THERMODYNAMICS OF λ-PCR PRIMER DESIGN AND EFFECTIVE RIBOSOME BINDING SITES

Emily Katherine Berg

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Biological Systems Engineering

Ryan S. Senger, Chair
R. Clay Wright
David R. Bevan

April 26, 2019
Blacksburg, Virginia

Keywords: λ-PCR, primer design, synthetic ribosome binding site (RBS), translation initiation rate (TIR), Gibbs free energy, relative fluorescence, cyan fluorescent protein (CFP)

THERMODYNAMICS OF λ-PCR PRIMER DESIGN AND EFFECTIVE RIBOSOME BINDING SITES

Emily Katherine Berg

ABSTRACT

Recombinant DNA technology has been commonly used in a number of fields to synthesize new products or generate products with a new pathway. Conventional cloning methods are expensive and require significant time and labor; λ-PCR, a new cloning method developed in the Senger lab, has a number of advantages compared to other cloning processes due to its employment of relatively inexpensive and widely available materials and time-efficiency. While the amount of lab work required for the cloning process is minimal, the importance of accurate primer design cannot be overstated. The target of this study was to create an effective procedure for λ-PCR primer design that ensures accurate cloning reactions. Additionally, synthetic ribosome binding sites (RBS) were included in the primer designs to test heterologous protein expression of the cyan fluorescent reporter with different RBS strengths. These RBS sequences were designed with an online tool, the RBS Calculator.

A chimeric primer design procedure for λ-PCR was developed and shown to effectively create primers used for accurate cloning with λ-PCR; this method was used to design primers for CFP cloning in addition to two enzymes cloned in the Senger lab. A total of five strains of *BL21(DE3)* with pET28a + CFP were constructed, each with the same cyan fluorescent protein (CFP) reporter but different RBS sequences located directly upstream of the start codon of the CFP gene. Expression of the protein was measured using both whole-cell and cell-free systems to determine which system yields higher protein concentrations. A number of other factors were tested to optimize conditions for high protein expression, including: induction time, IPTG concentration, temperature, and media (for the cell-free experiments only). Additionally, expression for each synthetic RBS sequence was investigated to determine an accurate method for predicting protein

translation. NUPACK and the Salis Lab RBS Calculator were both used to evaluate the effects of these different synthetic RBS sequences. The results of the plate reader experiments with the 5 CFP strains revealed a number of factors to be statistically significant when predicting protein expression, including: IPTG concentration, induction time, and in the cell-free experiments, type of media. The whole-cell system consistently produced higher amounts of protein than the cell-free system. Lastly, contrasts between the CFP strains showed each strain's performance did not match the predictions from the RBS Calculator. Consequently, a new method for improving protein expression with synthetic RBS sequences was developed using relationships between Gibbs free energy of the RBS-rRNA complex and expression levels obtained through experimentation. Additionally, secondary structure present at the RBS in the mRNA transcript was modeled with strain expression since these structures cause deviations in the relationship between Gibbs free energy of the mRNA-rRNA complex and CFP expression.

THERMODYNAMICS OF λ-PCR PRIMER DESIGN AND EFFECTIVE RIBOSOME BINDING SITES

Emily Katherine Berg

GENERAL AUDIENCE ABSTRACT

Recombinant DNA technology has been used to genetically enhance organisms to produce greater amounts of a product already made by the organism or to make an organism synthesize a new product. Genes are commonly modified in organisms using cloning practices which typically involves inserting a target gene into a plasmid and transforming the plasmid into the organism of interest. A new cloning process developed in the Senger lab, λ-PCR, improves the cloning process compared to other methods due to its use of relatively inexpensive materials and high efficiency. A primary goal of this study was to develop a procedure for λ-PCR primer design that allows for accurate use of the cloning method. Additionally, this study investigated the use of synthetic ribosome binding sites to control and improve expression of proteins cloned into an organism. Ribosome binding sites are sequences located upstream of the gene that increase the molecule's affinity for the rRNA sequence on the ribosome, bind to the ribosome just upstream of the beginning of the gene, and initiate expression of the gene. Tools have been developed that create synthetic ribosome binding sites designed to produce specific amounts of protein. For example, the tools can increase or decrease expression of a gene depending on the application. These tools, the Salis Lab RBS Calculator and NUPACK, were used to design and evaluate the effects of the synthetic ribosome binding sites. Additionally, a new method was created to design synthetic ribosome binding sites since the methods used during the design process yielded inaccuracies.

Each strain of *E. coli* contained the same gene, a cyan fluorescent protein (CFP), but had different RBS sequences located upstream of the gene. Expression of CFP was controlled via induction, meaning the addition of a particular molecule, IPTG in this system, triggered expression of CFP. Each of the CFP strains were tested with a variety of

conditions in order to find the conditions most suitable for protein expression; the variables tested include: induction time, IPTG (inducer) concentration, and temperature. Media was also tested for the cell-free systems, meaning the strains were grown overnight for 18 hours and lysed, a process where the cell membrane is broken in order to utilize the cell's components for protein expression; the cell lysate was resuspended in new media for the experiments. ANOVA and multiple linear regression revealed IPTG concentration, induction time, and media to be significant factors impacting protein expression. This analysis also showed each CFP strain did not perform as the RBS Calculator predicted. Modeling each strain's CFP expression using the RBS-rRNA binding strengths and secondary structures present in the RBS allowed for the creation of a new model for predicting and designing RBS sequences.

ACKNOWLEDGEMENTS

This work would not have been completed without the help and support from a number of individuals, and I thank them. First, I would like to thank Dr. Ryan Senger for giving me the opportunity to complete my master's degree in his laboratory. His endless support and guidance was beyond helpful throughout this project, and I learned so much during my time in the Senger lab, both during my graduate and undergraduate careers. I also would like to thank my additional committee members, Dr. Clay Wright and Dr. David Bevan, for their feedback and advice throughout my thesis.

I would also like to thank Imen Tanniche for training me during my undergraduate career and for all of her help and suggestions throughout my thesis- she is an invaluable member of the Senger lab, and I cannot thank her enough. I also would like to thank Bert Huttanus for his help throughout the project. I'd also like to say thank you to Bert and David Scherr for their help providing additional data for cloning and enzyme expression.

Lastly, I would like to say thank you to my family and friends for their words of encouragement throughout my thesis. I could not have completed my master's without their support, and I cannot thank them enough.

TABLE OF CONTENTS

1 INTRODUCTION 1
1.1 SALIS LAB RBS CALCULATOR 3
1.1.1 Predicting Protein Translation Rates 6
1.1.2 Designing Ribosomal Binding Sites for Targeted Translation Rates 6
1.2 NUPACK 7
2 λ-PCR PRIMER DESIGN 10
2.1 INTRODUCTION 10
2.1.1 Motivation 10
2.1.2 λ-PCR Protocol Overview 10
2.2 MATERIALS AND METHODS 13
2.2.1 Primer Design Procedure for CFP1-5 13
2.2.1.1 NEBuilder 13
2.2.1.2 NUPACK: Megaprimer Thermodynamics Calculations 14
2.2.1.3 NUPACK: Checking Homologous Regions 16
2.2.2 λ-PCR Primer Design for SAM Synthase and Formaldehyde Ferredoxin Oxidoreductase 17
2.2.2.1 SAM Synthase Primer Design 17
2.2.2.2 Formaldehyde Ferredoxin Oxidoreductase Primer Design 20
2.2.3 Strains, Plasmids and Genes 24
2.2.4 Strain Construction 24
2.2.4.1 Amplifying the Target Gene 25
2.2.4.2 λ-Exonuclease Digestion 27
2.2.4.3 Cloning Reactions 27
2.2.4.4 Chemical Transformation of pET28a + CFP plasmid into E. coli BL21(DE3) 27
2.2.4.5 Verifying Transformations 28
2.3 RESULTS AND DISCUSSION 29
2.3.1 Primer Design Results: NUPACK 29
2.3.1.1 Megaprimer Thermodynamic Calculations 29
2.3.1.2 Checking Sequences with Homology to Plasmid 32
2.3.2 CFP1-5 Strain Construction 35
2.3.2.1 Gene Amplification 35
2.3.2.2 Digestion and λ-PCR Products 37

