take so that a protein which folds a certain way inside the cell may choose an alternate path *in vitro*. Regardless of the path it takes, the native state remains the same as it is determined only by the amino acid sequence and not the environment. The concept of multiple pathways leading to the same endpoint can be visualized by a funnel-shaped landscape [11]. As opposed to chemical reactions which generally have reactants going to intermediates and products along a singular pathway, the funnel depicts proteins as having multiple pathways which can be smooth (**Figure 1-1 A**), indicating two-state folding, or rugged (**Figure 1-1 B**), which is indicative of intermediates and energy barriers [17].

As a protein travels down the funnel towards the native state, the number of available conformations for a particular energy starts to decrease [19, 20]. This is because the protein becomes more thermodynamically stable as it is making more native contacts and is therefore more compact with fewer degrees of freedom and fewer available conformations. The protein continues to pack and gain more tertiary structure until it maximizes stability and the singular

native state is reached. Proteins known as “fast-folders” which can fold on a millisecond time scale are predicted to have smooth funnels with no observable barriers (**Figure 1-1 A**), whereas larger, more complex proteins with more than two states will have funnels with a rough surface (**Figure 1-1 B**)[17].



**Figure 1-1.** Energy landscapes for protein folding. A smooth funnel (**A**) indicates barrier-less folding whereas rough surfaces (**B**) arise from pathways exhibiting intermediates and barriers. Reproduced with permission from [17].

### 1.3 Protein folding and misfolding in the cell

Spontaneous folding *in vitro* is likely to be different than folding *in vivo* since the route a protein will take down the folding funnel is highly dependent on environment. Proteins that are refolded *in vitro* begin folding at a state with all information from the amino acid sequence available and possibly some residual structure already in place, whereas proteins coming out of a ribosome receive folding information in a vectorial fashion and thus begin folding before the entire chain has been synthesized and released (cotranslationally). In bacterial cells ribosomes synthesize proteins at an average rate of 10 – 50 amino acids per second [21] and often the ribosomes aggregate into “polysomes” and thus several proteins are being synthesized simultaneously in very close proximity. In addition, other proteins and macromolecules such as

DNA and RNA are present at concentrations in the range of 300-400 mg/mL [22]. This “macromolecular crowding” can be both beneficial and detrimental to protein folding as it will favour the formation of compact states [23] but it is also the major contributing factor to increased rates of aggregation [10, 23]. It is therefore essential that emerging polypeptides either be protected from their surroundings until they can fold into their stable native state, or fold so quickly that they are unsusceptible to them.

As a protein chain emerges from the ribosome it will automatically try to find its lowest energy conformation and thus as mentioned, in some cases folding will occur co-translationally with N-terminal domains beginning to fold before the chain is completely synthesized [6, 17]. It should be noted however, that cotranslational folding is much more prevalent in eukaryotic cells due to the higher percentage of proteins containing multiple domains (a domain being a three-dimensional part of the protein structure which oftentimes can be folded and stable independently), and this type of folding is thought to have evolved along with the evolution of multi-domain proteins [6]. Bacterial translation is much faster than in eukaryotes (5 to 10 times) and this is hypothesized to be a constraint on co-translational folding, especially for slow-folding domains, therefore large multi-domain proteins which require cotranslational folding are likely to misfold and aggregate in *E. coli* [6]. Until all domains can interact properly to form the final native structure, hydrophobic surfaces that want to be buried between domains or within the core of the protein are often exposed while waiting for the completion of synthesis. The longer these surfaces are exposed, the higher the chance of aggregation between neighbouring polypeptides. In recombinant protein production this problem is magnified by the fact that the heterologous protein is the predominant polypeptide being translated and it has been hypothesized that aggregation is specific and more likely to occur between identical chains [10]. With these factors in mind it is easy to see why the probability of expressing and crystallizing single domain proteins is much higher than that of multi-domain proteins [1]. Bacterial cells are not as well

equipped as eukaryotic cells to produce massive multi-domain proteins, especially ones requiring extensive post-translational modifications such as disulfide bond formation. To assist folding in these cases the cell relies on molecular chaperones.

Molecular chaperones make up the cellular machinery whose purpose is to promote proper folding, refold partially unfolded proteins, dissolve aggregates and decompose irreversibly denatured peptides [3]. Many chaperones are labeled “heat shock proteins” (HSPs) because they are upregulated during stressful situations which lead to the accumulation of misfolded protein such as an increase in temperature or the over-expression of recombinant protein [24, 25]. Chaperones can be classified under three main types: holding, folding, and unfolding [26]. Generally they recognize hydrophobic residues or unstructured backbone regions as substrates, however different chaperones will interact with a chain during different stages of the folding process. Approximately 10 to 20 % of *E. coli* proteins will interact with the ribosome associated holding chaperone trigger factor (TF) for protection and to prevent premature folding [3, 27]. TF can bind to chains as short as 57 residues [28], however its targets are predominantly large multidomain proteins over 60 kDa [3]. Longer nascent chains may interact with DnaK and its cochaperones DnaJ and GrpE, although the substrate pools for DnaK and trigger factor do overlap [29]. DnaK targets peptides averaging 7 residues in length that are hydrophobic in their central region and have basic residues in the flanking region [30]. On average a region with these characteristics arises every 36 residues and is usually associated with buried  $\beta$ -strands in the native structure [30]. Proteins that require an isolated area away from the cytosol to fold properly (10 – 15% of newly synthesized *E. coli* proteins) must interact with the large GroEL-GroES chaperonin complex. This complex provides a sanctuary within its structure in which proteins can fold while protected from the cytosol [3, 24, 28]. A single peptide may need to interact with a chaperone (such as DnaK) or chaperonin (GroEL-GroES) several times during the folding process, or may be simultaneously be interacting with several of them at once [28]. For proteins

which have reached their native state but then become unfolded due to environmental stress, the small holdases IbpA and IbpB will bind to the partially unfolded proteins until the stress subsides and then pass them along to DnaK [3]. If all of the above methods fail and a protein does aggregate, a last-ditch effort in the form of ClpB, an ATPase of the Hsp100 family, will try to dissolve aggregates and transfer these proteins back to DnaK [3].

As a result of the increased understanding of the mechanisms behind protein misfolding, inclusion body formation and molecular chaperones, effective strategies for improving the folding and solubility of recombinant proteins in the cytosol of *E. coli* have been developed. The next section will cover some general strategies that have been proven effective in some cases towards this goal.

#### **1.4 Methods for enhancing recombinant protein solubility in *E. coli***

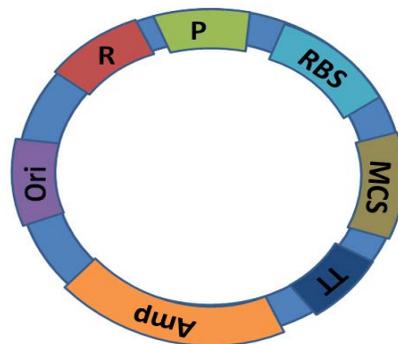
When attempting to optimize the soluble expression of recombinant proteins several factors should be considered, and these factors can be divided into two classes: Factors intrinsic to the protein and factors extrinsic to the protein. Extrinsic factors are those which alter the conditions around protein folding without altering the protein itself, including but not limited to: promoter strength (the efficiency with which mRNA is transcribed), culturing temperature, fusion partners and molecular chaperones. Intrinsic factors will alter either the nucleotide sequence or the amino acid sequence of the protein, and can include altering codon usage or protein engineering via rational mutation or irrational mutation.

##### **1.4.1 Extrinsic factors affecting protein solubility**

Slowing the production of protein is a common strategy for improving the solubility of recombinant proteins [31, 32]. The tendency for a protein to form inclusion bodies (defined here as insoluble aggregates of nonnative proteins [3]) is almost totally a consequence of overproduction and cannot be directly correlated to the size of the protein, relative

hydrophobicity, cysteine fraction, or subunit structure [33]. Production rate will be dependent on several factors including plasmid copy number, promoter strength, mRNA stability, and how efficiently translation is initiated. For most cases, these factors are determined by the choice of expression vector.

One of the benefits of using *E. coli* as an expression host is the vast array of compatible vectors to choose from. The majority are designed with the following features, outlined in **Figure 1-2**: a regulatory gene upstream from a promoter region, a ribosome binding site, a multiple cloning site, transcriptional and translational regions, an antibiotic resistance gene, and an origin of replication [2, 34, 35]. **Table 1-1** outlines the essential function of each feature and how it can affect protein solubility or expression. Generally the ribosome binding site, transcriptional terminators and stop codons are closely related between the various vectors and therefore are not the focus when deciding which vector to use. In terms of optimizing features to improve solubility, the primary targets will be plasmid copy number and promoter strength, but many vectors boast additional features targeted at improving solubility such as genes encoding N- or C-terminal fusion partners, or genes encoding chaperones.



**Figure 1-2.** Basic components of an *E. coli* expression vector: origin of replication (Ori), regulatory gene (R), promoter (P), ribosome binding site (RBS), multiple cloning site (MCS), transcriptional and translational termination region (TT), and antibiotic resistance gene (Amp), in this case a gene encoding a  $\beta$ -lactamase to confer resistance to ampicillin. Figure adopted from [35].

---

**Table 1-1:** Vector considerations for protein solubility

---

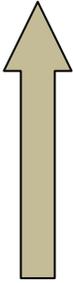
Vector Feature	Effect on Solubility and Expression
Origin of replication and antibiotic resistance gene	<ul style="list-style-type: none"><li>• Dictates copy number of vector and therefore gene dosage. A lower copy number vector may be used to slow protein expression, however for efficient protein production it is essential that all daughter cells maintain at least one copy of the vector. For this purpose an antibiotic resistance gene is a useful way to confer survival only to cells carrying the plasmid.</li></ul>
Regulatory gene	<ul style="list-style-type: none"><li>• Controls rate of transcription. Regulatory genes that can be gradually induced and keep basal transcription to a minimum are ideal as it is beneficial to have control over the rate of transcription so protein production is slowed. Differences in reg. genes: eg lacI vs arabinose (leaky vs strict control of transcription).</li></ul>
Promoter	<ul style="list-style-type: none"><li>• Strong promoters will cause rapid transcription of mRNA leading to high levels of protein. For proteins highly prone to aggregation, a weaker promoter may be better.</li></ul>
Ribosome binding site	<ul style="list-style-type: none"><li>• Initiation of translation can affect expression levels. Sequences at the 5' end of mRNA are critical in determining the efficiency of initiation of translation. The Shine-Dalgarno (SD) sequence and its spacing from the AUG initiation codon can be enhanced to promote efficient translation [36].</li></ul>
Transcriptional and translational termination sites	<ul style="list-style-type: none"><li>• A transcriptional terminator downstream from the coding region ensures plasmid stability by preventing transcription through the origin of replication. It also stabilizes mRNA by forming a stem loop at the 3' end. Translation is terminated by the presence of a stop codon. <i>E. coli</i> prefers the UAA codon however vectors sometimes have three consecutive stop codons to ensure translation is ceased [34].</li></ul>

---

The promoter region of the vector will have a direct impact on the solubility of recombinant proteins as it dictates the efficiency at which mRNA is transcribed. Strong promoters will bind RNA polymerase more strongly than weak ones, resulting in more successful initiations of transcription. Strength of the promoter is dependent on three main elements: a region 10 base pairs upstream from the start of transcription, a spacer, and another region 35 base pairs upstream from the start of transcription. Statistical analysis of over 300 *E. coli* promoter regions unveiled a consensus for these elements [37] and generally the strength or weakness of a promoter is determined by how closely its sequence matches this consensus

sequence, with some exceptions [38]. For instance, the *tac* promoter is considered to be very strong relative to other promoters derived from *E. coli*, and it differs from the consensus sequence only by the length of the spacer. The closely related *trc* promoter matches the consensus exactly but is 90% as active as the *tac* promoter. Under the control of these promoters recombinant protein levels may reach 15 -30% of the total cellular protein [29]. **Table 1-2** shows the *E. coli* consensus sequence [37] and other promoter sequences derived from *E. coli* which are commonly used in expression vectors.

**Table 1-2:** Promoter sequences in *E. coli*

Promoter	-10 Region	Spacer Length	-35 Region	Strength
Consensus	TATAAT	17	TTGACA	
<i>tac</i>	TATAAT	17	TTGACA	
<i>trc</i>	TATAAT	16	TTGACA	
<i>trp</i>	TAACTA	18	TTGACA	
lacUV5	TATAAT	18	TTTACA	
<i>lac</i>	TATGTT	18	TTTACA	

Another promoter that is commonly used but not native to *E. coli* is the T7 promoter which is the cornerstone of the popular series of pET expression vectors (Novagen). This promoter is derived from a bacteriophage and thus *E. coli* RNA polymerases will not recognize it. For transcription to occur, a plasmid bearing the T7 RNA polymerase gene or a host strain lysogenized with this gene must be used in conjunction. Because of its viral origin, the T7 promoter and T7 RNA polymerase system is very strong and fast, capable of rates of transcription in the range of 230 nucleotides per second [2]. This is approximately five times faster than *E.*

*coli* RNA polymerase which is capable of transcribing at a rate of about 50 nucleotides per second [2]. Under the T7 promoter a target protein can reach yields of 40 – 50% of total cellular protein [29], and for small, fast-folding single-domain proteins this is ideal. This extreme overproduction of protein will be deleterious if the target protein is prone to aggregation or if it becomes too much of a burden on the cell, leading to cell death [29, 39]. Because of these observations a primary strategy to reduce inclusion body formation and improve protein solubility is to use a weaker promoter or a strong one which can be gradually induced, to slow down the production of protein [39, 40].

In addition to vector considerations, another common way to slow the production of protein is to lower the growth temperature of the culture [32]. Lowering the temperature of the culture not only slows the production of protein via decrease in the rates of translation and transcription [41] but it also decreases the rate of aggregation and inclusion body formation [23], alters the kinetics of folding [36], and diminishes protease activity [26]. Taking it a step further, Mujacic and coworkers developed a vector utilizing the promoter for the cold-shock protein CspA for expression of toxic and proteolytically sensitive proteins [42]. Expression is induced by lowering the temperature of the culture to 15 or 23 °C and is well repressed at and above 37 °C. Using this vector the authors successfully expressed a TolAI- $\beta$ -lactamase fusion protein which was toxic and highly unstable when expressed under the T7 promoter at 37°C.

Another way to alter the environment *in vivo* to favour proper folding is to fuse recombinant proteins to a partner that expresses well, folds efficiently, and is highly stable. A number of fusion partners have been shown to effectively increase soluble recombinant protein expression in *E. coli*, including: glutathione-S-transferase (GST), maltose binding protein (MBP), thioredoxin, N-utilizing substance A (NusA), and small ubiquitin-related modifier (SUMO) [27, 29, 43, 44]. The mechanism behind how the enhancement of solubility occurs is still unclear but

it is hypothesized that in the case of MBP, the fusion protein acts as a chaperone and shields its partner from other nascent chains, thereby preventing aggregation [45]. It has also been suggested that since MBP itself requires chaperones to fold, it may effectively recruit chaperones into the vicinity of the passenger protein [29]. In a comparison study of the solubility enhancing potential of GST, MBP and thioredoxin, MBP was found to be the most effective at not just acting as a solubilizing agent, but increasing the amount of protein that reaches its biologically active native state [45]. A later study determined that both NusA and SUMO were more successful than MBP at enhancing expression and solubility of recombinant protein, however the authors felt SUMO was the overall best fusion protein because it has the attractive feature of having its own natural protease (SUMO protease) and therefore no protease site needs to be incorporated into the fusion construct [43].

Just as the chaperone-like characteristics of fusion proteins enhance the solubility of their partners, coexpression of molecular chaperones has also been an effective strategy to increase yields of soluble protein. As mentioned, certain chaperones will interact with a folding polypeptide at different stages, so the type of chaperone chosen for overexpression can be critical for the success of the experiment. Limited success has been achieved by the overexpression of individual chaperones, for example human growth hormone (HGH) showed a significant decrease in inclusion body formation and aggregation when expressed in the presence of elevated levels of DnaK but no effect was observed when GroESL was used in place of DnaK [46]. Similarly several examples of improved protein production as a consequence of co-overexpression of GroESL alone are in the literature (reviewed in [47]). More recently it was reported that to increase chances of success by overexpressing chaperones, more than one chaperone must be overexpressed at a time. De Marco *et al.* showed that coordinately overproducing four chaperones systems (DnaK/DnaJ, GroEL/GroES, ClpB, IbpA/IbpB) along with a recombinant

protein resulted in an increase in solubility for 70% of the proteins they studied (64 of them), with some showing an improvement in solubility of up to 42-fold [48].

In spite of the success of the techniques discussed above, some recombinant proteins simply cannot be expressed in a soluble form in the cytoplasm of *E. coli*. Deposition into inclusion bodies may be unavoidable, but can also be advantageous as it can be a route to ultra-pure protein. Inclusion bodies are highly homogeneous, with the recombinant protein comprising up to 90% of inclusion body material [49]. By resolubilizing the protein in a denaturant such as urea or guanidinium chloride and then refolding it by slowly removing the denaturant, one can recover properly folded, highly pure soluble proteins. Refolding yield is quite variable, however, and unfortunately usually as low as 15 to 25 % of the total protein that was in the inclusion body [50]. This yield will be dependent on many factors such as temperature and composition of refolding buffer [51]. Recently, it was reported that since inclusion bodies can contain a high amount of native secondary structure, dissolving under mild conditions to preserve as much of this structure as possible will lead to higher yields of biologically active protein [50].

The above techniques, although useful, need to be optimized for every target and this can be time consuming and costly. Another drawback is that even when an increase in the yield of soluble protein is observed in the cell lysate, the protein may still precipitate during purification and workup [52]. This is due to a fundamental problem with all of these methods: the intrinsic folding and stability of the protein remains unchanged. In essence, the protein has been pampered into folding correctly, but an overall increase in stability has not been achieved. The only way to alter a protein's fundamental ability to reach and remain in a native conformation is to alter the ultimate deciding factor of whether or not it will reach this state—the amino acid sequence.

### 1.4.2 Intrinsic factors affecting protein solubility

Intrinsic properties that affect protein solubility arise from both the amino acid sequence of the protein and its encoding gene. A gene's codon usage is highly specialized for an organism, with some organisms preferring certain codons for a specific amino acid over others, leading to variable levels of the available tRNAs. For instance, in *E. coli* the occurrence of some codons are very rare (less than 1%), as summarized in **Table 1-3**, and the presence of these codons in a heterologous protein can dramatically slow the rate of translation (up to 6-fold) [53]. Slowing of translation may or may not be a bad thing since it is hypothesized that there are regions of mRNA encoding protein domain boundaries which are “translationally slow”, and these may assist co-translational folding of individual domains [54]. On the other hand, a high abundance of rare codons can cause major problems such as frameshifting, hopping, and premature termination of translation [55]. To alleviate this problem the tRNAs for rare codons can be co-transcribed or alternatively the rare codons can be mutated to codons that more commonly used in the host organism either by site-directed mutagenesis or by entire gene synthesis [56]. Codon optimization has improved the yield of many proteins and the examples of this are nicely summarized in [56].

---

**Table 1-3** Rare codon usage in *E. coli*

---

Rare codons	Encoded amino acid	Frequency per 1000 codons
AGG/AGA	Arg	1.4/2.1
CGA	Arg	3.1
CUA	Leu	3.2
AUA	Ile	4.1
CCC	Pro	4.3
CGG	Arg	4.6
UGU	Cys	4.7
UGC	Cys	6.1
ACA	Thr	6.5
CCU	Pro	6.6
UCA	Ser	6.8
GGA	Gly	7.0
AGU	Ser	7.2
UCG	Ser	7.8
CCA	Pro	8.2
UCC	Ser	9.4
GGG	Gly	9.7
CUC	Leu	9.9

---

Adopted from [34].

Moving up a level and altering a protein's amino acid sequence to improve folding and stability is not as straightforward. Even when structural information is known for a target, predicting which residues to mutate to increase stability is often not successful, as oftentimes stabilizing mutations occur in unexpected places [57]. Success through rationalized mutagenesis

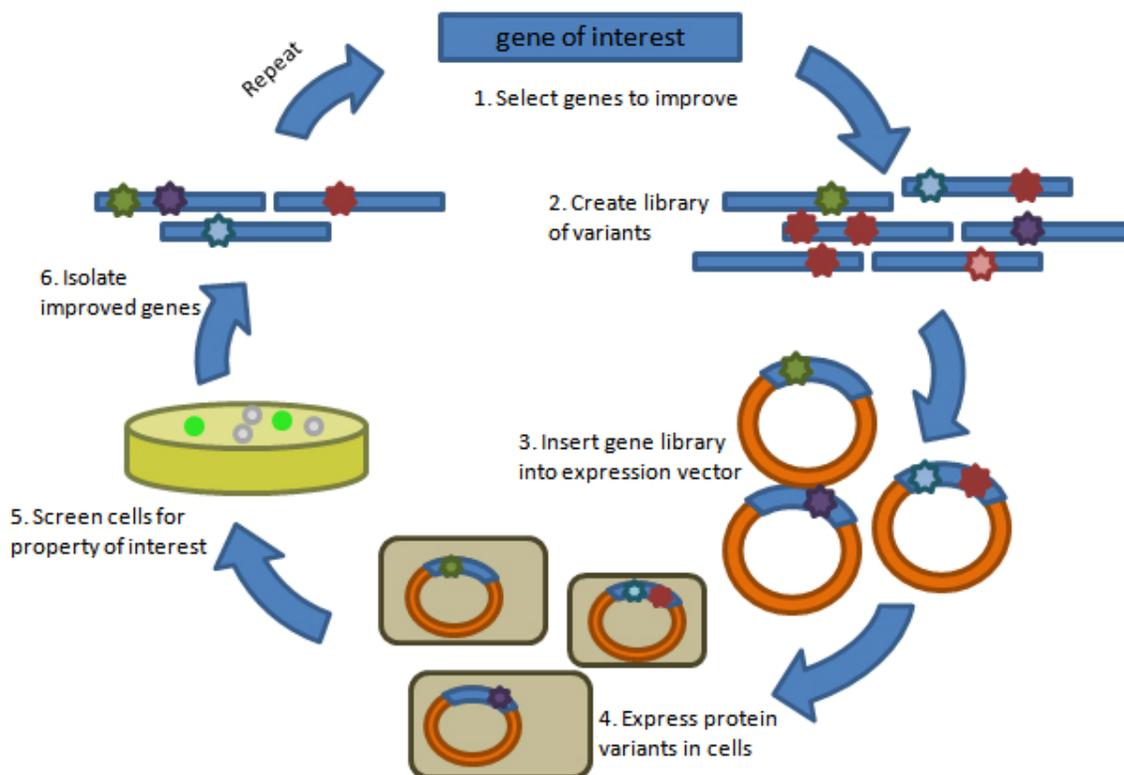
has been achieved with stabilization or rigidification via mutation to proline or incorporation of disulfide bonds, and also with helix optimizations, construction of salt bridges, and introduction of stabilizing aromatic interactions (reviewed in [58]). The improvement of software and bioinformatics has also led to the rational optimization of key stabilizing residues, either via sequence alignment software which allows identification of highly conserved “consensus” residues important for stability [59, 60], and/or through computer modeling software that tries to predict the thermodynamic effects of point mutations.

One of the main reasons that structurally-based rational design does not have a high success rate is because the rationalizations are based on features of the final native structure. In many cases it is the kinetic and thermodynamics of intermediate states that are of key importance in proper folding, and mutations affecting these states, perhaps by disfavouring off-pathway species, cannot be predicted by only looking at the end product [61]. To overcome this obstacle another method, termed directed evolution, can be used as it requires no knowledge of structure, function, or folding pathway. Certainly many highly interesting targets have no known function or structure because of the difficulty of getting enough soluble protein to work with. Proteins such as this are prime targets for directed evolution experiments.

### **1.5 Directed evolution as a strategy to improve protein solubility**

Directed evolution has become a powerful way of altering enzymes to become highly functional outside of their normal biological contexts. **Figure 1-3** outlines the basic steps involved in the laboratory evolution of enzymes. Once a target gene has been identified and cloned into an appropriate expression vector it is diversified through mutagenesis or recombination. This generates a pool of mutant genes that are subsequently cloned back into the expression vector and the resultant library is expressed upon transformation of bacterial cells. Selection or screening for the desired trait can then occur *in vivo* or *in vitro* depending on the trait

being improved, after which the genes encoding the improved variants are used as the parents for the next round of evolution. This cycle is repeated as many times as is necessary to achieve the desired result.



**Figure 1-3.** Basic steps involved in laboratory evolution of proteins. Adopted from [26].

The applications of this methodology are very broad and useful to many fields of research from pharmaceutical development to agriculture [62]. Laboratory evolution of enzymes has been used both to design new protein functions [63] and improve or alter existing ones such as enantioselectivity, substrate specificity, catalytic rate, thermostability and resistance to organic solvents. The field is very broad and has been reviewed several times [62-65]. This section will review diversification and selection/screening methods commonly used to evolve proteins with improved folding and/or solubility upon overexpression in *E. coli*, as well as some examples of

successful evolution experiments which have led to an improvement in the solubility of a heterologous protein.

### **1.5.1 Diversification**

Whereas screening or selections must be carefully designed for each evolution experiment, strategies for generating library diversity are generally applicable. These methods can be classified as truncation methods, random mutagenesis methods, or combinatorial methods and often a combination is used. Whatever method is used, it is essential that the sequence space is efficiently explored as the number of distinct variants in a library will always vastly outnumber the actual variants that can be selected or screened due to the limitations of current experimental methods [66]. To improve the chances of finding the ‘needle in the haystack’, it is essential that that library you are selecting from is of high quality (has a minimal amount of non-protein coding or “junk” DNA sequences) and not only a sufficient amount of diversity, but also the right type of diversity such that all potential beneficial mutations are accessible using the given protocol. Due to the biases inherent to random mutagenesis protocols and the redundancy of the genetic code, an average of only 3.14 to 7.40% of amino acid substitutions can be achieved per residue for a given protocol [67]. An interesting idea put forward by Tawfik et al. suggests that one way to maximize the amount of diversity that can be effectively screened is to mutate key residues back to a pre-determined consensus sequence prior to the start of diversification [68]. Back-to-consensus/ancestor mutations in TEM-1 $\beta$ -lactamase led to an increase in its thermodynamic and kinetic stability, thus endowing a greater tolerance for a broad range of deleterious mutations. The ability to withstand the destabilizing effects of diversification can make a protein more amenable to evolution [69, 70]. The next sections will discuss common diversification techniques used to generate libraries for directed evolution.

### **1.5.1.1 Truncation and fragmentation methods**

If the protein you are trying to evolve is very large or has many domains, it is sometimes useful to truncate or fragment the gene in an effort to find the soluble portions or domains [71]. Although it is not ‘diversification’ per se, it is proven to be an effective means of finding soluble portions of insoluble target proteins. Sometimes domain boundaries can be predicted ahead of time using sequence alignments and bioinformatic tools, however domain boundaries can have notoriously low sequence homology, making the task of predicting them difficult [72]. Alternatively, one can generate libraries of randomly truncated or fragmented versions of the gene and screen for soluble variants with a combinatorial approach. Several methods are available for randomly generating truncation and fragment libraries, including: enzymatic digestion of the gene with a non-specific enzyme such as exonuclease III or DNaseI, physical fragmentation via sonication or hydrodynamic shearing, combinatorial domain hunting, and tagged-PCR [72]. The last two methods are PCR-based.

With combinatorial domain hunting, a standard PCR is performed on the target gene with the regular dTTP nucleotide replaced with a dTTP/dUTP mixture [73]. A low-fidelity polymerase such as Taq will randomly incorporate dUTP along the gene in the final PCR product. The integrated dUTPs are then excised by uracil-DNA glycosylase generating abasic sites which are subsequently cleaved by endonuclease IV to generate single strand nicks in the DNA. Treatment with S1 nuclease will turn the single strand nicks into double strand breaks and this library of blunt-ended PCR products can be ligated directly into the screening vector.

Tagged-PCR generates fragments of random lengths by means of two subsequent PCR reactions. The primers for the first PCR contain defined 5'- sequences of about 15-20 bp (but not complementary to the target sequence) and random 3'- sequences of about 5-15 bp. During the first PCR the random 3'-ends anneal to potentially every possible position on the target gene,

allowing the polymerase to copy the template from various starting points [74]. The small random fragments are then amplified in a second PCR using two primers that match only the specific 5'-sequences of the first set of primers, generating PCR products that can be digested and ligated into an appropriate screening vector.

Non-random methods for generating truncated constructs are also possible, such as 'primer pair walking' in which multiple PCR reactions are performed with primers that are designed to anneal in various places along the gene. This method is advantageous in that all constructs will be in-frame and the identity of any positive clones will be known immediately. The major disadvantage lies in the high number of PCRs that need to be performed, making this strategy not amenable to high-throughput applications [72]. It should also be noted that truncation and fragmentation methods are not considered to be "evolutionary" methods as mutation or recombination is not the end goal. To attempt to improve the solubility of a target as a whole, random mutagenesis and/or recombination will be the main routes of diversification.

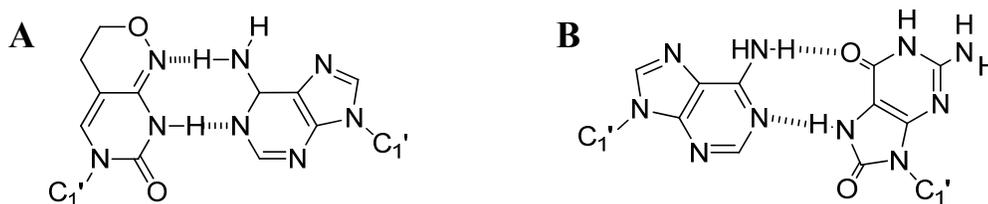
#### **1.5.1.2 Random mutagenesis methods**

The most popular methods for introducing random point mutations along genes are usually PCR based, but other methods involving physical or chemical mutagens have been used. UV irradiation and alkylating agents act by damaging DNA, causing it to be incorrectly replicated or repaired [75]. Another PCR-free method is to use mutator host strains in which the DNA repair pathways are disrupted, leading to vastly higher mutation rate compared to normal strains [76]. A drawback with these strategies is that they are non-specific—all DNA contained in the subjected cells will suffer damage, including chromosomal DNA. These processes can also be very slow, sometimes needing several passages through the hosts to incorporate one or two mutations per gene. For these reasons error-prone PCR (epPCR) has become the method of choice for most labs.

EpPCR is highly popular due to its simplicity. Taking advantage of the already low fidelity of *Taq* polymerase, a high rate of mutation is achieved by replacing the normally used  $Mg^{2+}$  cofactor with  $Mn^{2+}$  and upsetting the balance of the bases [77]. The level of mutation can be controlled by altering  $Mn^{2+}$  concentration, the number of cycles in the PCR reaction, or increasing the overall  $Mg^{2+}$  concentration. As mentioned before, a major problem with PCR-based mutagenic methods is that they are not unbiased in the types of mutations that can occur, and all potential mutations are not equally represented in the library [71, 75]. A library that is completely unbiased would mean that each amino acid could be substituted with any of the 19 others with equal probability. EpPCR accesses only 34% of this diversity [67]. This bias is a result of several factors, one being that in general, transition-type misincorporations (purine to purine or pyrimidine to pyrimidine) are highly favoured over transversions (purine to pyrimidine or vice versa) [67, 78]. Even though there are twice as many transversions possible than transitions, most epPCR methods that use *Taq* polymerase,  $Mn^{2+}$ , and unbalanced bases, have transition biases of up to 80% [67].

One method to try and even out the transversions to transitions ratio is to use base pair analogues such as the triphosphate derivatives of 6-(2-deoxy- $\beta$ -D-ribofuranosyl)-3,4-dihydro-8H-pyrimido-[4,5-C][1,2]oxazin-7-one (dPTP) or 8-oxo-2'-deoxyguanosine (d8-oxoGTP). In one study when dPTP was combined with the other dNTPs at equimolar amounts, A $\rightarrow$ G transitions accounted for 46.6 % of the mutations and T $\rightarrow$ C transitions accounted for 35.5% of the total mutations. As substrates for *Taq* polymerase dPTP is closest in kinetic properties to dTTP in terms of efficiency of incorporation, thus the transitions that occurred arose from dPTP base-pairing with dATP on either strand and then subsequent pairing of dGTP with the incorporated dPTP. Using d8-oxoGTP resulted in a majority of transversions, with A  $\rightarrow$  C accounting for 39 % and T $\rightarrow$ G accounting for 59 % of the total mutations. These mutations are also thought to occur upon misincorporation of the analogous base pair opposite of A. The base pairing of P with

A is shown in **Figure 1-4A** and 8-oxoG with A is shown in **Figure 1-4B**. Although both base pair analogs can replace T and base pair with A, dPTP has much faster kinetics than d8-oxodGTP and if used in equimolar amounts, dPTP will incorporate more often. To achieve an optimal transition to transversion ratio, the authors suggest that these two base pair analogs be used in conjunction, but their relative concentrations need to be adjusted to compensate for their differing kinetics [78].



**Figure 1-4.** Base analogs P (A) and 8-oxoG (B) pairing with A. Adopted from [78].

A similar strategy using base pair analogs is transversion-enriched sequence saturation mutagenesis (SeSam-Tv<sup>+</sup>) [79]. In the same way as above, it complements the transition bias of epPCR by incorporation of nucleotide analogs each with unique base pairing properties, making the type of base pairing highly tunable. The transition to transversion bias can be easily overcome by specifically choosing the types of analogs, and therefore the fraction and types of transversions for each gene. Another advantage with this method is the occurrence of consecutive nucleotide mutations in up to 16.7 % of the final gene pool.

Consecutive nucleotide changes are critical for conversion to occur between chemically diverse amino acids [67]. This is a type of bias that arises simply from the nature of the genetic code and the degeneracy of certain codons. The physiochemical and ambiguity reduction theory proposes that the majority of amino acids which differ by only one codon are chemically similar, thus buffering the potentially harmful effects of mutation [80]. A valine residue can be mutated

to a phenylalanine, leucine, isoleucine, alanine, aspartate, or glycine by a single point mutation, but to convert to any other amino acids two or three mutations are needed [75].

Obviously the codon bias is not something that can be altered but care can be taken when choosing a mutagenic protocol such that the bias of the mutagenic strategy complements the codon bias in a way to achieve the correct diversity. In the statistical analysis of 19 different random mutagenesis protocols it was found that transition-biased methods had a *lower* probability of introducing stop codons and helix-disrupting mutations (conversion to glycine or proline), therefore reducing the number of useless sequences generated [81]. This program, named mutagenesis assistant program (MAP, publicly available at <http://map.iu-bremen.de/MAP.html>) [67] could be highly beneficial when deciding which random mutagenesis protocol to use as the data will help determine, based on the nucleotide and amino acid sequences, what sort of codon biases will occur and which mutagenesis method is best to counteract these biases.

The easiest way to overcome any bias is to selectively maximize the diversity of only a few targeted residues in a semi-rational fashion. If important residues have been identified either through computational analysis or enrichment via DNA shuffling (more about this technique below), these positions in the amino acid sequence can be subjected to maximum diversity by the use of synthetic nucleotides. Synthetic oligonucleotides are short pieces of DNA which have one randomized codon flanked by two regions which anneal to the template sequence. These semi-random primers can be added to the fragment mix in a DNA shuffling reaction, and in this way a specific residue has a greater chance of being converted to any of the other 19 amino acids [82]. To ensure all 20 amino acids have an equal chance of incorporation at a specific position, 20 separate primers can be synthesized and used as a mixture [75].

One final issue worth mentioning with PCR-based random mutagenesis methods is the amplification bias which arises from the exponential nature of the PCR itself. Mutations that are

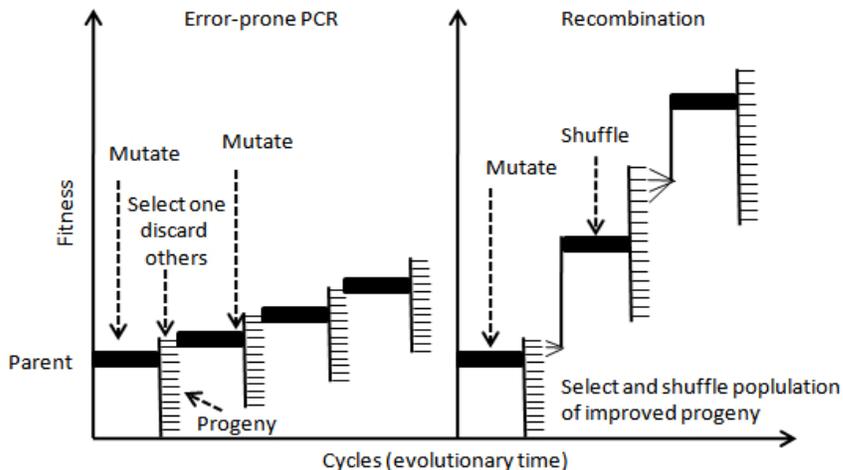
acquired early in the reaction will be over-represented in the final library and this poses a significant problem if these mutations are deleterious. One way to counteract this bias is to perform several sequential PCRs each with a reduced number of cycles [75]. For a detailed review of random mutagenesis methods, their biases, and how to choose an appropriate method prior to beginning an evolution experiment, see Ref [81].

### **1.5.1.3 Recombination methods**

In nature, evolution occurs not only as a result of acquiring mutations over time, but also from the bringing together of beneficial mutations and removal of deleterious ones through sexual recombination of genetic material. In 1994 Willem Stemmer was the first to introduce a method to mimic this powerful aspect of natural evolution and called it DNA shuffling [83, 84]. In his paper he describes a protocol in which a pool of homologous genes is digested into 10 – 50 base pair fragments and then recombined via a PCR in which they self-prime to produce full length genes comprising fragments from various parental genes. This reassembly step will produce a streak of DNA on an agarose gel indicating a mixture of various gene lengths is present. To achieve a single product of the correct size, this mixture is used as the template for a final PCR in which primers flanking the gene of interest are used to amplify only the product of the correct size. He successfully used this method to recombine homologous genes from the same family of enzymes as well as recombine a pool of genes differing by point mutations. As proof of principle he used this method to evolve TEM-1, and increased the minimum inhibitory concentration (MIC) of this  $\beta$ -lactamase 32, 000-fold against cefotaxime—from 0.02  $\mu\text{g}/\text{mL}$  to 640  $\mu\text{g}/\text{mL}$  [84].

Several advantages of recombination over recursive mutagenesis strategies have become apparent. Stemmer likens the differing strategies to “editing a manuscript by changing individual letters rather than by moving blocks of letters, words and sentences around” [85]. He even demonstrated this by taking the same wild type  $\beta$ -lactamase which he improved 32,000-fold using

DNA shuffling and evolving it via 3 rounds of epPCR. This resulted in an only 16-fold increase in the MIC over wild-type [84]. This jump in efficiency is partly a result of the ability to remove deleterious mutations at each recombination step, rather carrying them through and allowing them to accumulate. In this way optimal phenotypes can be reached much faster, as shown in **Figure 1-5** [4, 86, 87].



**Figure 1-5.** A graphical comparison of phenotypic optimization achieved by random mutagenesis methods and recombination methods. Adopted from [4].

In general, with recursive mutagenesis strategies generation of beneficial mutations is low compared to the incorporation of deleterious mutations [62] and it has even been estimated that 70 – 99% of a starting library generated by epPCR is made up of nonfunctional variants [87]. Each beneficial mutation must be found in a sequential step, and thus cannot be combined unless acquired one after the other. On the other hand, recombination strategies allow for the accumulation of several beneficial mutations and simultaneous expulsion of deleterious mutations in a single round, thus making leaps and bounds over a point-mutagenesis strategy. The removal of deleterious mutations can further be removed by shuffling the DNA with wild-type DNA in the final rounds of diversification.

Since the 1994 Stemmer article, DNA shuffling has taken the molecular evolution world by storm and several optimizations and new recombination procedures have been developed. One strategy, developed by Lorimer and Pastan employs  $Mn^{2+}$  as the metal ion cofactor in place of the normal  $Mg^{2+}$  during the DNaseI digestion step [88]. The advantage of using  $Mn^{2+}$  as the cofactor is that it will make double strand cuts on template DNA rather than single strand nicks. For isolated double-strand fragments less than 50 bp in length, single strand nicks can be deleterious as these fragments will separate upon denaturation in the first step of a PCR and produce single-stranded DNA much shorter than 50 bp. If this occurs they are no longer long enough to act as primers for *Taq* polymerase and are useless in a recombination step. Using  $Mn^{2+}$  ensures all fragments remain the same size upon denaturation and eliminates the need for gel purification [88]. This method was successfully used to recombine antibody single chain Fv sequences with fewer steps and a much lower rate of point mutations compared to the original shuffling protocol (0.2 % vs. 0.7%). To further reduce the occurrence of point mutations acquired during DNA shuffling Zhao and Arnold developed a protocol where the proof-reading polymerase *Pfu* was used in place of the normally used *Taq* polymerase during the recombination and amplification steps. This lowered the rate of point mutations to 0.05 % [89].

Arnold and coworkers also designed an alternate strategy to generate and combine DNA fragments using only PCR called staggered extension process (StEP) [90]. Rather than using a DNaseI digest to generate fragments, a PCR reaction in which the elongation time is very short and/or done at low temperatures is used. The starting gene pool in the PCR is combined with a single flanking primer to initiate the extension process. The extension is only given a few seconds to occur so that it is stopped before a full-length gene is made. During the next cycle this unfinished gene will separate from its template upon denaturation and anneal to a different parent before elongating again for a brief period. Thus when a full-length gene is made it is a result of copying several different parents. The number of recombination events that can occur is strictly

dependent on the elongation conditions thus several optimizing attempts may need to be made for each gene being evolved.

DNA shuffling and StEP both require that the genes being recombined have considerable homology at the points of cross-over (> 60%), thus limiting the sharing of genetic information to very closely related families of proteins. Computer simulations have shown that in the quest to create new protein folds, non-homologous recombination has a higher chance of success than iterative mutagenesis or homologous recombination [91]. A method deemed incremental truncation for the creation of hybrid enzymes (ITCHY) is a way of combining two genes independent of DNA sequence homology [92, 93]. In this method, the two genes (gene A and gene B) that are being recombined are first digested in opposite directions using exonuclease III. This will result in two pools of genes: one pool with 5' fragments of gene A and another pool with 3' fragments of gene B. The digested genes are extended with Klenow polymerase to have blunt ends, then ligated together into an appropriate screening vector. Using this methodology the authors generated chimeras of an *E. coli* and human GAR transformylase and found that the most active mutant discovered was a result of a crossover in a non-homologous region [92]. One major disadvantage of this method is that only one crossover can be achieved per gene. To overcome this Lutz and coworkers devised a method they called SCRATCHY, a combination of ITCHY and DNA shuffling [94]. With this method incremental truncation libraries are created as in ITCHY, then chimeric genes are selected for functionality and correct length. DNA shuffling is then used to recombine the selected chimeras to generate genes with multiple crossovers which may have occurred in homologous or non-homologous regions.

This brief description of library diversification techniques is by no means exhaustive, but was simply an attempt to outline some of the considerations that need to be made prior to undertaking an evolution experiment and examples of techniques used to overcome some

obstacles. In terms of improving recombinant protein solubility, the end goal is not to evolve a novel function or new fold, but to tweak the native structure or intermediate states so as to encourage rapid folding and stability. In this case simple and straight-forward diversification techniques such as epPCR and DNA shuffling have been proven effective time and time again. The major hurdle in an evolution experiment aimed at improving solubility is how to effectively screen or select improved variants from large libraries with a low number of “false-positives.” The next section will discuss screening and selection techniques that have been developed to identify library variants exhibiting improved folding or stability.

### **1.5.2 Selection and screening strategies for evolving proteins with improved folding and solubility**

Although the level and types of diversity introduced into a library are important considerations, implementation of diversity is usually straight-forward. The task of selecting or screening for variants exhibiting improved folding or solubility is usually where the bottleneck occurs. Selecting or screening for an improvement in soluble heterologous protein expression in *E. coli* can be divided into two general categories: reporter-based methods where the activity of a reporter is being monitored, or function-based screens where the function of the protein in question is used for selection [61]. Selections (as opposed to screening) subject all members of the library to the same conditions simultaneously and are usually based on tying the function of a protein to its phenotype, whether it be a metabolic function required for cell survival, or binding to a ligand. Functional proteins are almost always folded proteins, therefore when a specific function is observed it's usually safe to say that the protein is in, or near, its native folded state. Obviously this strategy is only effective when a function is known. Another drawback is that this type of selection can only distinguish between folded or unfolded polypeptides and gives no indication of the degrees of thermodynamic and kinetic stability for the positive clones. Distinguishing between one protein that is stable (folded) and another highly homologous protein

that is slightly more stable (or folded) requires additional selection pressure to be exerted either in the form of increased temperature of expression or addition of denaturants [61, 66]. The main advantage of selection methods is that they allow much larger libraries to be searched ( $10^9$  variants or more) compared to screening methods ( $10^3$  -  $10^6$  variants) [66].

Phage display, mRNA display and ribosome display can be classified as function-based selections. The selections are performed *in vitro* via binding the protein of interest to an immobilized ligand. Phage display libraries are limited in size by a transfection step (introducing DNA into *E. coli* cells), however mRNA and ribosome display are performed entirely cell-free and thus offer the advantage of searching libraries with more than  $10^{13}$  members [95]. With phage display, the gene encoding the protein of interest is inserted into the genome of a bacteriophage such that it is then expressed and “displayed” on the surface of the phage along with a coat protein. The phage is produced in *E. coli* and released into the supernatant where it is collected by centrifugation. Phage particles displaying the protein are isolated by binding of the protein to an immobilized ligand—a process called “panning”. Because the encoding genetic material is harboured within the phage particle, it is relatively simple to isolate the sequences of the improved variants. Genotype to phenotype linkage is crucial for any selection or screening strategy so that only the sequences of improved variants are isolated for the next round.