2.3.2.3 Colony PCR 40
2.3.3 Cloning Enzymes with λ-PCR: S-adenosylmethionine (SAM) Synthase and Formaldehyde Ferredoxin Oxidoreductase 42
2.3.3.1 Gene Amplification 42
2.3.3.2 λ-PCR Products 43
2.3.3.3 Colony PCR 44
3 SYNTHETIC RBS STUDIES 46
3.1 INTRODUCTION 46
3.2 MATERIALS AND METHODS 48
3.2.1 Design Procedure for Strains CFP1-3 48
3.2.1.1 Salis Lab RBS Calculator 48
3.2.1.2 NUPACK: Checking Megaprimer Thermodynamics 50
3.2.2 Design Procedure for Strains CFP4 and CFP5 51
3.2.2.1 Salis Lab RBS Calculator Modifications 51
3.2.2.2 CFP4 and CFP5 RBS Design 52
3.2.2.3 NUPACK 55
3.3.3 Further Analysis: Modeling Translation Initiation 55
3.3.3.1 NUPACK Thermodynamic Analysis of mRNA-rRNA Complex 55
3.3.3.2 Salis Lab RBS Calculator 56
3.3.4 Buffer and Media Preparation 58
3.3.5 *In vivo* Assays 58
3.3.5.1 Microplate Reader Assays: Constant IPTG Concentration 58
3.3.5.2 Microplate Reader Assays: Variable IPTG Concentration 58
3.3.5.3 Evaluating Fluorescence Levels 59
3.4 RESULTS AND DISCUSSION 59
3.4.1 Factors Impacting Cyan Fluorescent Protein Expression 59
3.4.1.1 Evaluating RBS Calculator Predictions with Plate Reader Assays 64
3.4.1.2 Determining Optimum IPTG Concentration and Time of Induction for Expression under T7 Promoter 67
3.4.1.3 Temperature Effects on Protein Expression 69
3.4.2 NUPACK: Complex Formation Results 71
3.4.2.1 Modeling RBS-rRNA Interactions in NUPACK 78
3.4.3 RBS Calculator Outputs for Each CFP Strain with Canonical and Non-Canonical 16S rRNA Sequences 87
4 CONCLUSION AND FUTURE DIRECTIONS 89

4.1 CONCLUSION 89
4.2 FUTURE DIRECTIONS 89
Appendix A: CELL-FREE STUDIES 91
A.1 INTRODUCTION 91
A.1.1 Cell-Free Metabolic Engineering 91
A.2 MATERIALS AND METHODS 92
A.2.1 Buffer and Media Preparation 92
A.2.2 Strain Construction 92
A.2.3 Microplate Reader Assays: Constant IPTG Concentration 92
A.2.4 Microplate Reader Assays: Variable IPTG Concentration 92
A.2.5 Sonication Procedure 93
A.3 RESULTS AND DISCUSSION 93
A.3.1 Strain Performance 94
A.3.2 Optimum media for protein expression 103
A.3.3 Optimum IPTG Concentration for Maximum Protein Expression 103
A.3.4 Cell-Free vs. *In Vivo* System 103
A.4: CONCLUSION 104
A.5: SUPPLEMENTARY DATA 106
Appendix B: R-STUDIO CODE FOR DATA ANALYSIS 114
Appendix C: OD AND FLUORESCENCE READING- DATA BY EXPERIMENT 126
C.1: Experiment in vivo 1 126
C.2: Experiment in vivo 2 128
C.3: Experiment in vivo 3 131
C.4: Experiment in vivo 4 134
C.5: Experiment in vivo 5 136
C.6: Experiment in vivo 6 137
C.7: Experiment in vivo 7 140
REFERENCES 142

LIST OF FIGURES

Figure 1: Initial System State in RBS Calculator. 5
Figure 2: Final System State in RBS Calculator..... 5
Figure 3. λ-PCR Cloning Steps..... 12
Figure 4: Flow Diagram showing the iterative algorithm of λ-PCR primer design 23
Figure 5: Gel electrophoresis of PCR 1 products in the λ-PCR protocol for CFP1-3. 36
Figure 6: Gel electrophoresis of PCR 1 products in λ-PCR protocol for CFP4 and CFP5. 37
Figure 7: Gel electrophoresis of PCR 2 products in the λ-PCR protocol for CFP1-3. 38
Figure 8: Gel electrophoresis of digestion and λ-PCR products for CFP4 and CFP5. 39
Figure 9: Gel electrophoresis of the colony PCR products for CFP1 and CFP2. 40
Figure 10: Gel electrophoresis of the colony PCR products for CFP2 and CFP3. 41
Figure 11: Gel electrophoresis of the colony PCR products for CFP4 and CFP5. 41
Figure 12: Gel electrophoresis of PCR 1 products in the λ-PCR protocol for SAM synthase and FOR enzymes. 43
Figure 13: Gel electrophoresis of the λ-PCR products (PCR 2) for the SAM synthase and FOR enzymes. 44
Figure 14: Gel electrophoresis of colony PCR products for SAM synthase. 45
Figure 15: Gel electrophoresis of the colony PCR products for the FOR enzyme. 45
Figure 16: RBS 4 and 5 Breakdown of Design Components [16] 54
Figure 17. 95% Confidence Intervals for Average Fluorescence for CFP1-3. 65
Figure 18: NUPACK graphics of the secondary structure of the mRNA for CFP2 at 30 and 37 °C 70
Figure 19: CFP1 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation 72
Figure 20: CFP2 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation 73
Figure 21: CFP3 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation 74

Figure 22: CFP1 (mRNA beginning at 3’ end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation..... 75
Figure 23: CFP2 (mRNA beginning at 3’ end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation..... 76
Figure 24: CFP3 (mRNA beginning at 3’ end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation..... 77
Figure 25. Diaphorase SDS-page gel. 79
Figure 26. SAM Synthase SDS-page gel. 80
Figure 27. Change in Gibbs free energy vs. expression 82
Figure 28. Change in Gibbs free energy vs. expression, without initial structures 83
Figure 29. Full mRNA transcript of CFP3 with the hairpin containing the RBS sequence outlined in the box. 84
Figure 30. Full mRNA transcript of CFP4 with the hairpin containing the RBS sequence outlined in the box. 85
Figure 31. Accessibility of RBS in an mRNA transcript vs. expression. 86

LIST OF TABLES

Table 1. Original Primer Sequences for CFP cloning by λ-PCR. 14
Table 2. Final CFP Primer Designs..... 16
Table 3. Original primer sequences for SAM synthase λ-PCR generated by NEBuilder.19
Table 4. Final primer sequences for SAM synthase λ-PCR generated by NEBuilder.... 20
Table 5. Original Primer Sequences for FOR Primer Design. 21
Table 6. Final Primer Sequences for FOR Primer Design. 22
Table 7. Primer Sequences for CFP Amplification. 26
Table 8. Primers Used in CFP Colony PCR..... 28
Table 9. NUPACK Output for Megaprimer Thermodynamics Calculations: Original CFP Forward Primer. 30
Table 10. NUPACK Output for Megaprimer Thermodynamics Calculations: Original CFP Reverse Primer..... 31
Table 11. NUPACK Output for Checking for Homologous Regions Between CFP Forward Primer and pET28a..... 33
Table 12. Initial CFP RBS Designs and Predicted TIR 50
Table 13. Final RBS Designs for CFP1-3 51
Table 14. Salis Lab Outputs for CFP4 and CFP5 54
Table 15. Relative fluorescence for each CFP strain, In Vivo Experiments 1-3 60
Table 16. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 4 & 5 61
Table 17. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 6 & 7 62
Table 18. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 6 & 7 with varying IPTG Concentrations 63
Table 19. RBS sequence simulated in NUPACK listed by protein expressed..... 81
Table 20. Salis Lab RBS Calculator Outputs for Each CFP Design Simulated with Both Potential 16S rRNA Sequences 88
Table 21. Experiment In Vitro 1 Results..... 95
Table 22. Experiment In Vitro 2 Results..... 97

Table 23. Experiment In Vitro 3 Results..... 99
Table 24. Experiment In Vitro 4 Results..... 101
Table 25. Experiment In Vitro 4 Results..... 102

1 INTRODUCTION

Recombinant DNA technology has been used to advance a number of industries forward, namely in food, health and medicine, energy, and applications that attempt to improve environmental quality. Plants have been genetically modified to improve yields and use resources more efficiently in suboptimal growing conditions, such as drought. Medicinal industries have used recombinant DNA methods to invent new drugs, such as therapeutic proteins, and produce new vaccines with efficient processes using bacteria.[1] Bacteria have been employed in bioremediation practices to successfully breakdown harmful pollutants.[2] Furthermore, microbes have been genetically altered, commonly with systems and synthetic biology and metabolic engineering practices, for a range of applications in energy, such as more sustainable biofuel and hydrogen production.[1] These changes to bacterial metabolism have been applied to synthesize enzymes for use as a product itself or assembled in a cascade reaction.[3] In turn, these enzyme cascades are used *in vivo* or assembled *in vitro* to convert biological material into a number of valuable products, including but not limited to fuels, pharmaceuticals, and commodity chemicals.[4] Metabolic engineering improves the microbial conversion of biological material into valuable products by engineering bacteria to metabolize molecules differently, commonly in a more efficient manner.[5] Metabolic engineering has usually been employed using *in vivo,* or whole cell, methods with a systems biology approach.[6]

Heterologous proteins are typically expressed in bacteria by cloning the source gene into a plasmid, transforming the plasmid into the host organism, inducing expression, and purifying the protein.[7] Optimizing the cloning process speeds the process along while reducing costs associated with protein expression since the enzymes required for cloning tend to be expensive. The following sections include overviews on common cloning practices as well as an introduction to a new cloning method with key advantages to conventional cloning.

Gibson Assembly

Two conventional methods of plasmid construction include Gibson Assemblies and T-A Cloning. With Gibson Assemblies, large primers are designed to extend the replicated sequences on the 5' and 3' end for both the target sequence and plasmid sequence; the primers for the plasmid are designed to be located at the target gene insertion site. After each sequence is replicated with PCR, the fragments are then digested with T5 exonuclease to create sticky ends from the overhangs on each segment. Phusion DNA polymerase then adds nucleotides to the regions where the segments were digested to synthesize complementary strands, and a Taq DNA ligase seals the strands together.[8] This method is particularly efficient because it requires one reaction aside from PCR and can be applied to assemble multiple genes together on one plasmid; however, this method limits plasmid modifications to restriction enzyme sites, and requires three enzymes and large primer designs, proving to be costly.[9]

T-A Cloning

Another method used for plasmid construction is T-A Cloning. In this cloning method, linearized plasmids with a 5' dT overhang are incubated with Taq DNA polymerase PCR products, containing a 3' dA overhang, to generate a plasmid with the inserted gene. The process requires a ligase to seal the final cloning product. This process requires little time and labor; however, plasmid modifications are limited to restriction enzyme sites, and cloning is non-directional. Additionally, it has been shown that these T-vectors degrade over time due since they exist in the linearized form in storage, significantly reducing reaction efficiency.[10]

A number of new advancements have risen in the biotechnology field to improve cloning efficiencies while saving resources. Novel cloning methods have been developed in an attempt to decrease the amount of time and labor required to create a new strain of bacteria, in particular. One of these advancements is a technique described as λ-PCR; this method of cloning requires a fraction of the time and work compared to conventional methods while using universal materials and open-source tools.[11] The labor-intensive part of this technique is the primer design portion due to all of the variables in the reactions that must be checked before proceeding in the laboratory. The work described in this

manuscript focuses on an algorithmic method for designing these primers. It also deals with advances in recombinant DNA technology through the discovery of a novel method for cloning and improving protein yields from heterologous expression with synthetic ribosomal binding site (RBS) usage. RBS designs were tested with *in vivo* and *in vitro* systems, and λ-PCR was used for strain construction with each RBS.[12] A number of tools were used throughout this thesis project to optimize primer designs for λ-PCR, including the Salis lab RBS Calculator[13] and NUPACK.[14] The following sections entail descriptions of the Salis lab tool as it was used during the project.

1.1 SALIS LAB RBS CALCULATOR

The Salis Lab at Pennsylvania State University developed the Ribosomal Binding Site (RBS) Calculator for the use of controlling protein expression through translation initiation in bacteria. The tool has a variety of applications revolving around the optimization of gene translation for finely-tuned expression levels. It is available at https://salislab.net/software/. The algorithm functions using a well-characterized thermodynamic model representing the change in total Gibbs free energy, $\Delta G_{total} = \Delta G_{final} - \Delta G_{initial}$, during the formation of the 30S ribosomal unit and messenger RNA (mRNA) complex.[13,15]

The final state of this model (ΔG_{final}) is described as the sum of the following components: the change in Gibbs free energy due to formation of the mRNA and 16S ribosomal RNA (rRNA) complex ($\Delta G_{mRNA\text{-}rRNA}$), the change in Gibbs free energy of the spacing sequence located at the 3' end of the RBS ($\Delta G_{spacing}$), the change in Gibbs free energy of the standby sequence located at the 5' end of the RBS ($\Delta G_{standby}$), and the change in Gibbs free energy associated with the pairing of the start codon to the initiator tRNA, methionine (ΔG_{start}).[16] $\Delta G_{spacing}$ quantifies the favorability of the distance between the end of the RBS and the start codon. Distances too short or too long have proven to decrease translation initiation considerably due to differences between the precise distance between the 30S ribosomal unit and the 16S ribosomal RNA and the distance between the start codon and the RBS binding region; as a result, $\Delta G_{spacing}$ was included in

the model's thermodynamics formula.[13,17,18] $\Delta G_{standby}$ represents the energy required to unfold and access the nucleotide sequence just upstream of the RBS since the 30S subunit binds to the standby sequence before moving downstream to the RBS; this is represented as a negative value in the thermodynamic calculations since it is either exposed and accessible ($\Delta G \approx 0$) or containing secondary structures that prohibit the 30S subunit from binding and require energy for unfolding the structures ($\Delta G < 0$).[13,15,19] A graphic illustration of the Gibbs free energies associated with the system's final state is shown below in Figure 2.