Ribosome and mRNA display are very similar to phage display, with the main difference being that production of the protein is done entirely *in vitro*. To satisfy the criteria of genotype to phenotype linkage both methods involve linking a protein’s encoding mRNA to the protein itself. With mRNA display, the mRNA is covalently linked to the polypeptide via a puromycin linker—an antibiotic which resembles tRNA but contains an amide linkage and is therefore resistant to hydrolysis. The ribosome stops when it reaches puromycin, and the unfinished polypeptide is released along with its covalently linked mRNA. Alternatively, with ribosome display the

mRNA and expressed polypeptide are non-covalently linked. The encoding mRNA is fused to a spacer sequence that does not have a stop codon, causing the ribosome to stall at the end of the polypeptide sequence. The stability of the complex is further enhanced by the addition of a high concentration of  $Mg^{2+}$ . The resulting protein-ribosome-mRNA (PRM) complexes displaying the ability to bind to a target ligand are isolated via panning or affinity chromatography. The mRNA is separated from the selected complexes via disruption with EDTA and then the released mRNA is converted to cDNA by reverse-transcriptase PCR (RT-PCR) [96].

As mentioned, these function-based screens need to be combined with other selection pressures to achieve an increase in stability of the protein variants. For instance, during the panning stage, the library can be washed with increasing concentrations of a chaotropic agent, thereby denaturing the least stable variants at each stage. Another option is to subject libraries to increasing temperatures either during expression or *in vitro* translation. In this way, only those stable enough to withstand the denaturant or increased temperature will retain their structure and bind their ligand (not applicable to ribosome display, but certainly mRNA display and possibly phage display to a certain extent because phage particles are relatively tough). Another interesting strategy that involves phage display but does not rely on protein function to bind a ligand is ProSIDE (**P**rotein **S**tability **I**ncreased by **D**irected **E**volution), where the selection is based on a phage's ability to remain infectious after exposure to a protease [97]. Target genes are cloned between two domains of a minor coat protein (gene-3-protein) and the construct is subjected to increasing amounts of protease and only those stable enough to resist proteolysis will keep the coat protein's domains together, allowing the phage to remain infectious.

Solubility assays that do not require structural or functional data are necessary since many interesting targets that need characterization do not have this associated data. If this is the case, reporter proteins are the best way to relay information about how the target protein is

behaving *in vivo*. Reporters must have an easily monitored function and are either fused directly to the target (fusion reporter) or activated due to a cellular response invoked by the overproduction of the target (stress reporter) [52].

Stress reporter methods for determining an improvement in protein folding and solubility rely on the cellular response to protein misfolding. Misfolding and aggregation of proteins triggers a stress response in cells and causes heat shock proteins such as chaperones and other ribosome associated proteins to be up-regulated [24]. A study conducted by Lesley *et. al.* determined specifically which proteins were induced to a greater extent in cells expressing proteins which misfold compared to cells in which soluble proteins were being expressed [98]. Based on their results they cloned the promoter region for the heat shock protein IbpAB in front of the gene for a  $\beta$ -galactosidase reporter. Proteins that misfold lead to greater induction of this gene and the consequent expression and higher activity of  $\beta$ -galactosidase. With this system they identified a soluble N-terminal domain of the large and insoluble protein (Rep68) by generating a fragment library (via DNase digest) and screening via reduced  $\beta$ -galactosidase activity. Because this system is a positive indicator of *misfolding*, an additional screen is needed to verify solubility in the wells not showing activity. One advantage of this method over a fusion-reporter method is that it avoids possible perturbation of solubility by the reporter protein.

Fusion reporter methods rely on the ability of the target protein to directly affect the activity or function of the fused reporter protein. In terms of folding and solubility, a protein that is insoluble and fused to a reporter protein should render the reporter protein insoluble (and hopefully inactive) as well. Proteins or peptides used as reporters of protein solubility include (but are not limited to) green fluorescent protein (GFP), chloramphenicol acetyltransferase (CAT), and the lacZ $\alpha$  fragment of *E. coli*  $\beta$ -galactosidase.

CAT confers resistance to chloramphenicol and it was shown that fusing CAT to insoluble proteins results in the cell's inability to resist chloramphenicol [99]. Plating a library of variants fused to CAT on plates with increasing concentrations of chloramphenicol effectively selected hybrids of human cytochrome P450 and a bacterial P450 that were more soluble than the wild-type human form [100].

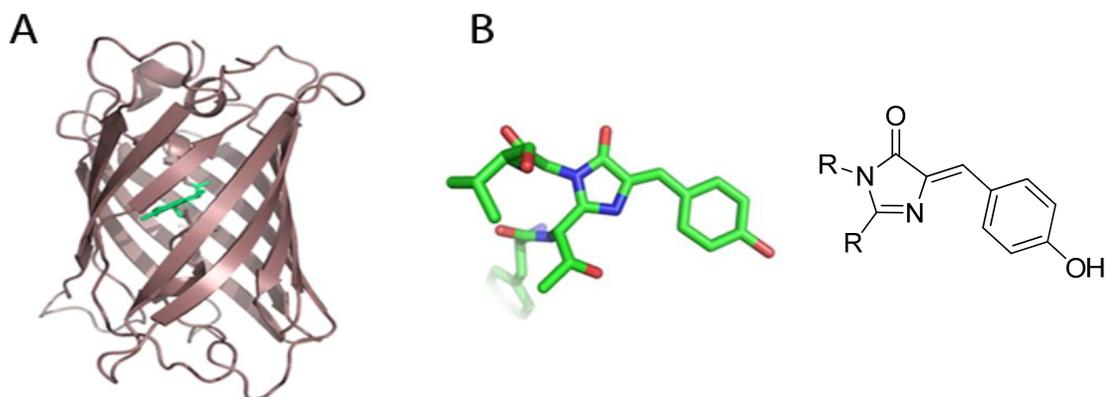
Fusion of a small 100 amino acid  $\alpha$  fragment of  $\beta$ -galactosidase ( $\text{lacZ}\alpha$ ) to a target protein is an example of a split-protein assay. Interaction of this small fragment with the much larger  $\omega$  fragment of  $\beta$ -galactosidase is critical for its activity, thus target protein aggregation will restrict the availability of the fused  $\alpha$  fragment to interact with the  $\omega$  fragment and activity will be diminished [101]. One advantage of this fusion partner is that because the  $\alpha$  fragment is a small peptide it is believed to affect the solubility of the target protein to a lesser degree than a full-protein fusion partner would. Also, because it's a split-reporter method, sensitivity is thought to be increased.

The most common fusion protein used to report folding *in vivo* is GFP. As GFP was the reporter of choice for this study, the next section will discuss the usage of GFP in this context in greater detail.

### **1.5.3 Green fluorescent protein as a folding reporter**

Green fluorescent protein (GFP), isolated from the jellyfish *Aequorea victoria*, has become an indispensable biotechnological tool. Its importance is highlighted both by the amount of research that has gone into improving and modifying it and also by its use in a broad array of applications from live cell imaging to determination of protein-protein interactions to of course, directed evolution. This protein has become such a valuable tool that its discoverers, Osamu Shimomura, Martin Chalfie, and Roger Y. Tsien received the 2008 Nobel Prize in Chemistry. The usefulness of GFP lies in its unique structure that confers high stability and the ability to

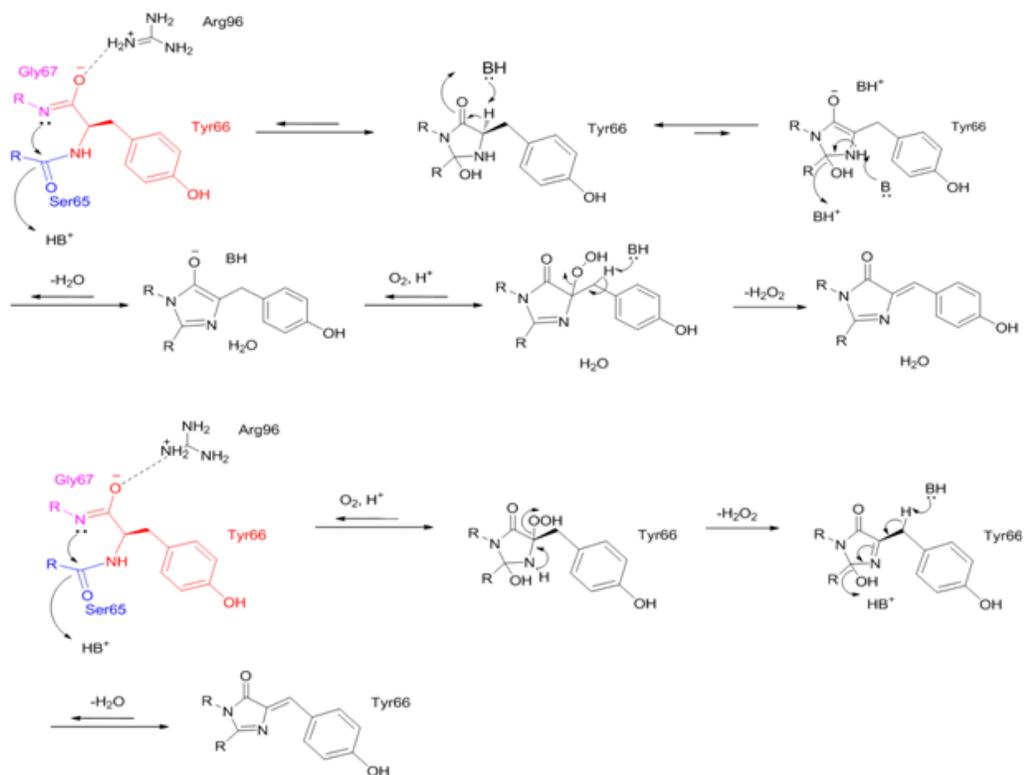
fluoresce even when expressed recombinantly in many different organisms and under extreme conditions. Structurally, as shown in **Figure 1-6A**, GFP consists of a barrel made up of 11  $\beta$ -strands with the chromophore (Figure 1-6B) located in the centre along a single  $\alpha$ -helix [102].



**Figure 1-6.** Structure of GFP (A) and its chromophore (B). PDB accession code: 1ema.

Chromophore formation occurs post-translationally via an oxidative reaction involving ring closure between the residues Ser65, Tyr66, and Gly67. The mechanism (**Figure 1-7**) involves three main steps: cyclization via bond formation between the nitrogen of Gly67 and the carbonyl of Ser65, dehydration of that same carbonyl, and oxidation to produce a C $\alpha$ -C $\beta$  double bond. The order of these steps has been under some debate, with cyclization-dehydration-oxidation (**Figure 1-7A**) being favoured by Barondeau et al. on the basis that the unfavourable cyclization product will be stabilized by the formation of the aromatic ring system upon dehydration while awaiting the very slow oxidation step [103]. On the other hand, Zhang et al. favour a cyclization-oxidation-dehydration mechanism (**Figure 1-7B**) which they verified *in vitro* by monitoring the kinetics of hydrogen peroxide formation during chromophore maturation [104]. They found that peroxide was generated prior to formation of fluorescence and also that reaction species analyzed from this point in the reaction showed mass loss of 2 Da upon tryptic digestion.

What is conclusive from these studies is that chromophore formation is strictly dependent on proper folding of GFP, and that the rate limiting step, regardless of which order the reaction occurs in, is the oxidation step. Studies where the chromophore was allowed to mature *in vivo* report the time constant for this step to be  $\sim 4$  hours [105], whereas when chromophore formation is induced *in vitro*, the time constant for the oxidation step was measured as  $\sim 34$  minutes [104]. The slow nature of the oxidation step is one of the main reasons that GFP is an excellent reporter of protein folding. It is assumed that aggregation of the protein of interest (should it occur) will happen on a much faster timescale than chromophore maturation. Thus, if the fusion is already deposited in an inclusion body prior to the completion of the essential oxidation step, fluorescence should not be observed.



**Figure 1-7.** Post-translational synthesis of GFP Chromophore. The top scheme (A) depicts a cyclization-dehydration-oxidation mechanism whereas the bottom scheme (B) depicts a cyclization-oxidation-dehydration mechanism.

This concept was validated by Waldo et al. in his 1999 *Nature Biotechnology* paper which introduced GFP as a folding reporter for proteins fused to its N-terminus [106]. Using 20 proteins as a dataset he showed that there was a correlation between the fluorescence of protein-GFP fusions and the solubility of the protein when expressed alone. Furthermore, he then applied this concept to directed evolution by dramatically improving the solubility of a C33T mutant of gene V protein and bullfrog H-subunit ferritin with four rounds of DNA shuffling, 3 rounds of backcrossing against wild-type DNA and screening of 10,000 clones per round.

Since the publishing of this groundbreaking paper, numerous examples of evolution experiments employing GFP as a folding reporter can be found in the literature. Several of these examples are summarized in **Table 1-4**. Besides its use in an evolution context, GFP has also been used as a marker of protein expression and solubility for the purposes of optimizing expression conditions [107, 108].

An interesting thing to note from **Table 1-4** is that the actual GFP used as the reporter varies between several mutants of wild-type GFP. Although highly expressed, wild type GFP will primarily be expressed in inclusion bodies in *E. coli* [109] and thus evolution of GFP has become a popular research area. This evolution has led both to variants exhibiting fluorescence in other colours such as cyan, blue, yellow, and orange and also to greater stability, brighter fluorescence and shifted excitation and emission maxima for GFP [110]. **Table 1-5** lists a number of GFP variants and their phenotypic differences. Generally, mutations occurring near the central helix will affect the excitation and emission maxima whereas mutations more distal to the chromophore affect folding [110]. The ‘superfolder’ GFP mutant is so robust that it is not an effective folding reporter since it folds correctly regardless of whether or not a protein fused to it misfolds [111].

**Table 1-4.** Literature examples of improvement of solubility via evolution and screening with a GFP folding reporter

<i>Example</i>	<i>Diversification Strategy</i>	<i>GFP Variant Used</i>	<i>Result</i>	<i>Reference</i>
Mouse Vav protein	Tagged random primer PCR	GFPuv/F64L	4 soluble domains identified	[112]
Dihydrofolate reductase	Site-specific mutagenesis and epPCR	GFPwt/S65A/V68L/S72A	3 to 6 fold increases in solubility over WT	[113]
Telomerase reverse transcriptase	Tagged random primer PCR	GFPuv/F64L	Identification of a soluble and critical domain	[114]
TEV protease	epPCR and DNA shuffling	eGFP	5.5x increase in purified product over parental	[115]
Creation of novel protein folds	Fragments encoding secondary structural	eGFP	Identification of four soluble species containing folded	[116]
RV2002 gene product from <i>Mycobacterium</i>	epPCR and DNA shuffling	GFPuv/S65T/F64L	Soluble mutant discovered and crystallized	[117]
Alzheimer's A $\beta$ 42 peptide	epPCR, Mutazyme <sup>TM</sup> , doped oligonucleotides	GFPuv/S65T/F64L	Identified 36 variants with a reduced tendency to aggregate	[118]

**Table 1-5** GFP Mutants and Their Properties

Mutant	Mutations*	Properties	Ref
Wild type	Phe <sup>64</sup> Ser <sup>65</sup> Phe <sup>99</sup> Met <sup>153</sup> Val <sup>163</sup> Ser <sup>30</sup> Tyr <sup>39</sup> Asn <sup>105</sup> Tyr <sup>145</sup> Ile <sup>171</sup> Ala <sup>206</sup>	Excitation max at 395 nm, emission max at 504 nm	[109]
Enhanced GFP (eGFP, GFPmut1)	Leu <sup>64</sup> Thr <sup>65</sup>	35-fold increase in fluorescence intensity, S65T mutation induces a red-shift such that excitation maxima is shifted to ~490 nm	[119]
GFPuv	Ala <sup>163</sup> Ser <sup>99</sup> Thr <sup>153</sup>	18-fold increase in brightness, same emission and excitation maxima as WT	[120]
Folding Reporter GFP	Leu <sup>64</sup> Thr <sup>65</sup> Ser <sup>99</sup> Thr <sup>153</sup> Ala <sup>163</sup>	Combines folding mutations of GFPuv and chromophore mutations of eGFP, thus it is both red-shifted and more intense	[106]
GFPuv, F64L	Leu <sup>64</sup> Ser <sup>65</sup> Thr <sup>153</sup> Ala <sup>163</sup>	Identical to folding reporter GFP minus the red-shifting mutation, thus retains the same emission and excitation maxima as WT, with much higher intensity	[112]
SuperfolderGFP	Leu <sup>64</sup> Thr <sup>65</sup> Ser <sup>99</sup> Thr <sup>153</sup> Ala <sup>163</sup> Arg <sup>30</sup> Asn <sup>39</sup> Thr <sup>105</sup> Phe <sup>145</sup> Val <sup>171</sup> Val <sup>206</sup>	54-fold increase in fluorescence intensity, not a good folding reporter as it is so robust it will fold regardless of the solubility of the fusion protein	[111]

\* only amino acid changing mutations are shown, all variants also contain silent mutations which are described in their respective references

The GFP folding reporter is not immune to the common problems associated with evolution experiments. It is commonly observed that “you get what you select for” [121] and this is certainly the case when selecting for increased fluorescence. Brighter fluorescence is really only an indication that whatever is fused to GFP is interfering with its fluorescence to a lesser degree than the wild type fused protein. This may occur not only because of an improvement of folding of the fusion protein, but also because of false-positive results which can arise from the creation of internal ribosome binding sites during diversification, leading to truncated versions of the target protein that do not interfere with the folding of GFP. Also, proteins which have slow

aggregation kinetics would not interfere with GFP folding, resulting in bright fluorescence even though the protein is no less prone to aggregation [122].

One way to combat these false positives comes in the form of a reporter that combines the sensitivity of split-protein methods with the easily detectable phenotype of GFP. Split-GFP, also developed by Waldo and coworkers, is a reporter in which the protein of interest is expressed as an N-terminal fusion of a small fragment of GFP (termed GFP-11 M3) which is comprised of amino acids 215 to 230 of GFP (encoding  $\beta$ -strand 11), and contains the mutations L221H, F223Y, and T225N [123, 124]. These mutations are essential in order to balance the fragment's deleterious effect on target protein solubility and still allow good complementation with the much larger fragment GFP 1-10 OPT. The larger fragment, which is expressed separately, consists of amino acids 1- 214 and encodes  $\beta$ -strands 1-10 and also contains several mutations (described in [123]) which improve both complementation and solubility when combined with the smaller fragment. Fluorescence will only be observed if the two fragments interact to fully form the GFP  $\beta$ -barrel. A main advantage of this technology is that by expressing the fusion and allowing it to fold prior to expressing the larger GFP 1-10 OPT fragment, complementation and therefore fluorescence cannot occur prior to aggregation. Also, depending on the type of GFP 11 tag used one could increase the stringency of selection as other forms of this tag were developed which perturb the solubility of the target to a greater degree.

Split-GFP also has some drawbacks, one being that it is possible for the tag to be buried within a properly folded and soluble structure, thus preventing fluorescence even though proper folding has been achieved. Another potential drawback of this and C-terminal GFP reporters is that they are not dynamic, meaning that once GFP has assembled and the chromophore has formed, it is irreversible even if the target aggregates afterwards [123]. Also, truncation artifacts can still be a problem with this system. To combat false-positives generated from truncation

artifacts, a third type of GFP reporter was developed by Waldo *et al* using circular permutation variants of GFP.

In this method the target protein is inserted into the sequence of a circularly permuted GFP variant [125]. This reporter is thought to lower the instances of false positives due to truncation artifacts since it is less likely that a truncated version of the gene will allow the separate halves of GFP to associate and fluoresce. It is also thought to be more sensitive to misfolded proteins since generally permutants are inherently more likely to misfold themselves [126].

To access different varieties of this reporter with varying degrees of sensitivity, several variants were constructed using different combinations of superfolder GFP and folding reporter GFP as the separated halves of the circularly permuted GFP. The most fluorescent circular permutants have the break point either between  $\beta$ -strands 8 and 9 (amino acids 172 and 173) or between  $\beta$ -strands 7 and 8 (amino acids 152 and 153), and so the authors tested both break points with all combinations of superfolder GFP (**Table 1-5**) and folding reporter GFP (**Table 1-5**) as the separate halves. As expected, permutants with both halves consisting of superfolder GFP were too robust to be used as folding reporters, and permutants with both halves containing only folding reporter mutations showed the lowest fluorescence when fused with misfolded protein. Interestingly, the authors showed that truncated versions of a protein that were brightly fluorescent when expressed as an N-terminal fusion to folding reporter GFP did not exhibit any fluorescence when fused into the permutant GFP reporter that they were trying to evolve. Using the permutant reporter they successfully evolved soluble, well-expressed variants of this protein (Rv0113) [125].

Because of the numerous successes in the literature proving that GFP can be used as a folding reporter for screening and identifying variants that have improved folding and solubility

(see **Table1-4** for a few examples), this method was combined with epPCR and DNA shuffling to evolve three proteins of high interest to our lab. One of the proteins to be discussed, PhnG—a member of the *E. coli* carbon-phosphorus-lyase operon, was an ideal target for this methodology as there is no known structural or functional data available for it. The other two proteins, RebG, a glycosyl transferase involved in the biosynthesis of rebeccamycin, and human 5-lipoxygenase, an enzyme which converts arachidonic acid to leukotriene A<sub>4</sub>, have assigned functions but their structures have not been solved. All three proteins primarily form insoluble aggregates when expressed in *E. coli*, and thus crystal structures have not been attainable.

High resolution structures of the proteins studied here would greatly further many areas of research. For 5LO, an enzyme implicated in many serious diseases such as Alzheimer's and cancer, a detailed structure could provide mechanistic details as well as aid in inhibitor design. A detailed structure of RebG would provide insights into protein folding pathways, and access to large quantities of this enzyme could serve as means to synthesize drug-like glycosylate indolocarbazole derivatives via biocatalysis. The structure of PhnG may indicate a possible function for this enzyme and provide detailed information regarding the steps involved in the degradation of organophosphonates which are of environmental concern. It is the hope of this study that evolution via error-prone PCR, DNA shuffling, and screening with the GFP folding reporter will result in the discovery of mutants of these proteins with the improved ability to fold correctly and remain soluble when expressed in *E. coli*, while retaining their respective functions. The next chapters will discuss the evolution of these three proteins, and also the considerations that need to be made while conducting an evolution experiment to ensure high-quality libraries and effective selections.

## 1.6 References

1. Edwards AM, Arrowsmith CH, Christendat D, Dharamsi A, Friesen JD, Greenblatt JF, Vedadi M: **Protein Production: feeding the crystallographers and NMR spectroscopists.** *Nat Struct Biol* 2000, **7**:970-972.
2. Sorensen HP, Mortensen KK: **Advanced genetic strategies for recombinant protein expression in *Escherichia coli*.** *J Biotechnol* 2005, **115**(2):113-128.
3. Baneyx F, Mujacic M: **Recombinant protein folding and misfolding in *Escherichia coli*.** *Nat Biotechnol* 2004, **22**(11):1399-1408.
4. Smith GP: **The progeny of sexual PCR.** *Nature* 1994, **370**:324-325.
5. Anfinsen CB: **Principles that Govern the Folding of Protein Chains.** *Science* 1973, **181**(4096):223-230.
6. Ellis RJ, Hartl FU: **Principles of protein folding in the cellular environment.** *Curr Opin Struct Biol* 1999, **9**:102-110.
7. Radford SE: **Protein folding: progress made and promises ahead.** *Trends Biochem Sci* 2000, **25**:611-618.
8. Dill KA: *Biochemistry* 1990, **29**(31):7133-7155.
9. Chen J, Stites WE: **Packing is a Key Selection Factor in the Evolution of Protein Hydrophobic Cores.** *Biochemistry* 2001, **40**(50):15280-15289.
10. Ellis RJ, Minton AP: **Protein Aggregation in Crowded Environments.** *Biol Chem* 2006, **387**:485-497.
11. Dill KA, Ozkan SB, Shell MS, Weikl TR: **The Protein Folding Problem.** *Annu Rev Biophys* 2008, **37**:289-316.
12. Finucane MD, Woolfson DN: **Core-Directed Protein Design. II. Rescue of a Multiply Mutated and Destabilized Variant of Ubiquitin.** *Biochemistry* 1999, **38**:11613-11623.
13. Daggett V, Fersht A: **The present view of the mechanism of protein folding.** *Nat Rev Mol Cell Biol* 2003, **4**(6):497-502.
14. Ghosh K, Ozkan SB, Dill KA: **The Ultimate Speed Limit to Protein Folding is Conformational Searching.** *J Am Chem Soc* 2007, **129**:11920-11927.

15. Ferguson N, Fersht AR: **Early events in protein folding.** *Curr Opin Struct Biol* 2003, **13**:75-81.
16. Fersht AR: **Nucleation mechanisms in protein folding.** *Curr Opin Struct Biol* 1997, **7**:3-9.
17. Dill KA, Chan HS: **From Levinthal to pathways to funnels.** *Nat Struct Biol* 1997, **4**(1):10-19.
18. Plaxco KW, Simons KT, Baker D: **Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins.** *J Mol Biol* 1998, **277**:985-994.
19. Dobson CM, Karplus M: **The fundamentals of protein folding: bringing together theory and experiment.** *Curr Opin Struct Biol* 1999, **9**:92-101.
20. Sridevi K, Lakshmikanth GS, Krishnamoorthy G, Udgaonkar JB: **Increasing Stability Reduces Conformational Heterogeneity in a Protein Folding Intermediate Ensemble.** *J Mol Biol* 2004, **337**:699-711.
21. Ignatova Z, Krishnan B, Bombardier JP, Marcelino AMC, Hong J, Gierasch LM: **From the Test Tube to the Cell: Exploring the Folding and Aggregation of a  $\beta$ -Clam Protein.** *Biopolymers* 2007, **88**:157-163.
22. Ellis RJ, Minton AP: **Join the crowd.** *Nature* 2003, **425**:27-28.
23. Kiefhaber T, Rudolph R, Kohler HH, Buchner J: **Protein Aggregation *in vitro* and *in vivo*: A Quantitative Model of the Kinetic Competition Between Folding and Aggregation.** *Biotechnology (N Y )* 1991, **9**:825-829.
24. Hoffmann F, Rinas U: **Stress Induced by Recombinant Protein Production in *Escherichia coli*.** *Adv Biochem Eng Biotechnol* 2004, **89**:73-92.
25. Frydman J: **Folding of Newly Translated Proteins In Vivo: The Role of Molecular Chaperones.** *Annu Rev Biochem* 2001, **70**:603-647.
26. Mansell TJ, Fisher AC, DeLisa MP: **Engineering the Protein Folding Landscape in Gram-Negative Bacteria.** *Curr Protein Pept Sci* 2008, **9**:138-149.
27. Sorensen HP, Mortensen KK: **Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*.** *Microb Cell Fact* 2005, **4**:1-8.
28. Hartl FU, Hayer-Hartl M: **Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein.** *Science* 2002, **295**(5561):1852-1858.

29. Baneyx F: **Recombinant protein expression in *Escherichia coli***. *Curr Opin Biotechnol* 1999, **10**(5):411-421.
30. Bukau B, Horwich AL: **The Hsp70 and Hsp60 Chaperone Machines**. *Cell* 1998, **92**:351-366.
31. Swartz JR: **Advances in *Escherichia coli* production of therapeutic proteins**. *Curr Opin Biotechnol* 2001, **12**:195-201.
32. Schein CH, Noteborn MH: **Formation of Soluble Recombinant Proteins in *Escherichia coli* is Favored by Lower Growth Temperature**. *Biotechnology (N Y)* 1988, **6**:291-294.
33. Schein CH: **Production of Soluble Recombinant Proteins in Bacteria**. *Biotechnology (N Y)* 1989, **7**:1141-1149.
34. Jana S, Deb JK: **Strategies for efficient production of heterologous proteins in *Escherichia coli***. *Appl Microbiol Biotechnol* 2005, **67**:289-298.
35. Makrides SC: **Strategies for Achieving High-Level Expression of Genes in *Escherichia coli***. *Microbiol Rev* 1996, **60**(3):512-538.
36. Schumann W, Ferreira LCS: **Production of recombinant proteins in *Escherichia coli***. *Genet Mol Biol* 2004, **27**:442-453.
37. Lisser S, Margalit H: **Compilation of *E.coli* mRNA promoter sequences**. *Nucleic Acids Res* 1993, **21**:1507-1516.
38. Brosius J, Erfle M, Storella J: **Spacing of the -10 and -35 Regions in the *tac* Promoter**. *J Biol Chem* 1985, **260**:3539-3541.
39. Mayer MR, Dailey TA, Baucom CM, Supernak JL, Grady MC, Hawk HE, Dailey HA: **Expression of human proteins at the Southeast Collaboratory for Structural Genomics**. *J Struct Funct Genomics* 2004, **5**:159-165.
40. Kopetzki E, Shumacher G, Buckel P: **Control of formation of active soluble or inactive insoluble baker's yeast  $\alpha$ -glucosidase PI in *Escherichia coli* by induction and growth conditions**. *Mol Gen Genet* 1989, **216**:149-155.
41. Shaw MK, Ingraham JL: **Synthesis of Macromolecules by *Escherichia coli* near the Minimal Temperature for Growth**. *J Bacteriol* 1967, **94**:157-164.
42. Mujacic M, Cooper KW, Baneyx F: **Cold-inducible cloning vectors for low-temperature protein expression in *Escherichia coli*: application to the production of a toxic and proteolytically sensitive fusion protein**. *Gene* 1999, **238**:325-332.

43. Marblestone JG, Edavettal SC, Lim Y, Lim P, Zuo X, Butt TR: **Comparison of SUMO fusion technology with traditional gene fusion systems: Enhanced expression and solubility with SUMO.** *Protein Sci* 2006, **15**:182-189.
44. LaVallie ER, McCoy JM: **Gene fusion expression systems in *Escherichia coli*.** *Curr Opin Biotechnol* 1995, **6**:501-506.
45. Kapust RB, Waugh DS: ***Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.** *Protein Sci* 1999, **8**:1668-1674.
46. Blum P, Velligan M, Lin N, Matin A: **DnaK-mediated alterations in human growth hormone protein inclusion bodies.** *Bio/Technology* 1992, **10**:301-304.
47. Hockney RC: **Recent developments in heterologous protein production in *Escherichia coli*.** *Trends Biotechnol* 1994, **12**:456-463.
48. de Marco A, Deuerling E, Mogk A, Tomoyasu T, Bukau B: **Chaperone-based procedure to increase yields of soluble recombinant proteins produces in *E. coli*.** *BMC Biotechnol* 2007, **7**:32.
49. Ventura S, Villaverde A: **Protein quality in bacterial inclusion bodies.** *Trends Biotechnol* 2006, **24**(4):179-185.
50. Singh SM, Panda AK: **Solubilization and Refolding of Bacterial Inclusion Body Proteins.** *J Biosci Bioeng* 2005, **99**(4):303-310.
51. Vallejo LF, Rinas U: **Strategies for the recovery of active proteins through refolding of bacterial inclusion bodies.** *Microb Cell Fact* 2004, **3**:11.
52. Waldo GS: **Genetic screens and directed evolution for protein solubility.** *Curr Opin Chem Biol* 2003, **7**:33-38.
53. Sorensen MA, Kurland CG, Pedersen S: **Codon Usage Determines Translation Rate in *Escherichia coli*.** *J Mol Biol* 1989, **207**:365-377.
54. Thanaraj TA, Argos P: **Ribosome-mediated translational pause and protein domain organization.** *Protein Sci* 1996, **5**:1594-1612.
55. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G: **Effects of Consecutive AGG Codons on Translation in *Escherichia coli*, Demonstrated with a Versatile Codon Test System.** *J Bacteriol* 1993, **175**:716-722.
56. Gustafsson C, Govindarajan S, Minshull J: **Codon bias and heterologous protein expression.** *Trends Biotechnol* 2004, **22**(7):346-353.

57. Polizzi KM, Chaparro-Riggers JF, Vazquez-Figueroa E, Bommarius AS: **Structure-guided consensus approach to create a more thermostable penicillin G acylase.** *Biotechnol J* 2006, **1**:531-536.
58. Eijsink VGH, Bjork A, Gaseidnes S, Sirevag R, Synstad B, van den Burg B, Vriend G: **Rational engineering of enzyme stability.** *J Biotechnol* 2004, **113**:105-120.
59. Steipe B, Schiller B, Pluckthun A, Steinbacher S: **Sequence statistics reliably predict stabilizing mutations in a protein domain.** *J Mol Biol* 1994, **240**:188-192.
60. Maxwell KL, Davidson AR: **Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects.** *Biochemistry* 1998, **37**:16172-16182.
61. Roodveldt C, Aharoni A, Tawfik DS: **Directed evolution of proteins for heterologous expression and stability.** *Curr Opin Struct Biol* 2005, **15**:50-56.
62. Yuan L, Kurek I, English J, Keenan R: **Laboratory-Directed Protein Evolution.** *Microbiol Mol Biol Rev* 2005, **69**:373-392.
63. Jäckel C, Kast P, Hilvert D: **Protein Design by Directed Evolution.** *Annu Rev Biophys* 2008, **37**:153-173.
64. Kaur J, Sharma R: **Directed Evolution: An Approach to Engineer Enzymes.** *Crit Rev Biotechnol* 2006, **26**:165-199.
65. Kuchner O, Arnold FH: **Directed evolution of enzyme catalysts.** *Trends Biotechnol* 1997, **15**:523-530.
66. Magliery TJ, Regan L: **Combinatorial approaches to protein stability and structure.** *Eur J Biochem* 2004, **271**:1595-1608.
67. Wong TS, Roccatano D, Zacharias M, Schwaneberg U: **A Statistical Analysis of Random Mutagenesis Methods Used for Directed Protein Evolution.** *J Mol Biol* 2006, **355**:858-871.
68. Bloom JD, Labthavikul ST, Otey CR, Arnold FH: **Protein stability promotes evolvability.** *Proc Natl Acad Sci USA* 2006, **103**:5869-5874.
69. Bershtein S, Goldin K, Tawfik DS: **Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins.** *J Mol Biol* 2008, **379**:1029-1044.
70. Gupta RD, Tawfik DS: **Directed enzyme evolution via small and effective neutral drift libraries.** *Nat Methods* 2008, **5**:939-942.

71. Hart DJ, Tarendeau F: **Combinatorial library approaches for improving soluble protein expression in *Eshcherichia coli***. *Acta Crystallogr D Biol Crystallogr* 2006, **D62**(1):19-26.
72. Prodromou C, Savva R, Driscoll PC: **DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries**. *Drug Discov Today* 2007, **12**:931-938.
73. Reich S, Puckey LH, Cheetham CL, Harris R, Ali AAE, Bhattacharyya U, Maclagan K, Powell KA, Prodromou C, Pearl LH, Driscoll PC, Savva R: **Combinatorial domain hunting: an effective approach for the identification of soluble protein domains adaptable to high-throughput applications**. *Protein Sci* 2006, **15**:2356-2365.
74. Grothues D, Cantor CR, Smith CL: **PCR amplification of megabase DNA with tagged random primers (T-PCR)**. *Nucleic Acids Res* 1993, **21**:1321-1322.
75. Neylon C: **Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution**. *Nucleic Acids Res* 2004, **32**:1448-1459.
76. Nguyen AW, Daugherty PS: **Production of randomly mutated plasmid libraries using mutator strains**. *Methods Mol Biol* 2003, **231**:39-44.
77. Caldwell RC, Joyce GF: **Randomization of genes by PCR mutagenesis**. *PCR Methods Appl* 1992, **2**:28-33.
78. Zacco M, Williams DM, Brown DM, Gherardi E: **An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues**. *J Mol Biol* 1996, **255**:589-603.
79. Wong TS, Roccatano D, Loakes D, Tee KL, Schenk A, Hauer B, Schwaneberg U: **Transversion-enriched sequence saturation mutagenesis (SeSaM-Tv<sup>+</sup>): A random mutagenesis method with consecutive nucleotide exchanges that complements the bias of error-prone PCR**. *Biotechnol J* 2008, **3**:74-82.
80. Di Giulio M: **The origin of the genetic code: theories and their relationships, a review**. *Biosystems* 2005, **80**:175-184.
81. Wong TS, Zhurina D, Schwaneberg U: **The Diversity Challenge in Directed Protein Evolution**. *Comb Chem High Throughput Screen* 2006, **9**:271-288.
82. Ness JE, Kin S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J: **Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently**. *Nat Biotechnol* 2002, **20**:1251-1255.

83. Stemmer WPC: **DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution.** *Proc Natl Acad Sci U S A* 1994, **91**:10747-10751.
84. Stemmer WPC: **Rapid evolution of a protein *in vitro* by DNA shuffling.** *Nature* 1994, **370**:389-391.
85. Stemmer WPC: **Searching Sequence Space: Using recombination to search more efficiently and thoroughly instead of making bigger combinatorial libraries.** *Biotechnology* 1995, **13**:549-553.
86. Zhang YX, Perry K, Vinci VA, Powell K, Stemmer WPC, del Cardayré SB: **Genome shuffling leads to rapid phenotypic improvement in bacteria.** *Nature* 2002, **415**:644-646.
87. Bacher JM, Reiss BD, Ellington AD: **Anticipatory evolution and DNA shuffling.** *Genome Biol* 2002, **3**:1021.1-1021.4.
88. Lorimer IAJ, Pastan I: **Random recombination of antibody single chain Fv sequences after fragmentation with DNaseI in the presence of Mn<sup>2+</sup>.** *Nucleic Acids Res* 1995, **23**:3067-3068.
89. Zhao H, Arnold FH: **Optimization of DNA shuffling for high fidelity recombination.** *Nucleic Acids Res* 1997, **25**:1307-1308.
90. Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH: **Molecular evolution by staggered extension process (StEP) *in vitro* recombination.** *Nat Biotechnol* 1998, **16**:258-261.
91. Bogarad LD, Deem MW: **A Hierarchical Approach to Protein Molecular Evolution.** *Proc Natl Acad Sci U S A* 1999, **96**:2591-2595.
92. Ostermeier M, Shim JH, Benkovic SJ: **A combinatorial approach to hybrid enzymes independent of DNA homology.** *Nat Biotechnol* 1999, **17**:1205-1209.
93. Ostermeier M, Nixon AE, Shim JH, Benkovic SJ: **Combinatorial Protein Engineering by Incremental Truncation.** *Proc Natl Acad Sci U S A* 1999, **96**:3562-3567.
94. Lutz S, Ostermeier M, Moore GL, Maranas CD, Benkovic SJ: **Creating Multiple-Crossover DNA libraries Independent of Sequence Identity.** *Proc Natl Acad Sci U S A* 2001, **98**:11248-11253.
95. Roberts RW: **Totally *in vitro* protein selection using mRNA-protein fusions and ribosome display.** *Curr Opin Chem Biol* 1999, **3**:268-273.

96. He M, Taussig MJ: **Ribosome display: Cell-free protein display technology.** *Brief Funct Genomic Proteomic* 2002, **1**:204-212.
97. Sieber V, Pluckthun A, Schmid FX: **Selecting proteins with improved stability by a phage-based method.** *Nat Biotechnol* 1998, **16**:955-960.
98. Lesley SA, Graziano J, Cho CY, Knuth MW, Klock HE: **Gene expression response to misfolded protein as a screen for soluble recombinant expression.** *Protein Eng* 2002, **15**:153-160.
99. Maxwell KL, Mittermaier AK, Forman-Kay JD, Davidson AR: **A simple in vivo assay for increased protein solubility.** *Protein Sci* 1999, **8**:1908-1911.
100. Sieber V, Martinez CA, Arnold FH: **Libraries of hybrid proteins from distantly related sequences.** *Nat Biotechnol* 2001, **19**:456-460.
101. Wigley WC, Stidham RD, Smith NM, Hunt JF, Thomas PJ: **Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein.** *Nat Biotechnol* 2001, **19**:131-136.
102. Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ: **Crystal Structure of the *Aequorea victoria* Green Fluorescent Protein.** *Science* 1996, **273**:1392-1395.
103. Barondeau DP, Kassmann CJ, Tainer JA, Getzoff ED: **Understanding GFP Chromophore Biosynthesis: Controlling Backbone Cyclization and Modifying Post-translational Chemistry.** *Biochemistry* 2005, **44**:1960-1970.
104. Zhang L, Patel HN, Lappe JW, Wachter RM: **Reaction Progress of Chromophore Biogenesis in Green Fluorescent Protein.** *J Am Chem Soc* 2006, **128**:4766-4772.
105. Heim R, Pracher DC, Tsien RY: **Wavelength mutations and posttranslational autoxidation of green fluorescent protein.** *Proc Natl Acad Sci U S A* 1994, **91**:12501-12504.
106. Waldo GS, Standish BM, Berendzen J, Terwilliger TC: **Rapid protein-folding assay using green fluorescent protein.** *Nat Biotechnol* 1999, **17**:691-695.
107. Cha HJ, Wu C-, Valdes JJ, Rao G, Bentley WE: **Observations of Green Fluorescent Protein as a Fusion Partner in Genetically Engineered *Escherichia coli*: Monitoring Protein Expression and Solubility.** *Biotechnol Bioeng* 2000, **67**:565-574.
108. Rucker E, Schneider G, Steinhäuser K, Löwer R, Hauber J, Stauber RH: **Rapid Evaluation and Optimization of Recombinant Protein Production Using GFP Tagging.** *Protein Expr Purif* 2001, **21**:220-223.

109. Tsien RY: **The green fluorescent protein.** *Annu Rev Biochem* 1998, **67**:509-544.
110. Shaner NC, Patterson GH, Davidson MW: **Advances in fluorescent protein technology.** *J Cell Sci* 2007, **120**:4247-4260.
111. Pédelacq J-, Cabantous S, Tran T, Terwilliger TC, Waldo GS: **Engineering and characterization of a superfolder green fluorescent protein.** *Nat Biotechnol* 2006, **24**:79-88.
112. Kawasaki M, Inagaki F: **Random PCR-based Screenig for Soluble Domains Using Green Fluorescent Protein.** *Biochem Biophys Res Commun* 2001, **280**:842-844.
113. Japrun D, Chusacultanachai S, Yuvaniyama J, Wilairat P, Yuthavong Y: **A simple dual selection for functionally active mutants of *Plasmodium falciparum* dihydrofolate reductase with improved solubility.** *Protein Eng Des Sel* 2005, **18**:457-464.
114. Jacobs SA, Podell ER, Wuttke DS, Cech TR: **Soluble domains of telomerase reverse transcriptase identified by high-throughput screening.** *Protein Sci* 2005, **14**:2051-2058.
115. van den Berg S, Löfdahl PA, Hård T, Berglund H: **Improved solubility of TEV protease by directed evolution.** *J Biotechnol* 2006, **121**:291-298.
116. Graziano JJ, Liu W, Perera R, Geierstanger BH, Lesley SA, Schultz PG: **Selecting Folded Proteins from a Library of Secondary Structural Elements.** *J Am Chem Soc* 2008, **130**:176-185.
117. Yang JK, Park MS, Waldo GS, Suh SW: **Directed evolution approach to a structural genomics project: Rv2002 from *Mycobacterium tuberculosis*.** *Proc Natl Acad Sci USA* 2003, **100**:455-460.
118. Wurth C, Guimard NK, Hecht MH: **Mutations that Reduce Aggregation of the Alzheimer's A $\beta$ 42 Peptide: an Unbiased Search for the Sequence Determinants of A $\beta$  Amyloidogenesis.** *J Mol Biol* 2002, **319**:1279-1290.
119. Cormack BP, Valdivia RH, Falkow S: **FACS-optimized mutants of the green fluorescent protein (GFP).** *Gene* 1996, **173**:33-38.
120. Cramer A, Whitehorn EA, Tate E, Stemmer WPC: **Improved Green Fluorescent Protein by Molecular Evolution Using DNA Shuffling.** *Nat Biotechnol* 1996, **14**:315-319.
121. Schmidt-Dannert C, Arnold FH: **Directed evolution of industrial enzymes.** *Trends Biotechnol* 1999, **17**:135-136.

122. Cabantous S, Pedelacq JD, Mark BL, Naranjo C, Terwilliger TC, Waldo GS: **Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from *Mycobacterium tuberculosis*.** *J Struct Funct Genomics* 2005, **6**:133-119.
123. Cabantous S, Terwilliger TC, Waldo GS: **Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein.** *Nat Biotechnol* 2005, **23**:102-107.
124. Cabantous S, Waldo GS: ***In vivo* and *in vitro* protein solubility assays using split GFP.** *Nat Methods* 2006, **3**:845-854.
125. Cabantous S, Rogers Y, Terwilliger T, Waldo GS: **New Molecular Reporters for Rapid Protein Folding Assays.** *PLoS One* 2008, **3**:e2387.
126. Topell S, Hennecke J, Glockshuber R: **Circularly permuted variants of the green fluorescent protein.** *FEBS Lett* 1999, **457**:283-289.

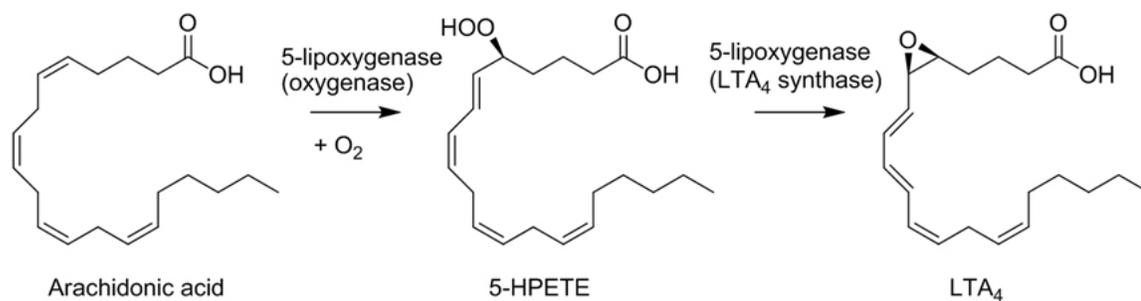
## Chapter 2

# Application of the Green-Fluorescent Protein Solubility Assay to the Directed Evolution of Human 5-Lipoxygenase and RebG

### 2.1 Introduction

This chapter will discuss the directed evolution of human arachidonate 5-lipoxygenase (5LO) (GenBank accession code BC143985.1) and *Lechevalieria aerocolonigenes* N-glycosyl transferase RebG (GenBank accession code BAC15749.1) in an attempt to improve their solubility upon expression in *E. coli*.