The Gibbs free energy of the complex's initial state ($\Delta G_{initial}$) is represented as the Gibbs free energy of mRNA transcript (ΔG_{mRNA}) secondary structure surrounding the start codon in the RBS Calculator. This total change in Gibbs free energy during mRNA-rRNA complex formation, ΔG_{total}, is then related to the translation initiation rate (TIR) using mathematical relationships determined by the Salis lab. TIR was used as the main output to represent translation success since it is the limiting step, preceding protein elongation.[15] Two functions of the RBS Calculator were used during the course of this project: the Prediction and Design tools.

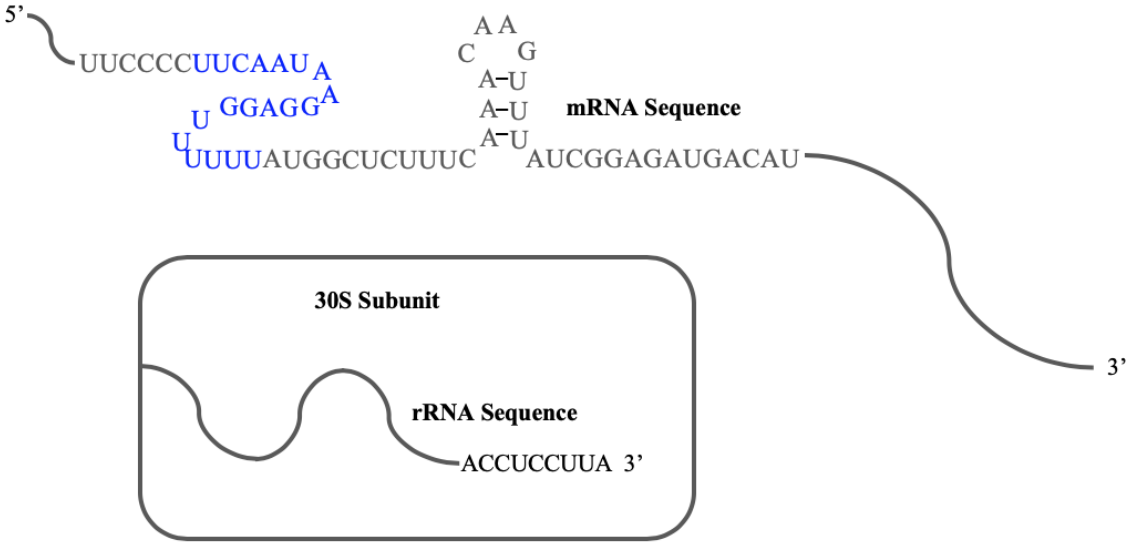


Figure 1: Initial System State in RBS Calculator.

The synthetic RBS is outlined in blue text. [15,16]

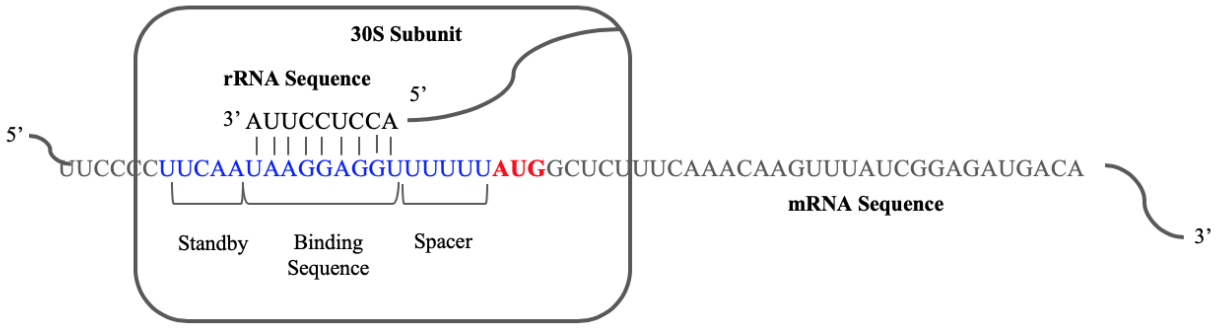


Figure 2: Final System State in RBS Calculator

Assembled mRNA-rRNA Complex.[15,16]

1.1.1 Predicting Protein Translation Rates

The RBS Calculator's Prediction tool requires the mRNA transcript and host organism for gene expression as inputs to the tool; the organism is selected from a list provided by the Prediction tool with rRNA sequences listed in parentheses. 16S rRNA sequences can be inputted to the tool as well if they are not provided in the list of organisms. The algorithm functions by using the thermodynamic relationship described previously, $\Delta G_{total} = \Delta G_{final} - \Delta G_{initial}$, to assess the success of the 30S ribosomal unit and 16S rRNA binding correctly to the mRNA transcript at the RBS/start codon for protein translation initiation. The TIR is computed using ΔG_{total} for each start codon the algorithm finds in the provided mRNA transcript (AUG or GUG). The output for this tool includes a list of potential start codon positions with their corresponding TIR, ΔG_{total}, $\Delta G_{mRNA\text{-}rRNA}$, $\Delta G_{spacing}$, $\Delta G_{standby}$, ΔG_{mRNA}, and links to the graphics displaying the mRNA molecular structure surrounding the start codon.[13, 16] The secondary structure surrounding the start codon must be considered while optimizing protein translation because these structures in regions upstream and downstream of the start codon control the efficiency of translation initiation.[20, 21]

1.1.2 Designing Ribosomal Binding Sites for Targeted Translation Rates

The design option of the RBS Calculator prompts the following inputs to the tool: the protein pre-sequence, the protein coding sequence, the targeted TIR, and the selection of the host organism used for protein expression.[13] The pre-sequence includes nucleotides upstream of where your gene will be inserted; for our purposes, we included 40 nucleotides upstream. The protein coding sequence contains at least the first 35 nucleotides of the target protein sequence with longer sequences preferred. The targeted TIR is chosen within a range of 0.001 to 100,000 arbitrary units (au) with "maximizing protein expression" provided as an option. The host organism is selected using the list of organisms built into the Design tool as described previously for the Prediction tool; additionally, a unique 16S rRNA sequence can be inputted by typing the sequence into the box. The algorithm runs through 4^35 potential RBS sequences using the inputs to the tool, computes ΔG_{total} with the thermodynamic model described previously, and relates

ΔG_{total} to the TIR using a logarithmic relationship. A modified version of this Design tool prompts for the inputs listed previously with additional constraints regarding the length and specific nucleotide sequences used in the designed RBS sequence.[15,16] The output for the Design tool includes the generated RBS sequence located between the provided pre-sequence and protein coding sequence. Additionally, the Design tool output includes the Prediction tool's outputs described previously that show the breakdown of the Gibbs free energies in the thermodynamic model.

1.2 NUPACK

RNA molecules naturally form structures both in vivo and in vitro at the primary, secondary, tertiary, and quaternary levels. Primary structure refers to the molecule's nucleotide sequence.22 Secondary structure is the structure that an individual molecule forms through base-pairing with itself to achieve the most stable free energy at equilibrium.23 Tertiary structure refers to the 3D structures the nucleic acid forms with itself, and quaternary structure refers to the structure that occurs with multiple RNA molecules interacting.22, 24 Each level of molecular structure is dependent on the previous level, and secondary structure is the strongest and most stable folding that occurs in a nucleic acid.23, 25 Secondary structure is important to consider during cloning reactions in the λ-PCR protocol, particularly in regions of homology between the plasmid and the megaprimer, since the folding that occurs on this level largely determines the behavior of the molecule in vitro and the success of the cloning reaction. During the design process, secondary structure formation was quantified and evaluated with NUPACK, a software available at http://nupack.org/partition/new.

NUPACK predicts structures forming within nucleic acids using thermodynamic models that quantify secondary structure formation with minimum free energy (MFE). The program does this with a number of built-in functions: i) calculating the partition function and MFE of the secondary structure for unpseudoknotted complexes, ii) calculating equilibrium concentrations for species involved in complex formation, iii) calculating base-pairs that occur at equilibrium with the partition function and concentrations

inputted by the user, and iv) designing nucleic acids that behave with a particular non-pseudoknot structure at equilibrium.[14]

The program includes a number of potential inputs: i) the nucleic acid sequence, DNA or RNA, ii) temperature in which the thermodynamic calculations are performed (or inputting a temperature range for the algorithm to compute the melting temperature), iii) number of strand species if the user is interested in any complexes that form, iv) maximum complex size and v) concentrations for each molecule inputted. The program performs the thermodynamic calculations based on the inputs included and displays the equilibrium concentrations for each strand inputted. Selecting each strand or interaction of strands reveals a graphic of the MFE structure at equilibrium, the free energy of the structure, and a graphic showing the pair probabilities forming at the chosen temperature. Additionally, graphics are produced to show the fractions of base pairs that occur at the input temperature. Downloading the generated data shows the probability of each base-pair forming at the temperature as well as the likelihood of every nucleotide in each strand being unpaired at equilibrium in the MFE secondary structure.[14]

Functions i) and iii) were used most often throughout this project to quantify the MFE of each nucleic acid, and any potential complexes, and quantifying the likelihood of base-pairs occurring using probabilities produced by the algorithm. Additionally, graphics displaying the interactions forming at equilibrium were used to investigate mRNA-rRNA complex formation.

This thesis investigates heterologous cyan fluorescent protein (CFP) expression in *Bl21(DE3)* using synthetic RBS sequences designed using the Salis Lab RBS Calculator. Chapter 2 outlines a new cloning method used in this project, λ-PCR, and provides a primer design algorithm created for accurate cloning with λ-PCR. Additionally, multiple case studies showing successful λ-PCR cloning with heterologous enzymes are included. Chapter 3 includes the process used to design synthetic RBS sequences with different expression strengths. Differences in the RBS Calculator's predictions for each RBS design and experimental results caused a number of methods to be used to investigate

factors affecting protein expression and potential improvements to the method used for designing RBS sequences. Appendix A covers the results of the cell-free protein expression experiments tested throughout the course of this project in order to assess the best system for maximizing protein expression in *E. coli.*

2 λ-PCR PRIMER DESIGN

2.1 INTRODUCTION

2.1.1 Motivation

λ-PCR is a new cloning method used for plasmid construction. It is advantageous to conventional cloning methods because it requires the design of one pair of chimeric primers, it is inexpensive compared to other commercial cloning kits, uses only two enzymes, and it significantly decreases time and labor required for strain construction compared to other cloning methods. Additionally, all tools required for its efficient design and use are widely available. [11,12]

2.1.2 λ-PCR Protocol Overview

λ-PCR was used throughout this project for heterologous protein expression. While the procedure efficiently constructs plasmids in a cost-effective and timely manner, ensuring the proper design of the chimeric primers is crucial for accurate plasmid construction. This chapter entails a step-by-step approach to designing chimeric primers for efficient plasmid construction by λ-PCR. Chimeric primers are primers in which the region towards the 5' end on both the forward and reverse primers contains homology to the plasmid near the insertion site, and the region towards the 3' end contains homology to the gene for gene amplification.[12] This particular case study describes the design process for heterologous cyan fluorescent protein (CFP) expression. This approach utilizes a tool developed by the Salis lab at Pennsylvania State University, the RBS Calculator, designed to optimize protein expression at desired rates by adding a synthetic ribosome binding site (RBS) as a spacer sequence in the forward primer.[13] This part can either be included or removed from the primer design protocol based on the needs of the user.

The λ-PCR protocol requires chimeric primers to be designed with sequences homologous to the target location on the plasmid and the target gene; these primers are designed so the 5' end consists of the sequence homologous to the plasmid while the 3' end contains the sequence homologous to the target gene. The λ-PCR process begins with

the phosphorylation of the 5' end of the reverse primer by T4 polynucleotide kinase. The target gene is then amplified using the unphosphorylated forward primer and phosphorylated reverse primer to generate dsDNA copies of the target gene with one 5' end phosphorylated; this is shown below with Step 1 in Figure 3.[12] This double-stranded product is treated with λ-exonuclease, an enzyme known to digest blunt DNA by the 5' phosphorylated end.[26] As a result, overnight digestion of the dsDNA will yield one single-stranded copy; this is the megaprimer used during the second PCR. This digestion is Step 2 in Figure 3 below. In the second PCR, Step 3 in Figure 3, this megaprimer binds to the plasmid with the ends homologous to the sequences surrounding the target location on the plasmid, forming an "**Ω**" shape, and inserts the target gene into the plasmid backbone by PCR with a universal reverse primer binding to the plasmid.[12] With efficient primer design, a new strain can be produced and verified with λ-PCR in a fraction of the time required for conventional cloning methods. The following section details factors that must be evaluated and accounted for in order for successful λ-PCR reactions to occur.

Chimeric primers were designed to clone the CFP gene into pET28a for expression in *E. coli BL21(DE3).* The following section details the procedure used to design these primers for λ-PCR since minor factors can lead to an unsuccessful cloning reaction or the synthesis of an undesirable plasmid. Section 2.3, Results and Discussion, includes the successful results of CFP cloning with graphics of each PCR product gel bands as well as case studies that show the success of λ-PCR in the Senger lab with enzyme expression in *BL21(DE3).*

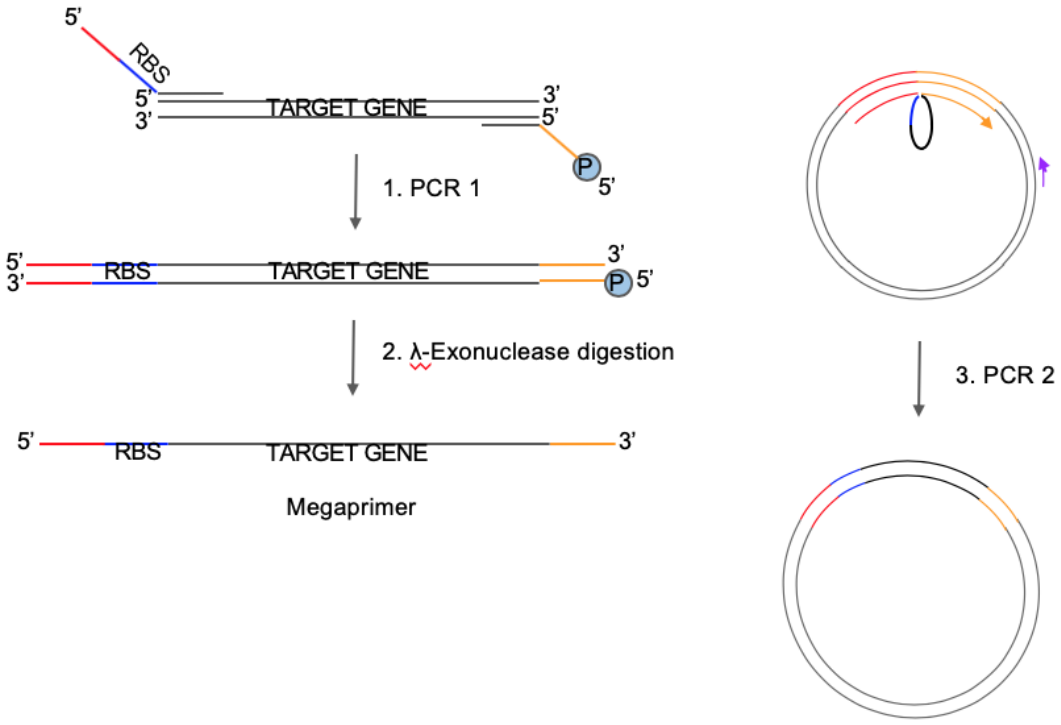


Figure 3. λ-PCR Cloning Steps.