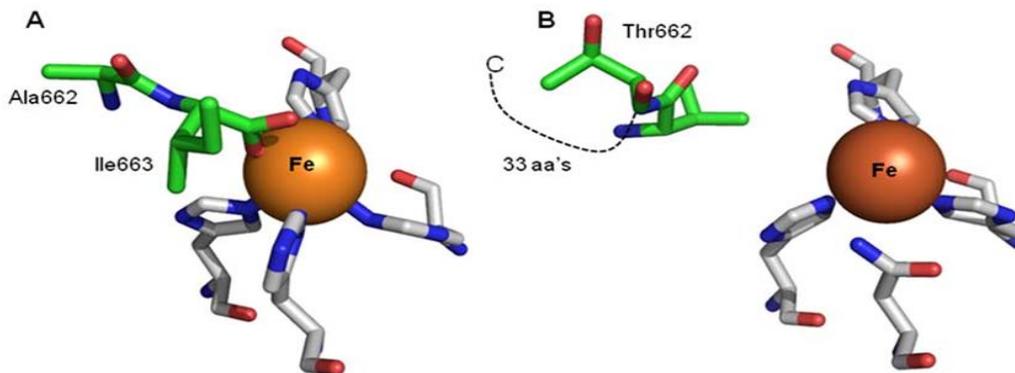
5LO is an extremely interesting target for solubility improvement as it is a widely studied protein, yet expression and purification of this protein is extremely difficult and no crystal structure is available. 5LO is involved in the biosynthesis of leukotrienes (LTs) and acts on arachidonic acid (AA) to synthesize leukotriene A<sub>4</sub> (LTA<sub>4</sub>). It exhibits two types of activities; it first acts as an oxygenase, using dioxygen and a non-heme iron cofactor to convert AA to 5(S)-hydroperoxy-6-*trans*-8,11,14-*cis*-eicosatetraenoic acid (5-HPETE), after which it catalyzes 5-HPETE to LTA<sub>4</sub> using its LTA<sub>4</sub> synthase activity (**Figure 2-1**) [1]. Leukotrienes are known to be involved in the body's inflammatory response as found in asthma and bronchitis [2], and the action of 5LO has also implicated in the development of cancer [3]. Furthermore, gene polymorphisms of 5LO have been linked to vascular diseases and Alzheimer's disease [4]. The many connections 5LO has with important and potentially harmful physiological processes make it the target of a substantial amount of research, thus a better-folding, more stable form of 5LO would be extremely useful. A crystal structure would take our understanding of this enzyme even further and would aid immeasurably in drug design.



**Figure 2-1.** Reaction scheme for 5LO. Adopted from [5].

5LO shares a 40% sequence identity with a crystallized rabbit reticulocyte homolog and 41% sequence identity with human 12LO in which a truncated version has also been crystallized. Based on the identity between these homologs, the structure of 5LO was modeled and shown to have a C-terminal catalytic domain that is highly helical and contains the iron cofactor, and a small N-terminal domain consisting of a C2-type  $\beta$ -sandwich [2]. The last four amino acids at the C-terminus are considered to be highly important for ensuring correct orientation of the C-terminus with the last residue, Ile663, acting as an iron ligand (via the  $\alpha$ -carboxylate) that is essential for catalysis [6]. Because of the importance of these last residues in maintaining proper structure and function, it may be a concern that fusing a C-terminal tag such as GFP to the enzyme could actually inhibit proper structure from forming. On the contrary however, analysis of the crystal structures of 12LO and rabbit reticulocyte 15LO show that lipoxygenase structure is highly tolerant to modification of the C-terminus. In the crystal structure of the rabbit enzyme Ile663 is resolved, but in the human 12LO structure it is not. Upon closer examination of the sequence of the crystallized human 12LO it becomes apparent that there is an additional 32 residues after the last residue of 12LO (Thr662 in this case). These residues match a sequence that arises from cloning the 12LO gene directly into a pET-28 vector since the stop codon comes after the DNA sequence encoding these residues rather than after the lipoxygenase coding sequence. Moreover, the Ile663 is mutated to Ser in this construct. The presence of this tag

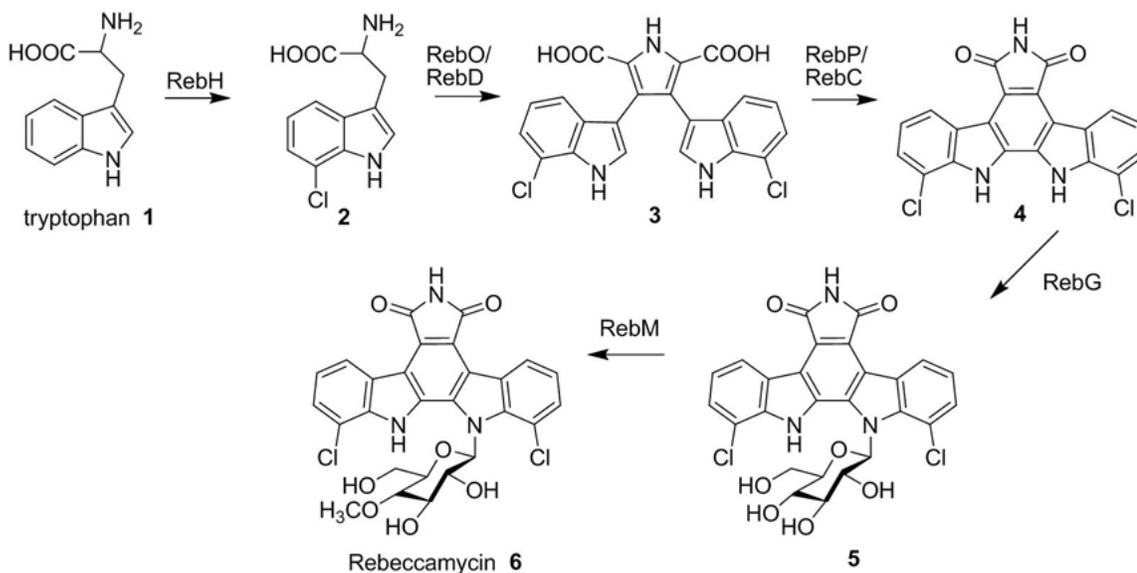
eliminates Ile663 as an iron atom ligand via the  $\alpha$ -carboxylate (since it is now part of a peptide bond). **Figure 2-2** shows a close-up of the critical iron-binding region of both enzymes and it is clear from the structures that the C-terminal residue of the 12LO mutant is not acting as a ligand for iron. Nevertheless, the 12LO structure overlaps the rabbit structure quite closely, even with the 33 amino acid tag attached. Also of interest is that Thr662 is solvent exposed, which is ideal if one wants to attach a tag to the enzyme. Thus, it seems that a tag on the C-terminus of 5LO is not likely to inhibit proper folding, and if it does inhibit iron from binding and subsequent loss of activity, there is a good chance that activity can be restored to the improved enzyme by removal of the tag.



**Figure 2-2.** Active sites of rabbit reticulocyte LO (A) (PDB accession code 2P0M) and human 12LO (B) (PDB accession code 3D3L). The C-terminal Ile662 is absent in the 12LO structure, which contains a modified C-terminus, but is visible in the rabbit 15LO structure.

A second target of interest is RebG, a member of the biosynthetic pathway responsible for the production of rebeccamycin (**Figure 2-3**) in *Lechevalieria aerocolonigenes*. Tryptophan (**1**) is halogenated by RebH to produce **2** which is then oxidatively dimerized by RebO and RebD to give the chlorinated chromopyrrolic acid **3**. Oxidative ring closure and decarboxylation facilitated by RebP and RebC generates **4**, upon which the glycosyltransferase RebG acts to give **5**, which is finally converted to rebeccamycin **6** via RebM mediated methylation of the

glucopyranosyl moiety [7, 8]. Rebecamycin's antitumor properties arise from its capacity to intercalate into DNA and stabilize a complex between the DNA and topoisomerase I [9, 10]. The stabilization of this DNA cleavage complex increases occurrence of single-stranded nicks on the DNA and the resultant damage induces cell death. It is known that the presence and stereochemistry of the sugar moiety are crucial for the antitumor action [11].



**Figure 2-3.** Rebecamycin biosynthetic pathway. Adopted from [8].

Glycosyltransferases encompass a vast superfamily of proteins of which relatively few have been crystallized. Out of the 72 families known in 2006, only 17 of the families had a crystal structure available for at least one member [12]. Because of the low sequence homology between family members, modeling of homologs based on the available crystal structures is a difficult task. Despite the low sequence homology, structural homology is quite high, with the vast majority of structures consisting of GT-A or GT-B-type topology. GT-B structures contain two Rossman domains—a commonly found  $\beta\alpha\beta\alpha\beta$  structural motif which is known to bind nucleotides. The two domains are separated by a linking region containing a catalytic site. GT-

A-type structures are comprised of an  $\alpha/\beta/\alpha$  sandwich which contains a 7-stranded  $\beta$ -sheet in the order of 3214657 with strand 6 being anti-parallel to the rest. The fact that this diverse group of proteins has high structural homology yet low sequence homology makes RebG an interesting target for protein folding studies. Directed evolution would allow for identification of key residues critical for folding and stability. Additionally, access to large quantities of stable and soluble RebG would provide access to a catalyst with excellent potential for glycosylating indolocarbozole derivatives, thereby providing access to new drug-like molecules [13]. As it stands now, RebG is non-soluble when expressed in *E. coli* and ends up primarily in inclusion bodies. It is not amenable to other solubility enhancing techniques such as expression under lower temperatures, expression as a fusion protein, co-expression of chaperones, or refolding after denaturation [8].

This chapter will discuss the attempts to improve the solubility of these two targets using directed evolution combined with the green-fluorescent protein (GFP) reporter assay. Diversification for the first round of evolution was accomplished with error-prone PCR (epPCR) and all subsequent rounds utilized DNA shuffling. Screening for improved clones was completed by fusion of the GFP folding reporter to the C-terminus of library members and scanning of LB-agar plates for the brightest colonies. This chapter will also summarize the implementation of the GFP folding reporter system and ways to optimize the protocols involved so that chances of success can be improved for any target. It will also discuss some important issues that may arise when screening libraries with this system, and how to avoid false-positive results.

## 2.2 Experimental Procedures and Methods

### 2.2.1 Materials

Oligonucleotides used for PCR were synthesized by Sigma-Genosys. *Taq* and *Vent* polymerases were purchased from New England Biolabs Canada, and all other polymerases used (PfuTurbo, PfuUltra, Herculase II) were purchased from Stratagene. Restriction enzymes, T4 DNA ligase, and calf intestinal alkaline phosphatase were purchased either from Fermentas or New England Biolabs. DNase I was purchased from Fermentas. DMSO, MnCl<sub>2</sub>, EDTA, Tris Base, NaCl, imidazole and ampicillin were purchased from Sigma Aldrich (Oakville, Ontario and US). IPTG was purchased from Invitrogen. Nucleospin plasmid purification kits (Macherey-Nagel) were ordered from MJS Biolynx, Inc. All other DNA purification kits were purchased from QIAGEN, as well as Ni-NTA resin. All cells (XL1-Blue, ElectroTen-Blue, and BL21) were purchased from Stratagene (supplied by VWR, Canada). Fisher Biosciences Canada supplied all media (Luria Bertani broth, Luria Bertani agar, glucose), 500 cm<sup>2</sup> plates (Corning), electroporation cuvettes (Eppendorf), and the additional dNTPs used for error-prone PCR (dTTP and dCTP). The pGFPuv cloning vector was purchased from Clontech, and the pProEx cloning vector was obtained from Invitrogen. All sequencing reactions were performed by either Robarts Research Institute (London, Ontario) or TCAG sequencing facility (The Hospital for Sick Children, Toronto, ON). LED lights (400 nm) were purchased from Super Bright LEDs, Inc.

## 2.2.2 Construction of the pProEx\_GFPuv screening vector

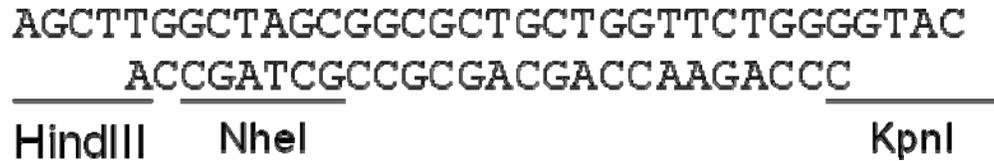
As outlined by Kawasaki *et al.* [14], the multiple cloning region of the vector pProEx (Invitrogen, discontinued) was modified to replace the original multiple cloning region with a new cloning region containing only *Hind*III, *Eco*RI and *Nhe*I sites, in that order (**Figure 2-6 A**). To do this two oligonucleotides were used: 5'-**AAGCTT**GCATGCCTGCAGGAATTC-GCTAGCTAG-3' and 5'-AGCTCTAGCTAGCGAATTCCTGCAGGCATGCA**AGCTT**-3'. The *Hind*III sites are in bold, the *Eco*RI sites are italicized and the *Nhe*I sites are underlined. The oligonucleotides were annealed by heating equimolar amounts at 95 °C for five minutes and then cooling on ice. **Figure 2-4** shows the two primers annealed together. The annealed primers were ligated into the *Ehe*I and *Hind*III sites on pProEx.

```
AAGCTTGCATGCCTGCAGGAATTC GCTAGCTAG
TTCGAACGTACGGACGTCCTTAAG CGATCGATCTCGA
HindIII                      EcoRI   NheI
```

**Figure 2-4.** Insert used for modification of the pProEx multiple cloning region.

pGFPuv (Clonotech) was modified prior to the insertion of GFPuv into pProEx such that changes were made both to the multiple cloning site upstream of GFPuv and to GFPuv itself. The cloning region immediately upstream of the sequence encoding GFPuv was altered by swapping the sequence between the *Hind*III and *Kpn*I sites on the original pGFPuv vector for a simpler sequence containing a new *Nhe*I site (**Figure 2-6 B**). The new sequence was created by mixing equimolar amounts of the oligonucleotides 5'-AGCTT**GGCTAG**CGGCGCTGCTGGTTCTGGGGTAC-3' and 5'-CCCAGAACCAGCAGCGCC**GCTAG**CCA-3' (*Nhe*I site in bold), heating them to 95 °C for five minutes, and then subsequently cooling on ice (annealed primers shown in **Figure 2-5**) The

annealed primers formed the insert that was ligated between the *HindIII* and *KpnI* sites of pGFPuv.



**Figure 2-5.** Insert used for modification of pGFPuv cloning region.

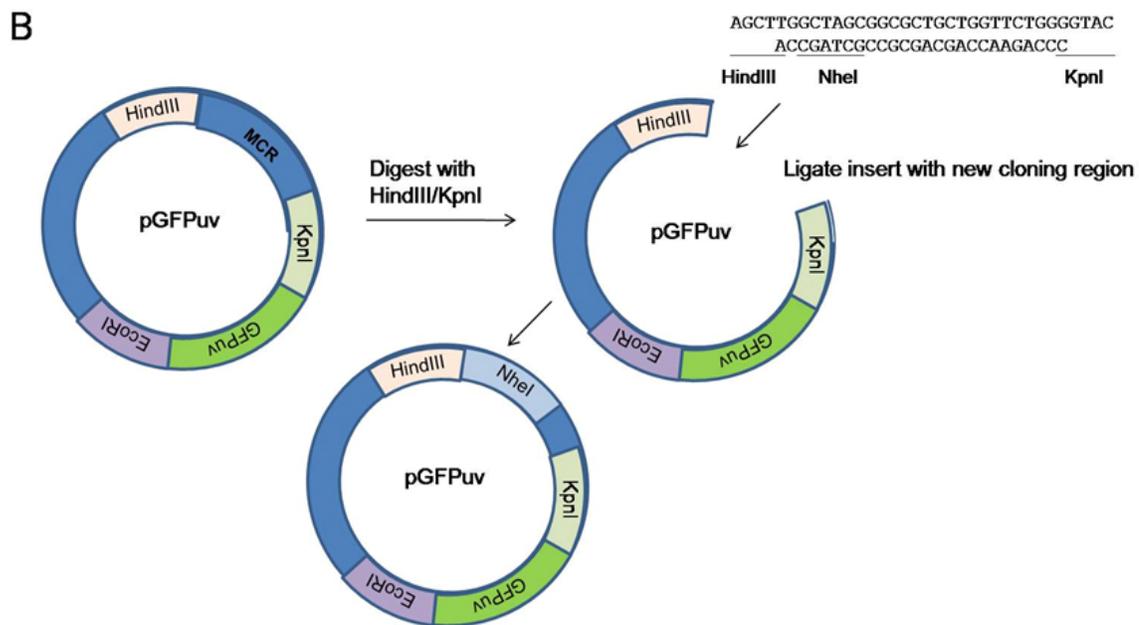
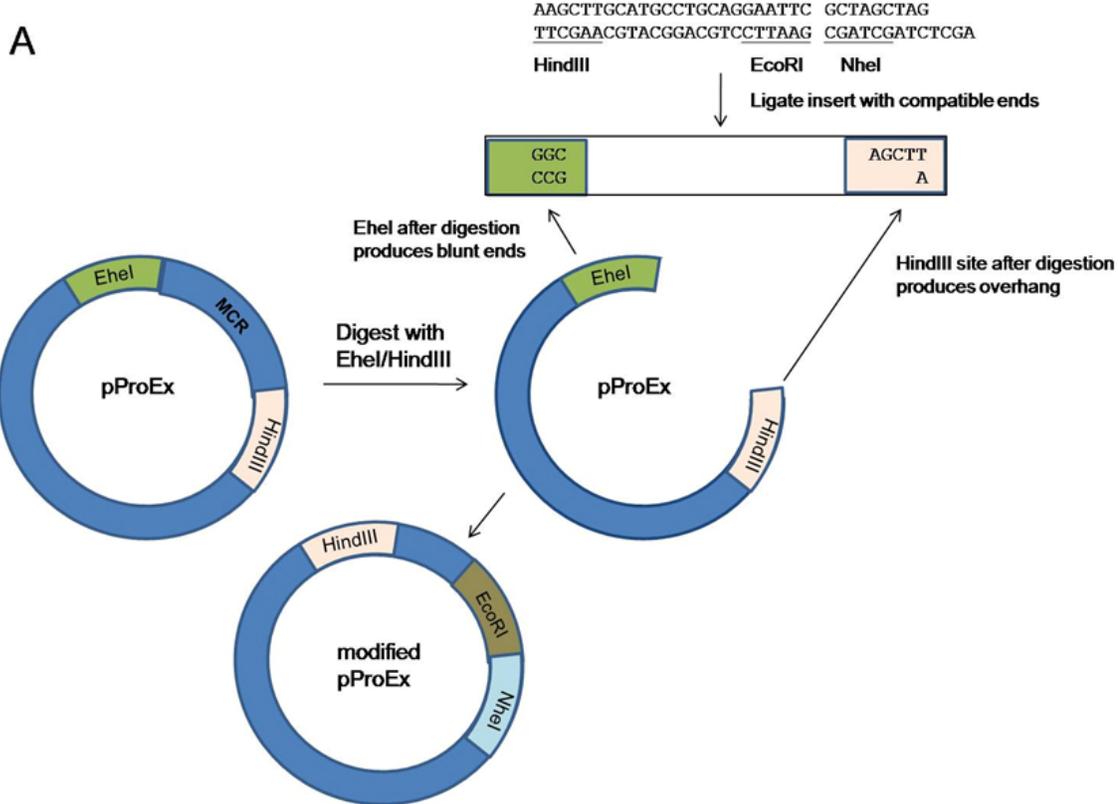
To incorporate the F64L mutation into GFPuv, 4-primer PCR mutagenesis was used. 5'-CGCCAAGCTTGCTAGCGGCGCTGCTGGTTCTGGGGTACCGGT-3' and 5'-GTTGGA**A**TTCAATTATTTGTA-3', the forward and reverse flanking primers, respectively, contained a *HindIII* restriction site (underlined) and an *EcoRI* site (bold). The internal forward primer, 5'-CAACACTTGTCACTACT**CTGT**CTTATGGTGTTC AATGC-3', and the internal reverse primer, 5'-AGCATTGAACACCATAAGAC**AG**AGTAGTGACAAGTGTTG-3', incorporated the F64L mutation (mutated codon in bold). The first PCR generated a small 260 bp fragment using the forward flanking primer and internal reverse primer, and a second PCR generated a larger 550 bp fragment using the internal forward and flanking reverse primers. The two fragments were assembled via PCR with the two flanking primers and the 260 and 550 bp fragments as the template to produce a single length gene of 769 bp (GFPuv plus the new multiple cloning region upstream of GFPuv). All PCRs were performed with *Taq* polymerase (New England Biolabs).

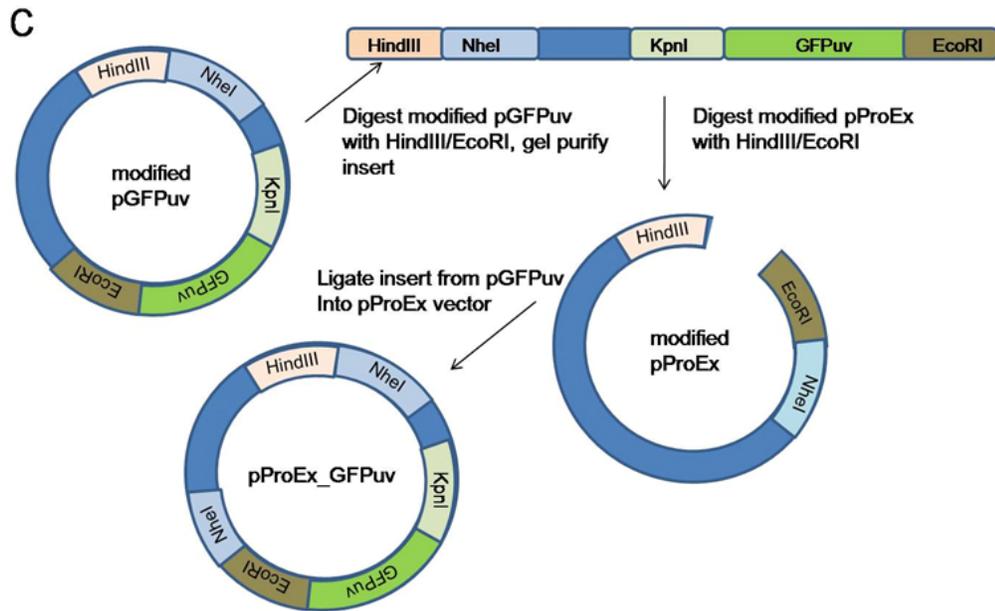
The mutated GFPuv along with the modified cloning region upstream of its initiation codon were digested out of pGFPuv using *HindIII* and *EcoRI* and ligated between the same two sites on the modified pProEx (**Figure 2-6 C**). The final construct includes a *HindIII* site and *NheI* site followed by a small spacer upstream of GFP, and an *EcoRI* site and *NheI* site after GFP. A

final feature added to this screening vector was the insertion of a short sequence of DNA encoding stop codons in all three frames in the middle of the *HindIII* restriction site. To create this insert, the two oligonucleotides 5'-AGCTTTGTAACTGAGTAA-3' and 5'-AGCTTTACTCAGTTAACAA-3' were annealed together by heating at 95 °C for ten minutes and cooling on ice.

To allow for directional cloning of library variants upstream of GFP, a second screening vector was constructed called pProEx\_GFPuv1, in which the *NheI* site following the *EcoRI* site at the end of GFP was eliminated with a silent mutation (GCTAGC to GCAAGC). To incorporate this mutation the QuikChange™ method (Stratagene) was employed with the following complementary forward and reverse primers: 5-AAATAATGAATTCGCAAGCTAGAGCTTGGCTG-3' (forward) and 5'-CAGCCAAGCTCTAGCTTGCGAATTCATTATTT-3' (reverse). Also, in this vector the stop codon sequence was inserted between the *HindIII* and *NheI* sites upstream of GFPuv rather than in between the *HindIII* site. To do this, the oligonucleotides 5'-AGCTTTGTAACTGAGTAG-3' and 5'-CTAGCTACTCAGTTAACAA-3' were annealed together and used as an insert for ligation between the *HindIII* and *NheI* sites.

All constructs were sequenced with the forward primer 5'-AGCGGATAACAATTCACACAGG-3', and the reverse primer 5'-ATCTTCTCTCATCCGCCAAAAC-3' to ensure that only the correct mutations were made.





**Figure 2-6.** Overview of pProEx\_GFPuv vector construction. MCR represents the multiple cloning regions on the respective vectors.

### 2.2.3 Cloning of wild type 5LO and RebG into pProEx\_GFPuv1

Wild type 5LO was PCR amplified out of its original vector pcDNA3-h5LOX (a generous gift from Professor Colin Funk, Queen's University) using PfuUltra™ polymerase and the primers 5'- AGGGCAAAGCTTATGCCCTCCTACACGGTCACCGTG-3' (forward, *Hind*III site underlined), and 5'- AGCGCCGCTAGCGATGGCCACACTGTTCGGAATC-3' (reverse, *Nhe*I site underlined). The 2022 bp product and pProEX\_pGFPuv1 were digested using *Hind*III and *Nhe*I, with calf intestinal alkaline phosphatase (CIAP) added to the vector digest. The wild type insert was ligated into pProEX\_pGFPuv1 using T4 DNA ligase and the ligation mixture was used to transform XL1-Blue cells which were plated on Luria-Bertani (LB) agar plates containing 100 µg/mL ampicillin. Resulting colonies were grown overnight in 4 mL cultures and plasmids purified with a NucleoSpin® plasmid DNA preparation kit to isolate the plasmid DNA. The plasmids were analyzed via restriction digest with *Hind*III and *Nhe*I to verify the presence of the

desired insert. Positive clones were then sequenced using forward primer 5'-AGCGGATAACAATTTACACAGG-3', internal primer 5'-ATCGATGCCAAATGCCACAA-3', and reverse primer 5'-ATCTTCTCTCATCCGCCAAAAC-3'.

RebG was PCR amplified from pET28a\_RebG (cloned previously in the Zechel lab by Dr. Anupam Bhattacharya) using the forward primer 5'-GGGCAAAGCTTATGGGCGCACGAGTGCTG-3' (*Hind*III site underlined) and reverse primer 5'-GCCGCTAGCGACGAGGCCCTCGATCAGG-3' (*Nhe*I site underlined), and PfuUltra™ polymerase. The PCR product was digested with *Hind*III and *Nhe*I and ligated into the previously digested screening vector pProEx\_GFPuv1. XL1-Blue cells were transformed with the ligation mixture and grown on LB-Agar plates containing 100 µg/mL ampicillin. Colonies were picked and grown overnight in a 4 mL culture to isolate the plasmid DNA. The plasmid DNA was analyzed via restriction digest and any clones producing an insert of 1263 bp were sequenced using the forward primer 5'-AGCGGATAACAATTTACACAGG-3' and the reverse primer 5'-ATCTTCTCTCATCCGCCAAAAC-3' to verify no mutations were incorporated during the cloning process.

#### **2.2.4 Mutagenic PCR of 5LO and RebG**

The mutagenic PCR protocol followed was based the procedure described by Caldwell and Joyce [15, 16] in which the mutagenic components consist of 0.5 mM MnCl<sub>2</sub>, 4.8 mM additional MgCl<sub>2</sub>, and 0.8 mM additional dCTP and dTTP. For a successful mutagenic PCR that provides sufficient desired product with minimal side products, both the Mg<sup>2+</sup> and Mn<sup>2+</sup> concentrations needed to be optimized. The reaction mix used for both 5LO and RebG is listed in **Table 2-1**. Not listed in the table are the template plasmid DNA and DMSO, for both of which 1 µL was added. The templates used were the wild type 5LO and RebG genes in pProEx\_GFPuv1.

**Table 2-1** Mutagenic PCR reaction mixes for 5LO and RebG

Reagent	5LO			RebG		
	Stock conc.	Volume ( $\mu$ L)	Final conc.	Stock conc.	Volume ( $\mu$ L)	Final conc.
ddH <sub>2</sub> O		34			31.5	
10 x Buffer*		5			5	
dNTPs	25 mM each	1	0.125 $\mu$ M	25 mM each	1	0.125 $\mu$ M
Extra dTTP	100 mM	0.5	1 mM	100 mM	0.5	1 mM
Extra dCTP	100 mM	0.5	1 mM	100 mM	0.5	1 mM
Extra MgCl <sub>2</sub>	50 mM	2	2 mM	50 mM	2	2 mM
MnCl <sub>2</sub>	10 mM	2.5	0.5 mM	4 mM	5	0.4 mM
For. Primer	10 $\mu$ M	1	0.2 $\mu$ M	10 $\mu$ M	1	0.2 $\mu$ M
Rev. Primer	10 $\mu$ M	1	0.2 $\mu$ M	10 $\mu$ M	1	0.2 $\mu$ M
Taq (NEB)	5000 U/mL	0.5	2.5 U	5000 U/mL	0.5	2.5 U
Total		48			48	

\*the 10x reaction buffer supplied by NEB included 100 mM Tris, 1.5 mM MgCl<sub>2</sub>, and 50 mM KCl

The PCR program used for RebG was 94 °C for 3 minutes, 30 x (94 °C for 30 seconds, 72 °C for 30 seconds, 72 °C for 2 minutes), 72 °C for 5 minutes. The program used for 5LO was 94 °C for 3 minutes, 25 x (94 °C for 30 seconds, 54 °C for 30 seconds, 72 °C for 2.5 minutes), finishing with 72 °C for 5 minutes.

### 2.2.5 DNA shuffling of 5LO and RebG

The DNase I digest for both 5LO and RebG were performed under the same conditions with the only difference being the length of digestion time. The gene pool collected from the previous round of evolution was used as a template for PCR amplification to generate enough

material for DNase I digestion. The PCR product was purified via a gel extraction kit (QIAGEN) and DNA concentration was estimated by measuring absorbance at 260 nm and using the formula

$$[\text{DNA}] = A_{260} \times 0.020 \text{ (}\mu\text{g/ml)}^{-1} \text{ cm}^{-1} \quad [17]$$

Approximately 2-5  $\mu\text{g}$  of DNA was used per DNase I digest. The total reaction volume was 50  $\mu\text{L}$  and consisted of 5  $\mu\text{L}$  10  $\times$  reaction buffer (100 mM Tris, 1 mM  $\text{CaCl}_2$ , pH 7.5), 5  $\mu\text{L}$  100 mM  $\text{MnCl}_2$  (final concentration 10 mM), and 37  $\mu\text{L}$  of purified PCR product and water. This mix was precooled at 15  $^\circ\text{C}$  for 10 minutes prior to adding 3  $\mu\text{L}$  of 0.1 U/ $\mu\text{L}$  DNase I. For 5LO, after the addition of DNase I the digest was allowed to continue for 1.5 minutes at 15  $^\circ\text{C}$ , and then terminated by addition of 1  $\mu\text{L}$  of 25 mM EDTA and incubating at 90  $^\circ\text{C}$  for 10 minutes. The same protocol was followed for RebG, with the only change being the length of digestion was 1 minute instead of 1.5 minutes. The digestions were done in triplicate and the fragments from each were combined during purification via QIAquick<sup>®</sup> nucleotide removal kit (QIAGEN) to remove fragments below 17 bp before use in the reassembly

The reassembly protocol used was the same for both RebG and 5LO. The reaction mix consisted of 10  $\mu\text{L}$  of purified fragments, 0.5  $\mu\text{L}$  dNTPs (25 mM each), 4  $\mu\text{L}$  5  $\times$  Herculase II reaction buffer (which provides a final  $\text{Mg}^{2+}$  concentration of 2 mM), 1  $\mu\text{L}$  DMSO, and 4.5  $\mu\text{L}$  of water for a final volume of 20  $\mu\text{L}$ . 1  $\mu\text{L}$  Herculase II was added during the first step for a “hot start.” The thermocycler program was: 96  $^\circ\text{C}$  for 3 minutes, 40 x (94  $^\circ\text{C}$  for 1 minute, 55  $^\circ\text{C}$  for 1 minute, 72  $^\circ\text{C}$  for 1 minute plus 5  $^\circ\text{C}$  per cycle), finishing with 72  $^\circ\text{C}$  for 7 minutes. Reassembled products were purified over a QIAquick<sup>®</sup> PCR purification spin column (QIAGEN).

To amplify full-length reassembled 5LO, a PCR was performed using the purified reassembly PCR products as the template and the same forward and reverse flanking primers used for cloning of the wild type DNA (Section 2.2.3). The mix consisted of 5  $\mu$ L 10x PfuTurbo reaction buffer, 1  $\mu$ L dNTPs (25 mM each), 1  $\mu$ L of 10  $\mu$ M each primer, 1  $\mu$ L DMSO, and 1  $\mu$ L of the purified reassembly products. The reaction was initiated by a ‘hot-start’ with 0.5  $\mu$ L of PfuTurbo polymerase added during the first step of the PCR program at 94 °C. The program used was 94 °C for 3 minutes, 30  $\times$  (94 °C for 30 seconds, 50 °C for 30 seconds, 72 °C for 2.5 minutes), finishing with 72 °C for 5 minutes.

The reaction mix for the amplification of RebG was identical with the only changes being that the primers used were the forward and reverse flanking primers described in the cloning of wild type RebG (Section 2.2.3), and the template used was the purified reassembly products from the reassembly of RebG. Also, the program used was 94 °C for 5 minutes, 30  $\times$  (94 °C for 30 seconds, 72 °C for 30 seconds, 72 °C for 1 minute), finishing with 72 °C for seven minutes.

The single products generated from the amplification of reassembled 5LO and RebG were isolated on a 1% agarose gel and purified using a QIAquick® Gel Extraction kit, then digested with *Hind*III and *Nhe*I for 1.5 hours. The digested and purified library was then ligated into the screening vector pProEx\_GFPuv1.

### **2.2.6 Rational truncation of lipoxygenases and tagged random-primer PCR**

Four lipoxygenase homologs were rationally truncated to match the truncated version of crystallized human 12LO (PDB accession code 3D3L). **Figure 2-7** shows a portion of the amino acid sequence alignment performed with Clustal W [18] to determine the starting codon (highlighted in yellow) for each truncated variant. Primers were designed to anneal along the portions of these genes, with a *Hind*III site flanking the new starting codon. The primers used for each gene are listed in **Table 2-2**. Amplification using these primers produced truncation

products with lengths of 1485 bp (8LO), 1473 bp (15LO), 1476 bp (12LO), and 1497 bp (5LO).

All products were ligated into the screening vector pProEx\_GFPuv1 to determine fluorescence.

```

human 5LO      LRDGRAKLARDDQIHILKQHRKELETRQKQYRMEWNPGFPLSIDAKCHKDLPRDIQFD 171
mouse 8LO      LREGAAKVSQDHHPTLQDQRQKELESRQKMYSWKTYIEGWPRCLDHETVKDLDLNLKYS 178
human 12LO     LPEGTARLPGDNALDMFQKHREKELKDRQQIYCWATWKEGLPLTIAADRKDDLPPNMRFH 167
human 15LO     LPEGTGRTVGEDPQGLFQKHREEELEERRKLYRWGNWKDGLILNMAGAKLYDLPVDERFL 167
short 12LO     -----

human 5LO      SEKGVDFVLNYSKAMENLFINRFMHMFQSSWNDFADFEKIFVKISNTISERVMMNHQEDL 231
mouse 8LO      AMKNAKLFKHAHSAYTELKVKGLLDRTG-LWRSLEMRRLFNFRKTPAAEYVFAHWQEDA 237
human 12LO     EEKRLDFEWTLKAGALEMALKRVYTLIS-SWNCLEDFDQIFWGQKSALAEKVRQCWQDDE 226
human 15LO     EDKRVDFEVS LAKGLADLAIAKDSLNLVT-CWKDLDDFNRIFWCGQSKLAERVRDSWKEDA 226
short 12LO     ---LDFEWTLKAGALEMALKRVYTLIS-SWNCLEDFDQIFWGQKSALAEKVRQCWQDDE 55
  
```

**Figure 2-7.** Alignment of human 5LO, human 15LO, and mouse 8LO with truncated human 12LO to determine starting residue of truncated variants. The starting residue is highlighted in yellow.

**Table 2-2.** Primers for truncation of lipoxygenases

Lipoxygenase	Starting codon	Forward Primer ( <i>Hind</i> III sites in bold)	Reverse Primer ( <i>Nhe</i> I sites italicized)
Human 5LO	V176	5'GGGCA <b>AGCT</b> TTGTGGACT TTGTTCTGAATTACTC-3'	5'- <i>CGCCGCTAGCGATGGCCAC</i> -3'
Mouse 8LO	A183	5'GGGCA <b>AGCT</b> TGCCAAAC TCTTCTTTAAAGC-3'	5' <i>CGCCGCTAGCGATGGAGACAC</i> TGTTCTCA-3'
Human 12 LO	L172	5'GGGCA <b>AGCT</b> TCTGGACT TTGAATGGACACTG-3'	5'- <i>CGCCGCTAGCGATGGTGAC</i> -3'
Human 15 LO	V172	5'GGGCA <b>AGCT</b> TTGTTGACT TTGAGTTTCGCTG-3'	5' <i>CGCCGCTAGCGATGGCCACACT</i> GTTTTCCA-3'

Tagged-random primer PCR to generate random truncated variants of 5LO was accomplished following the protocol by Jacobs *et al.* [19]. Two primers were used in two sequential PCRs. The first primer consisted of a 5'- 15 bp sequence encoding a *HindIII* restriction site, and a 3'- 15 bp randomized sequence that could potentially anneal to any portion of the 5LO gene (5'-GTATTTTTCAGGGCAAGCTTNNNNNNNNNNNNNNNN-3'). The second primer was identical to the first but without the random tail (5'-GTATTTTTCAGGGCAAGCTT-3'). The first PCR mix contained the random primer (100  $\mu$ M), 25 mM each dNTP, gel purified wild type 5LO template (no backbone present) and 10  $\times$  PfuUltra II reaction buffer, in a total reaction volume of 25  $\mu$ L. 1 mL of PfuUltra II polymerase was added during the first step of the thermocycler program at a temperature of 95  $^{\circ}$ C. The thermocycler program used was 10  $\times$  (95  $^{\circ}$ C for 1 minute, 40  $^{\circ}$ C for 3 minutes, 68  $^{\circ}$ C for 3 minutes), finishing with 68  $^{\circ}$ C for ten minutes. The products from this PCR were purified with a QIAquick<sup>®</sup> PCR purification kit, eluted in a final volume of 30  $\mu$ L. All 30  $\mu$ L was used in the second PCR which also included 5  $\mu$ L 10  $\times$  PfuUltra II reaction buffer, 0.5  $\mu$ M of a mixture of 25 mM each dNTP, 5  $\mu$ L of the second primer (100  $\mu$ M stock), and 8.5  $\mu$ L of water for a final volume of 49  $\mu$ L. 1  $\mu$ L of PfuUltra II polymerase was added during the first 95  $^{\circ}$ C step of the thermocycler program. The thermocycler program used was 30  $\times$  (95  $^{\circ}$ C for 1 minute, 55  $^{\circ}$ C for 1 minute, 68  $^{\circ}$ C for 1 minute), finishing with 68  $^{\circ}$ C for ten minutes. The products from this PCR were gel purified to isolate only those products between 300 and 1500 bp, and digested with *HindIII* for ligation into the screening vector pProEx\_GFPuv.

### **2.2.7 Screening of library variants for improved fluorescence**

To ensure the ligations were successful, control ligations were performed with water in place of insert along with the regular ligations. 5  $\mu$ L of the control ligation reaction and regular ligation reaction were each used to transform chemically competent XL1-Blue cells via heat-

shock and plated on LB agar plates containing 100 µg/mL ampicillin. A ratio of colonies on the control ligation plate versus the ligation plate of at least 1:20 was ensured before libraries were screened.

Successful ligations were purified with a QIAquick® PCR purification kit (QIAGEN) before transforming via electroporation into ElectroTen-Blue electroporation competent cells. Two µL of pure ligation mix was used to transform 50 µL of electrocompetent cells via electroporation at 1800 V to produce an average time constant of 4.8. One mL of LB was immediately added to the electroporated cells and the cells were recovered by shaking at 37 °C for 1.5 hours. After recovery cells were plated on 500 cm<sup>2</sup> LB agar plates containing 100 µg/mL ampicillin and grown overnight at 37 °C. In the morning the plates were removed from the incubator and allowed to rest for another 24 hours at 4 °C to ensure maximal fluorescence was reached. This procedure yielded libraries with no less than 20,000 and up to 50,000 colonies on one plate per transformation.

Libraries were viewed by eye under 400 nm LED lights wearing glasses with yellow lenses to filter out blue light and observe green fluorescence. Colonies which appeared to have increased fluorescence were streaked on another LB agar plate containing 100 µg/mL ampicillin. If more than 50 fluorescent colonies were visible, only 50 were initially picked. If less than 50 were visible, all were picked. Second and third transformations were necessary in some cases to isolate even 30 fluorescent colonies. After growing overnight at 37 °C, the fluorescence of the streaks was compared by eye to the fluorescence of a wild-type streak. Those streaks with enhanced fluorescence were further analyzed by colony PCR.

Any streaks showing an insert of the correct size after colony PCR were then purified to ensure only one cell with one plasmid was giving rise to the correct insert and increased fluorescence. To isolate selected clones, cells from each streak were grown overnight in 4 mL LB

cultures containing 1 % glucose and 100 µg/mL ampicillin, and plasmid DNA was isolated via NucleoSpin® Plasmid DNA kit. The plasmid DNA was diluted 10-fold, and 1 µL used to transform chemically competent XL1-Blue cells which were plated on LB agar plates containing 100 µg/mL ampicillin, and grown overnight at 37 °C. The resulting colonies were examined for homogeneity of fluorescence. If cells exhibiting two or more types of fluorescence intensity were observed, the most fluorescent colony was picked to be grown overnight. The plasmid DNA isolated from this colony was analyzed via restriction digest to ensure the presence of the correct insert. If insert was present the plasmid was checked for purity again by transforming chemically competent XL1-Blue cells and ensuring homogeneous fluorescence of the resultant colonies. 1 µL of each pure plasmid was combined and used as the gene pool for the subsequent round of evolution.

### **2.2.8 Expression and determination of excitation and emission maxima of GFPuv**

pProEx\_GFPuv without the insertion of the stop codon sequence between the *HindIII* restriction sites was used to transform BL21 cells. The cells were grown overnight at 37 °C on LB agar plates containing 100 µg/mL ampicillin to produce single fluorescent colonies. One colony was picked to inoculate a 4 mL culture containing LB, 1 % glucose and 100 µg/mL ampicillin, which was shaken overnight at 250 rpm, 37 °C. 0.5 mL of this preculture was used to inoculate a 50 mL culture containing LB, 1 % glucose and 100 µg/mL. The 50 mL culture was grown at 37 °C, 250 rpm until an OD<sub>600</sub> of 0.6 was reached, at which time it was induced with 1 mM isopropyl β-D-galactoside (IPTG). After induction the culture was grown for an additional 4 hours at 37 °C, 250 rpm. Cells were harvested via centrifugation at 3000 g for 20 minutes, at which point they were either stored at -20 °C or immediately lysed. Cells were lysed with an EmuSiFlex cell homogenizer (Avestin, Inc., Ottawa) and the supernatant was purified via immobilized metal ion affinity chromatography (IMAC) over Ni-NTA resin (QIAGEN). The

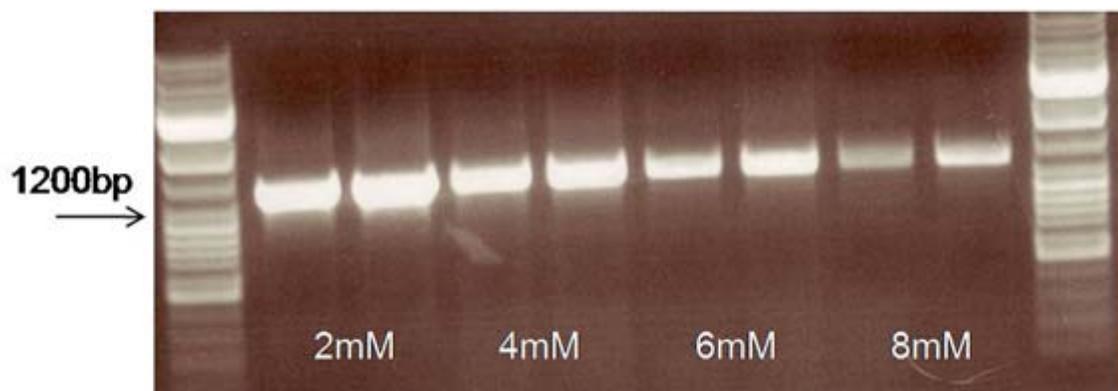
supernatant was bound to the column in a buffer containing 25 mM Tris, 300 mM NaCl and 10 mM imidazole, pH 7.2. The protein was eluted from the column with buffer comprised of 25 mM Tris, 300 mM NaCl and 500 mM imidazole, pH 7.2.

The fluorescence of pure GFPuv was measured on a Photon Technology International (PTI) fluorimeter, and data collected with FeliX32 software (PTI). The excitation maximum was scanned over a range of 300 to 460 nm with the emission maximum set to 508. The emission maximum was scanned over a range of 450 to 600 nm with the excitation maximum set to 399.

## 2.3 Results and Discussion

### 2.3.1 Optimization of error-prone PCR

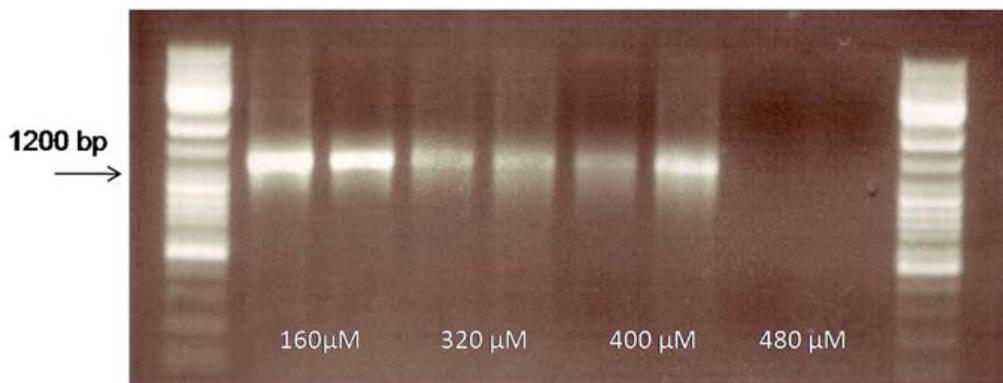
To ensure a sufficiently high mutation rate for epPCR it is essential that a low-fidelity polymerase is used, typically *Taq* polymerase. The additional mutagenic components of an epPCR ( $Mn^{2+}$ , excess  $Mg^{2+}$ , unbalanced nucleotides) are likely to have a negative effect on the yield and specificity of the reaction, so prior to starting an epPCR it is a good idea to ensure that the 'standard' PCR is already optimized for the selected template. *Taq* polymerase activity is dependent on the presence of  $Mg^{2+}$  ions [20]; however the optimal concentration required varies for different templates and when starting a PCR with a new target, a variety of  $Mg^{2+}$  concentrations should be checked. **Figure 2-8** shows the standard PCR reaction of RebG with varying amounts of  $MgCl_2$  added. The epPCR protocol used in this study called for 4.8 mM additional  $MgCl_2$  to be added on top of the usual 1.5 to 2 mM [16], but as shown in the figure, as the amount of additional  $MgCl_2$  increases, product yield decreases. Thus 2 mM was chosen as the optimal concentration to use in addition to the regular amount of  $MgCl_2$  prior to adding the other mutagenic components. Because the buffer supplied with *Taq* polymerase already included  $MgCl_2$  at a concentration of 1.5 mM, the total amount of  $MgCl_2$  in the reaction mix was 3.5 mM. The amount of additional  $Mg^{2+}$  could be increased again later if it is discovered that the reaction is tolerant to  $Mn^{2+}$ .



**Figure 2-8.** Effect of  $Mg^{2+}$  concentration on PCR yield of RebG. Each reaction also contained 1.5 mM of  $Mg^{2+}$  as supplied by the reaction buffer.

Increasing the concentration of  $MnCl_2$  had an even more marked effect on the final product yield of the RebG PCR, as shown in **Figure 2-9**. The chosen protocol suggested an initial concentration of 0.5 mM  $MnCl_2$  in the reaction [16,17], and clearly with RebG as a template, a  $MnCl_2$  concentration that high would not result in a successful PCR even when using the previously optimized additional  $Mg^{2+}$  concentration of 2 mM. Thus a balance must be made between sufficient amounts of mutagenic components to achieve an appropriate mutation rate, yet not so much that the yield is considerably diminished. In the case of RebG a final  $Mn^{2+}$  concentration of 0.4 mM was chosen as it was the closest concentration to the recommended concentration of 0.5 mM that still produced a product. 5LO epPCR trials indicated that a satisfactory yield could be obtained using 0.5 mM of  $MnCl_2$ . It should also be noted that prior to discovering the optimal  $Mg^{2+}$  and  $Mn^{2+}$  concentration it was found that the addition of 1  $\mu$ L of DMSO to a 50  $\mu$ L reaction resulted in a drastic improvement of both specificity and yield. DMSO is a common PCR additive and although not essential, it has been shown in many cases to improve the specificity of a PCR [21]. Specificity can also be increased by the length of the primers used and the annealing temperature chosen for the cycling reaction. Care must also be

taken when choosing the elongation time, as some polymerases will copy much slower than others, and an elongation time too short will give a lower yield.



**Figure 2-9.** Effect of  $Mn^{2+}$  concentration on the yield of a RebG mutagenic PCR.

The protocol chosen for this study was analyzed by Wong *et al.* in their MAP (mutagenesis assistant program) analysis of 19 mutagenic PCR protocols and it was found to be biased towards transitions over transversions, yet introduced a lower number of stop codons and helix disrupting mutations (to proline or glycine) than a non-biased protocol would [22]. The amounts and types of mutations incorporated using this protocol were analyzed and are summarized in **Table 2-3**. The 5LO reaction introduced more mutations than in the RebG case, probably due to its ability to withstand a higher  $Mn^{2+}$  concentration, producing an average of 5 mutations over the length of the 2022 bp gene. Of the total 25 5LO mutations analyzed, only one transversion was found. Also, most mutations ended up causing amino acid changes, with only five out of the twenty five being silent. For the five clones analyzed from the RebG mutagenic PCR, only two showed mutations, with the other three being identical to wild type. Only one amino acid changing mutation was found for all five clones, and four silent mutations. The protocol followed, which called for more  $Mg^{2+}$  and  $Mn^{2+}$  than was actually used, reported a mutation rate of  $7 \times 10^{-3}$  per nucleotide per duplication. The error rate for both 5LO and RebG were higher than this reported value, with the 5LO mutagenic PCR producing an error rate of  $6 \times$

$10^{-2}$  per nucleotide per duplication (25 cycles), and the RebG mutagenic PCR producing an error rate of  $2 \times 10^{-2}$  per nucleotide per duplication (30 cycles).

**Table 2-3 Analysis of mutation load after epPCR for 5LO and RebG**

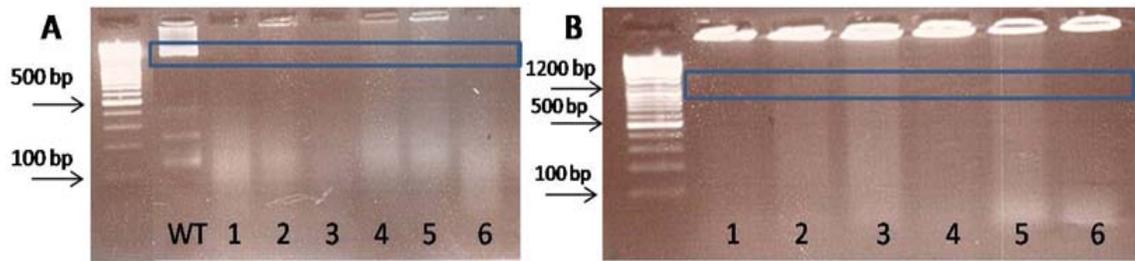
Clone	Number of transitions (A↔G)	Number of transversions (C↔T)	Silent mutations	Non-silent mutations
5LO-1	4	0	1	3
5LO-2	5	1	1	5
5LO-7	2	0	0	2
5LO-8	8	0	3	5
5LO-13	5	0	0	5
Reb-1	2	1	2	1
Reb-2	0	0	0	0
Reb-3	0	2	2	0
Reb-5	0	0	0	0
Reb10	0	0	0	0

### 2.3.2 Optimization of DNase I digestion

Successful DNA shuffling requires high stringency, especially during the DNase I digest. It was shown by Maheshri *et al.* that both the average fragment size (AFS) and the concentration of fragments going into the reassembly step are crucial determinants of the quality of recombination [23]. In their comparison of GFP reassembly under four different conditions they observed that using a low fragment concentration (8 ng/μL) and larger AFS (50 bp, minimum of 50 bp), produced the most reassembled product of the correct size (~ 750 bp). Alternatively, increasing the concentration of fragments (120 ng/μL) and decreasing the AFS (45 bp, minimum

of 25 bp), greatly increased the amount of products smaller and larger than 750 bp, both experimentally and computationally. Stemmer also made the observation in his original DNA shuffling paper that when recombining fragments 100 – 200 bp in size a single product can be achieved, but when recombining fragments 10 – 50 bp, some products of the correct size as well as “products of heterogeneous molecular weights” are observed [24]. We also found that the polymerase used for both the recombination and amplification PCRs had a profound effect on the outcome of DNA shuffling.

Prior to reassembling digested 5LO and RebG, trials to find the optimal digestion time were run on both templates. Digestion protocols are quite varied in the literature, however depending on the chosen method, conditions for reassembly and frequency of crossovers can be somewhat controlled. **Figure 2-10 A** shows the time-trials for 5LO and **Figure 2-10 B** shows the trials for RebG. Interestingly, even after a 15 second digestion, no full length template is evident on either agarose gel, indicating the digestion occurs very rapidly. For RebG, two different concentrations of DNase I (0.05 U and 0.3U) were tried, and the results are similar for both concentrations, although it seems the intensity of the fragments below 100 bp is highest for the longest digestion time at both concentrations of DNase I. RebG is shorter than 5LO (1263 bp vs. 2022 bp) and thus digestion under the same conditions produces fragments much smaller than fragments from the digestion of 5LO. Length of digestion should therefore be reduced for shorter templates to ensure a sufficient amount of fragments long enough to be used for recombination remain after purification.

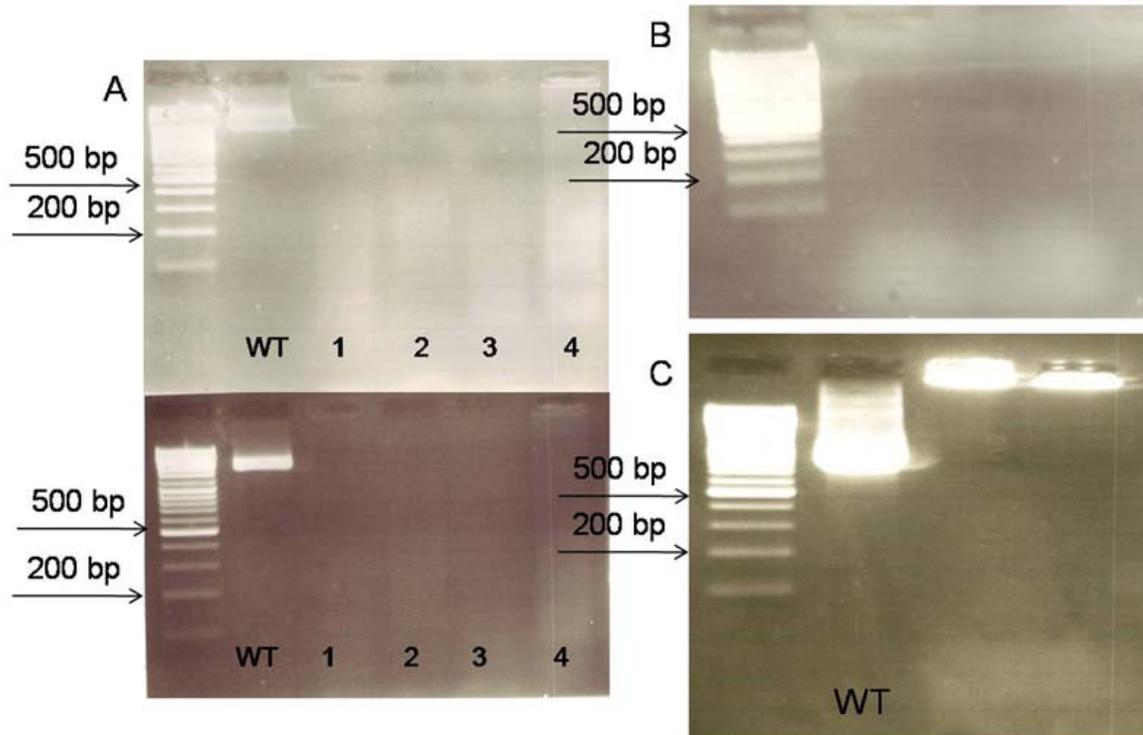


**Figure 2-10.** Dnase I digestion time trials for 5LO (A) and RebG (B). For the digestion of 5LO, 0.3 U of DNaseI was used and the digestion time for each sample increases for lanes 1-6 in the order of 15 s, 30 s, 45 s, 1.0 min, 1.5 min, and 2.0 min. For RebG (B), samples in lanes 1, 2, and 3 were digested with 0.05 U of DNase I for 0.25, 0.50, and 1.0 min respectively. Lanes 4, 5, and 6 were digested with 0.3 U of DNase I for 0.25, 0.50, and 1.0 min respectively. All digestions were performed at 15°C and terminated by the addition of 1  $\mu$ L of 25 mM EDTA and incubation at 90 °C for ten minutes. The box indicates where the full-length template would be.

The fact that the digestion is very rapid has been noted several times in the literature; Lorimer and Pastan observed that when using only 0.0003 U of DNase I on 4.5  $\mu$ g of template at 15 °C for 1 minute, the template appears to be completely digested [25]. Continuing the digest for up to 20 minutes resulted in only a slight decrease in the size of the fragments and slight increase in the intensity of the fragments on an agarose gel, leading them to conclude that DNase I is much less active on smaller fragments and thus the timing of the digest is not a critical factor. Because of these results, the usual gel purification step to select the size of fragments to use for reassembly was omitted and replaced with a simple gel extraction to remove any metal ions and fragments smaller than 16 bp. Although the size of the fragments does not seem to change much on a gel, leaving the fragments to digest for 20 minutes would not be a good idea, as DNase I can act on fragments as small as 12 bp to generate products around 6 bp [26]. Fragments of this size will be removed upon the purification step, leaving no DNA available for recombination.

Based on the time trial results, a digestion protocol was chosen for 5LO and RebG using 0.3 U of DNaseI and 1.5 minutes of digestion time for 5LO and 1 minute of digestion time for

RebG at 15°C. This length of time was chosen to ensure no full-length template remained, yet not so over-digested such that the fragments become too small to be used in recombination. **Figure 2-11 A** shows the fragments generated from the digestion of 5LO using this protocol, before and after purification. In order to actually see the fragments on a gel, exposure time on the imaging system needed to be increased. The image captured under regular exposure conditions is shown directly beneath the overexposed image. As the figure shows, under these digestion conditions the fragments appear most intense just below 200 bp. This was also observed when conducting the time trials (**Figure 2-10 A**). The fragments from the digestion of RebG (**Figure 2-11 B,C**) are much smaller (< 100 bp), and interestingly much of the fragments seem to aggregate in the wells of the agarose gel (“glowing wells”). This is also evident on the gel showing the digestion time-trials of RebG (**Figure 2-10 B**). This phenomenon can occur for many reasons such as precipitation of the DNA, or a DNA concentration that is too high prior to electrophoresis. In this case it seems to be a result of using too much template for the digestion producing a very high fragment concentration. It did not occur when template that had been gel purified was used (gel purification generally reduces the final yield of pure DNA) (**Figure 2-11 B**), but it did occur when product that was column purified was used (generally high yield of purified product) (**Figure 2-11 C**). Also this problem does not seem to be associated with reagents or reaction conditions as the same conditions/reagents were used for the RebG and 5LO time trials, and it did not occur for 5LO. Finally, it is not a template-specific problem as it did also occur for PhnG on one occasion.

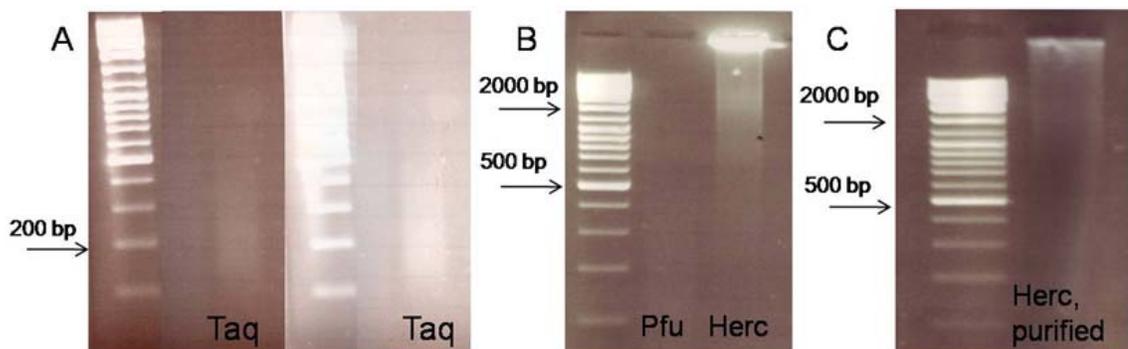


**Figure 2-11.** Fragments from the digestion of 5LO (A) and RebG (B ,C). For 5LO (A), lanes 1, 2 and 3 show fragments from three separate digestion reactions and lane 4 is the combined fragments after purification. The top and bottom images are from the same gel under different exposure conditions. B shows fragments after digestion of RebG with no product stuck in the wells, and C also shows fragments from the digestion of RebG, but with much of the product stuck in the wells. C also shows the wild type DNA used prior to digestion.

### 2.3.3 Optimization of recombination and amplification

The reassembly step was found to be highly dependent on the polymerase used. **Figure 2-12 A-C** shows the drastic effect changing the polymerase had on the reassembly and amplification of 5LO. *Taq* polymerase was used for the first attempt at recombination (**Figure 2-12 A**) and resulted in a smear on an agarose gel with the highest intensity located very low, around 200 bp. For an efficient and specific recombination, the highest intensity on a gel should centre on where the full length gene would be located. In this instance, because the recombination seemed to be incomplete, amplification off of the recombination products with

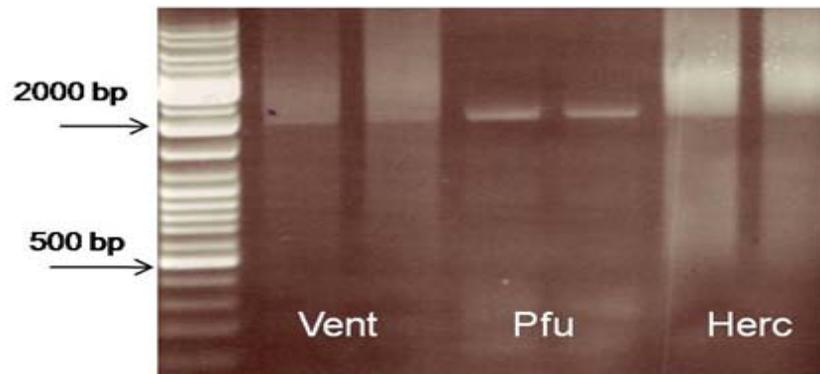
flanking primers was unsuccessful and no product could be detected on a gel. For the second attempt, PfuTurbo and Herculase II polymerases were tried with varied results. With PfuTurbo, recombination resulted again in a smear of low size (**Figure 2-12 B**), but recombination with Herculase II polymerase resulted in the complete opposite. Most of the products from the Herculase II recombination became stuck in the well of the agarose gel and made a smear of very high average length. Even after purification of this product over a QIAquick® PCR purification column, the product remained caught in the well (**Figure 2-12 C**). Both the Herculase II and PfuTurbo reactions were performed with fragments from the same digestion, thus fragment quality or concentration cannot be considered a reason for the discrepancy in this case. Because the products resulting from the Herculase II recombination did encompass the size of 5LO, amplification using flanking primers was tried using this recombination product as the template.



**Figure 2-12.** Recombination of 5LO with Taq (A), Pfu (B), and Herc (B) polymerase. The image showing Taq recombination (A) is shown under both regular and overexposed conditions.

**Figure 2-13** shows the products of the amplification reaction with three different polymerases—Vent, Pfu, and Herculase II. Interestingly, the Vent and Herculase II polymerases could not specifically amplify one product of the correct molecular weight out of the mixture of recombination products, but rather produced another mixture in which the smears showed the

highest intensity around the 2000 bp mark which indicates an enrichment of the correct full length product. Only Pfu could successfully amplify a singular product with a molecular weight matching that of wild type 5LO. As a result, a DNA shuffling protocol was designed for this template using Herculase II as the polymerase for recombination and Pfu as the polymerase for amplification.

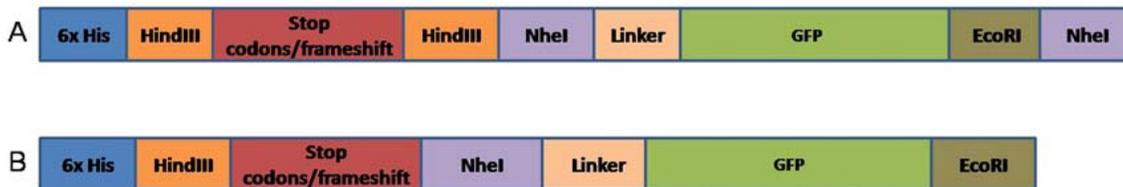


**Figure 2-13.** Amplification of recombined 5LO with Vent, Pfu, and Herculase II polymerases

#### 2.3.4 Screening vector construction

Two versions of the screening vector were constructed. In one construct (pProEx\_GFPuv, **Figure 2-14 A**) the library variants are cloned in the *Hind*III site upstream of the gene encoding GFPuv, and the GFPuv gene can be quickly removed via digest with *Nhe*I if it is of interest to express the target protein alone. Directional cloning is not an option with this construct as *Hind*III is the only available restriction site for insertion of the target gene. In the second construct (pProEx\_GFPuv1, **Figure 2-14 B**), the *Nhe*I site at the 3'-end of GFPuv has been eliminated, thus the library can be directionally cloned between the *Hind*III and *Nhe*I sites upstream of GFPuv. The disadvantage with this construct is that to express the target protein alone, it must be digested out of the screening vector and ligated into a modified pProEx vector that does not contain the gene encoding GFPuv. Since our screening methodology required selecting the brightest clones from libraries that were restricted in size to the number of clones

that can be screened on LB Agar plates, directional cloning was desired and pProEx\_GFPuv1 was used, thereby preventing half of the library from being ligated into the vector in the wrong direction. Both vectors have N-terminal hexahistidine tags for protein purification and the inclusion of a short sequence of DNA containing stop codons in all three frames between the cloning sites. This stop codon sequence is necessary to prevent the expression and fluorescence of GFPuv from occurring unless this sequence has been replaced with another insert during the ligation step. The promoter used in pProEx is the *trc* promoter which is much weaker than the bacteriophage T7 promoter, thereby slowing the production of protein and decreasing the chances of protein aggregation [27]. A final feature incorporated into both vectors is the F64L GFP folding mutation. A double mutant of GFP was discovered that included both the F64L folding mutation and the red-shift mutation S65T, and was shown to have a fluorescence intensity 35-fold greater than wild type GFP [28]. The screening vector construction was based on the screening vector designed by Kawasaki *et al* [14].



**Figure 2-14.** Screening vector constructs pProEx\_GFPuv (A) and pProEx\_GFPuv1 (B). pProEx\_GFPuv (A) does not allow for directional cloning, but allows for easy removal of GFP, whereas pProEx\_GFPuv1 (B) does allow for directional cloning, but the insert would have to be digested out of this vector and ligated into a separate vector lacking GFPuv to be expressed alone.

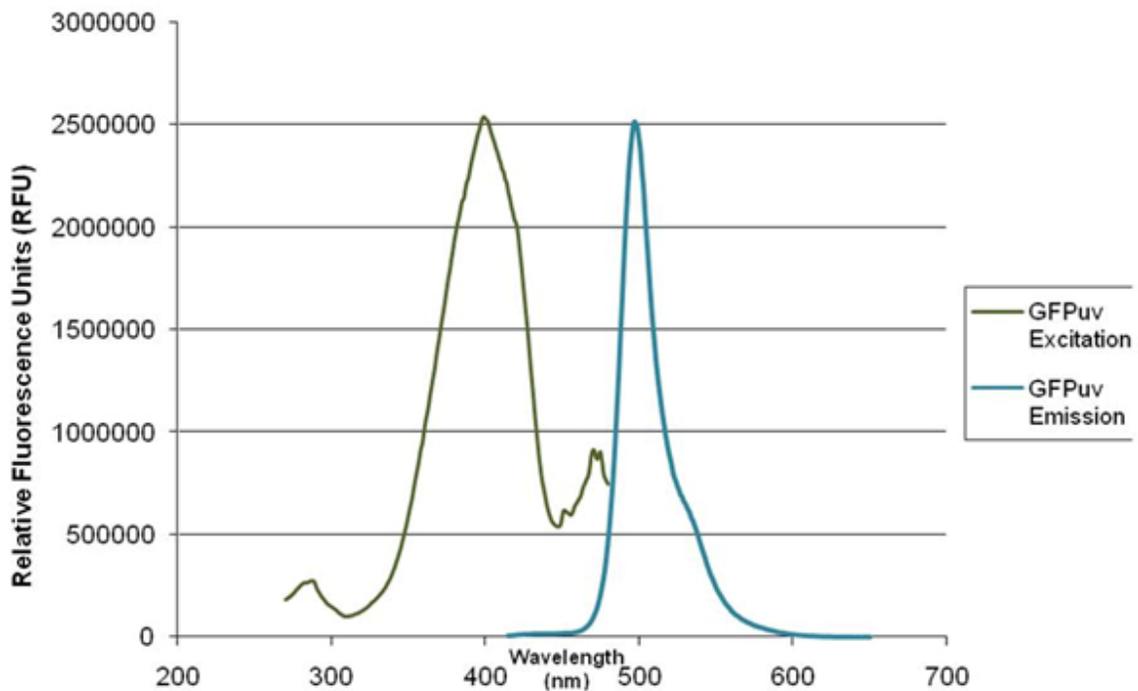
### 2.3.5 Considerations for effective screening

In the literature, screening is often presented as a relatively pain-free process in which the brightest colonies are easily selected by eye off of solid media, or by fluorescence activated cell

sorting (FACS). In reality, screening may not be straightforward and the importance of using the highest stringency when selecting and analyzing “improved” clones cannot be overstated. This point is illustrated by the several attempts at finding more fluorescent variants of 5LO- and RebG-GFP fusions.

For the first round of evolution on 5LO and RebG, epPCR was used to generate diversity and subsequently library variants were ligated into the screening vector construct pProEx\_GFPuv1 (**Figure 2-14 B**). Because it is necessary to screen a large number of clones at one time, electroporation was chosen as the method of choice for transforming the library into competent cells since electroporation is known to have vastly greater transformation efficiency (by at least one order of magnitude) than any other chemical transformation methods [29]. A large number of highly fluorescent clones were visible upon the first screen. Electroporation typically yielded ~25000 clones per 500 cm<sup>2</sup> plate and for the first round of evolution 50 fluorescent clones could easily be picked from just one plate. RebG proved to be more difficult in terms of selection than 5LO, as the wild type protein exhibits much higher basal fluorescence when fused to the GFP folding reporter. By eye it proved to be difficult to distinguish clones that were more fluorescent than the wild type fusion. Wild type 5LO exhibits almost no basal fluorescence and distinguishing clones that were brighter was much easier.

It is very important to carefully select the type of lighting used when performing the selections. The folding reporter we used was measured to have an excitation maximum of 399 nm and an emission maximum of 508 nm (**Figure 2-15**). This reporter is identical to GFPuv (a reported excitation maximum at 395 nm and emission maximum at 510 nm) purchased from Clontech except for the incorporated F64L mutation. Had the red-shifting S65T mutation of EGFP also been incorporated, the excitation maxima would be shifted to around 488 nm [30].

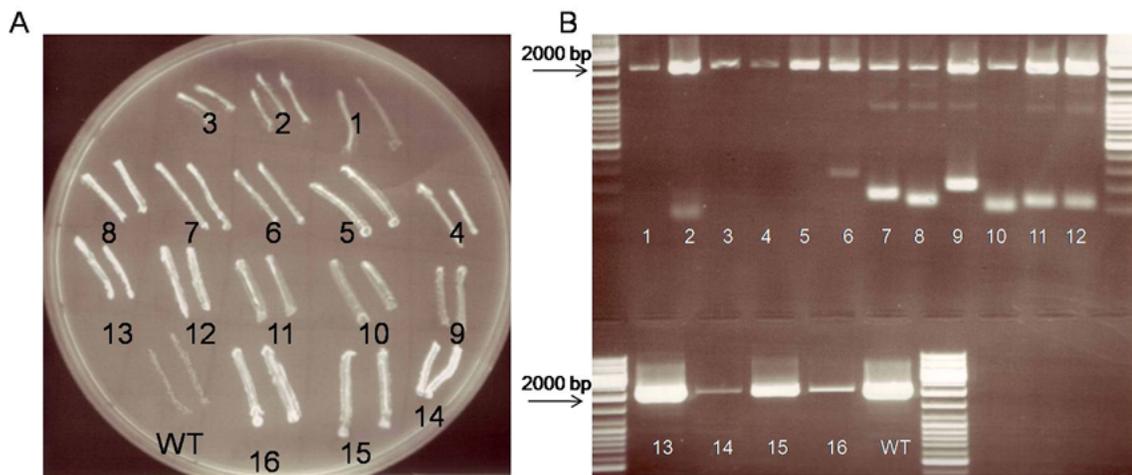


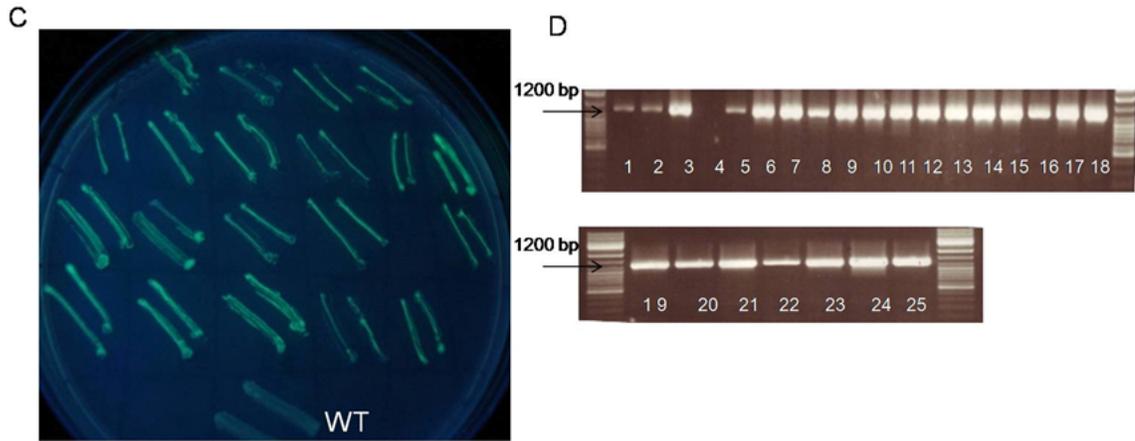
**Figure 2-15.** Determination of the excitation and emission maxima for the GFP folding reporter.

The first selections were performed by illuminating the plates from below on a transilluminator (Fisher Scientific) used in the lab for gel extractions which has a wavelength of 360 nm. Fluorescence is clearly evident when exciting at this wavelength, however due to the long periods of time needed to keep the plates illuminated while screening, it was discovered that most of the cells would become damaged to the point of cell death from being irradiated with short wavelength light. If basal fluorescence is high, screening the plates could take up to one hour, thus a less damaging light needs to be used to ensure the colonies picked remain viable and can be propagated once selected. To overcome this obstacle, a screening apparatus was constructed in which the plates are illuminated from above using a combination of less harsh LED lights that have a wavelength of 400 nm and glasses with yellow lenses which effectively filter out light of wavelengths between 190 and 480 nm, making the colonies appear less blue and more green. Using this combination allows effective detection of fluorescent colonies under

conditions mild enough to be able to irradiate the cells for long periods of time. It should be noted that maximal fluorescence would be observed after the plates were kept overnight at 4 °C as opposed to immediately after removal from the 37 °C incubator, consistent with the slow oxidative formation of the GFP chromophore [31].

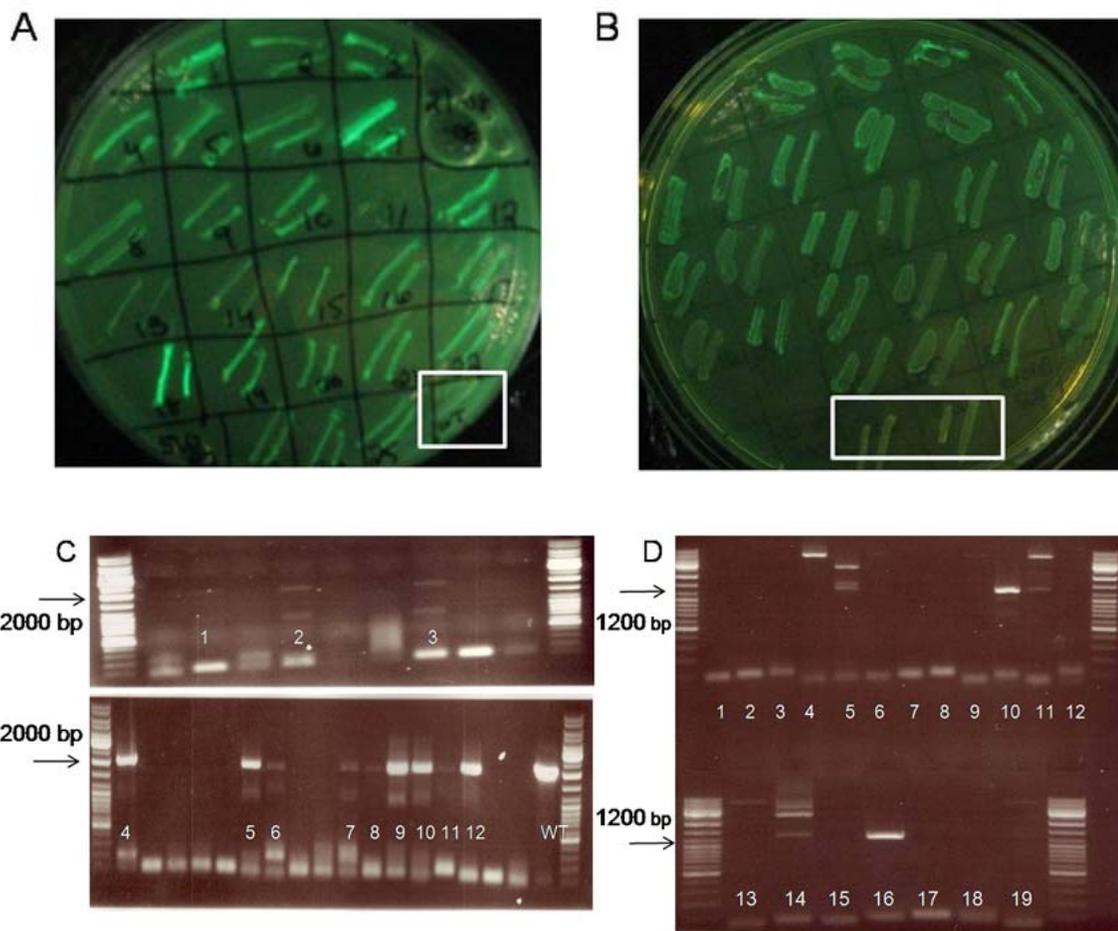
The first round of selections seemed to be very promising. Many clones were found for both 5LO and RebG that showed fluorescence brighter than the wild-type fusion and 50 were picked for each. Because of the stop codon sequence used to prevent false-positive readings that would arise from a construct lacking an insert, it was assumed that all fluorescent colonies found must have an insert present. To ensure the insert present was the correct one, colony PCR was performed on all of the selected clones, and as **Figure 2-16** shows, every one showed an insert of the correct size for both 5LO and RebG (minus lane 4 in **Figure 2-16 D**). From these results it was assumed that an improvement of fluorescence had been made for all of the selected clones and the PCR products from the colony PCRs were combined so that the mixture of templates could be shuffled in the next round of evolution.





**Figure 2-16.** Streaks and colony PCR of clones selected from round 1 of 5LO and RebG. 5LO streaks are shown in A and the colony PCR from each streak is shown in B. Examples of the fluorescence of colonies found in round 1 of RebG evolution are shown in C and the colony PCR from the streaks is shown in D. Not every clone that was selected is shown. Clones in A were imaged in a GeneGenius gel dock system (Syngene) whereas the clones in B were imaged by digital camera while being illuminated from above with 400 nm light.

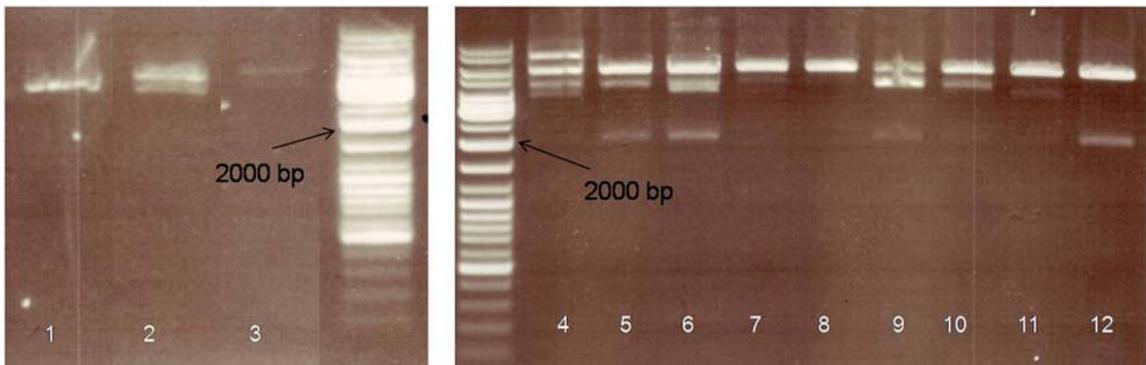
The second round of evolution for both RebG and 5LO is where unexpected results began to surface. Fluorescent colonies were still abundant (**Figure 2-17 A,B**) after DNA shuffling and ligation into the screening vector, however as **Figure 2-17 C,D** shows, only about half produced a product of the correct size after colony PCR for 5LO (numbered 1-12 on the gels) and only about 4 out of 19 (# 10, 11, 14, 16) showed an insert for RebG. The 5LO clones showing an insert were chosen for further analysis.



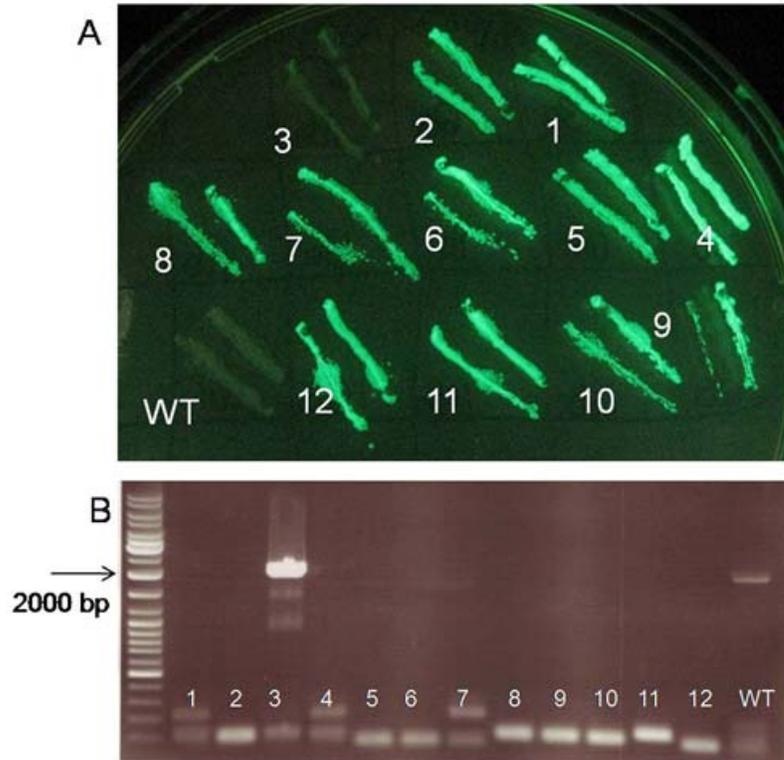
**Figure 2-17.** Examples of the fluorescence obtained from clones picked from round 2 for 5LO (A) and RebG (B). Wild type fluorescence is indicated by the white box. Colony PCR on some of the variants picked for 5LO (C) and RebG (D) is also shown.

For 5LO a total of twelve clones showed an insert from colony PCR and had varying degrees of fluorescence, thus cells from each streak were grown overnight in a 4 mL culture and the corresponding plasmids isolated. Restriction digest analysis of the plasmids gave varied results with some of the plasmids showing insert (**Figure 2-18**, # 12), some showing none (**Figure 2-18**, #1, 2, 3, 4, 7, 8, 11), and some having insert, but with more than two bands present indicating that perhaps a mixture of plasmids was present (**Figure 2-18**, # 5, 6, 10). Because some looked like mixtures, all 12 samples were retransformed into XL1-Blue cells, to determine

whether or not the cells would exhibit homogeneous fluorescence or would be a mixture of fluorescent and non-fluorescent cells. All samples were mixtures, except for one which was made up entirely of non-fluorescent cells. To purify the fluorescent colonies away from the non-fluorescent, fluorescent clones were picked off of each of the mixture plates and streaked onto a new plate (**Figure 2-19A**). This purification step resulted in streaks that were much brighter compared to the streaks from the initial selection, except for the one which came from the plate of only non-fluorescent cells (#3). The purified streaks were then analyzed by colony PCR and only the wild type streak and the other non-fluorescent streak showed an insert of the correct size (**Figure 2-19B**).



**Figure 2-18.** Restriction digest analysis of 5LO round 2 clones showing an insert from colony PCR. Although evident in the colony PCR, some clones lacked the insert upon restriction digest analysis. A single insert is visible for clone 12, and clones 5, 6, and 10 show insert but appear as mixtures. For all other clones no insert is visible.



**Figure 2-19.** 5LO round 2 purified clones (A) and colony PCR of the clones (B). Purification of the mixed plasmids resulted in streaks much brighter than the non-purified streaks, yet only the non-fluorescent one (lane 3) and wild-type (WT) contained an insert as detected by colony PCR.

To understand why streaks not showing insert were so fluorescent, five clones for both RebG and 5LO were sent in for sequencing. The results from the sequencing showed that small inserts had been ligated in place of the full length insert, mainly a result of “primer dimer”, or the forward and reverse primers annealing together and becoming extended during the PCR. These dimers have the proper restriction sites to become ligated into the screening vector and are small enough so as not to inhibit proper folding of GFP, leading to high fluorescence. These inserts are evident as bright bands at the bottom of the lanes in **Figure 2-19**. The sequences of the inserts are shown in **Figure 2-20**. For RebG (**Figure 2-20 A**), only one type of insert was found which is mainly comprised of the forward and reverse primers and also contains two base pairs of unknown origin. These two cytosine base pairs do not follow the primers in the forward or

reverse directions. For 5LO three different types of small inserts were found (**Figure 2-20 B**), one consisting mainly of the forward and reverse primers annealed together, another consisting of a portion of the reverse primer, and a third containing regions of the forward primer, the sequence of 5LO immediately following the forward primer, and a short sequence matching a region on GFP. Prior to ligation into the screening vector, all full length 5LO and RebG inserts were gel purified to try an avoid the insertion of these ‘primer-dimers’, however the gel purification was not completely effective in removing this type of contamination.

## A

RebG Forward Primer: 5'- **GGGCAAGCTTATGGGCGCACGAGTGCTG**-3'

RebG Reverse Primer: 5'-**GCCGCTAGCGACGAGGCCCTCGATCAGG**-3'

Sequence of small insert:

5'-**ATGGGCGCACGAGTGCTG**CCCCTGATCGAGGGCCTCGTC-3'

M G A R V L P L I E G L V

## B

5LO Forward Primer: 5'- **AGGGCAAGCTTATGCCCTCCTACACGGTCACCGTG**-3'

5LO Reverse Primer: 5'- **AGCGCCGCTAGCGATGGCCACACTGTTTCGGAATC**-3'

Sequence of small inserts:

1) 5'- **ATGCCCTCCTACACGGTCACCGTG**GCCACTCCGAACAGTGTGGCCATC-3'

M P S Y T V T V A T P N S V A I

2) 5'-**GCCCGCAGCGCC**-3'

A R S A

3) 5'-**ATGCCCTCCTACACGGTCACCGTG**GCCACTGGCAGCCAGTGGTTCGCCGGC**CCATATTAC**

M P S Y T V T V A T G S Q W F A G P Y Y

**TACTTGTCCCCTGAC**-3'

Y L S P D

**Figure 2-20.** Sequencing data from RebG (A) and 5LO (B) clones exhibiting very bright fluorescence but no insert. The primers used for RebG amplification are shown in A, with the forward primer being in red, and the reverse primer being in purple. The dimerized primer inserts are shown in B and coloured according to whether its sequence matches the forward or reverse primer. The underlined portions represent sequences along the gene that immediately follow the forward or reverse primer. The primers and insert for 5LO (B) are coloured in the same way, with the orange sequence representing a section annealing to GFP. Sequences in black that are not underlined have an unknown origin.

These results led to the conclusion that any of the fluorescent colonies found up until this point could be mixtures containing one proper length fusion exhibiting no improvement in fluorescence and another fusion with a short insert which allowed proper folding of GFP, or colonies which only contained only the short insert. The latter would not produce a product from colony PCR and the former would, but a mixture of the two in cells would exhibit increased fluorescence at the colony level. . Because it is desirable to screen as many colonies as possible on one plate, it is not uncommon for colonies to be so close together that it is virtually impossible to select only one without fear of contamination from a neighbouring colony. Also, because the clones you are selecting are sometimes very highly fluorescent, it can be difficult to see the non-fluorescent neighbouring colonies and you may not be aware that you are picking more than one colony. Another possible reason for the mixtures may be from two plasmids transforming a single cell. As Goldsmith *et al.* pointed out, in cases where you are maximizing the amount of plasmid DNA used to transform cells such as for library generation in directed evolution experiments, the occurrence of double transformants is greatly increased [35]. They observed that even when using two plasmids with different origins of replication and antibiotic resistance genes, double transformants contaminated up to 10% of the total population. In any case, it is critical to be sure that all clones you have isolated are a product of one cell containing only one plasmid. Only then is it safe to use this clone in the next round. If purity is not checked it is

possible to spend a great deal of time going through all the rounds only to realize that all of your fluorescent clones at the end are a result of contamination.

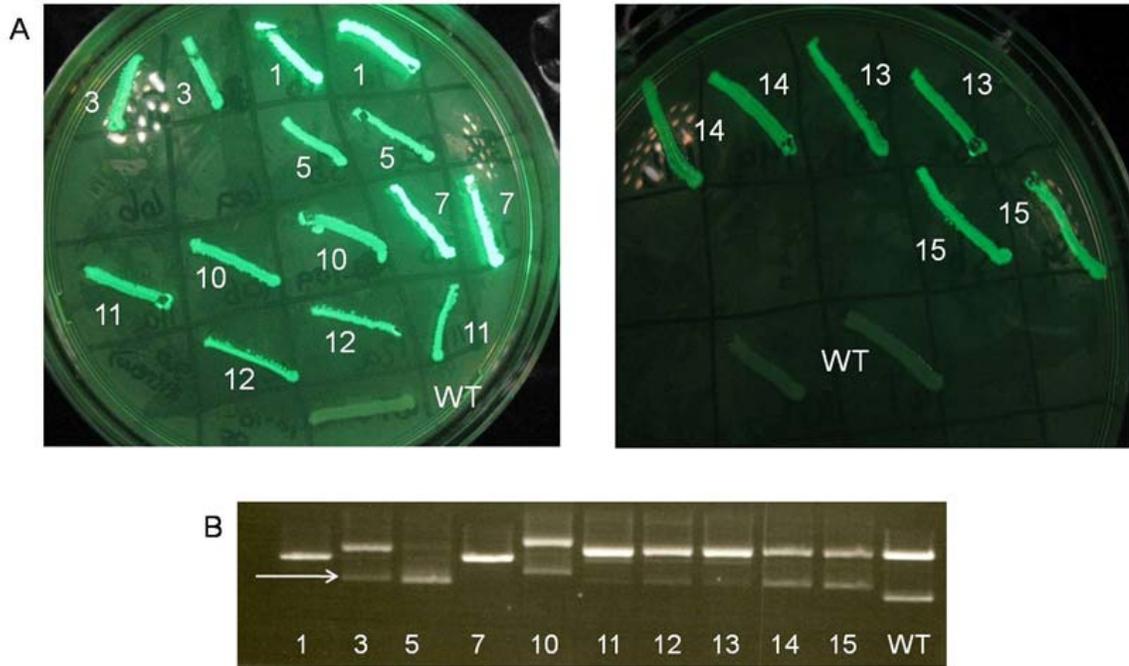
Because of these findings it was unclear whether the colony PCR products from the first round were actually from mutants exhibiting improved fluorescence, or from wild type contamination. It was thus decided that the first round for both 5LO and RebG should be repeated and restriction digest analysis rather than colony PCR would be used to analyze any improved clones.