The schematic above shows each reaction that occurs in the λ-PCR process. Step 1, PCR 1, is the amplification of the target gene with overhangs for the regions of homology on the plasmid. Step 2, λ-exonuclease digestion, generates the megaprimer by digesting the 5' phosphorylated strand of the dsDNA. Step 3, PCR 2, inserts the target gene into the plasmid using the regions of homology and a universal reverse primer.[12]

2.2 MATERIALS AND METHODS

2.2.1 Primer Design Procedure for CFP1-5

2.2.1.1 NEBuilder

The first step of the λ-PCR primer design procedure is to select a location on the plasmid for gene insertion; for the course of this project, this insertion site was originally chosen to be located 57 nucleotides downstream of the *lac* operon. This insertion site was chosen since its close proximity to the T7 promoter while including the native RBS upstream of the site enabled high expression of the gene, and N-terminal His tags were avoided with this precise location in pET28a. Additionally, regions directly upstream of this location, near the NcoI restriction site, are adenine-thymine heavy; as a result, the primer designs would need to be much longer to achieve a similar annealing temperature to regions with more cytosine and guanine. The CFP designs required a synthetic RBS in the forward primer for protein expression studies, as described in Chapter 3, so inserting the gene at the NcoI restriction site would require longer primers and directly increase the cost of the primers. After the site was selected, the plasmid design with the gene insertion was created in New England Biolab's NEBuilder Assembly Tool®, available at https://nebuilder.neb.com/#!/. With this tool, the location for gene insertion on the plasmid was selected by inputting the bases flanking the insertion site; for the plasmid orientation used in this design, these bases were 5080 and 5081. NEBuilder output the primer sequences based on the design inputs with their corresponding annealing and melting temperatures; the sequences for the original primer sequences are shown in Table 1 below.

Table 1. Original Primer Sequences for CFP cloning by λ-PCR.
RBS not included in the forward primer. The lowercase letters represent the portion of the primer that binds to the plasmid in the PCR 2 in Figure 3, and the capitalized letters represent the portion of the primer involved in PCR 1 in Figure 3.

Primer	**Sequence**	**Annealing Temperature °C**
CFP_Forward	*gagatataccatgggcagcaATGGCTCTTTCAAACAAG*	57
CFP_Reverse	*ctgtgatgatgatgatggcttTCAGAAAGGGACAACAGAG*	61 65 (PCR 2)

After generating the primer sequences from NEBuilder and selecting the insertion site on the plasmid, the final construct (plasmid + target gene + RBS) was checked to ensure the open reading frame (ORF) was detected; any changes to the ORF could impact protein expression by altering the amino acid sequence downstream of the shift, potentially changing the protein's structure or function.[27]

2.2.1.2 NUPACK: Megaprimer Thermodynamics Calculations

During the course of λ-PCR primer design, NUPACK (http://www.nupack.org) was used to evaluate the thermodynamics of the single-stranded DNA megaprimer and ensure secondary structure formation is minimized in the regions homologous to the insertion site on the plasmid.[12] The DNA option was selected in NUPACK for megaprimer analysis. The temperature used for the simulation was the annealing temperature of the primers for the plasmid since this is the temperature in which the homologous regions on the plasmid and megaprimer bind together; during this design process, the chosen annealing temperature for the primer regions with homology to the plasmid was 65°C, so the calculations were done at this temperature. The megaprimer sequence was input under "Strand Species." The megaprimer listed 5' to 3' includes the portion of the forward primer homologous to the plasmid, the entire gene sequence, and the reverse complement of the portion of the reverse primer homologous to the plasmid. After

ensuring the entire megaprimer sequence was included correctly, the thermodynamics of the sequence were analyzed.

The output for NUPACK included a graphic illustrating any hairpins forming in the megaprimer along with the free energy of the secondary structure as well as a graphic of the base index showing the free energy of the strand with a logarithmic relationship. Downloading the data shows a breakdown of nucleotide pairs bonding with the corresponding probability of that bond forming at equilibrium. Additionally, probabilities listed with a "-1" to the right of the nucleotide identifier represent the probability of that nucleotide being unpaired at equilibrium at the temperature chosen by the user; this was the focus of the NUPACK thermodynamic calculations for the single-stranded megaprimer. Using the probabilities of each nucleotide being unpaired at equilibrium, it was verified that the regions on the 5' and 3' ends of the megaprimer with homology to the plasmid are unpaired. Higher probabilities are desirable for a successful design, so these designs were optimized with probabilities as large as possible to ensure the megaprimer binds correctly to the plasmid. The thermodynamics of the original megaprimer were shown to form unfavorable secondary structures in regions with homology to the plasmid; accordingly, these structures can prevent the megaprimer in the second PCR from binding correctly.

Optimizing primer design so the ends of the megaprimer are unpaired at equilibrium, quantified in NUPACK with high probabilities associated with nucleotides in the regions with homology to the plasmid, was important for accurate λ-PCR. As a result, the insertion site was shifted in the NEBuilder Assembly Tool®, thus generating a new megaprimer; note that changing the insertion site by a few nucleotides severely impacted NUPACK calculations. The new megaprimer was simulated in NUPACK to verify the homologous regions on the megaprimer were unpaired at equilibrium, providing insight into the success of the second PCR. This process was repeated until NUPACK results revealed a megaprimer with minimal secondary structure.

NUPACK calculations showed one region of homology having a low probability of secondary structure formation at equilibrium and one region forming secondary structure that prevents the megaprimer from binding correctly, so the sequence forming secondary structure was adjusted by cutting out a portion of the plasmid. This was done by shifting the primers with poor NUPACK results upstream or downstream so that the regions of homology on the primers are not directly located next to each other on the plasmid; note that any nucleotides located between the chosen regions of homology, or downstream of the forward primer and upstream of the reverse primer, will not be replicated for the final plasmid. The CFP design was optimized by moving the reverse primer to prevent hairpin structures from forming in the megaprimer; as a result, the sequences for the forward and reverse primers were updated and are displayed in the table below.

Table 2. Final CFP Primer Designs

The lowercase letters represent the portion of the primer involved in PCR 2 in Figure 3, and the capitalized letters represent the part of the primer annealing to the gene in PCR 1 in Figure 3.