To ensure the primer based inserts would no longer be a source of contamination, agarose gels of all gel purified inserts prior to ligation were run for longer periods of time, allowing complete separation of the small contaminant inserts and the full length mutant inserts. Due to this increased stringency with insert preparation, the number of fluorescent colonies visible on a plate of around 25000 colonies went down dramatically during the second attempt at round 1. For 5LO, screening of three 500 cm<sup>2</sup> plates (~ 75000 colonies) resulted in only 30 colonies that were believed to have brighter than wild type fluorescence, however upon examination of the resultant streaks only a few of the colonies picked actually ended up being brighter<sup>1</sup>. These streaks were all found to be mixtures and were purified by retransforming the plasmid DNA into XL1-Blue cells and re-picking only the fluorescent colonies. Two fluorescent colonies were picked for each plasmid during the purification step to ensure that the resulting streaks would exhibit equal fluorescence, which they did (**Figure 2-21 A**). Notably, out of the ten purified streaks, 1 and 7 appear to be the brightest. Restriction digest analysis showed that these two clones were different from the rest as they appeared to lack an insert. All other clones contained

---

<sup>1</sup> It was a common occurrence that colonies picked upon initial screening that were thought to have increased fluorescence, did not actually appear more fluorescent when analyzed as a streak the next day. This is due to human error resulting from long screening times and fatigue. Also, sometimes it is hard to tell by eye whether a colony is actually brighter and in these cases the ambiguous colony is picked anyways, but may not actually exhibit an improvement in fluorescence.

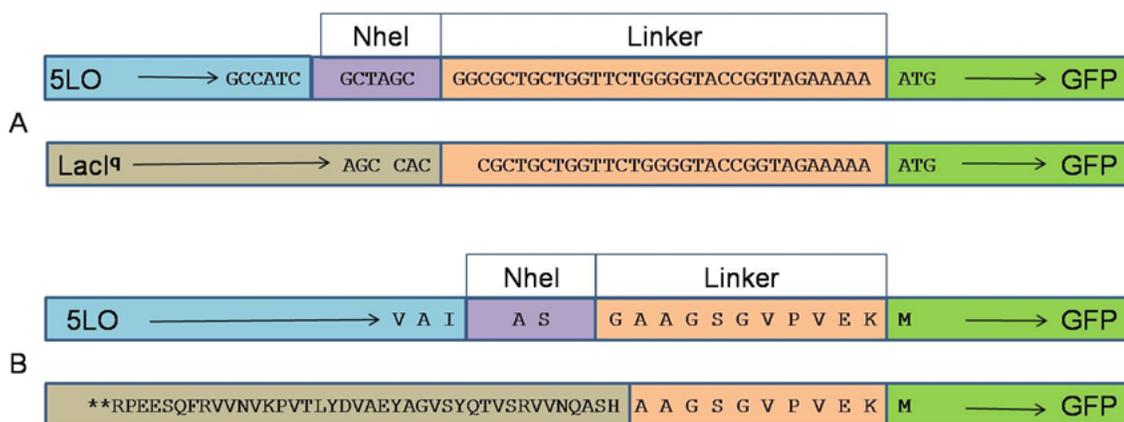
inserts that are larger than the expected wild-type gene (**Figure 2-21 B**). These latter clones were sequenced to determine the identity of the larger insert and the cause of the increased fluorescence.



**Figure 2-21.** Streaks of fluorescent clones found during the second attempt at round 1 (A) and restriction digest analysis of those clones (B). Clones represented in lanes 3, 5, 10, 11, 12, 13, 14, and 15 show insert larger than WT (indicated by the white arrow), and clones represented in lanes 1 and 7 appear to lack an insert.

Sequencing of the clones with the larger insert indicated that these were likely the result of vector contamination. The large insert consists of a sequence of DNA arising from the screening vector, specifically part of the *lacI<sup>q</sup>* gene. It was initially thought that the primers for 5LO were annealing along a portion of the screening vector and amplifying a 3000 bp sequence as a side-product, which was then ligated between *HindIII* and *NheI* in the screening vector instead of the correct insert. This is not the case as the same anomaly was present when fellow lab members (Dr. Daria Trofimova and 4<sup>th</sup> year student Shelley McArthur) were screening

variants of 15LO. This insert was also seen when screening for RebG. It is highly unlikely that three different sets of primers could amplify the same 3000 bp side product, and also that such a large insert which doesn't encompass a complete gene could still allow proper folding of GFP. Evidence that this is a vector rather than PCR issue is also supported by sequencing data, which is summarized in **Figure 2-22**. The DNA sequence of the normal vector is shown on the top in **Figure 2-22 A**, and the mutant vector's sequence is shown on the bottom. For the mutant, the complete GFP gene is intact and most of the linker sequence, however prior to where the *NheI* site should be on the vector, the sequence is replaced with the beginning of the *lacI<sup>f</sup>* gene. Upon translation of the mutant vector it is evident that two-stop codons are in frame with GFP about 50 amino acids upstream of GFP's initiating methionine (**Figure 2-22 B**). There are no other methionine residues between these two features so the ribosome must bind somewhere between them, however it is unclear where the ribosome binds exactly as there is no obvious Shine-Dalgarno sequence. The presence of a Shine-Dalgarno sequence is not critical for ribosome binding, but it may have an effect on how accessible the initiation codon is on the mRNA [32]. Accessibility of the start codon has been hypothesized to be the dominant factor associated with ribosome binding and this may be the mechanism behind GFP's translation in this case. The fact that there is no Shine-Dalgarno sequence is possibly the reason behind the lower fluorescence of these mutants compared to mutants with small inserts ligated in the cloning region, as translation may not be initiated as effectively. The sequencing data also indicates that the normal ribosome binding sequence on the vector seems to have been lost, as the forward sequencing reactions failed every time. This can be attributed to the loss of the sequence which the forward sequencing primer would bind to. This sequence encompassed the vector's ribosome binding site.



**Figure 2-22.** Sequence comparisons of the normal and mutant screening vectors. The nucleotide sequences for the normal pProEx\_GFPuv1 vector (top) and mutant vector (bottom) are shown in **A**. The amino acid sequences are shown in **B**.

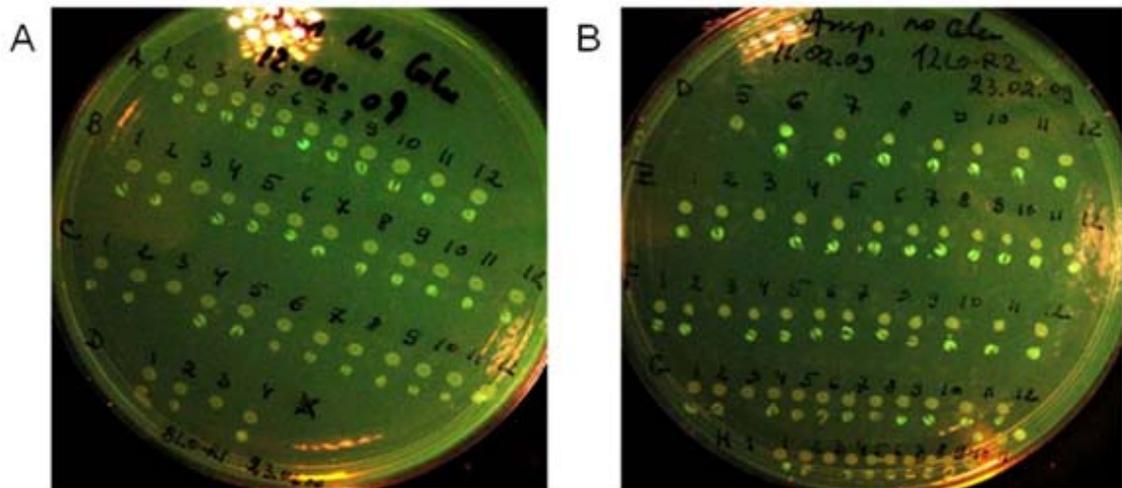
The source of the vector contamination is still unclear, but considering that only eight clones out of around 75000 exhibit it, the occurrence is quite low (around 0.01%). There may be some portions of the vector that are susceptible to “star activity” or regions in which restriction enzymes can cut even though they are not identical to a restriction site. This activity normally occurs when the digest is being performed under non-standard conditions such as a high amount of enzyme, a high pH, or the presence of organic solvents. No extreme conditions were used in this case, yet the possibility of star activity cannot be ruled out since the placement of restriction sites along the vector has changed (they cannot be found in the sequencing data). This mutant vector was probably not found in the first attempts at round 1 and round 2 because the small inserts were the major contaminant, and if the occurrence rate of the mutant vector was the same as it is here (0.01%), it is unlikely that the 2 or 3 containing it would be picked. Also, because the clones with the mutant vector contamination are not as bright as the clones with small insert contamination, they would not have been selected for this reason. Only two clones with small insert contamination were found during the second attempt at round 1 (**Figure 2-21 B**, # 1 and 7), showing that the increased stringency during gel purification effectively reduced this type of

contamination. This is also evident by the drastic reduction in fluorescent colonies upon screening.

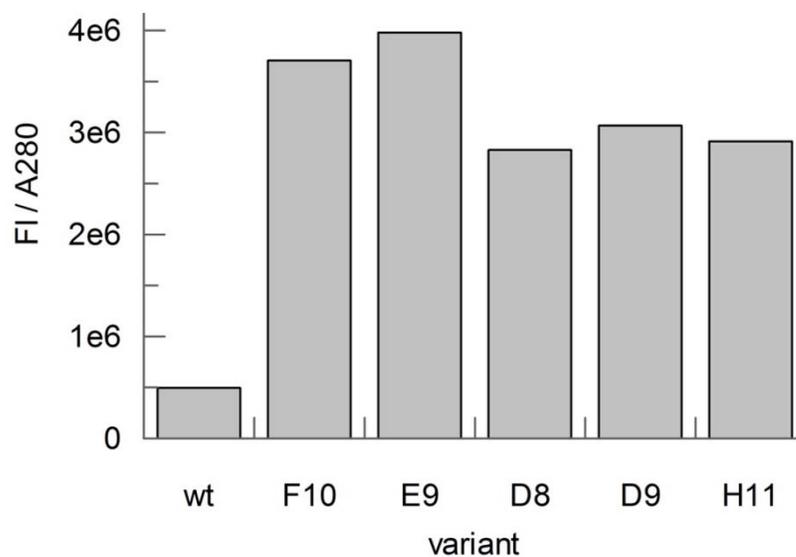
The major issue with these targets is the fact that no clones other than those with some sort of contamination could be found that exhibited increased fluorescence. If indeed an improvement in folding had been accomplished then those clones would have also been selected, and would have been analyzed along with the contaminant clones. The whole process was repeated a third time for both RebG and 5LO with the same results—all bright clones were all a result of vector contamination, except for a small fraction containing small inserts. The fact that no other viable clones could be found means one of three things: 1) These targets are highly resistant to evolution and a much larger number of clones needs to be analyzed to find any improvements or 2) the diversification technique needs to be optimized to obtain the correct type of diversity, or 3) the improvements are so slight that they cannot be detected by eye, but perhaps success may be achieved via FACS.

Interestingly, human 12-lipoxygenase which shares a 41% sequence homology with 5LO, and mouse 8-lipoxygenase which shares a 42% sequence identity with 5LO, have shown some improvement in fluorescence by Dr. Daria Trofimova (postdoc in Zechel lab) using the evolution protocols and screening vector described here. These targets both encountered many of the contamination results as with 5LO, however after purifying the fluorescent colonies away from the non-fluorescent ones, several pure clones were found that exhibited fluorescence greater than wild type and also contained the appropriate insert. One main difference between the protocol used for these lipoxygenases and 5LO is that the final clones were transformed via heat shock into XL1-Blue cells. After the initial transformation via electroporation the cells were washed off the 500 cm<sup>2</sup> plate with LB and grown overnight to isolate the plasmid DNA the next day. This plasmid DNA library was then used to transform chemically competent XL1-Blue cells for library selection. This protocol seems to reduce the occurrence of double transformants. The

fluorescence of some of these clones is displayed in **Figure 2-23**. Fluorescence measurements on the soluble fractions from lysates of round two 12LO mutants showed an up to 8-fold improvement in fluorescence for some clones (**Figure 2-24**). Expression and solubility analysis by SDS-PAGE for these mutants to quantitatively determine the improvement in solubility are pending.



**Figure 2-23.** Selected mutants from the first round of evolution on mouse 8LO (A) and the second round of evolution on human 12LO. For both, the spot representing the wild type fusion is shown above each mutant spot.

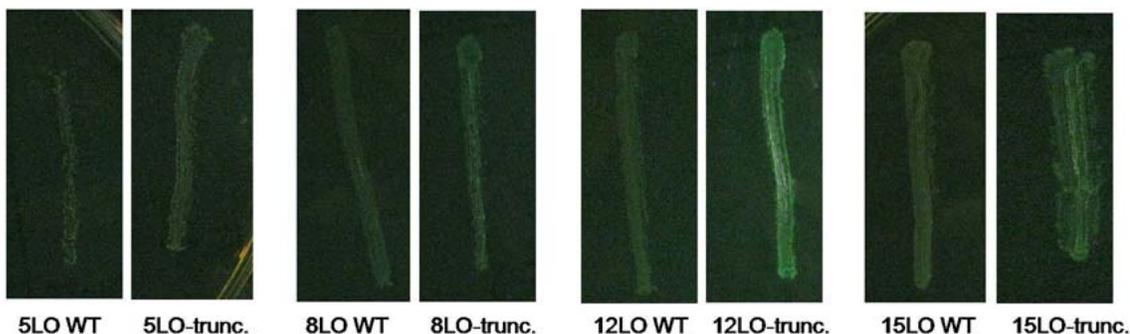


**Figure 2-24.** Fluorescence measurements on the soluble fraction of 12LO mutant cell lysates. Fluorescence measurements ( $\lambda_{em} = 395$ ) were normalized by dividing by absorbance at 280 nm, a measurement of total protein concentration.

### 2.3.6 Truncation strategies to improve solubility of 5LO

Diversification of the full-length 5LO gene did not give rise to clones with improved fluorescence, thus two truncation methods were tried based on the knowledge that a truncated version of 12LO has been crystallized. The first truncation method involved rationally truncating LO variants to match the length of the truncated version of 12LO which has been crystallized. The truncated variant of each lipoxygenase (human 12-LO, human 5-LO, human 15-LO, and mouse 8-LO) was cloned into pProEx\_GFPuv1 and streaked alongside the full-length wild type gene, also in pProEx\_GFPuv1 (**Figure 2-25**). It is already known that the truncated version of 12LO should give rise to more soluble material thus an improvement in fluorescence for this truncated variant was expected and was observed. Our truncated 12LO construct differs from the crystallized construct only by the placement of the stop codon, which is directly after Ile662 in our case. Truncated 5LO did not show a discernable difference in fluorescence over wild type, however truncated 8LO and 15LO did appear slightly brighter. Fluorescence measurements and

SDS-PAGE analysis of soluble fractions for these variants are needed to discern if an improvement in solubility has been made.

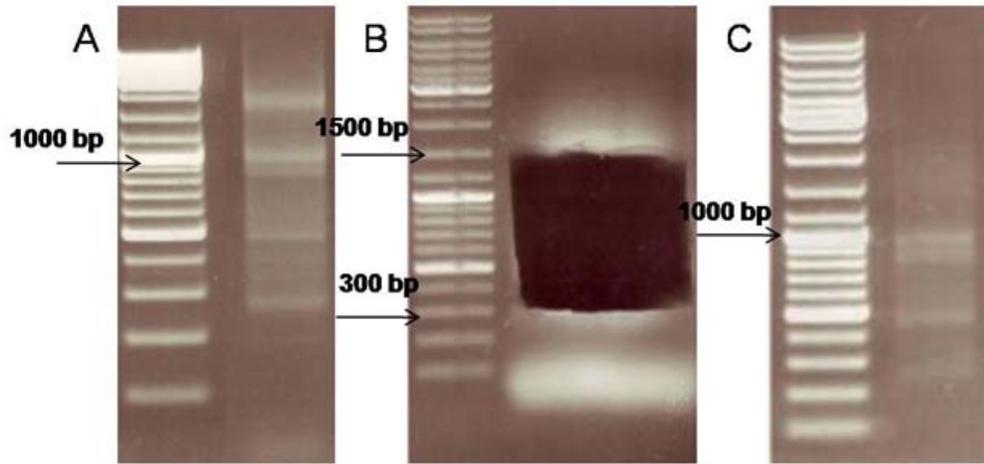


**Figure 2-25.** Fluorescence comparison between truncated lipooxygenase fusions and the corresponding full-length wild type fusions.

In a second attempt to find soluble truncated variants of 5LO, tagged random primer PCR was used (TP-PCR) (see Chapter 1, section 1.5.1.1). With this technique the ‘random primers’ can theoretically anneal to any portion along the length of the 5LO gene, however due to the random nature of this protocol, inserts can be amplified out of frame and/or backwards. Thus there is a 1 in 6 chance that an amplified insert will be inserted in the correct orientation and in frame with the original gene. There is also a 2 in 3 chance the gene will not be in frame with GFP. The latter point does not influence selections however, as any inserts not in frame with GFP will not fluoresce and will therefore not be selected.

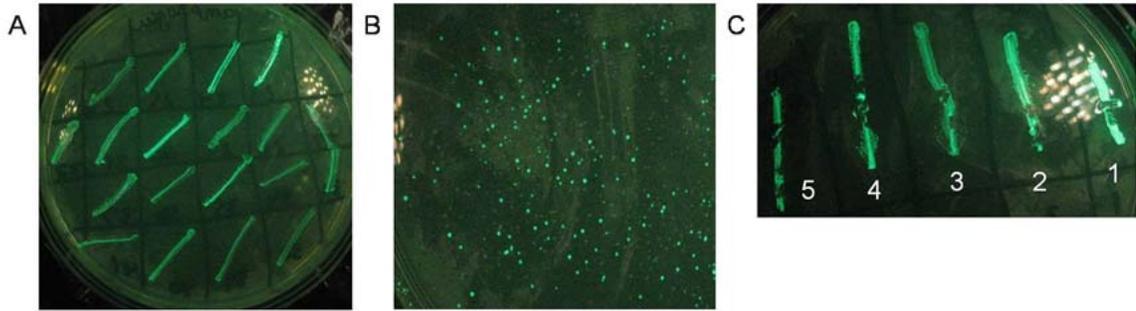
Tagged-random primer PCR should produce products of various lengths that appear as a smear on a gel. **Figure 2-26 A** shows the products obtained after TP-PCR. The smear shows that certain length products were amplified more than others (four brighter bands on the gel), most likely due to the nature of some of the random primers used and their higher affinity for certain regions on the gene. It could also be that these portions of the gene are GC-rich, and have a higher enthalpy of binding with some of the primers. Products 300 – 1500 bp in length were

selected via gel purification for ligation into the screening vector pProEx-GFPuv (**Figure 2-26 B, C**).

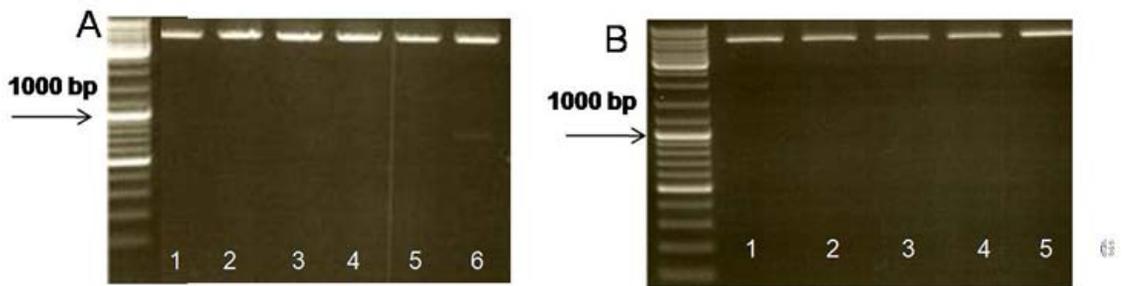


**Figure 2-26.** Products from tagged random primer PCR on 5LO. (A) shows the unpurified products, (B) shows the sizes selected for gel purification (300 – 1500 bp), and (C) shows the products after gel purification.

Many fluorescent colonies were evident upon screening one 500 cm<sup>2</sup> plate of the truncated library and 40 were initially selected for restriction digest analysis. As **Figure 2-27 A** shows, these clones were highly fluorescent. 20 of the 40 selected were grown overnight in 4 mL cultures to isolate the plasmid DNA, then analyzed by restriction digest. Of the twenty plasmids isolated, only 1 showed an insert of about 800 bp (**Figure 2-28 A**, lane 6). To verify plasmid purity, five were transformed (including the one showing insert) and all were found to be mixtures (**Figure 2-27 B** shows an example). Bright colonies were reselected from the mixture plates, and **Figure 2-27 C** shows the fluorescence of the clones after purification. **Figure 2-28 B** shows the products of restriction digest of the five purified clones, and it is clear that the 800 bp insert was purified away and this clone no longer showed an insert on a gel (**Figure 2-28B**, lane 5).



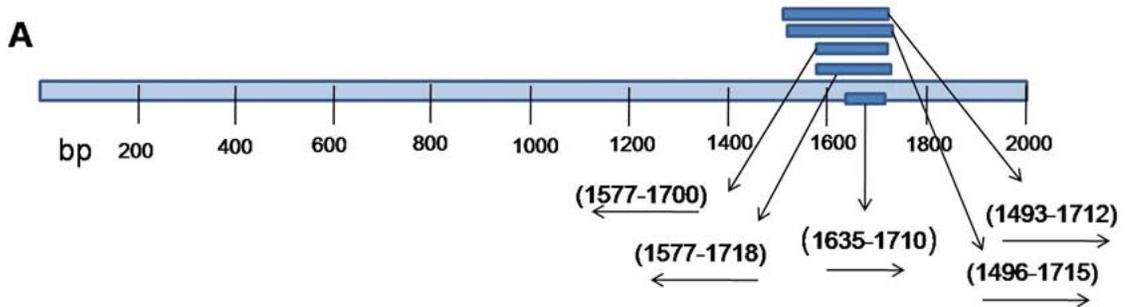
**Figure 2-27.** Fluorescence of clones selected after tagged random-primer PCR. (A) shows unpurified streaks, (B) shows an example of the heterogeneous colonies obtained from initial selections, and (C) shows the purified clones.



**Figure 2-28.** Restriction digest analysis of unpurified clones selected after TP-PCR (A), and after purification (B). The faint band in lane 6 of gel A (highlighted by the white arrow) represents an unpurified clone which initially contained an insert, but was lost after purifying away the non-fluorescent cells.

To ensure that some sort of contamination was not occurring again from inserts generated from the primers or from vector contamination, the five purified clones were sequenced, and the data is summarized in **Figure 2-29 A**. All five clones contained inserts too short to be observed on an agarose gel that arose from the 5LO gene, ensuring that contamination was not an issue. Interestingly, all five clones arose from the same region on the gene, indicating that this portion of the sequence has a high affinity for some of the random primers. Upon a

closer look at this area of the gene, it is clear that it is indeed GC-rich. Unfortunately, all five inserts were either out of frame with 5LO and encoded sequences not matching that of the lipoxygenase, or were inserted backwards, as was seen in two cases.



**B**

1) 5'-caccgcctccgcccagcagccgcggtcaacttcggccagtacgactgggtgctcctggat  
H R L R P A R R G Q L R P V R L V L L D  
cccccaatgcgcccc-3'  
P Q C A P

2) 5'-gggtgttggtggaggaccgagctgcaggacttcgtgaacgatgtctacgtgtacggcat  
G V G G G P G A A G L R E R C L R V R H  
gcggggcccgaagtccctcaggcttcccccaagtccggtcaagagccgggagcagctgtcgga  
A G P Q V L R L P Q V G Q E P G A A V G  
gtacctgaccgtggatcttcaccgcctccgcccagcagccgcggtcaacttcggcca  
V P D R G D L H R L R P A R R G Q L R P  
gtacgactgggtgctcctggatcccccaatgcgccccccac-3'  
V R L V L L D P Q C A P H

3) 5'-gggggaggaggaccgagctgcaggacttcgtgaacgatgtctacgtgtacggcatgcg  
G G G G P G A A G L R E R C L R V R H A  
gggcccgaagtccctcaggcttcccccaagtccggtcaagagccgggagcagctgtcggagta  
G P Q V L R L P Q V G Q E P G A A V G V  
cctgaccgtggatcttcaccgcctccgcccacacgcgcccgggtcaacttcggccagta  
P D R G D L H R L R P S R R G Q L R P V  
cgactgggtgctcctggatcccccaatgcgcccccaaccaa-3'  
R L V L L D P Q C A P N Q

- 4) 5'-gggggggggttggggatccaggagcaccagtcgtactggccgaagttgaccgcggtgc  
           G G G L G I Q E H Q S Y W P K L T A A C  
 tgggcgaggcggtgaagatcaccacggtcaggtactccgacagctgctcccggctcttg  
           W A E A V K I T T V R Y S D S C S R L L  
 accgacttgggg-3'  
           T D L G
- 5) 5'-gggggtggttgggggcgccattggggatccaggagcaccagtcgtactggccgaagttgacc  
           G V V G G A L G I Q E H Q S Y W P K L T  
 gggcgctgctgggcgaggcggtgaagatcaccacggtcaggtactccgacagctgctcc  
           A A C W A E A V K I T T V R Y S D S C S  
 cggctcttgaccgacttgggg-3'  
           R L L T D L G

**Figure 2-29.** (A) DNA sequences giving rise to fluorescent clones after TP-PCR. The numbers in brackets indicate the base pairs on the gene the sequences encompass and the arrows under these numbers indicate which direction the insert was cloned. (B) The sequences of the inserts. The top sequence is the DNA sequence, and underneath is the encoded amino acid sequence.

As only five clones were purified and sequenced, it remains to be seen whether or not TP-PCR can be used as a viable method of finding soluble truncated variants of 5LO. On the positive side, contamination from mutant vector or ‘primer inserts’ did not give rise to fluorescence, and every sequence found that did result in higher fluorescence was the result of an amplification off of the 5LO gene. On the other hand, every sequence found was amplified from the same region on the gene, indicating that the GC-richness of this area could actually inhibit the amplification of ‘random’ portions, as it seems to be exhibiting a selective pressure. This is probably due to a larger negative enthalpy of binding in this region. Another thing to note is that even though inserts 300 – 1500 bp in length were selected, all of the inserts found were much smaller than that. Re-purifying the TP-PCR products and selecting larger inserts may be

necessary to prevent selection of random inserts that are too small to affect the folding of GFP. Larger numbers of clones need to be selected and purified to determine overall if TP-PCR is a viable option for finding soluble variants of 5LO.

## 2.4 Conclusions

A basic rule that underlies any laboratory evolutionary process is that “you get what you select for” [33]. It is therefore extremely important that prior to starting an evolution experiment much thought has gone into the design of the screening or selection process. The method of choice must select for a trait specifically, such that false positive results are weeded out during the selection process. With the GFP reporter system, this means that an increase in fluorescence should only be observed as a result of an improvement of folding of the target protein. A problem with the GFP reporter system is that high fluorescence can be achieved not only from an improvement in folding, but also from fusion of short truncation artifacts upstream of GFP or due to slow aggregation of the target protein [34].

In this study, the problem of small inserts giving rise to increased fluorescence was encountered, along with other forms of unexpected contamination. The glycosyltransferase RebG and human arachidonate 5-lipoxygenase were chosen to undergo directed evolution in an attempt to improve their folding and solubility upon expression in *E. coli*.

The diversification procedures needed optimization in order to successfully result in a PCR product. Error-prone PCR requires the divalent cations  $Mg^{2+}$  and  $Mn^{2+}$  at concentrations so high they may be inhibitory to the reaction's success. RebG could only withstand a total of 3.5 mM  $Mg^{2+}$  and 0.4 mM  $Mn^{2+}$ , whereas 5LO could withstand 3.5 mM  $Mg^{2+}$  and 0.5 mM  $Mn^{2+}$ . Perhaps because of its ability to withstand the higher concentration of  $Mn^{2+}$ , the 5LO error-prone PCR incorporated more mutations on average than the RebG error-prone PCR. For 5LO, an average 5 mutations per 2022 bp gene was achieved, with 80% causing amino acid changes and 20% being silent. RebG averaged only 1 mutation per 1263 bp gene, with 20% causing amino acid changes and 80% being silent. Transitions greatly outnumbered transversions in a ratio of 13:2.

DNase I digestion needed to be optimized in terms of length of digestion time for each target. For the recombination and amplification steps, the polymerase chosen had the greatest effect on outcome. Herculase II was the only polymerase which could successfully produce a smear of the correct molecular weight from the recombination of 5LO. Successful amplification off of the smear of high molecular weight was achieved with PfuTurbo, which was the only polymerase able to selectively amplify a single product of the correct size.

Screening of libraries proved to be a difficult task for these two targets. Four attempts were made for each target to identify improved clones (initial attempts at round 1 and round 2, and round 1 repeated two more times) but all clones found were the result of some sort of contamination. The fluorescent clones either contained plasmids with small inserts rather than the correct full-length insert, or contained a mutant vector in which a portion of the *lacI<sup>d</sup>* gene has replaced the cloning region. It was also discovered that many of the streaks were in fact mixtures of cells containing plasmids with full length inserts and other cells containing plasmids with short inserts. These mixtures were likely a result of two colonies being picked at one time during the screening process or the presence of two plasmids in one cell which is often encountered when transforming cells with large amounts of DNA, such as in library generation.

After the several attempts at evolving both 5LO and RebG, which involved screening of 30-50 colonies per attempt, no viable clones could be found with improved fluorescence. These proteins may need to be screened via FACS so that larger libraries can be searched, or perhaps diversified with other methods that allow greater diversity to be achieved (perhaps by a method that allows consecutive nucleotide changes). However, two other lipoxygenases, mouse 8LO and human 12LO have shown improvements in fluorescence using the techniques described in this chapter, with some clones from the second round of 12LO evolution showing an 8-fold increase in fluorescence of the soluble fraction of cell lysates.

Truncation methods to find soluble portions of lipoxygenases did show some promise. Rationally truncating 5LO did not result in visible improvement of fluorescence, but the truncated versions of mouse 8LO and human 15LO did appear brighter than their full-length wild type counterparts. As expected, the truncated version of human 12LO was much brighter than its full-length version. Tagged random primer PCR of 5LO successfully found inserts generated from the random amplification of regions on the 5LO gene, however these inserts were not as 'random' as hoped. Sequence analysis of five of the isolated clones showed they all arose from the same region of the 5LO gene, and none of them were in frame with the original gene. Further studies need to be conducted to try and isolate larger fragments arising from the TP-PCR that are in frame and in the correct orientation.

Much was learned about the optimization of diversification techniques and the high level of stringency required during the screening and subsequent analysis of selected clones. Hopefully the record of this will prevent others from falling victim to 'false-positive' results arising from contamination, and can act as a guideline for troubleshooting the diversification process. Indeed, false-positive results arising from contamination is a wide-spread problem and has led to the retraction of several significant evolution papers due the false presence of activity from wild type contamination [36, 37]. Awareness of these potential problems aided in the successful evolution of PhnG, as will be discussed in the next chapter.

## 2.5 References

1. Peters-Golden M, Brock TG: **5-Lipoxygenase and FLAP**. *Curr Drug Targets Inflamm Allergy* 2002, **1**:99-109.
2. Rådmark O, Werz O, Steinhilber D, Samuelsson B: **5-Lipoxygenase: regulation of expression and enzyme activity**. *Trends Biochem Sci* 2007, **32**:332-341.
3. Furstenberger G, Krieg P, Muller-Decker K, Habenicht AJR: **What are cyclooxygenases and lipoxygenases doing in the driver's seat of carcinogenesis?** *Int J Cancer* 119, **119**:2247-2254.
4. Manev H, Manev R: **5-Lipoxygenase (ALOX5) and FLAP (ALOX5AP) gene polymorphisms as factors in vascular pathology and Alzheimer's disease**. *Med Hypotheses* 2006, **66**:501-503.
5. Werz O: **5-Lipoxygenase: Cellular biology and Molecular Pharmacology**. *Curr Drug Targets Inflamm Allergy* 2002, **1**:23-44.
6. Okamoto H, Hammarberg T, Zhang Y-, Persson B, Watanabe T, Samuelsson B, Radmark O: **Mutation analysis off the human 5-lipoxygenase C-terminus: Support for a stabilizing C-terminal loop**. *Biochim Biophys Acta* 2005, **1749**:123-131.
7. Sánchez C, Zhu L, Braña AF, Salas AP, Rohr J, Méndez C, Salas JA: **Combinatorial biosynthesis of antitumor indolocarbazole compounds**. *Proc Natl Acad Sci USA* 2005, **102**:461-466.
8. Zhang C, Albermann C, Fu X, Peters NR, Chisholm JD, Zhang G, Gilbert EJ, Wang PG, Van Vranken D. L., Thorson JS: **RebG- and RebM- Catalyzed Indolocarbazole Diversification**. *Chembiochem* 2006, **7**:795-804.
9. Bailly C, Riou J-, Colson P, Houssier C, Rodrigues-Pereira E, Prudhomme M: **DNA Cleavage by Topoisomerase I in the Presence of Indolocarbazole Derivatives of Rebeccamycin**. *Biochemistry* 1997, **36**:3917-3929.
10. Yamashita Y, Fujii N, Murakata C, Ashizawa T, Okabe M, Nakano H: **Induction of Mammalian DNA Topoisomerase I Mediated DNA Cleavage by Antitumor Indolocarbazole Derivatives**. *Biochemistry* 1992, **31**:12069-12075.
11. Bailly C, Qu X, Graves DE, Prudhomme M, Chaires JB: **Calories from carbohydrates: energetic contribution of the carbohydrate moiety of rebeccamycin to DNA binding and the effect of its orientation on topoisomerase I inhibition**. *Chem Biol* 1999, **6**:277-286.
12. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A: **Structures and mechanisms of glycosyltransferases**. *Glycobiology* 2006, **16**:29R-37R.
13. Salas JA, Mendez C: **Indolocarbazole antitumor compounds by combinatorial biosynthesis**. *Curr Opin Chem Biol* 2009, **13**:152-160.

14. Kawasaki M, Inagaki F: **Random PCR-based Screening for Soluble Domains Using Green Fluorescent Protein.** *Biochem Biophys Res Commun* 2001, **280**:842-844.
15. Caldwell RC, Joyce GF: **Randomization of genes by PCR mutagenesis.** *PCR Methods Appl* 1992, **2**:28-33.
16. Weissensteiner T, Griffin HG, Griffin A: **PCR Technology Current Innovations.** 2004, :392.
17. Sambrook J, Russell DW: **Quantitation of Nucleic Acids.** In *Molecular Cloning A Laboratory Manual. Volume 3.* Edited by a. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001:A8.19-A8.21.
18. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**:2947-2948.
19. Jacobs SA, Podell ER, Wuttke DS, Cech TR: **Soluble domains of telomerase reverse transcriptase identified by high-throughput screening.** *Protein Sci* 2005, **14**:2051-2058.
20. Steitz TA: **DNA Polymerases: Structural Diversity and Common Mechanisms.** *J Biol Chem* 1999, **274**(25):17395-17398.
21. Hung T, Mak K, Fong K: **A specificity enhancer for polymerase chain reaction.** *Nucleic Acids Res* 1990, **18**:4953.
22. Wong TS, Roccatano D, Zacharias M, Schwaneberg U: **A Statistical Analysis of Random Mutagenesis Methods Used for Directed Protein Evolution.** *J Mol Biol* 2006, **355**:858-871.
23. Maheshri N, Schaffer DV: **Computational and experimental analysis of DNA shuffling.** *Proc Natl Acad Sci USA* 2003, **100**:3071-3076.
24. Stemmer WPC: **DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution.** *Proc Natl Acad Sci U S A* 1994, **91**:10747-10751.
25. Lorimer IAJ, Pastan I: **Random recombination of antibody single chain Fv sequences after fragmentation with DNaseI in the presence of Mn<sup>2+</sup>.** *Nucleic Acids Res* 1995, **23**:3067-3068.
26. Sutton DH, Conn GL, Brown T, Lane AN: **The dependence of DNase I activity on the conformation of oligodeoxynucleotides.** *Biochem J* 1997, **321**:481-486.
27. Mayer MR, Dailey TA, Baucom CM, Supernak JL, Grady MC, Hawk HE, Dailey HA: **Expression of human proteins at the Southeast Collaboratory for Structural Genomics.** *J Struct Funct Genomics* 2004, **5**:159-165.
28. Cormack BP, Valdivia RH, Falkow S: **FACS-optimized mutants of the green fluorescent protein (GFP).** *Gene* 1996, **173**:33-38.

29. Dower WJ, Miller JF, Ragsdale CW: **High efficiency transformation of *E. coli* by high voltage electroporation.** *Nucleic Acids Res* 1988, **16**:6127-6145.
30. Shaner NC, Patterson GH, Davidson MW: **Advances in fluorescent protein technology.** *J Cell Sci* 2007, **120**:4247-4260.
31. Reid BG, Flynn GC: **Chromophore formation in green fluorescent protein.** *Biochemistry* 1997, **36**:6786-6791.
32. Nakamoto T: **A unified view of the initiation of protein syntheses.** *Biochem Biophys Res Commun* 2006, **341**:675-678.
33. Schmidt-Dannert C, Arnold FH: **Directed evolution of industrial enzymes.** *Trends Biotechnol* 1999, **17**:135-136.
34. Cabantous S, Pedelacq JD, Mark BL, Naranjo C, Terwilliger TC, Waldo GS: **Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from *Mycobacterium tuberculosis*.** *J Struct Funct Genomics* 2005, **6**:133-119.
35. Goldsmith M, Kiss C, Bradbury ARM, Tawfik DS: **Avoiding and controlling double transformation artifacts.** *Protein Eng Des Sel* 2007, **20**:315-318.
36. Altamirano MM, Blackburn JM, Aguayo C, Fersht AR: **Retraction: Directed evolution of new catalytic activity using the  $\alpha/\beta$ -barrel scaffold.** *Nature* 2002, **417**:468.
37. Zeytun A, Jeromin A, Scalettar B., Waldo GS, Bradbury AR: **Retraction: Fluorobodies combine GFP fluorescence with the binding characteristics of antibodies.** *Nat. Biotechnol* 2004, **22**:601.

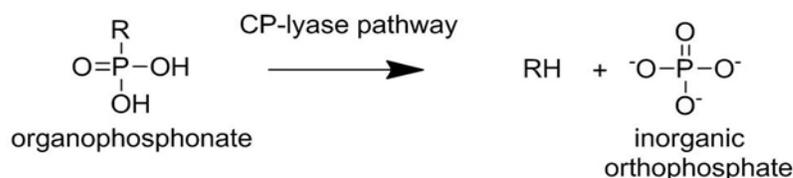
## Chapter 3

### Directed Evolution of *Escherichia coli* Phosphonate Metabolism Protein

#### PhnG

##### 3.1 Introduction

PhnG is one of 14 proteins expressed from the *phn* operon found in *E. coli*. This 14 gene operon—*phnCDEFGHIJKLMNOP*, is responsible for the degradation of organophosphonates to inorganic orthophosphate (Pi) and the corresponding hydrocarbon (**Figure 3-1**). Organophosphonates are characterized by the presence of a highly stable carbon-phosphorus (CP) bond which is resistant to chemical hydrolysis, thermal degradation, photolysis, and the action of phosphatases [1]. Because of its ability to break this CP bond, the *phn* operon is also known as the CP-lyase pathway. As Pi is the preferred phosphorus source for most microorganisms, this pathway is only activated when the natural abundance of Pi is low [2].



**Figure 3-1.** The CP-lyase pathway acts on organophosphonates to cleave the CP bond and produce a hydrocarbon and orthophosphate.

Due to the inherent stability of the CP bond, concern has arisen over the accumulation of synthetic and highly bioactive organophosphonates in various ecosystems. Synthetic organophosphonates are widely used as detergent additives, herbicides, antibiotics, flame retardants, and chemical warfare nerve agents, and their environmental fate is not fully

understood [3]. A potential strategy to eliminate phosphonates from the environment is biodegradation. Several microorganisms have the ability to break down phosphonates as a survival mechanism to acquire phosphorus when Pi levels are low. Apart from the CP-lyase pathway, a few other enzymes are known to break down specific phosphonates by acting on carbonyl groups  $\beta$  to the phosphorus centre, such as phosphonatase, phosphonoacetate hydrolase and phosphonopyruvate hydrolase [4, 5]. These enzymes have relatively specific substrates compared to the broad range of organophosphonates the CP-lyase pathway is capable of degrading. Thus understanding the functions and mechanisms of the enzymes in the CP-lyase family is a critical step towards being able to biodegrade harmful phosphonates in the environment.

Currently there is only a limited understanding of most of the enzymes in this pathway. Mutagenic studies conducted by Metcalf and Wanner unveiled that of the 14 proteins in the pathway, only seven of them—PhnG to PhnM, are critical for CP bond cleavage, with the other seven taking on transport, regulatory, or accessory roles [6]. Of the seven enzymes critical for CP bond cleavage, only PhnH has been crystallized, but has yet to be assigned a function [7]. PhnM has sequence similarity to the membrane component of a protein binding-dependent transport system, and the high hydrophobic content of PhnE suggests that it may also be membrane bound [8]. PhnP has also been crystallized and was discovered to have phosphodiesterase activity [9]. Overall the proteins in this pathway are thought to form a CP-lyase membrane associated complex [6] with PhnCDE being involved in transport, PhnF and PhnO possibly having regulatory roles, PhnG to PhnM being critical for CP bond cleavage, and PhnN and PhnP possibly being accessory proteins for the CP-lyase activity [8].

PhnG is the first protein in the group of proteins essential for CP bond cleavage. It is a small protein consisting of only 150 amino acids and weighing 16.5 kDa. It shares high homology with PhnG proteins from other bacteria however no structural data is available for any

of the homologs and no function can be assigned based on sequence similarities. Because so little is known about this protein, it was an ideal target for solubility enhancement via directed evolution. Previous efforts at expression with this protein were unsuccessful, thus to it was chosen to undergo diversification and selection with the GFP folding reporter in an attempt to enhance its solubility upon expression in *E. coli* and possibly acquire a crystal structure to provide further insights in to the mechanism of the important CP-lyase pathway.

## 3.2 Experimental Procedures and Methods

### 3.2.1 Materials

Oligonucleotides used for PCR were synthesized by Sigma-Genosys. Taq polymerase and mixed dNTPs were purchased from New England Biolabs Canada. Restriction enzymes, T4 DNA ligase, and calf intestinal alkaline phosphatase were purchased either from Fermentas or New England Biolabs. DNase I was purchased from Fermentas. DMSO, MnCl<sub>2</sub>, CaCl<sub>2</sub>, EDTA, Tris Base, Tris-HCl, NaCl, n-dodecyl-β-D-maltoside, dithiothreitol, the molecular weight marker kit for gel filtration chromatography, imidazole and ampicillin, were purchased from Sigma Aldrich (Oakville, Ontario and US). IPTG was purchased from Invitrogen. Nucleospin plasmid purification kits (Macherey-Nagel) were ordered from MJS Biolynx, Inc. All other DNA purification kits were purchased from QIAGEN, as well as Ni-NTA resin. All cells (XL1-Blue and ElectroTen-Blue) were purchased from Stratagene (supplied by VWR, Canada). Fisher Biosciences Canada supplied all media (Luria Bertani broth, Luria Bertani agar, glucose), 500 cm<sup>2</sup> plates (Corning), electroporation cuvettes (Eppendorf), urea, and the additional dNTPs used for error-prone PCR (dTTP and dCTP). Dialysis cassettes were purchased from Pierce Protein Research Products (a division of Thermo Scientific). Superdex<sup>TM</sup> 200 resin was purchased from GE Healthcare and the Tricorn 10/300 column from Amersham Biosciences. The pGFPuv cloning vector was purchased from Clontech, and the pProEx cloning vector was obtained from Invitrogen. All sequencing reactions were performed by either Robarts Research Institute (London, Ontario) or TCAG sequencing facility (The Hospital for Sick Children, Toronto, ON). LED lights (400 nm) were purchased from Super Bright LEDs, Inc.

### 3.2.2 Cloning of wild type *phnG* into pProEx\_GFPuv1

Wild type *phnG* was PCR amplified from a pQI\_PD\_PhnG vector (cloned previously in the Zechel lab by Shu-Mei He) using Taq polymerase, and the primers 5'-AGGGCAAGCTTATGCACGCAGATACCGCGAC-3' (forward, *Hind*III site underlined), and 5'-AGCGCCGCTAGCTGCGTTGTCTCCGCGAACCATC-3' (reverse, *Nhe*I site underlined). The 450 bp product and pProEX\_pGFPuv1 (vector construction described in Chapter 2 section 2.2.2, designed by Kawasaki *et al.* [10]) were digested using *Hind*III and *Nhe*I, with calf intestinal alkaline phosphatase added to the digest. The wild type insert was ligated into the screening vector using T4 DNA ligase and the ligation mixture was used to transform XL1-Blue cells which were plated on Luria-Bertani (LB) agar plates containing 100 µg/mL ampicillin. The resulting colonies were grown overnight in 4 mL cultures and plasmids purified with a NucleoSpin® plasmid DNA preparation kit to isolate the plasmid DNA. The plasmids were analyzed via restriction digest with *Hind*III and *Nhe*I to verify insertion of the correct gene, and then sequenced using the forward primer 5'-AGCGGATAACAATTTACACAGG-3' and reverse primer 5'-ATCTTCTCTCATCCGCCAAAAC-3'.