Primer	**Sequence**	**Annealing Temperature °C**
Forward	*tgagcggataacaattccccATGGCTCTTTCAAACAAG*	57
Reverse	*gctgctgtgatgatgatgatgaTCAGAAAGGGACAACAGAG*	61 65 (PCR 2)

2.2.1.3 NUPACK: Checking Homologous Regions

Another factor checked before proceeding with the λ-PCR protocol was verifying the megaprimer will not anneal to the plasmid in undesirable regions. This was checked by simulating complex formation between the homologous regions on the megaprimer with the plasmid sequence; note that NUPACK has a limit on strand length, so the plasmid sequence was analyzed in multiple simulations. In NUPACK, "DNA" was chosen with the annealing temperature for the second PCR inputted, and "2" was selected for "Number of strand species" and "Maximum complex size." Any appropriate strand

concentration can be selected in NUPACK since the outputs of concern, free energy of secondary structure and probabilities of nucleotides being unpaired at equilibrium, are not impacted by strand concentrations; these factors are primarily dependent on the nucleotide sequences.[14] 100 μM was chosen as the concentration for each strand in the NUPACK calculations. Each primer was run against the plasmid sequence used in the design steps as well as the reverse complement of this sequence to ensure the primer would not bind to either side of the plasmid. Each primer sequence should only bind in one particular location with all simulations. The output of the probability data shows the likelihood of a bond forming at equilibrium between two nucleotides for each potential nucleotide base pair. This was checked to ensure the chimeric primers bind correctly to the plasmid at the annealing temperature since any nonspecific binding can lead to the synthesis of the wrong plasmid. NUPACK showed a very low probability for any undesirable binding to occur; therefore, we proceeded forward with primer design for the first three CFP strains.

2.2.2 λ-PCR Primer Design for SAM Synthase and Formaldehyde Ferredoxin Oxidoreductase

λ-PCR has been used to clone in a number of projects in the Senger lab. S-adenosylmethionine synthase and formaldehyde ferredoxin oxidoreductase are two enzymes that were successfully cloned into the pET28a plasmid and transformed into competent *E. coli BL21(DE3)* cells for heterologous protein expression. The following sections include details on the primer design process used for each of the enzymes.

2.2.2.1 SAM Synthase Primer Design

SAM synthase had a similar target insertion site on pET28a compared to CFP, located at the NcoI restriction site (about 47 nucleotides downstream of the *lac* operon). This gene did not require a synthetic RBS, so there were fewer limitations regarding the insertion site location in pET28a, and consequently primer lengths, for this design. The NEBuilder Assembly Tool® generated the first set of primers shown in Table 3 below for this design

procedure. These designs were run in NUPACK to observe any issues with the megaprimer's secondary structure formations at the annealing temperature in PCR 2, 65 °C.

Table 3. Original primer sequences for SAM synthase λ-PCR generated by NEBuilder. Capitalized letters correspond to the part of the primers annealing to the gene in PCR 1 in Figure 3. Lowercase letters represent the part of the primers annealing to the plasmid during PCR 2 in Figure 3.

Primer	**Sequence**	**Annealing Temperature** °C
Forward	*ccctctagaaataattttgtttaactttaagaaggagatataATGAGAAG ACTTTTCACC*	56
Reverse	*atggctgctgcccatTTACATATTGAAAGCTCTTTTC*	55 65 (PCR 2)

NUPACK results revealed secondary structures forming at 65 °C at equilibrium at the 3' end of the megaprimer that could prevent PCR 2 from occurring correctly. Additionally, the total length of the forward primer was 60 nucleotides, making the primer more expensive. As a result the primers were redesigned with NEBuilder by selecting a new insertion site in pET28a until the megaprimer remained unpaired at equilibrium and each primer was under 60 nucleotides. Once desirable NUPACK results were received for the megaprimer analysis, the regions of the primers with homology to the plasmid were run in NUPACK against the entire pET28a sequence, with a maximum complex size of 2 selected, to ensure the primers would only bind to the targeted locations at the annealing temperature in PCR 2 (Figure 3). The final primer designs for SAM synthase are shown below in Table 4. The reverse primer was moved downstream to prevent any secondary structures from forming in the region annealing to the plasmid as best as possible. The final reverse primer anneals near the EcoRI and BamHI restriction sites, nearly 100 nucleotides downstream from the 3' end of the forward primer. This results in the portion of the plasmid downstream from the forward primer and upstream of the reverse primer getting cut out of the final plasmid during PCR 2 in Figure 3. This 100-nucleotide region does not contain any elements necessary for a functional cloning product, pET28a + SAM synthase, so this was acceptable.

Table 4. Final primer sequences for SAM synthase λ-PCR generated by NEBuilder. Capitalized letters correspond to the part of the primers annealing to the gene in PCR 1 in Figure 3. Lowercase letters represent the part of the primers annealing to the plasmid during PCR 2 in Figure 3.

Primer	**Sequence**	**Annealing Temperature** °C
Forward	*agaaataattttgtttaactttaagaaggagatataccATGAGAAGACT TTTCACC*	56
Reverse	*acggagctcgaattcggaTTACATATTGAAAGCTCTTTTC*	55 65 (PCR 2)

2.2.2.2 Formaldehyde Ferredoxin Oxidoreductase Primer Design

The insertion site on pET28a for formaldehyde ferredoxin oxidoreductase (FOR) was selected to be the same location on pET28a as the CFP designs since a synthetic RBS was included in the forward primer; therefore, the adenine-thymine heavy region near the NcoI restriction site was undesirable for primer design. The primer design tool with NEBuilder generated the initial set of primers shown in Table 5 below. These designs were run in NUPACK to check the secondary structures forming within the megaprimer at the annealing temperature for PCR 2, 65 °C.

Table 5. Original Primer Sequences for FOR Primer Design.
Lowercase letters represent the part of the primer involved in PCR 2 (Figure 3), and capitalized letters represent the part of the primer involved in PCR 1 (Figure 3).

Primer	**Sequence**	**Annealing Temperature °C**
Forward	*ctttaagaaggagatataccATGAATGTAAAGATGGTGG*	56
Reverse	*atggctgctgcccatTCACTCAAGGTTTGTAAC*	56 65 (PCR 2)

NUPACK thermodynamics calculations revealed significant secondary structures forming in the megaprimer that would prevent the megaprimer from binding to pET28a correctly; therefore, the insertion site was adjusted in NEBuilder to generate a megaprimer without significant secondary structures. The majority of these structures were present in the reverse primer, so the insertion site was modified by cutting out part of the plasmid by moving the reverse primer farther downstream. The final primer design moved the reverse primer about 30 nucleotides downstream to prevent secondary structure formation, resulting in 30 nucleotides being cut out of the final plasmid product, pET28a + FOR. The part of the primers participating in PCR 2 were run against the entire pET28a sequence in NUPACK, with a maximum complex size of 2, to ensure nonspecific binding did not occur at the annealing temperature for PCR 2. These results were acceptable, so cloning proceeded with the designs shown below in Table 6.

Table 6. Final Primer Sequences for FOR Primer Design.

Lowercase letters represent the part of the primer involved in PCR 2 (Figure 3), and capitalized letters represent the part of the primer involved in PCR 1 (Figure 3).