### 3.2.3 Error-prone PCR of *phnG*

The mutagenic PCR protocol followed was based the procedure originally published by Caldwell and Joyce [11, 12]. The reaction mix used is outlined in **Table 3-1**. Not listed in the table are the template plasmid DNA and DMSO, for both of which 1 µL was added. The template used was the wild type *phnG* in pProEx\_GFPuv1.

**Table 3-1 Mutagenic PCR Reaction Mix for PhnG**

Reagent	RebG		
	Stock conc.	Volume (μL)	Final conc.
ddH <sub>2</sub> O		31.5	
10 x Buffer*		5	
dNTPs	25 mM each	1	0.125 μM
Extra dTTP	100 mM	0.5	1 mM
Extra dCTP	100 mM	0.5	1 mM
Extra MgCl <sub>2</sub>	50 mM	2	2 mM
MnCl <sub>2</sub>	5 mM	5	0.5 mM
For. Primer	10 μM	1	0.2 μM
Rev. Primer	10 μM	1	0.2 μM
Taq (NEB)	5000 U/mL	0.5	2.5 U
Total		48	

\*the 10x reaction buffer supplied by NEB included 100 mM Tris, 1.5 mM MgCl<sub>2</sub>, and 50 mM KCl

The PCR program used was 94 °C for 4 minutes, 30 × (94 °C for 30 seconds, 58 °C for 30 seconds, 72 °C for 1.5 minutes), finishing with 72 °C for 5 minutes.

### 3.2.4 DNA shuffling of *phnG*

DNase I digestion of *phnG* was performed following the protocol described in Chapter 2, section 2.2.5, with a digestion time of 1 minute.

The reassembly reaction mix consisted of 20 μL of pure fragments, 0.5 μL dNTPs (25 mM each), 2.5 μL 10 × Taq reaction buffer (provides a final Mg<sup>2+</sup> concentration of 2 mM), 0.5 μL DMSO and 1 μL additional 25 mM MgCl<sub>2</sub> for a final volume of 25 μL. The thermocycler program was: 95 °C for 5 minutes, 40 × (95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30

seconds), finishing with 72 °C for 7 minutes. Reassembled products were purified over a QIAquick® PCR purification spin column (QIAGEN).

To amplify full-length reassembled PhnG, a PCR was performed using the purified reassembly PCR products as the template and the same forward and reverse flanking primers used for cloning of the wild type DNA (Section 3.2.2). The mix consisted of 5 µL 10x Taq polymerase reaction buffer, 1 µL dNTPs (25 mM each), 1µL of 10 µM each primer, 1 µL DMSO, 1µL of the purified reassembly products, and 39 µL of water for a final volume of 49 µL. 1 µl of Taq polymerase was added during the first step of the PCR program at 94 °C. The program used was 94 °C for 3 minutes, 25 × (94 °C for 30 seconds, 60 °C for 30 seconds, 72 °C for 1 minute), finishing with 72 °C for 7 minutes.

The single product generated from the amplification of reassembled PhnG was isolated on a 1% agarose gel and purified using a QIAquick® gel extraction kit (QIAGEN), then digested with *Hind*III and *Nhe*I for 1.5 hours. The digested and purified library was then ligated into the screening vector pProEx\_GFPuv1.

For the backcrossing round, all steps performed were identical to those for DNA shuffling, with only difference being the template preparation prior to DNase I digestion. To prepare the template for DNaseI digestion, the concentrations of wild type *phnG* and the mixed pool of pure plasmids isolated from round three were estimated by measuring their absorbance at 260 nm. Concentrations were calculated using the equation:

$$[\text{DNA}] = A_{260} \times 0.020 \text{ (}\mu\text{g/ml)}^{-1} \text{ cm}^{-1} \quad [13]$$

Based on these calculations the wild type and round 3 mixed plasmid concentrations were equalized and combined in a ratio of 40  $\mu$ L wild type plasmid to 1  $\mu$ L mixed round three plasmids.

### **3.2.5 Selection and purification of library variants with improved fluorescence**

Selection and purification of library variants of PhnG with improved fluorescence was performed following the protocol described in Chapter 2, section 2.2.7.

### **3.2.6 Expression of PhnG-GFP wild type and clone B6**

The plasmids harbouring the *phnG* wild type and clone B6 fusions were used to transform XL1-Blue cells. The cells were grown overnight at 37 °C on LB agar plates containing 100  $\mu$ g/mL ampicillin to produce single fluorescent colonies. One colony was picked to inoculate a 25 mL culture containing LB, 1% glucose and 100  $\mu$ g/mL ampicillin, which was shaken overnight at 250 rpm, 37 °C. 10 mL of this preculture was used to inoculate a 1 L culture containing LB, 1% glucose and 100  $\mu$ g/mL. The 1L culture was grown at 30 °C, 250 rpm until an OD<sub>600</sub> of 0.6 was reached. Just prior to reaching an OD<sub>600</sub> of 0.6, the temperature of the culture was reduced to 15 °C. When the culture reached an OD<sub>600</sub> of 0.6 and a temperature of 15 °C it was induced with 1 mM IPTG. After induction the culture was grown for an additional 24 hours at 15 °C, 250 rpm. Cells were harvested via centrifugation at 3000 g for 20 minutes, at which point they were either stored at -20 °C or immediately lysed.

### **3.2.7 Purification of PhnG-GFP wild type and clone B6**

Cells harvested after expression were lysed with an EmusiFlex cell homogenizer (Avestin Inc., Ottawa) and centrifuged at 40000 g for 30 minutes. The supernatant was purified via immobilized metal ion affinity chromatography (IMAC) over Ni-NTA resin. The supernatant was bound to the column in a buffer containing 25 mM Tris, 300 mM NaCl and 10 mM

imidazole, pH 7.2. The protein was eluted from the column with buffer comprised of 25 mM Tris, 300 mM NaCl and 500 mM imidazole, pH 7.2, and collected in 1 mL fractions.

### **3.2.7.1 Purification in the presence of detergent and DTT**

The same protocol was followed as for the standard purification described above with the only difference being the addition of 0.1 mM n-dodecyl- $\beta$ -D-maltoside (DDM) and 2 mM dithiothreitol (DTT) to both purification buffers.

### **3.2.7.2 Purification under denaturing conditions**

To purify the soluble fractions under denaturing conditions, the supernatant collected after cell lysis and centrifugation was denatured by diluting it in 8 M urea such that the final concentration of urea was 6 M. This solution was slowly mixed for 3 hours at room temperature then centrifuged at 20,000 g for 30 min to remove any particulate matter. The supernatant resulting from this centrifugation was purified as per the same protocol as the standard purification described above, with the addition of 8 M urea to both purification buffers.

### **3.2.8 Resolubilization, purification and refolding of PhnG-GFP wild type and clone B6 inclusion bodies**

Insoluble material collected as a pellet after centrifugation of the cell lysates was resuspended via vigorous vortexing into 10 mL of buffer containing 25 mM Tris, 300 mM NaCl, 10 mM imidazole, and 8 M urea, pH 7.2. This suspension was slowly mixed at room temperature for 4 hours after which any material that did not solubilize was pelleted via centrifugation at 20,000 g for 30 minutes. The supernatant resulting from this centrifugation was purified via IMAC purification as described above, with 8 M urea added to both purification buffers.

The concentration of the purified material was approximated via measurement of absorbance at 280 nm and dividing by the sum of the molar extinction coefficients of PhnG (41830  $M^{-1}cm^{-1}$ , calculated using the online ProtParam program found at

<http://ca.expasy.org/tools/protparam.html> [14]) and GFPuv ( $\epsilon_{280} = 20600 \text{ M}^{-1}\text{cm}^{-1}$  [15]). Based on these values the concentration of both the wild type and clone B6 PhnG-GFP fusions were normalized to ensure both solutions exhibited the same absorbance at 280 nm. The two solutions of equal concentration were placed in separate dialysis cassettes and then suspended in a beaker of stirring buffer (25 mM Tris, 300 mM NaCl, 2 mM DTT, pH 7.2) at 4 °C overnight. In the morning, the solutions were removed from the dialysis cassettes and centrifuged at 15000 rpm for 30 minutes. The concentration of the supernatants was determined via measuring absorbance at 280 nm and using the combined extinction coefficients of PhnG and GFPuv ( $62430 \text{ M}^{-1}\text{cm}^{-1}$ ).

### **3.2.9 Fluorescence measurements of PhnG-GFP wild type and clone B6**

To compare the fluorescence of whole cells between the clones isolated from each round, 4 mL overnight cultures were grown for each clone containing LB and 100  $\mu\text{g}/\text{mL}$  ampicillin. In the morning, the cells were pelleted via centrifugation at 3000 g for 20 minutes to separate them from the culturing medium. The pellet was resuspended in 10 mL of buffer (25 mM Tris, pH 7.2) and diluted until an approximate  $\text{OD}_{600}$  of 2 was reached. The cell suspension was immediately transferred to a quartz cuvette and fluorescence was measured on a Photon Technology International (PTI) fluorimeter, and data collected with FeliX32 software (PTI). The excitation maximum was scanned over a range of 300 to 460 nm with the emission maximum set to 508. The emission maximum was scanned over a range of 450 to 600 nm with the excitation maximum set to 399. All fluorescence measurements were normalized by dividing the fluorescence value by the  $\text{OD}_{600}$  of the respective samples. Fluorescence of the supernatants obtained after cell lysis and centrifugation were normalized by dividing the fluorescence value by the total protein concentration as determined by Bradford assay.

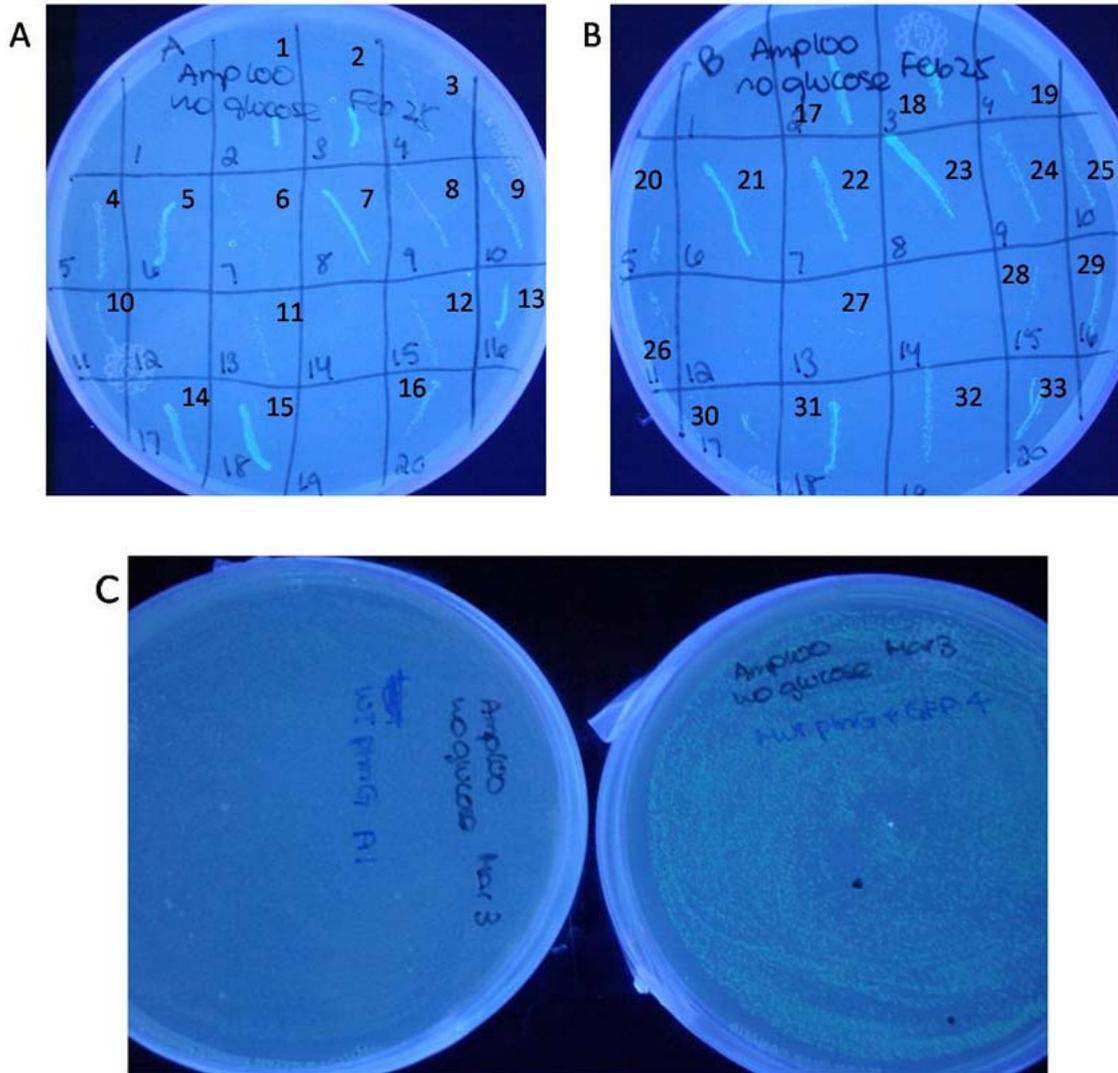
### **3.2.10 Size exclusion chromatography**

Both native protein isolated after purification and refolded protein were analyzed by size exclusion chromatography to determine the oligomeric state of the fusions. The Tricorn 10/300 size exclusion column was packed with Superdex 200 resin according to the protocol provided by the column manufacturer (Pharmacia). It was calibrated using a basic molecular weight marker kit containing the standards blue dextran (2000 kDa),  $\beta$ -amylase (200 kDa), alcohol dehydrogenase (150 kDa), bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), and cytochrome c (12.4 kDa). The running buffer used was 50 mM Tris-Cl, 100 mM KCl, pH 7.5, and the flow rate was 1 mL/min. For the samples that had been purified into buffer containing detergent and DTT, the size exclusion running buffer was also supplemented with equal concentrations of detergent and DTT.

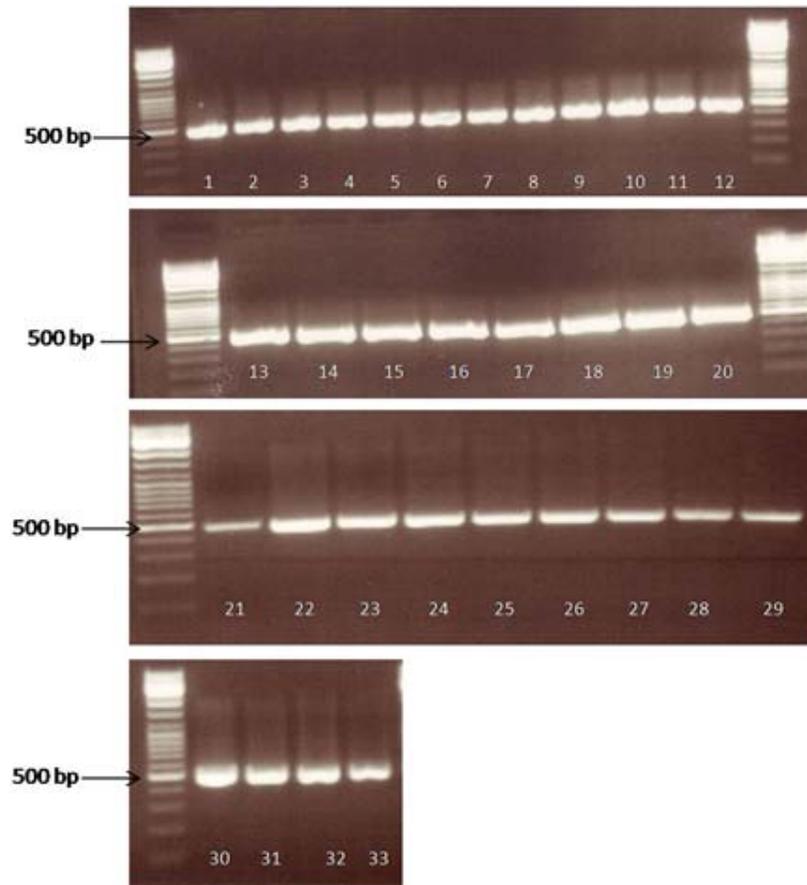
### 3.3 Results and Discussion

#### 3.3.1 Mutagenic PCR on *phnG*—round 1

Forty colonies were selected from the library generated from the first round of evolution on PhnG using error-prone PCR as the diversification technique. Perhaps because of the short length of the template (450 bp), the PCR reaction was highly tolerant to the mutagenic components of error-prone PCR and a high product yield was obtained using the full amounts of all mutagenic components. Two 500 cm<sup>2</sup> plates of approximately 25000 colonies each were needed to be able to select 40 colonies of high fluorescence, however due to the selections being performed with short wavelength light (360 nm), many of the clones did not grow (**Figure 3-2 A, B**). The clones that did grow were re-labelled 1 – 33, and plasmid DNA was isolated from each streak. Each plasmid was checked for purity by using it to transform XL1-Blue cells and examining the resulting colonies for homogeneity of fluorescence. The resulting colonies were also compared to a plate of cells harbouring a plasmid carrying the gene for the wild type PhnG-GFP fusion to ensure an improvement in fluorescence had been achieved (**Figure 3-2 C**). Any plasmids that were identified as mixtures were purified by growing a single colony from the mixture plate in a 4 mL culture overnight to isolate the plasmid DNA the next day. After purification, all plasmids were found to contain the correct insert as determined by PCR amplification using the purified plasmid as a template (**Figure 3-3**).



**Figure 3-2.** Clones picked from round 1 of evolution of PhnG (A and B) and comparison of mutant 4 and wild type PhnG colonies (C).



**Figure 3-3.** PCR amplification of purified mutant genes from round 1 of PhnG evolution

Fifteen of the mutant plasmids were sent in for sequencing to determine the types and amount of mutations acquired from error-prone PCR. Many of the variants that were sequenced were found to contain identical mutations. **Table 3-2** shows the overall mutation rate and the average amount of transitions and transversions encountered for the 15 mutants. The PCR was highly mutagenic, averaging 3.3 mutations per 450 bp gene, and, surprisingly, a high transition to transversion ratio of 2:1 was achieved. A transversion rate of 33% is slightly higher than the transversion rate of 20% achieved for the 5LO mutagenic PCR (Chapter 2). **Table 3-3** shows the resulting amino acid mutations, and how many identical clones were found for each. Out of the 15 sequenced, seven unique mutants were discovered. Of interest is Val47, which was mutated to either to Ala or to Glu in 11 of the 15 clones sequenced. The Ala residue at position 25 was also

highly selected, being mutated to either Val or Ser in three clones. A Leu67Pro mutation was discovered in 4 clones. Three identical clones containing a Val47Ala mutation and Tyr81Ser mutation were discovered, as well as three containing only a Val47Glu mutation. By eye, the brightest streaks appeared to be clones 6B, 8B, and 18B (**Figure 3-2, B**). Fluorescence was not quantified for the round 1 mutants. These mutants all contained Val47Glu mutations, with 8 and 18 (identical to R1-7 in **Table 3-3**) also containing a mutation at Ala25 to either Val (clone 18B) or Ser (clone 8B).

**Table 3-2 Mutation analysis of phnG after error-prone PCR**

	Silent Mutations	Total Mutations	Transitions	Transversions
Maximum per 450 bp	3	5		
Minimum per 50 bp	0	2		
Average per 450 bp	1.4	3.3	2.2	1.1

**Table 3-3 Amino acid mutations from selected round 1 PhnG mutants**

Clone	# selected	Mutations		
R1-2	3		V47A	Y81S
R1-3	2		V47A	A70T
R1-6	3		V47E	
R1-7	2	A25V	V47E	F142L
R1-8	1	A25S	N31S	V47E
R1-15	2		L67P	
R1-19	1		L67P	Q90R
R1-20	1		N31S	L67P

Homologs of PhnG were aligned to identify highly conserved residues and to verify whether any of the selected mutations found in round 1 matched the overall consensus. Mutations which convert an amino acid back to the corresponding amino acid found in the protein's consensus sequence have been shown on several occasions to have a stabilizing effect on the final structure of the protein [16]. As PhnG homologs have a very high sequence identity, only those with the lowest identity were chosen for the alignment. The alignment of PhnG homologs is shown in **Figure 3-4 A**, and the unique mutants are aligned in **Figure 3-4 B**. For the homolog alignment, strictly conserved residues are shown in yellow and the residues highlighted in green represent positions where mutations were found in the round 1 mutants. The highly selected mutations at Val47 target a highly conserved residue, with only one of the homologs not having a Val at this position. A homolog from *R. eutropha* contains Ala at this position, which is a mutation found in the R1 mutants. The Leu67Pro mutation also targets a highly conserved position, however interestingly every homolog except for *E. coli* PhnG has a Val at this position. Also interesting is the Ala70Thr mutation. In the homologs, 4 contain a cysteine at this position, with the other three (including *E. coli*) having Ala here. Mutation to Thr at this position converts the aliphatic Ala to a residue capable of hydrogen bonding that is more similar to the conserved cysteine residue.

```

P. gallaeciensis      MNSEKSMTAGENDHANR KAWMGLLATAADAKALQDLWQ---NYGCNPDHT 47
R. bacterium         -MTEKTDTA-----R RARMGLVAKAPPARLAALMA---GVEV-PGFD 37
L. vestfoldensis    MQATIDSQAA-----R KGWLGLLAKAPAAKLAQLWA---AANITPTHH 40
O. antarcticus      MNMMSDPNAA-----R KGWLGLLAKSPATEVARLWL---DLKIEPAHS 40
R. nubinhibens      -MTKK-----R QTWMLLARAPSERVIALWD---GIGKAPEFS 34
R. eutropha         MQDTTIADAS-----AARAGWLRILALAQPDALDAAYAQLSGQVLPAYR 45
E. coli             MHADTAT-----R QHWMSVLHLSQPAELAA RLN---ALNITADYE 37
Consensus           AARKGWMGLLAKAPA LAALW GL I P H

P. gallaeciensis    WLRPPEVGGVMVQGRMGASGAPFNLGEMTVTRCALT LAD---GTVGHGY 93
R. bacterium        WLRAPEVGGVMVGRMRMGGTGAPFNLGEMTVTRCALR LAD---GEVGHGY 83
L. vestfoldensis    VLRAPEIIGAVMVRGRAGAVGAAFNLGEMSVTRASVRLAD---GTIGHGY 86
O. antarcticus      VLRTPEIIGVMVRGRAGAVGAAFNLGEMTVTRASVK LAD---GTVGHGY 86
R. nubinhibens      WARMPETGGVMVGRMRMGGTGDAFNMGCEVTVTRCALR LAD---GTTGHAY 81
R. eutropha         LLRKPEAGMAMVRGRAGGTGAQFNLGEVSVTRCAIV LADASAGSTAGVAY 95
E. coli            VIRAAETGLVQIQARMGGTGERFFAGDATL TRAVRLTD---GTLGYSW 83
Consensus           VLR PEIGVMVGRMRMGGTGA FNLGEMTVTRCALRLAD GTVGHGY

P. gallaeciensis    VQGRSKLQAEATAAKVDALMQ-TDAAEEVHRRVLSPLQAAKHTRKMSRAAK 142
R. bacterium        VQGRDKAHAERAAALVDALMQ-TDRAEAVQAQVLDPLAEAAALTAKATRAAK 132
L. vestfoldensis    VQGRDRTHALHAALI DALMQ-TDAAGQVDRAILSPLRAAAADRQTARAAK 135
O. antarcticus      VQGRGKDHAMHAALVDALMQ-TAAATAIEADLLTPLRIAMKQGKTNRRAAK 135
R. nubinhibens      VQGRSRRHAEIAALADALMQ-TDEAAT IETGLLDPLQREEEARRARRAAK 130
R. eutropha         VQGRGTRHAEQAALVDALMQRADWHQRVDRDTVLAPLAQAHAARAANRAGV 145
E. coli            VQGRDKCHAERCALI DALMQQSRHFQNLSETLIA PLDADRMARIAARQAE 133
Consensus           VQGR K HAE AALVDALMQ TD A V LLAPL A RKA RAAK

P. gallaeciensis    AAATKVEFFTMVRGED- 158
R. bacterium        AAATKVDEFFTMVRGED- 148
L. vestfoldensis    AAATKVDEFFTMVRGED- 151
O. antarcticus      AAATKVDEFFTMVRGED- 151
R. nubinhibens      AAATKVDEFFTMVRGEDA 147
R. eutropha         AAQTRVEFFTMVRGED- 161
E. coli            VNASRVDEFFTMVRGDNA 150
Consensus           AAATKVDEFFTMVRGED

```

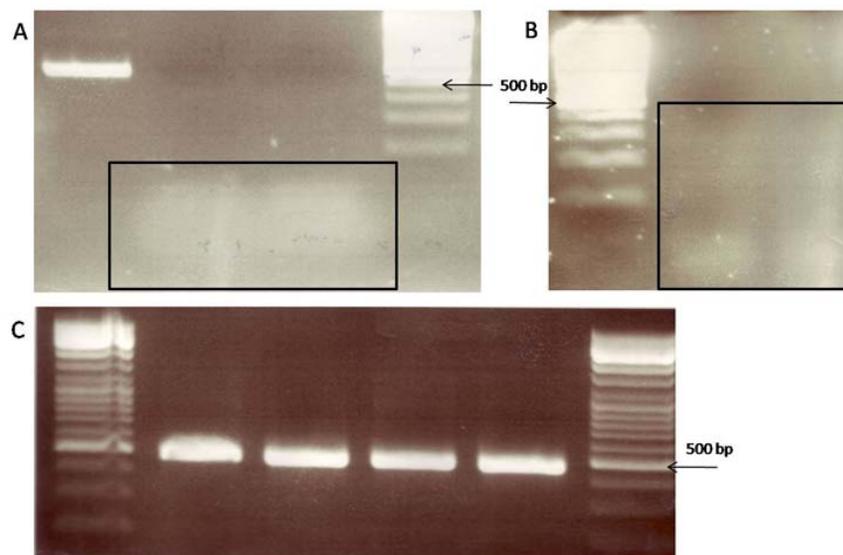
**Figure 3-4.** Multiple sequence alignment of PhnG homologs. Homologs are from *P. gallaeciensis* (gi: 163738993), *O. antarcticus* (gi: 254436493), *R. bacterium* (gi: 84685142), *R. nubinhibens* (gi: 83950680), *R. eutropha* (gi: 73538016), *L. vestfoldensis* (gi: 84516853), and *E. coli* (gi: 536945). Strictly conserved residues are highlighted in yellow, and positions mutated in round 1 mutants are highlighted in green. Alignments were performed using ClustalW [17].

	25	31	47	
R1_2	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLA	Q	QIQARMGGTGERF	60
R1_3	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLA	Q	QIQARMGGTGERF	60
R1_7	MHADTATRQHWMSVLAHSQPAELAV	R	LNALNITADYEVIRAAETGLE	QIQARMGGTGERF 60
R1_8	MHADTATRQHWMSVLAHSQPAELAS	R	LNALSITADYEVIRAAETGLE	QIQARMGGTGERF 60
R1_6	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLE	Q	QIQARMGGTGERF	60
R1_15	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLV	Q	IQARMGGTGERF	60
R1_19	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLV	Q	IQARMGGTGERF	60
R1_20	MHADTATRQHWMSVLAHSQPAELAARLNALS	I	TADYEVIRAAETGLV	QIQARMGGTGERF 60
phnG	MHADTATRQHWMSVLAHSQPAELAARLNALNITADYEVIRAAETGLV	Q	IQARMGGTGERF	60
	***** .*****			
	67	70	81	90
R1_2	FAGDATLTRA	A	VRLTDGTLG	S
R1_3	FAGDATLTRT	A	VRLTDGTLG	S
R1_7	FAGDATLTRA	A	VRLTDGTLG	S
R1_8	FAGDATLTRA	A	VRLTDGTLG	S
R1_6	FAGDATLTRA	A	VRLTDGTLG	S
R1_15	FAGDAT	P	TRA	A
R1_19	FAGDAT	P	TRA	A
R1_20	FAGDAT	P	TRA	A
phnG	FAGDATLTRA	A	VRLTDGTLG	S
	***** **:*****:*****:*****			
	142			
R1_2	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_3	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_7	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_8	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_6	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_15	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_19	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
R1_20	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
phnG	ADRMARIAARQA	EVNASRVDF	FTMVRGDNA	150
	*****:*****			

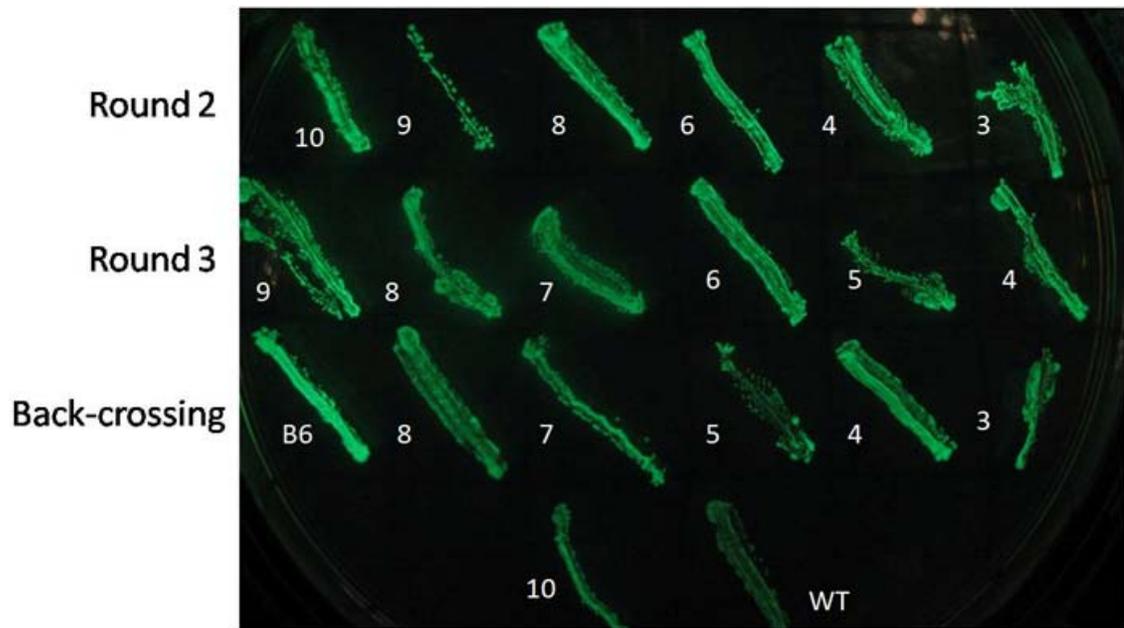
**Figure 3-5.** Alignment of unique PhnG mutants from round 1.

### 3.3.2 DNA shuffling of *phnG*—rounds 2 and 3 and back-crossing

To enrich the positive mutations discovered during the first round of evolution, DNA shuffling was used as the diversification strategy for the second and third rounds. The 33 purified plasmids isolated from the first round were combined in equal volumes for use as the gene pool in the second round. **Figure 3-6** shows example gels from the DNase I digestion, reassembly, and amplification of *phnG*. For each round, 40 colonies were picked and verified for purity before being carried onto the next round. Six of the brightest clones from each round were sequenced to determine if any mutations were becoming enriched or removed. **Figure 3-7** shows the fluorescence of the 6 sequenced clones from the 2<sup>nd</sup>, 3<sup>rd</sup>, and back-crossing rounds. As is evident from these streaks, fluorescence was already so high after the second round it was difficult to tell by eye whether or not an improvement in fluorescence was actually made in the third or backcrossing rounds. One clone, labeled B6 in **Figure 3-7** was clearly the brightest, and will be discussed in greater detail in the next sections.



**Figure 3-6.** DNA shuffling of *phnG*. Fragments from the DNase I digestion are shown in **A**, reassembly of the fragments is shown in **B**, and amplification of the reassembled gene is shown in **C**.



**Figure 3-7.** Fluorescence of the 6 brightest clones from rounds 2 and 3, and the backcrossing round.

To determine if in fact an improvement in fluorescence had been achieved, whole cell fluorescence measurements were made on all sequenced clones. Although Waldo concluded in his 1999 *Nature Biotechnology* paper that there was a “moderate correlation” between whole cell fluorescence measured using a fluorimeter and the amount of protein in the soluble fraction [18], later work examining this correlation definitively concluded that whole cell fluorescence measurements are a reliable way to predict the amount of soluble protein present in cells [19]. Cells from each of the clones shown in **Figure 3-7** were grown overnight in 4 mL cultures, centrifuged and then resuspended in Tris-NaCl buffer. Optical density of each suspension was determined, and all fluorescence measurements were normalized by dividing the fluorescence value by the optical density of the cell suspension. Mutations found in these clones, as well as their relative fluorescence are shown in **Table 3-4**.

**Table 3-4.** Amino acid mutations for selected clones from rounds 2 and 3, and back-crossing of PhnG evolution

<i>Sample</i>		<i>Mutations</i>	<i>Relative Fluorescence</i> (wild type = 1)
Round 2	R2-3	A25V, V47E F142S	1.18
	R2-4	A25V V47E F142L T5I	1.70
	R2-6	L67P Q90R	1.60
	R2-8	A25V V47E F142L T44A F61L	0.98
	R2-9	A25V V47E F142L	1.17
	R2-10	A25V V47E	1.15
Round 3	R3-4	A25V V47E	1.12
	R3-5	V47A E43G A70T	1.10
	R3-6	V47A F142L A70T	0.75
	R3-7	L67P Q90R T66A T114A	2.21
	R3-8	A25T V47E Q90R K89E	1.19
	R3-9	V47A F141L E43G A70T	1.26
Back-crossing	BC-3	V47A Q90R R59H A70T	1.13
	BC-4	V47E	1.62
	BC-5	V47E A70T	1.29
	BC-7	L67P	1.46
	BC-8	T44A	1.17
	BC-10	A25V V47E H17Y	0.82
	BC-6	V47A A70T D120G	6.68

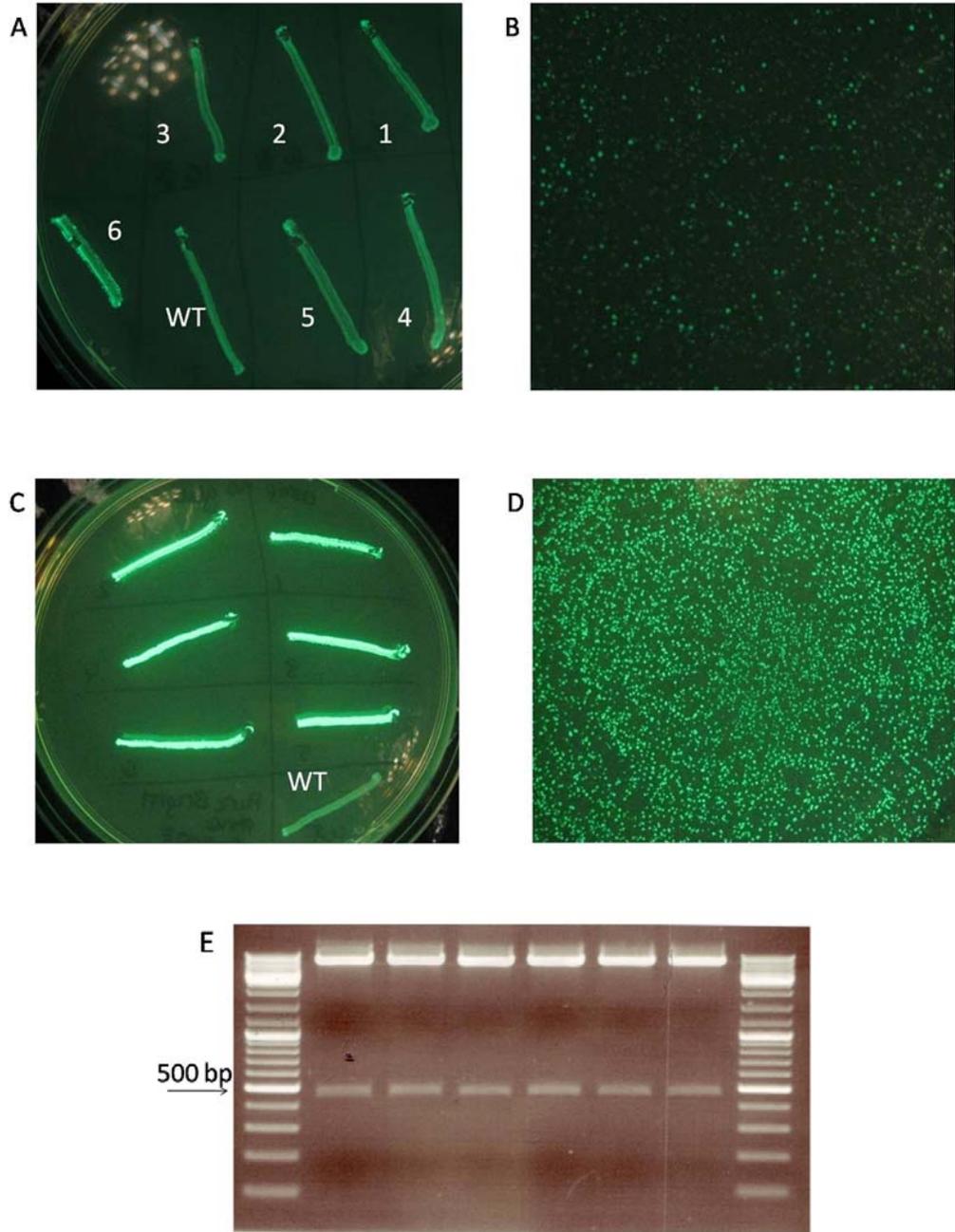
Sequence analysis from the selected clones showed a definite enrichment of the mutations at Val47 to either Glu or Ala, as this position was selected in 73% of the clones sequenced. Ala25 was also enriched greatly, but perhaps the most surprising mutation is Leu67Pro, which was found in all three rounds in some of the brightest mutants from each round (R2-6, R3-7, BC-7). Phe142Leu was found twice in the round 1 mutants, but appeared in four of the six mutants analyzed in round 2. Most mutations at this position are Phe142Leu, but one case it was Phe142Ser. Also, in one case an identical mutation was found directly next to this position—Phe141Leu. Ala70Thr was found to be greatly enriched in the third round appearing in four out of six clones. The brightest clone by far, B6, contained both the Val47Ala mutation and the Ala70Thr mutation, but also a unique Asp120Gly mutation. It is assumed that this last mutation is the reason for the dramatic increase in fluorescence as clone BC-5 is almost identical but lacks the Asp120Gly mutation and has similar fluorescence to wild type. The next brightest clone found, R3-7 contained four mutations: Leu67Pro Gln90Arg, Thr66Ala and Thr114Ala. The last two mutations were unique to this clone. The second brightest clone from round 2 also contained the Leu67Pro and Q90R mutations. Apart from the three missense mutations found in clone B6, there were two other silent mutations: an isoleucine (ATC-ATA), and a leucine (CTG-TTG). Neither of these mutations converts the natural codons to those which are less rare. In fact, the ATA codon for isoleucine is one of the rarest found in *E. coli* [20]. Because this clone was the only one that showed any dramatic improvement in fluorescence over the wild-type fusion, it was decided that the solubility and folding properties of this clone should be examined in greater detail.

### **3.3.3 Solubility and fluorescence analysis of clone B6**

Upon initial selection of clone B6 it was obvious that it was the brightest clone found but it also appeared to be non-homogeneous (**Figure 3-8 A**). To verify whether or not this clone was

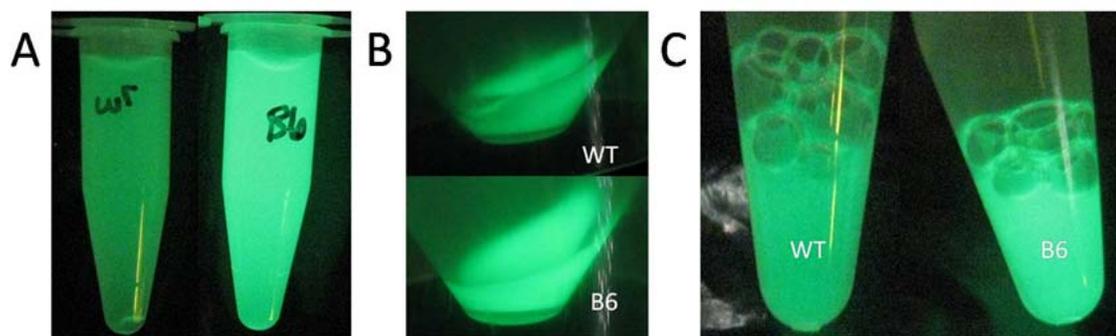
comprised of homogeneous cells the plasmid DNA was isolated and used to transform XL1-blue cells. The resulting colonies had two levels of fluorescence, indicating that the streak was a mixture (**Figure 3-8 B**). Even so, when the plasmid DNA was analyzed by restriction digest, only one insert of the correct size was observed, which could indicate that even though two levels of fluorescence were observed, both types of cells contained a full length insert with one giving rise to a properly folded gene product and the other giving rise to a protein which is prone to aggregation. An alternative possibility was that the highly fluorescent clone contained a random short insert that does not interfere with the folding of GFP, as had been observed in the cases of RebG and 5LO (see Chapter 2).

Six highly fluorescent cells from the mixture plate were selected and streaked on a new plate and all showed equal fluorescence which was much brighter than that observed for the wild type PhnG-GFP fusion (**Figure 3-8 C**). Restriction digest analysis on the plasmid DNA isolated from each of the pure streaks showed that all contained the correct 450 bp insert (**Figure 3-8 E**). As a final check of purity, the plasmids from each of the pure streaks were used to transform fresh XL1-Blue cells, and the resultant colonies exhibited homogeneous fluorescence (**Figure 3-8 D**). Sequencing was performed after the mutant was purified.

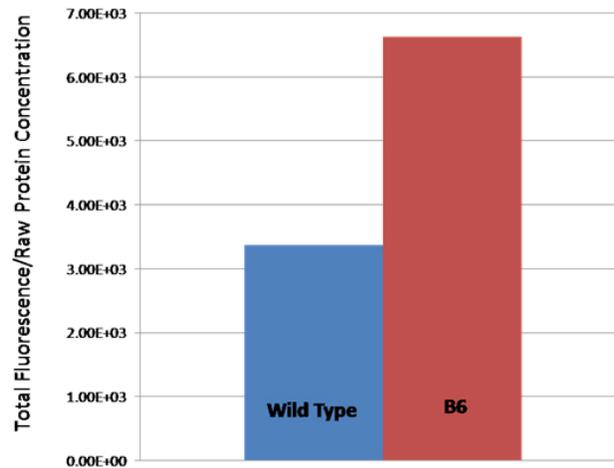


**Figure 3-8.** PhnG clone B6 before and after purification. The unpurified streak is shown as number 6 in **A**. **B** shows the mixture of cells obtained from transformation of XL1-Blue cells with plasmid DNA isolated from the impure clone 6. Pure clones obtained from streaking individual fluorescent colonies from the mixture plate is shown in **C**, restriction digest analysis of these pure streaks is shown in **E**, and the resulting homogeneous cells arising from transforming XL1-Blue cells with DNA obtained from the pure streaks is shown in **D**.

Clone B6 was expressed and the whole cells were measured to have a fluorescence approximately 7 times greater than that of cells expressing the wild type PhnG-GFP fusion—far brighter than any other clone found in the back-crossing round or any other round (**Table 3-4**). As a preliminary step to ensure that the increase in fluorescence occurred in the soluble fraction of the cells, 50 mL test cultures were grown and lysed via B-Per II bacterial protein extraction reagent (Pierce Protein Research Products). Whole cells from this test expression are shown in **Figure 3-9 A**. **Figure 3-9 B** and **C** show the pellets and supernatants after cell lysis and centrifugation. Even though the whole cells were measured to have fluorescence over six times that of wild type, the supernatant of clone B6 was measured to only have fluorescence twice that of wild type (**Figure 3-10**). In all three cases (whole cells, pellet, and supernatant) the B6 mutant appeared brighter, with the most noticeable difference being between the whole cells. From these results it was hard to determine whether or not an improvement in folding was achieved or in fact just an improvement in expression, as it appeared that there was more fluorescent material in both the soluble and insoluble fractions for the B6 mutant.



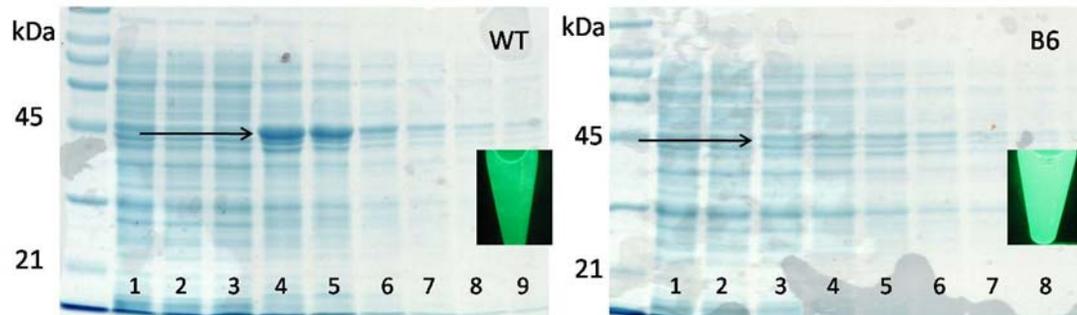
**Figure 3-9.** Test expression of PhnG wild type and clone **B6**. Whole cells are shown in **A**, the insoluble fractions are shown in **B**, and the soluble fractions are shown in **C**.



**Figure 3-10.** Fluorescence of *E. coli* supernatants containing wild type PhnG-GFP and clone B6. Total protein concentration was determined by Bradford assay.

To test whether or not more soluble material could be isolated for PhnG clone B6 compared to wild type PhnG after purification, another expression was performed and protein from the soluble fractions were purified via immobilized metal affinity chromatography (IMAC) under standard conditions. Cells were lysed into Tris-NaCl buffer containing 25 mM imidazole, and the soluble fraction was bound to a column of Ni-NTA resin. Fractions were collected upon elution with Tris-NaCl buffer containing 500 mM imidazole. SDS-PAGE of the fractions collected for the wild type and clone B6 fusions are shown in **Figure 3-11**. Even though the supernatant containing the clone B6 fusion appeared much brighter than the supernatant of the wild-type fusion, it was observed during purification that the mutant protein did not bind to the column. The eluent from the column was just as fluorescent as the original sample that it was loaded onto the column. The column itself appeared only mildly fluorescent after loading clone B6 (less fluorescent than the flow-through). Surprisingly, the Ni-NTA resin appeared more fluorescent after loading the wild type sample. The SDS-PAGE gels of the fractions collected for both wild type and clone B6 confirmed that only the wild-type fusion bound to the column and

could therefore be eluted upon loading the column with buffer containing a high amount of imidazole.



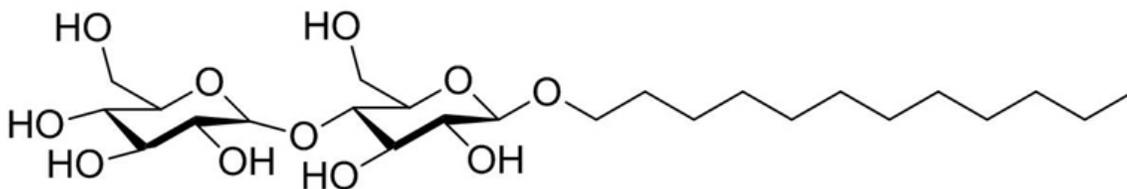
**Figure 3-11.** SDS-PAGE of wild type PhnG-GFP (A) and clone B6 PhnG-GFP (B) fractions after IMAC purification under standard conditions. Fluorescence of each sample prior to purification is shown. A small amount of wild-type protein bound to the column; however no significant amount of mutant protein was captured.

Because the supernatant for the clone B6 fusion was markedly more fluorescent than the supernatant of the wild type fusion, it was clear that protein was present, but perhaps not binding to the column due to an inaccessibility of the hexa-histidine tag. Indeed, the fact that the wild type protein could bind more easily to the column than the mutant protein indicated that a change in folding had occurred. The wild type PhnG-GFP fusion may allow better access to the histidine tag whereas the structure of clone B6 may effectively bury it. Further testing was required to fully reveal if there was more protein in the soluble phase and if it exhibited an improvement in folding ability or if it was simply expressed at a higher level.

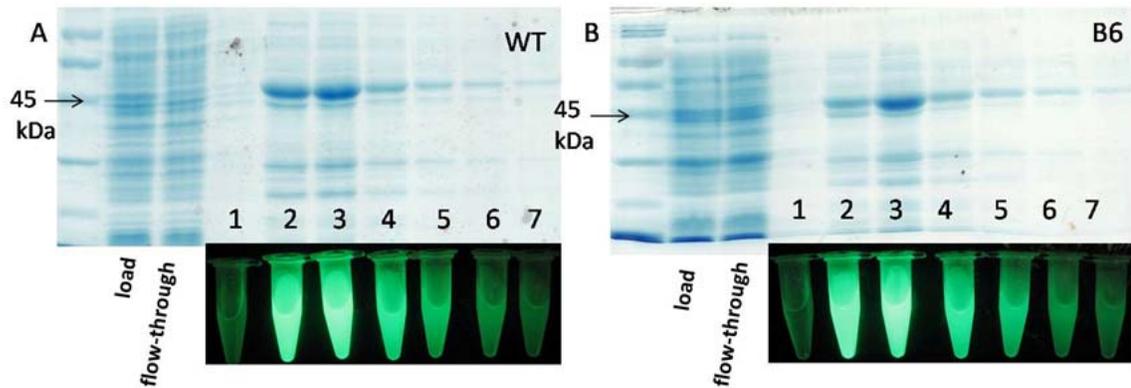
### 3.3.4 Purification of PhnG wild type and clone B6 fusions in the presence of detergent

To analyze the effect detergent would have on the binding of both PhnG wild type and clone B6 to a Ni-NTA purification column, another expression was performed and the harvested cells were lysed into buffer containing 0.1 mM of the non-ionic detergent n-dodecyl- $\beta$ -D-maltoside (DDM) (**Figure 3-12**) and 2 mM dithiothreitol (DTT). **Figure 3-13** shows the SDS-

PAGE gels of the fractions obtained after purification under these conditions. Adding the detergent and DTT did slightly improve the ability of the mutant to bind to the nickel-NTA column, but the amount of wild type PhnG that bound remained similar to that which bound without the presence of detergent or DTT (**Figure 3-11**). The amount of the clone B6 fusion that did bind to the column was still less than that for the wild type fusion, as was evident on the SDS-PAGE gels. However, the flow-through for clone B6 was still more fluorescent than the wild type flow-through, meaning that much of the protein was still did not bind to the column. If the hexahistidine tag of the clone B6 protein was buried within its interior or hidden by the formation of soluble aggregates, the addition of a non-ionic detergent such as DDM could aid in allowing the tag to bind to the nickel-NTA column more effectively [21]. The fact that adding the detergent did improve binding in the mutant's case only indicates that the histidine tag has become slightly more accessible, and does not give information as to whether the tag was hidden in the interior of the protein or hidden by soluble aggregates.



**Figure 3-12.** Structure of n-dodecyl- $\beta$ -D-maltoside.

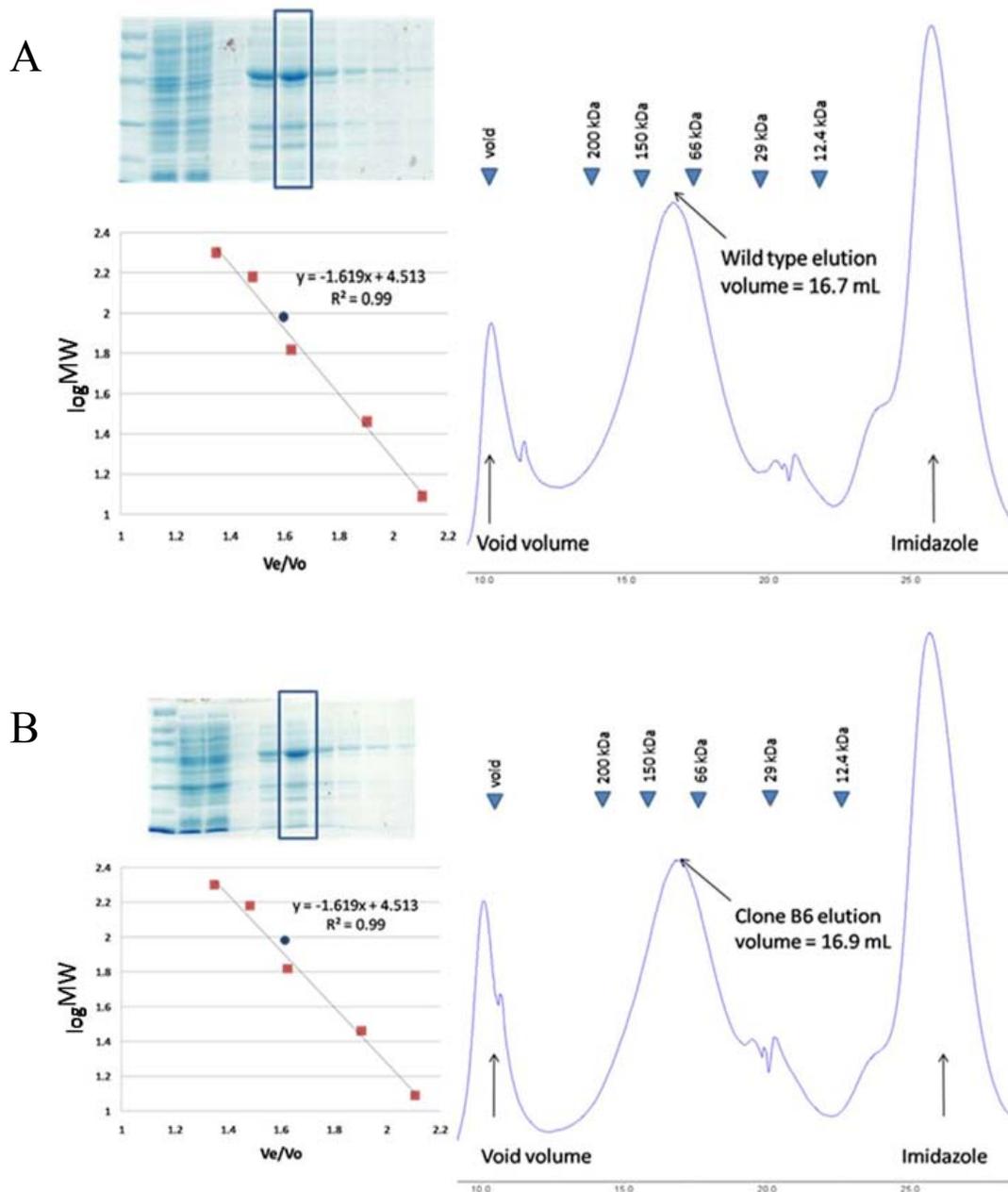


**Figure 3-13.** SDS-PAGE of PhnG wild type (A) and clone B6 (B) after purification in the presence of DDM and DTT. The fractions, fluorescing under 400 nm light, are shown directly under their respective lanes of the SDS-PAGE gel.

### 3.3.5 Size exclusion chromatography of PhnG wild type and clone B6 fusions

To examine whether there was a difference in oligomeric state for the wild type and the clone B6 PhnG-GFP fusions, the third fraction from both purifications shown in **Figure 3-13** were analyzed by size exclusion chromatography. Even though the samples were still partially impure, a peak should be evident on a size exclusion chromatogram as the strongest band present in both samples did represent the protein of interest. The chromatograms are shown in **Figure 3-14**. The samples had almost identical chromatograms, with the wild type fusion eluting at a volume of 16.7 mL, and clone B6 eluting at 16.9 mL. The elution volumes of the standards used for column calibration are shown on each chromatogram and produced a linear relationship when plotted against the log of their respective molecular weights. The calibration curve is shown inset on each chromatogram, along with the SDS-PAGE gel highlighting the fraction that underwent analysis. The wild type and clone B6 fusions, including his-tags have a calculated molecular weight of 47.7 and 47.6 kDa, respectively. According to the calibration of the column (void volume = 10.45 mL), the wild type protein's elution volume (16.7 mL) corresponds to a molecular weight of 84.3 kDa, and clone B6's elution volume (16.9) corresponds to a molecular

weight of 78.5 kDa. These values are approximately double the actual molecular weights of the fusions, indicating that they exist as dimers in solution. A data point plotting the log of the dimeric weight of each protein against their elution volumes is shown as a blue circle on the calibration curves for each chromatogram. It should be noted that GFP itself has been crystallized as a dimer, although the dimerization is thought to be an artifact of crystallization conditions rather than an inherent property of GFP itself [22]. In fact, the interface between the dimers of GFP is considered to be fairly weak and the monomer is presumed to be the predominant form. However, it has been noted that GFPuv has a greater propensity to dimerize compared to wild type GFP, but the actual conditions under which this occurs is unclear [23]. It is therefore inconclusive as to whether the dimers observed from size exclusion chromatography are a result of PhnG or GFP dimerization.

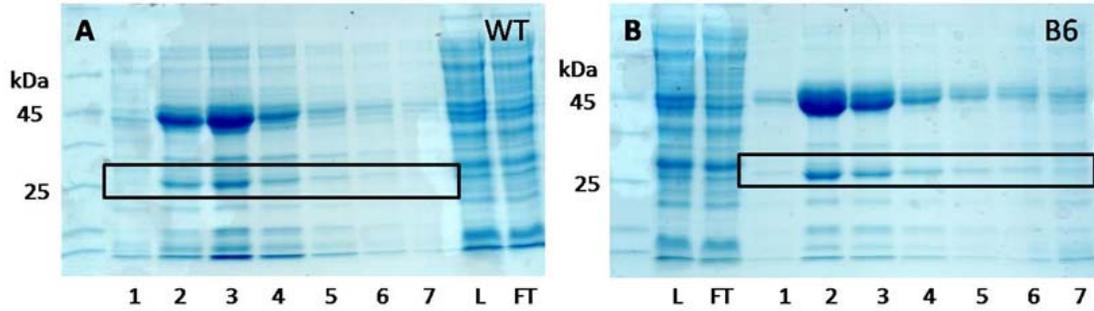


**Figure 3-14.** Size exclusion chromatograms of the semi-pure PhnG wild type (A) and clone B6 (B) fusions in DDM and DTT. The elution volume of the calibration standards is indicated along the top. The standards used were blue dextran (2000 kDa),  $\beta$ -amylase (200 kDa), alcohol dehydrogenase (150 kDa), bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), and cytochrome c (12.4 kDa). The calibration curve based on the elution volumes of the standard proteins is shown in inset, with the point representing the elution of PhnG indicated by the blue circle.

### **3.3.6 Purification under denaturing conditions and refolding of PhnG wild type- and clone B6-GFP fusions**

In order to be able to fully compare the yields of both proteins in the soluble phase, the soluble portions were denatured in 8 M urea and purified under denaturing conditions. Denaturing the proteins should fully expose the histidine tag whether it be hidden in the interior of the protein or hidden due to aggregation. To determine if an improvement in folding has occurred, the denatured proteins were then refolded to compare the refolding yields. If the increase of fluorescence for clone B6 is indeed a result of an improvement of folding, then more properly folded material should be recovered in a refolding experiment.

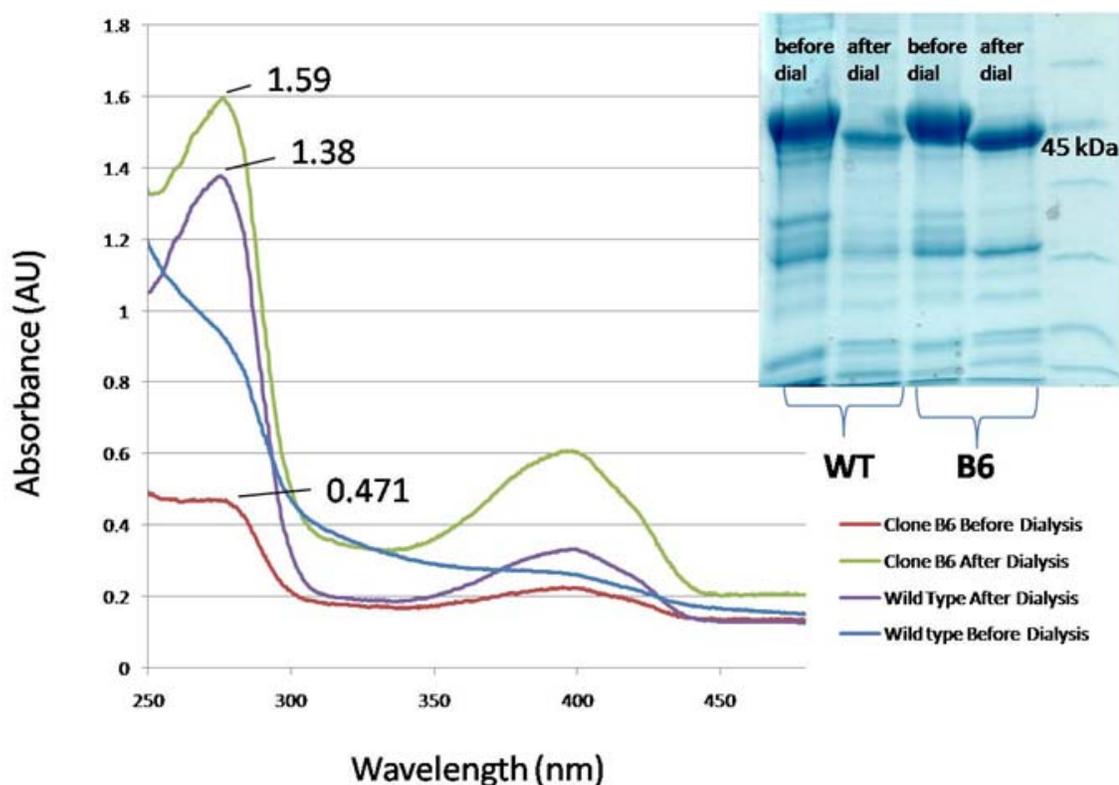
Equal volumes of the wild type and clone B6 supernatants were diluted in 8 M urea such that the final concentration of urea was 6 M. The denatured supernatants were then purified via IMAC under denaturing conditions and the fractions from the purification are shown in **Figure 3-15**. Interestingly, after denaturation the mutant appears to have more material in the soluble fraction compared to wild type, as shown on the right in **Figure 3-15B**. Also surprising was the fact that even though the samples were completely denatured, the flow-through coming off of the column was still highly fluorescent indicating that still some of the material did not bind. This fluorescent material is assumed to be GFP which has been separated from PhnG due to proteolysis. If PhnG is not as resistant to proteolysis as GFP, then there is a chance that some of it has been degraded, leaving the folded and intact GFP in solution. This GFP would not have a his-tag, and therefore would not be able to bind to the column. GFP on its own has a mass of 26.8 kDa, and there is a strong band on both gels in this region (highlighted by the black box).



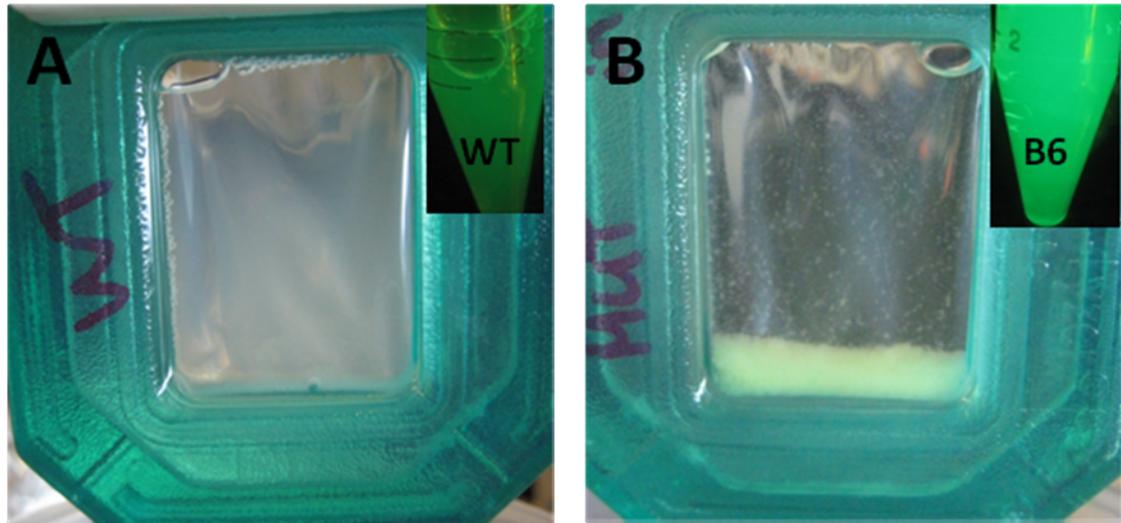
**Figure 3-15.** SDS-PAGE analysis of phnG wild type and clone B6 supernatants after IMAC purification under denaturing conditions. The box indicates the possible presence of GFP

Fractions 2 to 4 from each purification were combined and the UV-visible spectra determined. Interestingly, the wild type fusion did not exhibit a defined peak at 280 nm, even in 8 M urea, and as a result the yields could not be compared quantitatively. The combined fractions were instead compared qualitatively by running them on an SDS-PAGE gel. Because both samples appeared to have similar concentrations, the samples were used directly for refolding via dialysis. **Figure 3-16** shows the absorbance spectra for PhnG wild type and clone B6 GFP fusions before and after dialysis with the SDS-PAGE gel of the before and after samples inset. The pre-dialysis samples were diluted 10 times prior to measuring absorbance and thus the peaks are smaller than the scans taken after dialysis. After dialysis, both the wild type and clone B6 samples gave defined peaks at 280 nm. From the gel it looks as though the combined fractions before dialysis are relatively equal in concentration (despite the appearance of the absorbance spectra), but after dialysis there is clearly more material in the clone B6 sample, as is evident both by the  $A_{280}$  and the intensity of the PhnG-GFP band on the SDS-PAGE gel. Taking baselines into account and using the combined extinction coefficients of PhnG and GFPuv totaling  $62430 \text{ M}^{-1} \text{ cm}^{-1}$ , the pre-dialysis concentration of the clone B6 fusion was approximately  $54 \mu\text{M}$ . The concentration of the solutions after refolding were  $23 \mu\text{M}$  for clone B6 and  $20 \mu\text{M}$  for wild type. Another interesting observation arising from dialysis is that the structure of the inclusion bodies

appears to be different for the two variants, as shown in **Figure 3-17**. The wild type precipitate had a gel consistency, with no defined crystalline particles whereas the mutant precipitate did appear as individual flakes. The difference in precipitates is another indication that a change in folding has occurred. Also shown in **Figure 3-17** is the fluorescence of the refolded material after the precipitate had been removed by centrifugation, with the mutant being clearly more fluorescent.

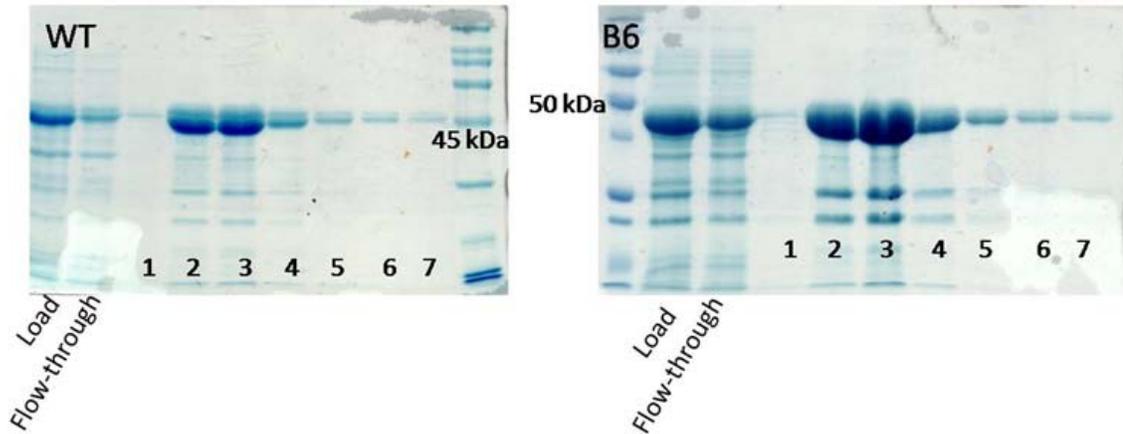


**Figure 3-16.** Absorbance spectra and SDS-PAGE analysis of wild type PhnG-GFP and clone B6 PhnG-GFP denatured in 6 M urea prior to refolding, and refolded PhnG-GFP fusions following dialysis



**Figure 3-17.** Precipitates of PhnG wild type (A) and clone B6 (B) fusions after dialysis

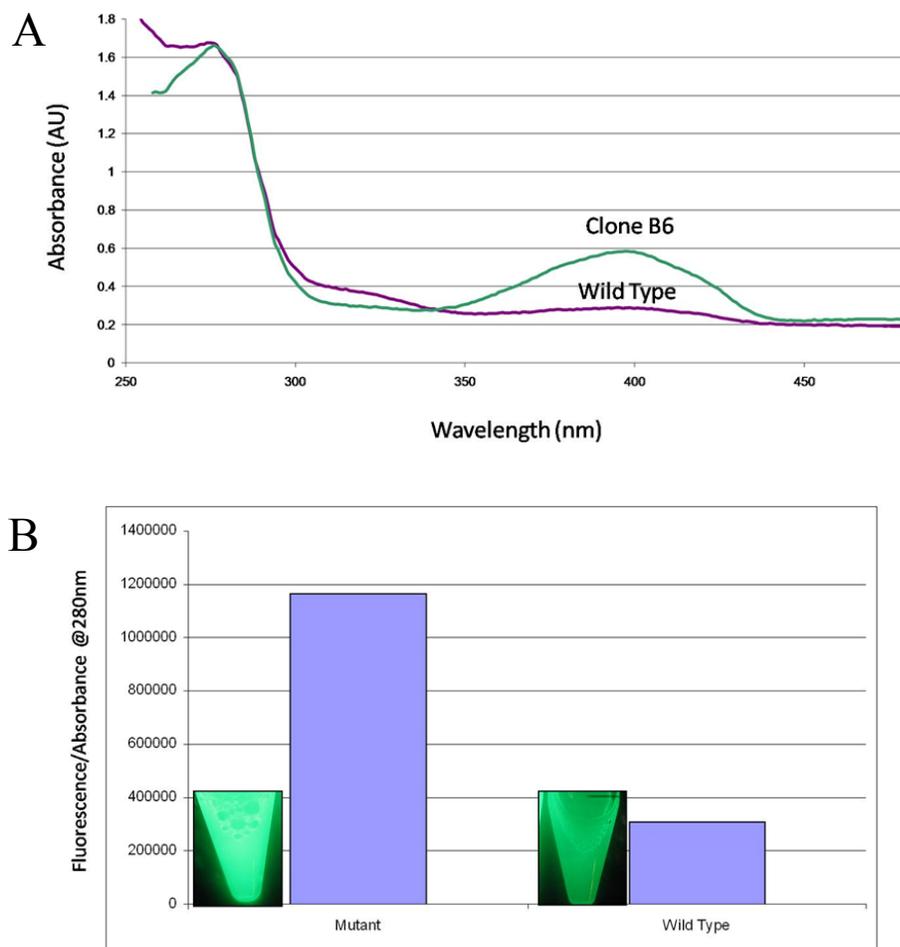
It was still necessary to obtain quantitative results comparing the refolding yields of the denatured wild type and clone B6 fusions to verify that clone B6 was producing more refolded material, thus the inclusion bodies from both expressions were also resolubilized in 8 M urea and purified via IMAC under denaturing conditions. **Figure 3-18** shows the fractions obtained from both purifications. The clone B6 pellet clearly had more protein than the wild type pellet and was also more fluorescent than the wild type pellet. This indicates that an improvement in overall expression has been achieved as more material was observed in both the soluble and insoluble fractions for this mutant.



**Figure 3-18.** SDS-PAGE gel of fractions obtained from denaturing purification of wild type and clone B6 inclusion bodies

To give both wild type and clone B6 equal chances of success during refolding, their concentrations were equalized, as shown in **Figure 3-19 A**. Again, the peak at 280 nm for the wild type fusion is not as defined as the peak for the clone B6 fusion. Aside from the peak at 280 nm, another peak at 399 nm, the excitation maximum for GFP, was unexpectedly much larger for the mutant than for the wild type even at equimolar concentrations of both fusions. Fluorescence measurements showed that at these concentrations the mutant was still almost 4 times as fluorescent, shown in **Figure 3-19 B**. The reason for the differing fluorescence between the two samples is most likely a reflection of the success or failure of the initial folding events at the cellular level. GFP's structure is extremely stable, thus once it is folded it will remain folded and fluorescent even if its fusion partner becomes denatured in the presence of urea [24]. The fact that more of the material in the clone B6 pellet is fluorescent could indicate that initially more of the protein reached a properly folded state, leading to more properly folded GFP. Over time, some of this material could have aggregated and ended up in an inclusion body, with no effect on the structure of GFP. On the other hand, the pellet for the wild type did not have as much folded GFP in it, indicating that during expression it was not folding as well. Therefore, even though both

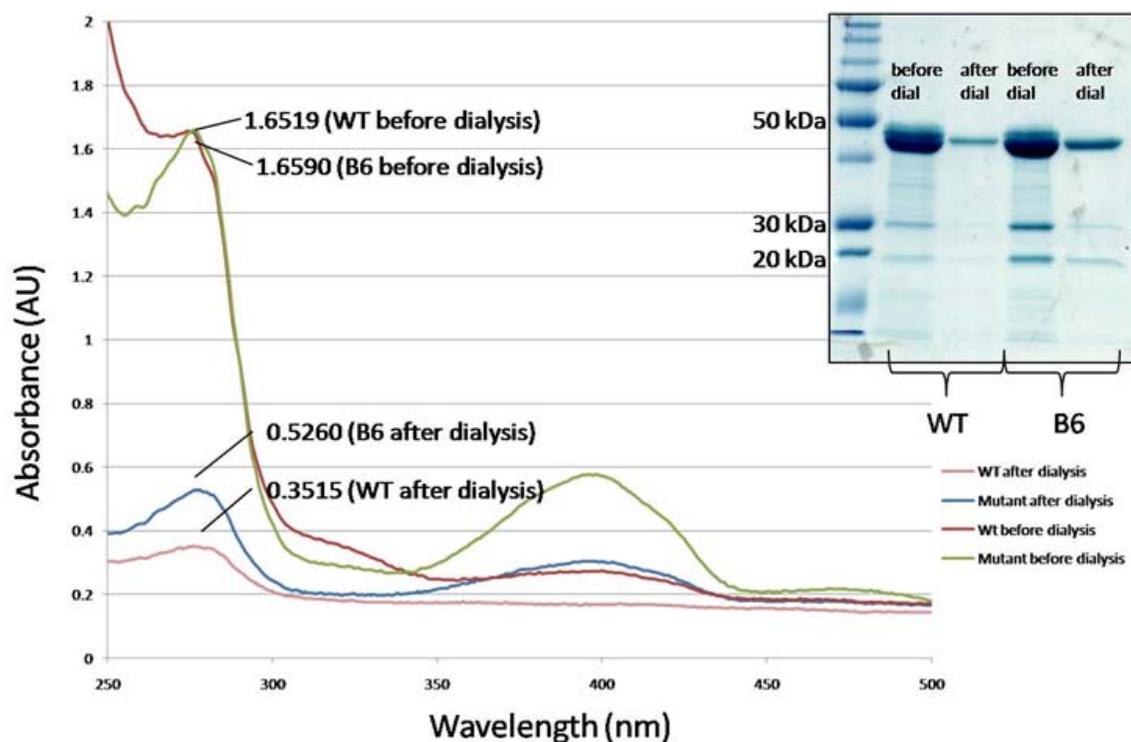
samples contain the same amount of protein, more of the clone B6 inclusion body material was derived from protein that was initially folded correctly and fluorescent, but the wild type inclusion body is mostly derived from protein that never reached a properly folded state and is thus less fluorescent. Thus fluorescence alone is not a sufficient indicator of protein content once that protein has been deposited in an inclusion body.



**Figure 3-19.** Equalized concentrations of denatured wild type and clone B6 fusions (A) and their relative fluorescence (B).

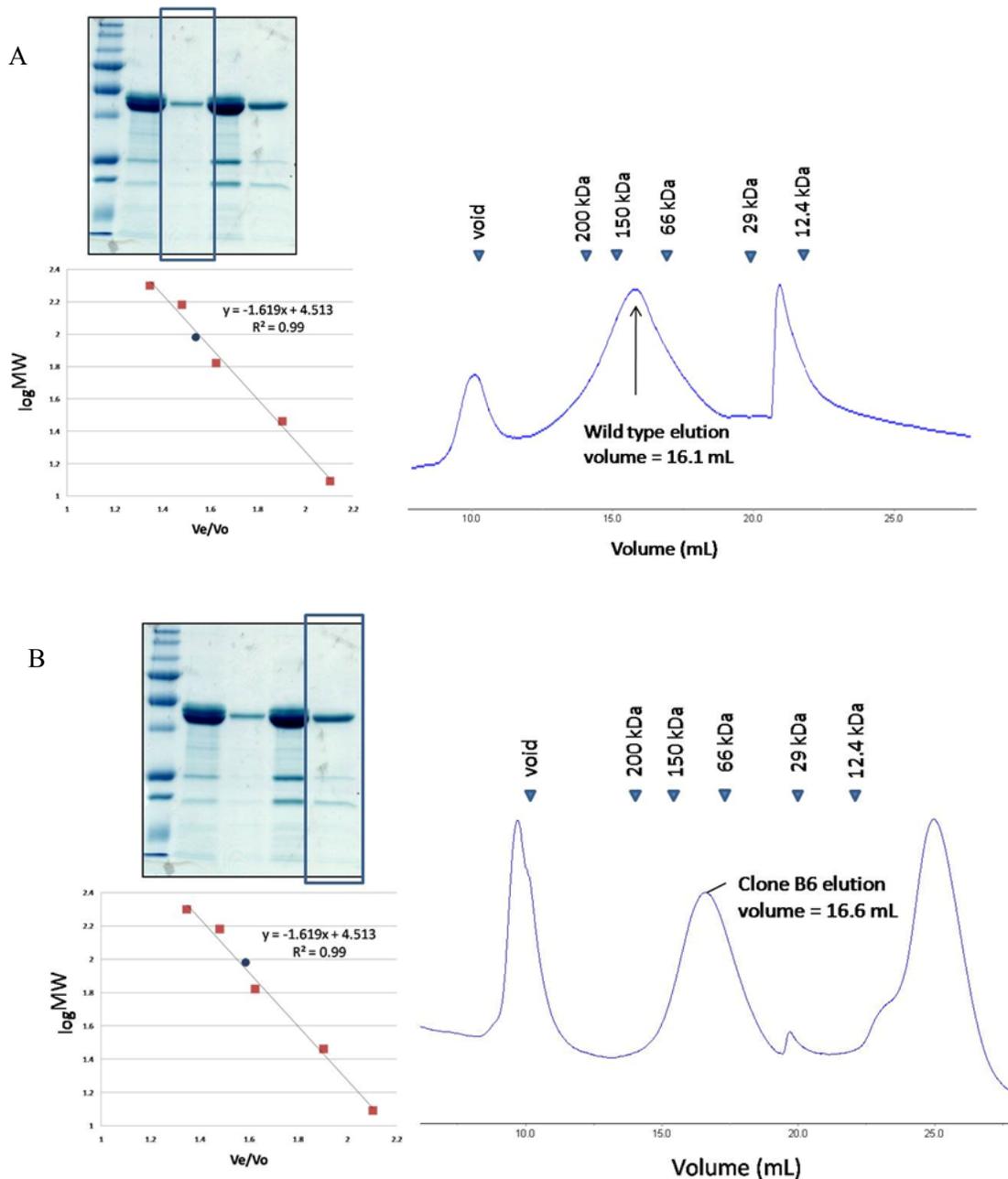
The two samples of equimolar concentration were then refolded via dialysis under identical conditions. The precipitates after dialysis appeared to have the same properties as the

precipitates generated from the refolding of the denatured soluble material, with the wild type appearing cloudy and clone B6 appearing flaky (**Figure 3-17**). **Figure 3-20** shows the absorbance spectra of both samples after dialysis. Again, clone B6 produced more refolded, soluble material as is evident both from the absorbance scans and on the SDS-PAGE gel. Taking baselines into account and using the PhnG and GFPuv combined extinction coefficients of  $41830 \text{ M}^{-1}\text{cm}^{-1}$  and  $20600 \text{ M}^{-1}\text{cm}^{-1}$  respectively, totaling  $62430 \text{ M}^{-1}\text{cm}^{-1}$ , the concentrations of the wild type and clone B6 samples prior to dialysis were both  $23 \mu\text{M}$ . The concentration of the refolded material after dialysis was  $3.2 \mu\text{M}$  for wild type and  $5.7 \mu\text{M}$  for clone B6. Thus wild type PhnG-GFP had a refolding yield of 14%, and clone B6 PhnG-GFP had a refolding yield of 25%. This refolding experiment was repeated twice more, with clone B6 producing more refolded material in each case.



**Figure 3-20.** Absorbance spectra and SDS-PAGE gels of wild type PhnG-GFP and clone B6 PhnG-GFP resolubilized inclusion bodies pre-and post-dialysis.

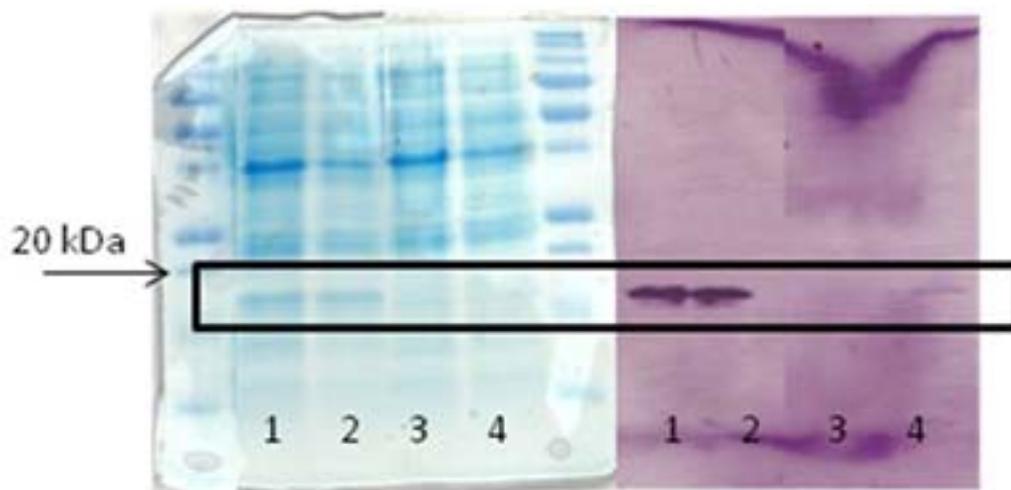
The samples refolded from resolubilized inclusion bodies were much more pure than samples obtained after native protein purified via IMAC (**Figure 3-13**) and also more pure than the samples refolded from the denatured soluble fractions (**Figure 3-16**), thus both samples were analyzed by size exclusion chromatography to ensure that the refolded material retained the same oligomeric state as the samples purified under native conditions. **Figure 3-21 A** shows the size exclusion chromatogram for the wild type refolded material, and **Figure 3-21 B** shows the chromatogram for the clone B6 refolded material. The wild type refolded fusion eluted at 16.1 mL, corresponding to a molecular weight of 104 kDa, whereas the clone B6 fusion eluted at 16.6 mL, corresponding to a molecular weight of 87.1 kDa. Using these elution volumes and the log of the dimeric molecular weight, both samples were plotted as blue circles on the calibration curves shown on each chromatogram. These elution volumes were similar to those determined for the native protein purified with buffers containing DDM and DTT (**Figure 3-14**) indicating that the refolded structure retains the same oligomeric state as the native structure.



**Figure 3-21.** Size exclusion chromatograms of refolded PhnG wild type (A) and clone B6 (B) fusions. The elution volume of the calibration standards is indicated along the top. The standards used were blue dextran (2000 kDa),  $\beta$ -amylase (200 kDa), alcohol dehydrogenase (150 kDa), bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), and cytochrome c (12.4 kDa). The calibration curve based on the elution volumes of the standard proteins is shown in inset, with the point representing the elution of PhnG based on the dimeric weight of 95.4 kDa indicated by the blue circle.

### 3.3.7 Expression of PhnG wild type and clone B6 as non-fusions

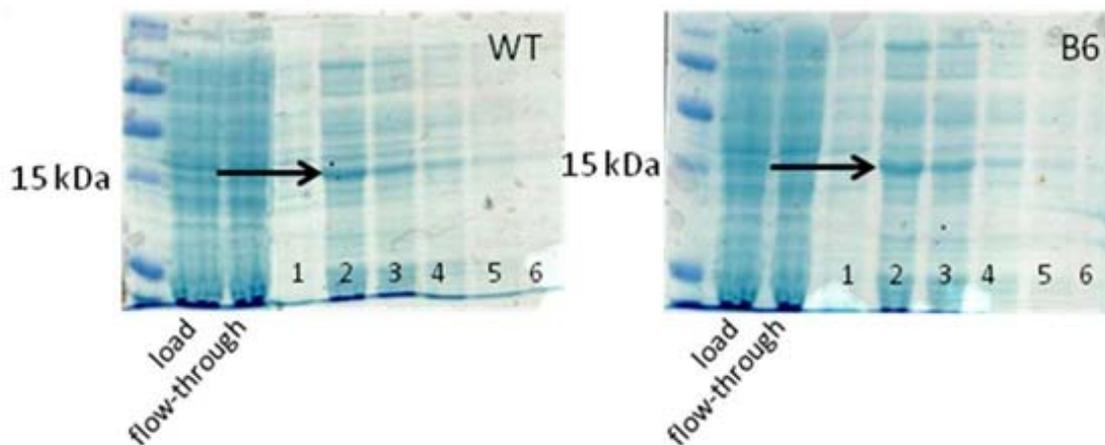
To determine solubility levels of both wild type PhnG and clone B6 without the GFP tag, the genes for each were sub-cloned into another pProEx vector which did not contain the gene for GFPuv. Using these plasmids, both were expressed under identical conditions as for the fusions and purified via IMAC. Initial purification under standard conditions gave the same results as with the fusions (**Figure 3-11**)—some material bound to the column for the wild type protein and almost nothing for clone B6 (**Figure 3-22**). Because the bands for both proteins were so faint on the SDS-PAGE gel, the presence of both wild type and clone B6 proteins was confirmed by western blot. However, as is consistent with previous observations of an inaccessible his-tag for clone B6, only a faint band was visible for this clone on the western blot (**Figure 3-22**, lane 4 on blot).



**Figure 3-22.** SDS-PAGE and western blot analysis of wild type PhnG and clone B6 as non-fusions. Lanes 1 and 2 on the SDS-PAGE gel and blot represent wild type fractions obtained after standard IMAC purification, and lanes 3 and 4 on both represent clone B6.

In a second attempt to capture the proteins as non-fusions, DDM and DTT were used in the purification buffers in hopes of being able to aid more native soluble material bind to the Ni-

NTA column. The detergent did improve the binding of clone B6 slightly, as is consistent with the result obtained for the clone B6-GFP fusion. **Figure 3-23** shows the fractions collected from both purifications. Compared to the fractions from the fusion purifications under the same conditions (**Figure 3-13**), the non-fusion fractions appear to not only have less material, but also to be less pure. Based on these results it can be said that GFP also acts as a solubility-enhancing tag by keeping more of the protein in the soluble phase, perhaps by partially blocking hydrophobic portions on the surface of PhnG and inhibiting aggregation. This observation has been made previously in the literature for both N- and C-terminal GFP fusions [25, 26]. The expression and purification of these two variants were repeated twice more, and from all three expressions the material recovered is shown in **Figure 3-23**. Because of this it was concluded that in order to sufficient material for biochemical or structural analysis, a PhnG-GFP fusion should be used, specifically one that has been refolded from a clone B6 inclusion body as this was the method that produced the most material pure enough for crystallographic studies.



**Figure 3-23.** SDS-PAGE gels of fractions collected from IMAC purification of PhnG wild type and clone B6 as non-fusions. The mass of PhnG is 16.5 kDa and is indicated by the black arrows.

### 3.4 Conclusions

PhnG, a member of the carbon-phosphorus lyase pathway in *Escherichia coli* was chosen as a target to undergo diversification and selection via the GFP folding reporter [18] in an effort to improve its expression, folding and solubility in *E. coli*. The first round of evolution using error-prone PCR successfully discovered 33 mutant-fusions with fluorescence greater than the wild type fusion. Sequencing of some of the mutants revealed highly selected residues, namely Val47, Ala25 ad Leu67.

The brightness of the clones was not visibly increasing from the second to third rounds, thus back-crossing in the presence of  $40 \times$  excess of wild type *phnG* was performed after the third round. One clone, labeled B6, isolated in the back-crossing round was shown to have the most dramatic increase in fluorescence ( $6.8 \times$ ) compared to any other clone found and was chosen to undergo folding and solubility analysis. This clone contained two silent mutations and the mutations V47A, A70T, and D120G. The D120G mutation was unique to only this clone. The next brightest clone, found in round 3, was only  $2.2 \times$  as bright as wild type and did not have any identical mutations to clone B6

There was a smaller increase in the fluorescence between the wild type and clone B6 lysate supernatants ( $2 \times$ ), and purification of clone B6 proved to be difficult. To capture any of the clone B6 on a Ni-NTA column, purification buffers needed to be supplemented with 0.1 mM DDM and 2 mM DTT. Even with detergent present, less material was captured for clone B6 than for wild type despite its higher fluorescence. It was observed that most of the material did not bind to the column, as the flow-through material was still highly fluorescent. Size exclusion chromatography on the wild type and clone B6 fusions revealed that both eluted at a volume corresponding to dimers. It is unclear whether the dimerization is a result of dimers forming between GFP or between PhnG.

The inability for clone B6 to bind to the Ni-NTA purification column was concluded to be an issue with the accessibility of the his-tag due to burial within the structure of PhnG. The inaccessibility of the tag was more severe for clone B6, indicating perhaps it has a tighter fold. This was confirmed by denaturing the proteins and purifying them under denaturing conditions. Only after being fully denatured could more clone B6 be captured compared to wild type.

Refolding experiments were performed to determine if an improvement in folding had been made for clone B6. Dialysis of equal concentrations of the wild type and clone B6 fusions under identical conditions resulted in refolding yields of 25 % for clone B6 and 14% for wild type. Another interesting observation arising from dialysis was the appearance of the precipitates that formed. The precipitate of the wild type fusion had a cloudy and globular appearance whereas the clone B6 precipitate appeared as distinct flakes. It was also observed that upon resolubilization and purification of the inclusion bodies of both wild type and clone B6 fusions, the clone B6 fusion had more material present. Size exclusion chromatography of the refolded material gave similar results to the size exclusion results of the native material—both fusions eluted as dimers. Interestingly, in both the native and refolded cases, clone B6 eluted at a slightly larger volume (smaller molecular weight) which supports the theory that the structure has tightened up.

In conclusion, the higher fluorescence of clone B6 accurately indicated that more protein was in the soluble phase compared to wild type for this fusion. This increase in soluble material is probably due to both an improvement in expression, as there was more material in both the soluble and insoluble phases, but also do to an improvement in folding as was demonstrated by refolding studies. Unfortunately, due to the tertiary structure of PhnG, the his-tag is not accessible and thus only a small amount of native material in the soluble phase could be captured via IMAC chromatography. The his-tag for clone B6 was even less accessible than for wild type, making it much more difficult to purify. It was also discovered that GFP acts as a solubility-

enhancing tag for PhnG as it was shown that expression of both wild type and cone B6 without GFP leads to smaller yields of protein in the soluble phase when purified under standard conditions and in the presence of detergent and DTT. Therefore, in terms of acquiring material for activity assays and crystallographic studies, resolubilizing and refolding PhnG-GFP inclusion bodies would yield the most material of high purity.

### 3.5 References

1. Yakovleva GM, Kim S-, Wanner BL: **Phosphate-independent expression of the carbon-phosphorus lyase activity of *Escherichia coli***. *Appl Microbiol Biotechnol* 1998, **49**:573-578.
2. Quinn JP, Kulakova AN, Cooley NA, McGrath JW: **New ways to break an old bond: the bacterial carbon-phosphorus lyases and their role in biogeochemical phosphorus cycling**. *Environ Microbiol* 2007, **9**:2392-2400.
3. Schowanek D, Verstraete W: **Phosphonate Utilization by Bacterial Cultures and Enrichments from Environmental Samples**. *Appl Environ Microbiol* 1990, **56**:895-903.
4. McGrath GW, Wisdom GB, McMullan G, Larkin MJ, Quinn JP: **The purification and properties of phosphonoacetate hydrolase. A novel carbon-phosphorus bond-cleavage enzyme from *Pseudomonas fluorescens* 23F**. *Eur J Biochem* 1995, **234**:225-230.
5. Morais MC, Zhang G, Zhang, W. Olsen, D. B., Dunaway-Mariano D, Allen KN: **X-Ray Crystallographic and Site-directed Mutagenesis Analysis of the Mechanism of Schiff-base Formation in Phosphonoacetaldehyde Hydrolase Catalysis**. *J Biol Chem* 2004, **279**:9353-9361.
6. Metcalf WW, Wanner BL: **Mutational Analysis of an *Escherichia coli* Fourteen-Gene Operon for Phosphonate Degredation, Using *TnphoA*' Elements**. *J Bacteriol* 1993, **175**:3430-3442.
7. Adams MA, Luo Y, Hove-Jensen B, He S-, van Staalduinen LM, Zechel DL, Jia Z: **Crystal Structure of PhnH: an Essential Component of Carbon-Phosphorus Lyase in *Escherichia coli***. *J Bacteriol* 2008, **190**:1072-1083.
8. Kononova SV, Nesmeyanova MA: **Phosphonates and Their Degradation by Microorganisms**. *Biochemistry (Moscow)* 2002, **67**:220-233.
9. Podzelinska K, He S-, Wathier M, Yakunin A, Proudfoot M, Hove-Jensen B, Zechel DL, Jia Z: **Structure of PhnP, a Phosphodiesterase of the Carbon-Phosphorus Lyase Pathway for Phosphonate Degredation**. *J Biol Chem* 2009, **284**:17216-17226.
10. Kawasaki M, Inagaki F: **Random PCR-based Screenig for Soluble Domains Using Green Fluorescent Protein**. *Biochem Biophys Res Commun* 2001, **280**:842-844.
11. Caldwell RC, Joyce GF: **Randomization of genes by PCR mutagenesis**. *PCR Methods Appl* 1992, **2**:28-33.
12. Weissensteiner T, Griffin HG, Griffin A: **PCR Technology Current Innovations**. 2004, :392.

13. Sambrook J, Russell DW: **Quantitation of Nucleic Acids.** In *Molecular Cloning A Laboratory Manual. Volume 3.* Edited by a. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2001:A8.19-A8.21.
14. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **Protein Identification and Analysis Tools on the ExPASy Server.** In *The Proteomics Protocols Handbook.* Edited by Walker JM. New York, NY: Humana Press; 2005:571-607.
15. Fukuda H, Arai M, Kuwajima A: **Folding of Green Fluorescent Protein and the Cycle3 Mutant.** *Biochemistry* 2000, **30**:12025-12032.
16. Bershtein S, Goldin K, Tawfik DS: **Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins.** *J Mol Biol* 2008, **379**:1029-1044.
17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**:2947-2948.
18. Waldo GS, Standish BM, Berendzen J, Terwilliger TC: **Rapid protein-folding assay using green fluorescent protein.** *Nat Biotechnol* 1999, **17**:691-695.
19. Hedhammar M, Stenvall M, Lonneborg R, Nord O, Sjolín O, Brismar H, Uhlen M, Ottosson J, Hober S: **A novel flow cytometry-based method for analysis of expression levels in Escherichia coli, giving information about precipitated and soluble protein.** *J Biotechnol* 2005, **119**:133-146.
20. Jana S, Deb JK: **Strategies for efficient production of heterologous proteins in Escherichia coli.** *Appl Microbiol Biotechnol* 2005, **67**:289-298.
21. Kneusel RE, Crowe J, Wulbeck M, Ribbe J: **Procedures for the Analysis and Purification of His-Tagged Proteins.** In *The Nucleic Acid Protocols Handbook.* Edited by Rapley R. Totowa, NJ: Humana Press; 2000:931.
22. Palm GJ, Zdanov A, Gaitanaris GA, Stauber R, Pavlakis GN, Wlodawer A: **The structural basis for spectral variations in green fluorescent protein.** *Nat Struct Biol* 1997, **4**:361-365.
23. **Properties of GFP and GFP Variants.** In *Living Colors Manual.* Clontech Laboratories, Inc.; 2001:8.
24. Alkaabi KM, Yafea A, Ashraf SS: **Effect of PH on the thermal- and chemical-induced denaturation of GFP.** *Appl Biochem Biotechnol* 2005, **126**:149-156.
25. Wu C-, Cha HJ, Rao G, Valdes JJ, Bentley WE: **A green fluorescent protein fusion strategy for monitoring the expression, cellular location, and separation of biologically active organophosphorus hydrolase.** *Appl Microbiol Biotechnol* 2000, **54**:78-83.

26. Japrun D, Chusacultanaichai S, Yuvaniyama J, Wilairat P, Yuthavong Y: **A simple dual selection for functionally active mutants of *Plasmodium falciparum* dihydrofolate reductase with improved solubility.** *Protein Eng Des Sel* 2005, **18**:457-464.

## Chapter 4

### Conclusions

Directed evolution using error-prone PCR and DNA shuffling (for diversification), and the GFP folding reporter (for screening) was used in an attempt to improve the folding and solubility of three diverse target proteins—RebG, 5LO, and PhnG

Screening of libraries proved to be a difficult task for 5LO and RebG. Four attempts were made for each target to identify improved clones (initial attempts at round 1 and round 2, and round 1 repeated two more times) but all clones found were the result of some sort of contamination. The fluorescent clones either contained plasmids with small inserts rather than the correct full-length insert, or contained a mutant vector in which a portion of the *lacI*<sup>f</sup> gene has replaced the cloning region. It was also discovered that many of the streaks were in fact mixtures of cells containing plasmids with full length inserts and other cells containing plasmids with short inserts. These mixtures were likely a result of two colonies being picked at one time during the screening process or the presence of two plasmids in one cell which is often encountered when transforming cells with large amounts of DNA, such as in library generation.

After the several attempts at evolving both 5LO and RebG, which involved screening of 30-50 colonies per attempt, no viable clones could be found with improved fluorescence. These proteins may need to be screened via FACS so that larger libraries can be searched, or perhaps diversified with other methods that allow greater diversity to be achieved (perhaps by a method that allows consecutive nucleotide changes).

The third target, PhnG, was more successful. One clone, labeled B6, isolated in the back-crossing round was shown to have the most dramatic increase in fluorescence (6.8 ×) compared to any other clone found and was chosen to undergo folding and solubility analysis. Several

differences became apparent between the clone B6 and wild type PhnG-GFP fusions, namely appearance of their respective precipitates after refolding, and a greater difficulty to purify clone B6 via IMAC purification. The inability for clone B6 to bind to the Ni-NTA purification column was concluded to be an issue with the accessibility of the his-tag due to burial within the structure of PhnG. The inaccessibility of the tag was more severe for clone B6, indicating perhaps it has a tighter fold. This was confirmed by denaturing the proteins and purifying them under denaturing conditions. Only after being fully denatured could more clone B6 be captured compared to wild type.

Refolding experiments were performed to determine if an improvement in folding had been made for clone B6. Dialysis of equal concentrations of the wild type and clone B6 fusions under identical conditions resulted in refolding yields of 25 % for clone B6 and 14% for wild type. Size exclusion chromatography of the native and refolded fusions gave similar results—all eluted as dimers

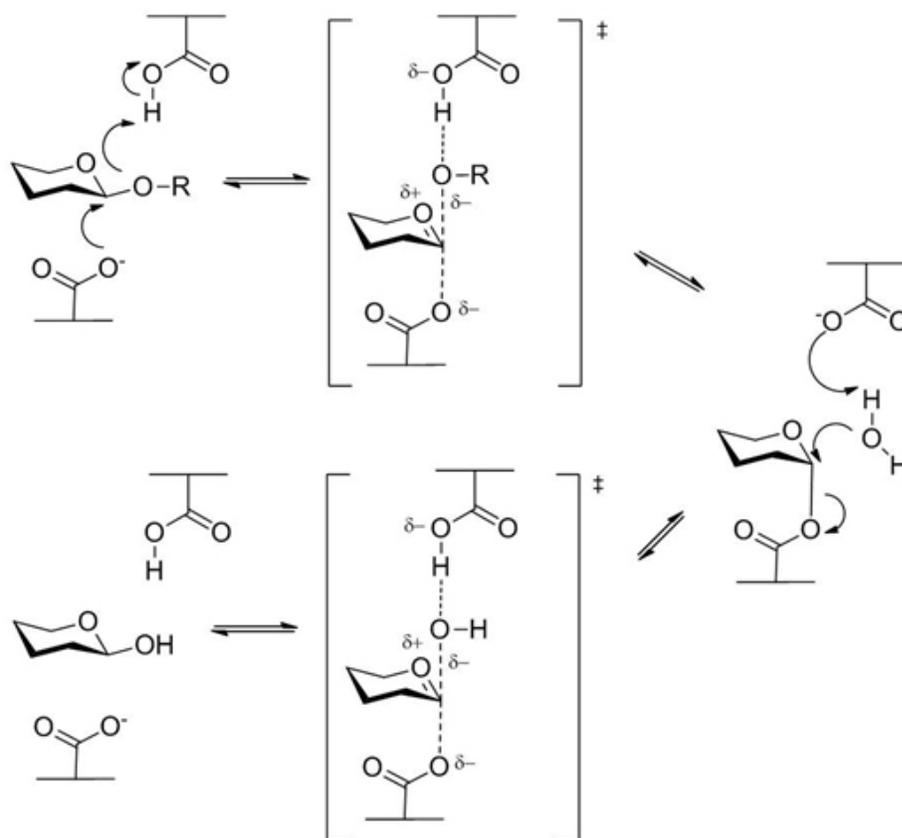
The higher fluorescence of clone B6 accurately indicated that more protein was in the soluble phase compared to wild type for this fusion. This increase in soluble material is due to both an improvement in expression, as there was more material in both the soluble and insoluble phases, but also do to an improvement in folding as was demonstrated by refolding studies. GFP was also discovered to be an effective solubility-enhancer for PhnG, as the wild type fusion exhibited more material in the soluble phase than had previously been seen for wild type PhnG when expressed alone.

## Appendix A

# Mutagenesis and Kinetic analysis of *TmGH1* in Preparation for Atomic Force Microscopy

### A.1 Introduction

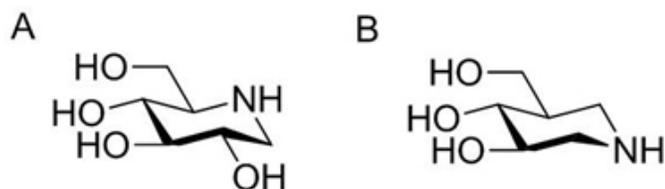
*Thermotoga maritima*  $\beta$ -glucosidase, herein called *TmGH1* is a member of the GH1 family of glycosyl hydrolases. Beta-glucosidases are encompassed by the important glycosyl hydrolase superfamily of enzymes which catalyze the hydrolysis of glycosidic bonds in a fashion that either retains or inverts the stereochemistry at the anomeric carbon [1]. Polysaccharides can have a vast variety of conformations with varying stereochemistry, thus to specifically target one glycosidic bond, an equally wide variety of glycosidases are required, as is evident by the 110 families that make up this superfamily [2]. *TmGH1* and other GH-1  $\beta$ -glycosidases cleave glycosidic bonds in a two-step process which results in the retention of the anomeric configuration (**Figure 4-1**). Firstly, an enzymatic general acid/base (Glu166 in *TmGH1*) provides the proton which assists in the departure of the leaving group, while an enzymatic nucleophile (Glu351 in *TmGH1*) attacks the anomeric carbon to form a covalent glycosyl-enzyme linkage. This linkage is then hydrolyzed via nucleophilic attack from a water molecule and results in a product with the same stereochemistry as the starting material [1, 3].



**Figure A-1.** Mechanism of *TmGH1*. Adopted from [3].

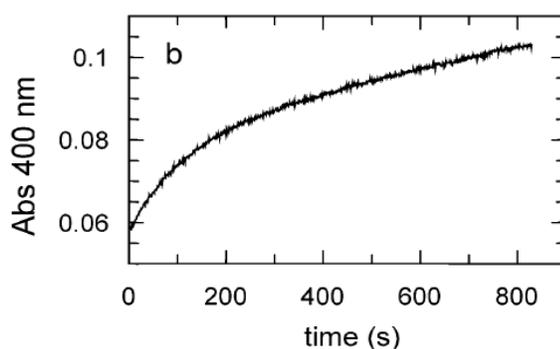
Inhibiting the action of glycosidases has become the goal of many drug therapies associated with major diseases such as cancer, HIV, and hepatitis B and C [4]. One well-studied class of inhibitors of high efficacy are iminosugars, such as 1-deoxynojirimycin and isofagomine (**Figure 4-2**) which were initially thought to be transition state mimics, however further investigation revealed that this may not be the case. This was concluded on the basis that isofagomine was deemed a more potent inhibitor due to more favourable entropy of binding than 1-deoxynojirimycin [5]. Entropic factors are known to influence transition state binding to a considerably lesser extent than enthalpic factors [20], and the favourable enthalpies of binding for both inhibitors were similar. Insights into the more favourable and dominant entropy of binding for isofagomine were revealed from the crystal structures of each of the inhibitors bound to the

enzyme. Isofagomine was shown to have a less constrained conformation in the active site as well as the absence of ordered water molecules [5].



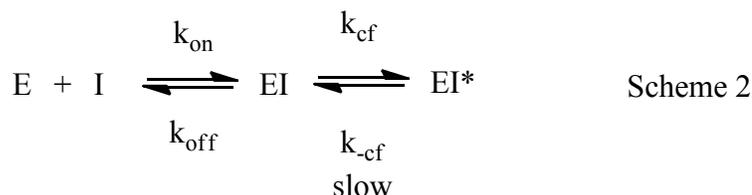
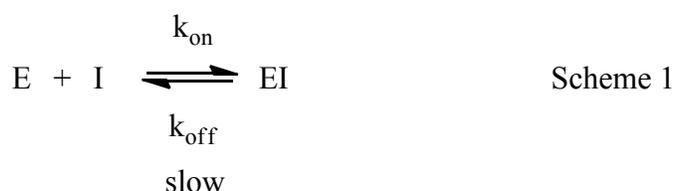
**Figure A-2.** Structure of  $\beta$ -glucosidase inhibitors 1-deoxynojirimycin (A) and isofagomine (B).

An interesting observation from the study of these inhibitors was the “slow-onset” binding that was exhibited by isofagomine. Slow-onset inhibition is exhibited by non-linear steady-state rates in the presence of inhibitor, indicating that the inhibitor does not reach full potency until after a certain length of time (around 200 s in **Figure 4-3**). This was observed regardless of inhibitor concentration, thus the slow-onset is not a result of insufficient amounts of inhibitor. Slow-onset inhibition is also observed with a sweet almond  $\beta$ -glucosidase [6] and also with nojirimycin, the parent compound of 1-deoxynojirimycin. Interestingly, nojirimycin also has a potency 50 times that of 1-deoxynojirimycin.



**Figure A-3.** Slow-onset inhibition of *TmGH1* by isofagomine. Copied with permission from [5].

Two schools of thought prevail on the mechanism behind slow-onset inhibition. For the sweet almond case it has been thought to arise from slow association between the inhibitor and enzyme (Scheme 1) [6], however others have suggested that a conformational change is taking place by the enzyme, initiated by the binding of the inhibitor [7]. With the latter case it is hypothesized once bound, the enzyme slowly rearranges itself around the inhibitor to form a ‘high affinity complex’ (EI\*) (Scheme 2) [8]. If a conformational change is taking place in which the enzyme is “tightening” around an inhibitor, it is the hope that the force required to unfold *TmGH1* in the place of the inhibitor would be measurably different from the force required to pull it apart with no inhibitor present. These forces can be observed by the use of single molecule force spectroscopy (SMFS).



Studies examining single molecule unfolding and refolding have been extensively reviewed several times [9-11]. Atomic force microscopy (AFM) has proven a valuable technique in which to study unfolding events at a single-molecule level, and the unfolding pathways for several proteins such as titin [12] and T4 lysozyme have been examined. Unfolding data obtained from these experiments has given important insights into stabilizing interactions with a protein’s structure.

In brief, to study singular proteins via AFM, a pure protein is covalently bound to a surface coated with gold or other specialized substrate. The cantilever of the AFM is brought down into contact with the protein-coated surface which causes the protein to adsorb to the tip of the cantilever. The surface is then pulled away from the cantilever tip with sub-nanometer precision, and if a protein is adsorbed to the tip it will resist separation between the tip and the surface. This resistant force is measured via deflection of the cantilever tip [10]. Proteins with a high degree of resistance to unfolding will result in a larger force required for separation. Thus, in terms of *TmGH1* it would be expected that the force required to pull the protein apart with no inhibitor bound would be measurably less than the force required with inhibitor present, if the binding induces a conformational change. It is hypothesized that if a conformational change to a ‘high-affinity complex’ is taking place, then this complex would require more force to become completely denatured.

Slow onset inhibition has also been studied by measuring changes in tryptophan fluorescence. Pandhare *et al.* showed that binding of proteinaceous alkaline protease inhibitor (API) to proteinase K caused a conformational change resulting in major changes in tryptophanyl fluorescence emission maxima [13]. The fluorescence data was correlated to kinetic data which together presented a strong case predicting the enzyme-inhibitor complex (EI) isomerized to a new enzyme-inhibitor complex (EI\*), which after another length of time underwent a conformational change to a “conformationally locked” complex (EI\*\*) from which the inhibitor could not dissociate.

Ligand-protein and inhibitor-protein interactions have also been studied with AFM by immobilizing the ligand/inhibitor of interest to the cantilever tip and the protein to the substrate. The tip bearing the ligand is brought down to the protein coated surface and allowed to bind. When the cantilever is brought up, the ligand’s unbinding force can be measured. Using this

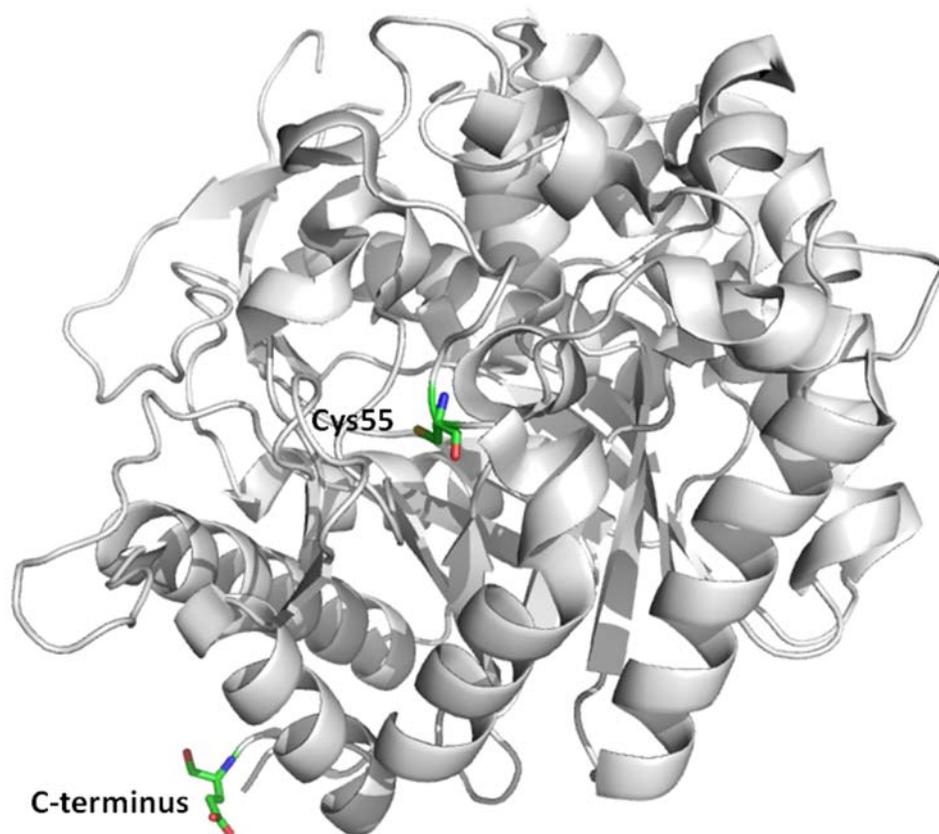
method, Schwesinger *et al.* revealed that a ligand's unbinding force is directly proportional to the negative logarithm of its "off rate" [14].

An alternate method involving covalently attaching each end of a protein to both the AFM cantilever tip and substrate via cysteine residues has also been successfully used to examine inhibitor binding. Wang *et al.* used AFM to detect conformational changes of bovine carbonic anhydrase B in the presence of the inhibitor *p*-aminomethylbenzene sulfonamide [15]. They covalently attached one end of the protein to an amino functionalized silicon surface and the other end to an AFM cantilever tip, and measured the stretching length required to disrupt the covalent linkage. It was found that the native monomer could be stretched to 13 nm under physiological conditions before disruption of the covalent networking, however in the presence of inhibitor this length increased to 28 nm. This increased length was attributed to the inhibitor increasing the "stretchable parts of the enzyme". They hypothesized that inhibitor binding decreased the lability of the central region of the enzyme but increased the thermal factor on the peripheral regions. It is this latter point that would lead to a longer stretching distance.

In another example, Kedrov and coworkers reported AFM-detected structural changes that occurred in the antiporter NhaA upon inhibitor binding [16]. They showed that the binding of the inhibitor 2-aminoperimidine caused the enzyme to enter a 'new energy minimum' resulting from a stabilized  $\alpha$ -helix with reduced conformational energy. The structural flexibility of this enzyme is important for its activity, thus the lower conformational energy is thought to aid in the inhibitory action.

This chapter will describe the preliminary stages of this project which is to be undertaken in collaboration with Dr. Hugh Horton in the Department of Chemistry at Queen's University. The preliminary stages involved preparation of *TmGH1* to incorporate a Cys residue at its C-terminus to aid in binding the protein to the substrate, and also to eliminate the only other Cys

residue in the interior of the protein (Cys55) so that the protein only binds at its C-terminus (**Figure 4-4**). Kinetic analysis of the mutant and wild type enzymes to verify that the mutagenesis did not alter the activity of the double Cys mutant compared to wild type is also discussed.



**Figure A-4.** Crystal Structure of *TmGH1*. The C-terminus where an additional cysteine was added (not shown) is highlighted, as well as the internal Cys55 which was mutated to Ser.

## **A.2 Experimental Procedures and Methods**

### **A.2.1 Materials**

Oligonucleotides used for PCR were synthesized by Sigma-Genosys. Vent polymerase and mixed dNTPs were purchased from New England Biolabs Canada. PfuTurbo polymerase was purchased from Stratagene. Restriction enzymes, T4 DNA ligase, and calf intestinal alkaline phosphatase were purchased from Fermentas. DMSO, Tris Base, Tris- NaCl, 2-nitrophenyl- $\beta$ -D-glucopyranoside, imidazole, sodium phosphate, and kanamycin were purchased from Sigma Aldrich (Oakville, Ontario and US). IPTG was purchased from Invitrogen. Nucleospin plasmid purification kits (Macherey-Nagel) were ordered from MJS Biolynx, Inc. All other DNA purification kits were purchased from QIAGEN, as well as Ni-NTA resin. All cells (XL1-Blue and BL21-DE3) were purchased from Stratagene (supplied by VWR, Canada). Fisher Biosciences Canada supplied all media (Luria Bertani broth, Luria Bertani agar, glucose). The pET\_28a cloning vector was purchased from Novagen. All sequencing reactions were performed by Robarts Research Institute (London, Ontario).

### A.2.2 Mutagenesis of wild type *TmGH1*

*Thermotoga maritima*  $\beta$ -glucosidase, *TmGH1* (GenBank accession code CAA52276) was mutated to replace the internal Cys residue at position 55 with a Ser residue and insert an additional Cys residue at the C-terminus of the protein. Four primers were used to accomplish both mutations in 3 sequential PCRs (PCR1, PCR2 and PCR3): Forward flanking Primer A: 5'-CAGCCATATGGCTAGCAACGTGAAAAAG-3' incorporated an *NheI* restriction site (underlined), internal Primer B: 5'-GTTGTAGTGGT**TCGGAGGCCACATCT**-3' and internal Primer C: 5'-GAGATGTGGCCTCC**GACCACTACAAC**-3', complementary primers which incorporated the TCG–TCC (Cys55Ser) mutation (in bold), and reverse flanking Primer D: 5'-GCCGCA**AAGCTTTTAGCAGTCTTCCAG**-3' which inserted the additional Cys residue (in bold) and a *HindIII* restriction site (underlined). All primers were designed to have  $T_m$  values within 3 °C of each other. The conditions for each PCR reaction are outlined in **Table 4-1**, with the template being a wild type pET\_28a\_ *TmGH1* clone (a generous gift from Professor Gideon Davies, York Structural Biology Laboratory, York, UK). PCR1 generated a 126 bp fragment, PCR2 a 1200 bp fragment, and the assembly PCR3 produced the final mutated *TmGH1* gene (*TmGH1*-Mut, 1360 bp). The *TmGH1*-Mut insert was digested at 37 °C for 1.5 hours with *NheI*/*HindIII* and ligated using T4 DNA ligase into a pET-28a vector containing a hexa-histidine sequence at the N-terminus and kanamycin resistance gene. Both DNA strands of *TmGH1* wild type and *TmGH1*-Mut genes were sequenced and the expected sequences were verified.

**Table A-1:** PCR conditions for the mutagenesis of *TmGH1*

Reaction	Primers	Template	Polymerase	Program
PCR1	A +B	<i>TmGH1</i> wild type in pET-28a	Pfu Turbo	30 × (cycles of 98 °C, 30 s; 65 °C, 30 s; 72 °C, 60 s), finish with 72 °C, 5 min.
PCR2	C + D	<i>TmGH1</i> wild type in pET-28a	Vent	Same as PCR1
PCR3	A +D	PCR product from PCRs 1 and 2	Vent	Same as PCR1

### A.2.3 Expression and purification of wild type *TmGH1* and *TmGH1*-Mut

The same protocol was followed for wild-type *TmGH1* and *TmGH1*-Mut. BL21-DE3 cells were transformed with plasmid DNA and grown overnight at 37 °C on solid LB-Agar media containing 1% glucose and 50 µg/mL kanamycin. A single colony was selected and grown overnight at 37 °C, 250 rpm in a 10mL culture comprising LB media, 1% glucose and 50 µg/mL kanamycin. 8 mL of this culture was used to inoculate an 800 mL culture which was grown under the same conditions and induced with 0.5 mM IPTG once an OD<sub>600</sub> of 0.6 was reached, and subsequently grown overnight. Cells were harvested by centrifugation at 3000 g for 20 minutes. Cell pellets were resuspended in buffer containing 50 mM Tris, 300 mM NaCl and 10 mM imidazole and lysed via cell rupture with an Emulsiflex cell homogenizer (Avestin Inc, Ottawa). The lysate was centrifuged at 40000 g for 30 minutes and the soluble fraction collected and placed in a heating bath at 75 °C for 20 minutes. All denatured material was removed via centrifugation at 15000 rpm for 20 min. The supernatant was purified further via immobilized metal ion affinity chromatography (IMAC) using nickel-NTA agarose resin. Fractions were analyzed by SDS-

PAGE to verify purity and concentrated using a Centricon centrifugal filter device (Millipore). The concentrated protein was then exchanged into sodium phosphate buffer (50 mM, pH 7) with a PD-10 desalting column (GE Healthcare) prior to kinetic analysis.

#### **A.2.4 Kinetic analysis of wild type *TmGH1* and *TmGH1*-Mut**

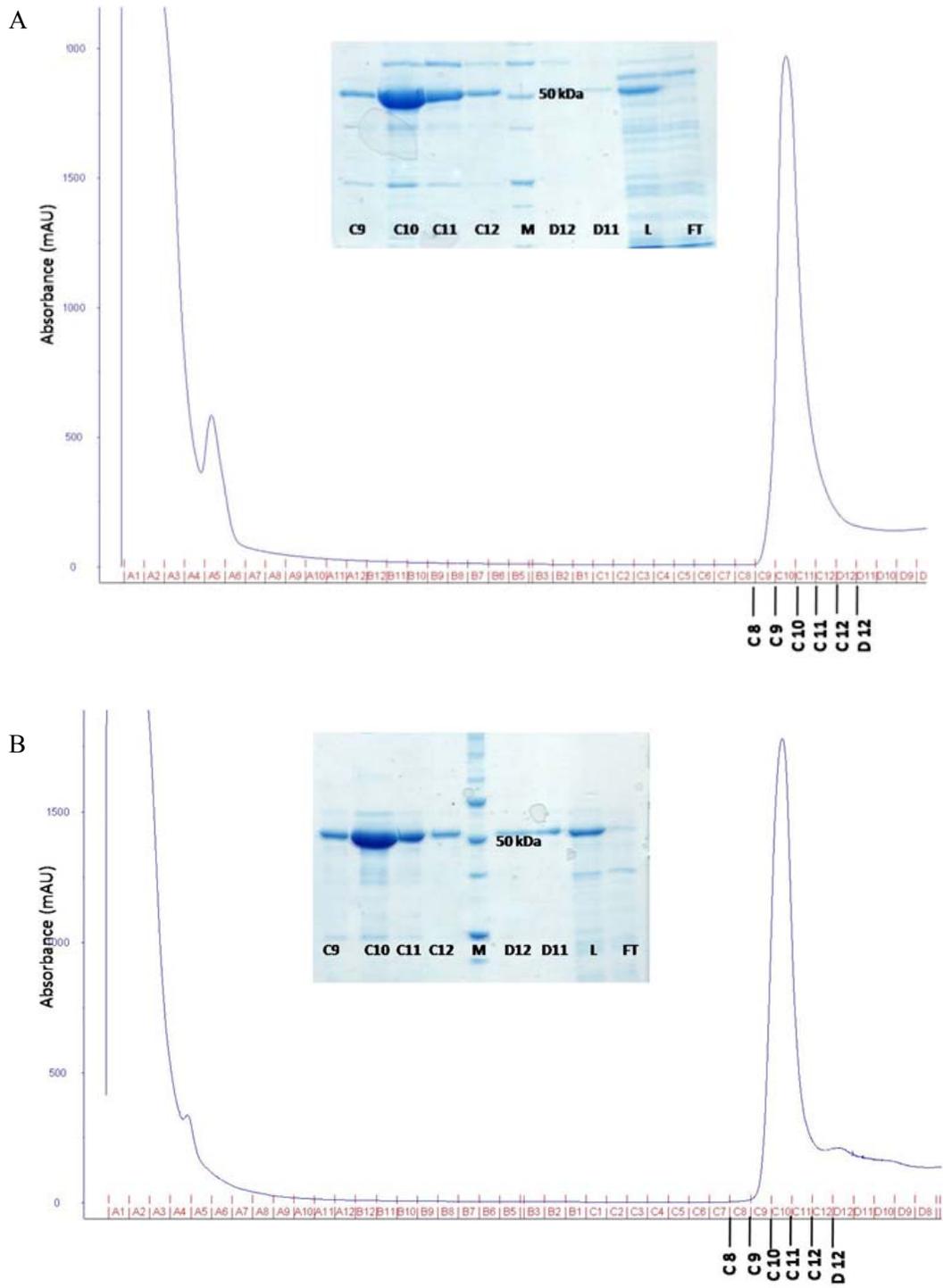
The same protocol was followed for both *TmGH1* wild-type and *TmGH1*-Mut. Activity was measured as the change in absorbance at 400 nm using 2-nitrophenyl- $\beta$ -D-glucopyranoside as a substrate. All reactions were performed in 50 mM sodium phosphate buffer, pH 7.0, containing 1 mg/mL BSA with a final reaction volume of 1 mL. Enzyme concentrations were determined by measuring absorbance at 280 nm and dividing by the calculated extinction coefficient of 121 240  $M^{-1} cm^{-1}$  [17]. Initial rates were calculated using enzyme concentrations of 35 nM and 92 nM for *TmGH1* wild-type and *TmGH1*-Mut respectively and were measured over substrate concentrations of 0.191 mM to 3.82 mM and 0.166 mM to 4.98 mM for *TmGH1* wild-type and *TmGH1*-Mut respectively. Initial rate data was plotted and curve-fitted to the Michaelis-Menten equation (**Equation 1**) using Grafit 6.0 (Erithacus Software Limited, UK). The  $k_{cat}$  value was determined using the reported extinction coefficient for 2-nitrophenyl- $\beta$ -D-glucopyranoside of 2170  $M^{-1} cm^{-1}$  [18].

$$V = V_{max}[S]/(K_M + [S]) \quad \text{Eq 1}$$

## **A.3 Results and Discussion**

### **A.3.1 Expression and purification of wild type *TmGH1* and *TmGH1*-Mut**

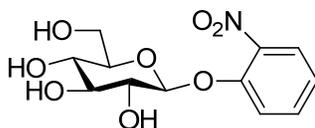
Both wild type and mutant *TmGH1* were expressed in soluble form in *E. coli*. The SDS-PAGE gel of the fractions obtained after IMAC purification of the wild type is shown in **Figure 4-5 A**, with the mutant fractions shown in **Figure 4-5 B**. The bands on the SDS-PAGE gels migrated in agreement with the predicted molecular weight of 53,957 Da. For both, Fractions C9, C10, and C11 were combined and concentrated to 1 mL, then diluted into 2.5 mL sodium phosphate buffer. This generated enzyme stock solutions of 7.1  $\mu\text{M}$  for the wild type enzyme and 9.2  $\mu\text{M}$  for *TmGH1*-Mut.



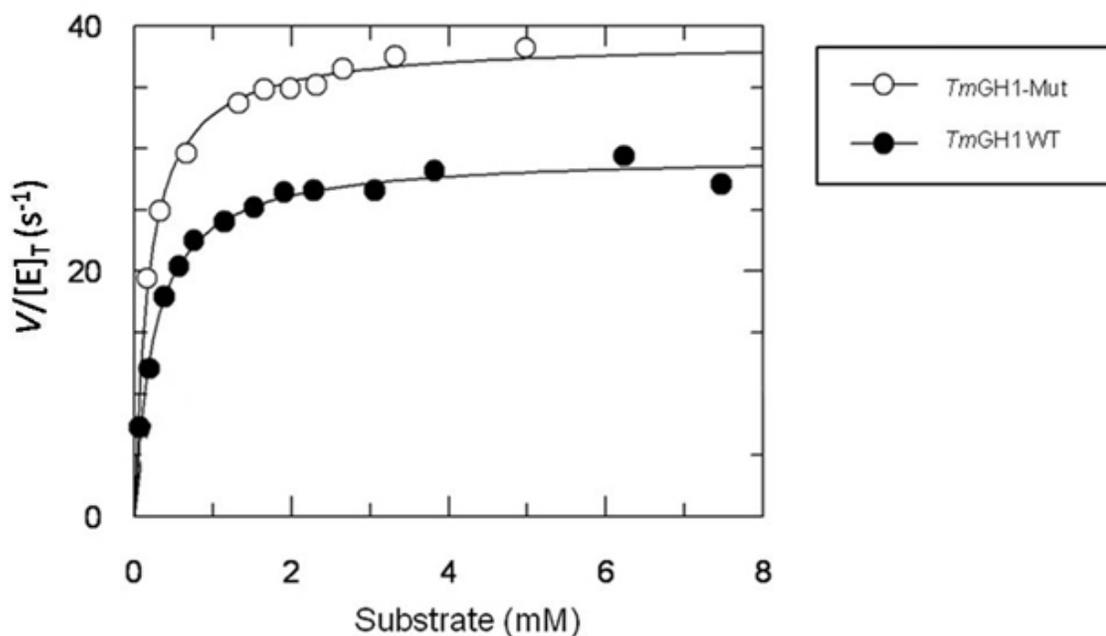
**Figure A-5.** Purification of wild type *TmGH1* (A) and *TmGH1*-Mut (B). Chromatograms show the absorbance measured by FPLC during IMAC purification. The fractions on the gel correspond to the fractions collected during purification, as shown on the chromatograms. In each case, fractions C9, C10, and C11 were collected and concentrated for further use.

### A.3.2 Kinetic analysis of wild type *TmGH1* and *TmGH1*-Mut

$K_M$  and  $k_{cat}/K_M$  values were determined by plotting substrate concentration against initial rate, using the substrate 2-nitrophenyl- $\beta$ -D-glucopyranoside (**Figure 4-6**). **Figure 4-7** shows substrate concentration plotted against rate over total enzyme concentration ( $V/[E]_T$ ) for *TmGH1* wild type ( $[E]_T = 71 \text{ nM}$ ) and *TmGH1*-Mut ( $[E]_T = 92 \text{ nM}$ ). Using half of the total amount of amount of wild type enzyme resulted in half of the  $V_{max}$ . For wild type,  $V_{max}/[E]_T$  was  $(6.3 \pm 0.07) \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$  for 71 nM enzyme and  $(6.5 \pm 0.1) \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$  for 35 nM enzyme, as expected for pseudo 1st order conditions where  $[S] \gg [E]_T$ , which is required for steady-state kinetic analysis. The values for the mutant were determined using a total enzyme concentration of 92 nM. Kinetic parameters obtained for both enzymes are summarized in **Table 4-2**. Previous kinetic data determined using 2,4-dinitrophenyl- $\beta$ -glucoside yielded a  $k_{cat}$  of  $42 \pm 1 \text{ s}^{-1}$  and  $K_M$  of 0.41 mM [5].



**Figure A-6.** Structure of 2-nitrophenyl- $\beta$ -D-glucopyranoside



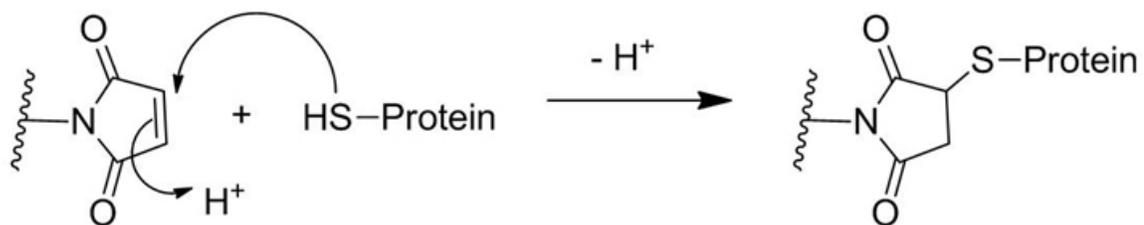
**Figure A-7.** Plots of 2-nitrophenyl- $\beta$ -D-glucopyranoside concentration vs. rate over total enzyme concentration for *TmGH1* wild type and *TmGH1*-Mut. Wild type data is represented by black circles and *TmGH1*-Mut by white circles.

**Table A-2.** Kinetic parameters obtained for *TmGH1* wild type and *TmGH1*-Mut

	$K_M$ (mM)	$k_{cat}$ ( $s^{-1}$ )	$k_{cat}/K_M$ ( $s^{-1} M^{-1}$ )
Wild type <i>TmGH1</i>	$0.21 \pm 0.02$	$30 \pm 0.3$	$(1.43 \pm 0.1) \times 10^5$
<i>TmGH1</i> -Mut	$0.18 \pm 0.01$	$39 \pm 0.4$	$(2.17 \pm 0.1) \times 10^5$

As preliminary tests to ensure the protein could bind to a modified PMMA surface, Geoff Nelson of Dr. Hugh Horton's group at Queen's University conducted AFM experiments to confirm the linkage. His data using the double Cys mutant was reported in detail in his MSc. thesis [19]. In his work he showed evidence that *TmGH1*-Mut could successfully bind to a poly (methylmethacrylate) (PMMA) surface modified with sulfo-EMCS ([N- $\epsilon$ -Maleimidocaproyloxy]sulfosuccinimide ester) linkers. The maleimide can react with SH groups to

form stable thioether linkages (**Figure 4-8**). This sets the stage for analyzing the unfolding behavior of TmGH1 in the presence of isofagomine, or inhibitor ‘pulling’ experiments, using AFM in a time resolved fashion.



**Figure A-8.** Reaction of maleimide with the thiol group of a cysteine residue to form a thioether linkage.

#### A.4 Conclusions

A double cysteine mutant of *TmGH1* was constructed for the purpose of analyzing it via AFM in an attempt to measurably detect any conformational changes that may occur upon inhibitor binding. The insertion of a cys residue at the C-terminus was required to covalently attach the protein to the AFM substrate, and the conversion of an internal cysteine to serine was done as a precautionary measure to ensure the protein only bound the surface at its C-terminus. The two mutations were accomplished via 4-primer PCR mutagenesis.

The wild type and mutant protein were expressed and purified under identical conditions, giving comparable yields of soluble protein. Kinetic analysis confirmed that the double mutant exhibited the same activity as wild-type. For the substrate 2-nitrophenyl- $\beta$ -D-glucopyranoside the  $K_M$  and  $k_{cat}/K_M$  for the wild type protein were  $0.21 \pm 0.2$  mM and  $(1.43 \pm 0.1) \times 10^5$  s<sup>-1</sup>·M<sup>-1</sup> respectively, and the  $K_M$  and  $k_{cat}/K_M$  values for the double mutant were  $0.18 \pm 0.01$  mM and  $(2.17 \pm 0.1) \times 10^5$  s<sup>-1</sup>·M<sup>-1</sup> respectively. Because the mutant showed no loss in activity compared to wild type it was deemed suitable for analysis via AFM.

## A.7 References

1. Davies G, Henrissat B: **Structures and mechanisms of glycosyl hydrolases.** *Structure* 1995, **3**:853-859.
2. Lairson LL, Henrissat B, Davies GJ, Withers SG: **Glycosyltransferases: Structures, Functions, and Mechanisms.** *Annu Rev Biochem* 2008, **77**:521-555.
3. Gloster TM, Meloncelli P, Stick RV, Zechel DL, Vasella A, Davies GJ: **Glycosidase Inhibition: An Assessment of the Binding of 18 Putative Transition-State Mimics.** *J Am Chem Soc* 2007, **129**:2345-2354.
4. Caines MEC, Hancock SM, Tarling CA, Wrodnigg TA, Stick RV, Stutz AE, Vasella A, Withers SG, Strynadka NCJ: **The Structural Basis of Glycosidase Inhibition by Five-Membered Iminocyclitols: The Clan A Glycoside Hydrolase Endoglycoceramidase as a Model System.** *Angew Chem Int Ed* 2007, **46**:4474-4476.
5. Zechel DL, Boraston AB, Gloster T, Boraston CM, Macdonald JM, Tilbrook DMG, Stick RV, Davies GJ: **Iminosugar Glycosidase Inhibitors: Structural and Thermodynamic Dissection of the Binding of Isofagomine and 1-Deoxynojirimycin to  $\beta$ -Glucosidases.** *J Am Chem Soc* 2003, **125**:14313-14323.
6. Lohse A, Hardlei T, Jensen A, Plesner IW, Bols M: **Investigation of the slow inhibition of almond  $\beta$ -glucosidase and yeast isomaltase by 1-azasugar inhibitors: evidence for the 'direct binding' model.** *Biochem J* 2000, **349**:211-215.
7. Tanaka A, Ito M, Hiromi K: **Equilibrium and kinetic studies on the binding of gluconolactone to almond beta-glucosidase in the absence and presence of glucose.** *J Biochem (Tokyo)* 1986, **100**:1379-1385.
8. Frieden C, Kurz LC, Gilbert HR: **Adenosine Deaminase and Adenylate Deaminase: Comparative Kinetic Studies with Transition State and Ground State Analogue Inhibitors.** *Biochemistry* 1980, **19**:5303-5309.
9. Zhuang X, Rief M: **Single-molecule folding.** *Curr Opin Struct Biol* 2003, **13**:88-97.
10. Fisher TE, Marszalek PE, Fernandez JM: **Stretching single molecules into novel conformations using the atomic force microscope.** *Nat Struct Biol* 2000, **7**:719-724.
11. Neuman KC, Nagy A: **Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy.** *Nat Methods* 2008, **5**:491-505`.
12. Li H, Linke WA, Oberhauser AF, Carrion-Vasquez M, Kerkvliet JG, Lu H, Marszalek PE, Fernandez JM: **Reverse engineering of the giant muscle protein titin.** *Nature* 2002, **418**:998-1002.

13. Pandhare J, Dash C, Rao M, Deshpande V: **Slow tight binding inhibition of proteinase K by a proteinaceous inhibitor: conformational alterations responsible for conferring irreversibility to the enzyme-inhibitor complex.** *J Biol Chem* 2003, **278**:48735-48744.
14. Schwesinger F, Ros R, Strunz T, Anselmetti D, Guntherodt HJ, Honegger A, Jermutus L, Tiefenauer L, Pluckthun A: **Unbinding forces of single antibody-antigen complexes correlate with their thermal dissociation rates.** *Proc Natl Acad Sci USA* 2000, **97**:9972-9977.
15. Wang T, Arakawa H, Ikai A: **Force Measurement and Inhibitor Binding Assay of Monomer and Engineered Dimer of Bovine Carbonic Anhydrase B.** *Biochem Biophys Res Commun* 2001, **285**:9-14.
16. Kedrov A, Appel M, Baumann H, Ziegler C, Muller DJ: **Examining the Dynamic Energy Landscape of an Antiporter upon Inhibitor Binding.** *J Mol Biol* 2008, **375**:1258-1266.
17. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: **Protein Identification and Analysis Tools on the ExPASy Server.** In *The Proteomics Protocols Handbook*. Edited by Walker JM. New York, NY: Humana Press; 2005:571-607.
18. Kempton JB, Withers SG: **Mechanism of *Agrobacterium*  $\beta$ -Glucosidase: Kinetic Studies.** *Biochemistry* 1992, **31**:9961-9969.
19. Nelson GW: **Covalently Bonded Protein Surfaces on Poly(Methyl Methacrylate): Characterization by X-ray Photoelectron Spectroscopy and Atomic Force Microscopy.** 2008.
20. Wolfenden R, Snider MJ: **The Depth of Chemical Time and the Power of Enzymes as Catalysts.** *Acc. Chem. Res.* 2001, **34**:938-945.