Primer	**Sequence**	**Annealing Temperature °C**
Forward	*ctttaagaaggagatataccATGAATGTAAAGATGGTGG*	56
Reverse	*gcaccaggccgctgctgtgaTCACTCAAGGTTTGTAAC*	56 65 (PCR 2)

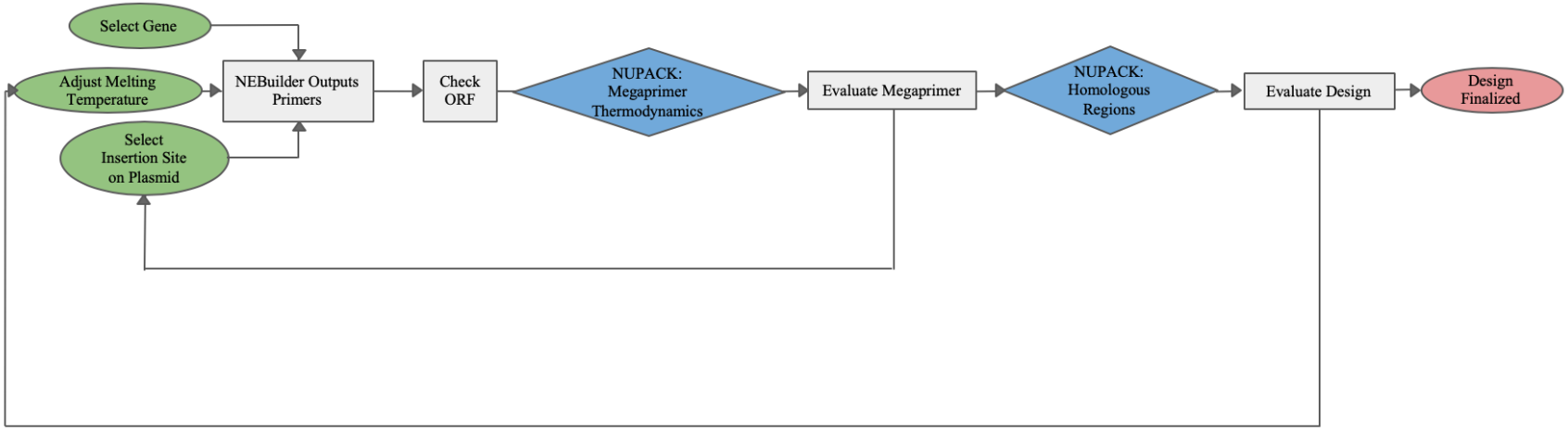


Figure 4: Flow Diagram showing the iterative algorithm of λ-PCR primer design

2.2.3 Strains, Plasmids and Genes

E. coli BL21(DE3) was chosen as the host for heterologous protein expression since *E. coli* is the most characterized bacterial species, has a fast growth rate, and can be genetically modified with ease. Additionally, *BL21(DE3)* was used since this particular strain of *E. coli* has been optimized to produce high yields of protein by deleting multiple proteases, the *Lon* and *OmpT* genes, and has the gene for T7 polymerase expression in its genome.[28] pET28a was the plasmid used for heterologous expression since it contains the inducible T7 promoter and *lac* operon, allowing for control of T7 RNA polymerase expression, and is known to produce high yields of protein sequences immediately downstream of the promoter.[29] The cyan fluorescent protein (CFP) is encoded by the *AmCyan* gene, and protein expression was easily quantified with its known excitation and emission wavelengths, 458 and 489 nm, respectively.[30] SAM synthase was cloned from the source organism *Thermotoga maritima*, and formaldehyde ferredoxin oxidoreductase was cloned from *Pyrococcus furiosus.*

2.2.4 Strain Construction

The design procedure for each of the chimeric primers used in this protocol is described in section 2.2. The reporter gene, cyan fluorescent protein (CFP), was cloned from strains previously sequenced and confirmed to contain the *AmCyan* gene. The *AmCyan* gene used as the template in the first PCR was originally cloned from the plasmid, pAmCyan (Takara Bio USA; Mountain View, CA). The first PCR in the protocol used chimeric primers designed with regions to amplify the *amcyan* gene and with regions of homology to pET28a. All primers were produced by Integrated DNA Technologies (Coralville, IA). *E. coli BL21(DE3)* (New England Biolabs; Ipswitch, MA) was the strain used in all transformations. The same protocol detailed in the following sections was followed for cloning SAM synthase and formaldehyde ferredoxin oxidoreductase using the final primer designs and corresponding annealing temperatures outlined in sections 2.2.2.1 and 2.2.2.2.

2.2.4.1 Amplifying the Target Gene

The first step of the protocol is to phosphorylate the reverse primer for the first reaction. This was done with the following protocol: 0.5 uL reverse primer (100 µM), 0.5 µL 10X T4 DNA ligase buffer (New England Biolabs), 0.5 µL T4 polynucleotide kinase (New England Biolabs), and 3.5 µL molecular biology grade water for a 5 µL reaction. This reaction mixture was incubated at 37 °C for one hour and deactivated for 10 minutes at 75 °C. The first PCR was performed to amplify the target gene with the chimeric primers to generate overhangs with homology to pET28a. The 25 µL PCR reaction consisted of 0.5 µL forward primer, 1 µL phosphorylated reverser primer, template DNA (~50 ng/µL), 12.5 µL Q5 ® High-Fidelity 2X Master Mix (New England Biolabs), and 11 µL molecular biology grade water; the reaction was done with 30 cycles, and the primers used in this PCR are described in the previous chapter. The PCR program was designed with the following procedure: i) denaturation at 98 °C for 30 seconds, followed by 30 cycles of ii) denaturation at 98 °C for 10 seconds, iii) annealing at 57 °C for 30 seconds, and iv) extension at 72 °C for 30 seconds (30 seconds/kb). The last step in the reaction was the final extension at 72 °C for 2 minutes.[12] The PCR product band was visualized on a 1% agarose gel with ethidium bromide. After verifying the product size, it was purified using a GeneJET PCR purification kit (ThermoFisher Scientific; Waltham, MA) so the DNA concentration was between 50 and 100 ng/µL; product concentrations were measured using Nanodrop.

Table 7. Primer Sequences for CFP Amplification.

The lowercase portion represents regions with homology to the plasmid that anneal to the plasmid during the cloning reaction (second PCR). The underlined portion represents the RBS sequence in each strain. The capitalized portion is the part of the primer that anneals to the gene during the first PCR. Annealing temperature for all primer designs in PCR 2 is 65 °C.

Primer	**Sequence**	**Annealing Temperature (°C)**
CFP 1 Forward	*tgagcggataacaattcccc<u>CTTTTGGTTTCAG</u>ATGGCTCTTTCAAACAAG*	57
CFP 2 Forward	*tgagcggataacaattcccc<u>TGCAACTTTTCGGTTTCAGAC</u>ATGGCTCTTTCAAACAAG*	57
CFP 3 Forward	*tgagcggataacaattcccc<u>CAACTTTTCGGTTTCAGAC</u>ATGGCTCTTTCAAACAAG*	57
CFP 4 Forward	*tgagcggataacaattcccc<u>TTCAATAAGGAGGTTTTTT</u>ATGGCTCTTTCAAACAAG*	57
CFP 5 Forward	*tgagcggataacaattcccc<u>CCCCCCTACGGTTAAAAAA</u>ATGGCTCTTTCAAACAAG*	57
CFP Reverse	*gctgctgtgatgatgatgatgaTCAGAAAGGGACAACAGAG*	61

2.2.4.2 λ-Exonuclease Digestion

The purified form of the product was digested overnight in the following 10 µL reaction: 6 µL PCR product, 0.5 µL λ-exonuclease enzyme (New England Biolabs), 1 µL 10X λ-exonuclease buffer (New England Biolabs), and 2.5 µL molecular biology grade water. This reaction was incubated overnight at 37 °C and then deactivated at 75 °C for 10 minutes.[12] The digestion product was visualized on a 1% agarose gel with ethidium bromide to verify the double-stranded DNA had been degraded. All agarose gels visualizing PCR and digestion products can be found in section 2.3.1.

2.2.4.3 Cloning Reactions

The cloning reactions consisted of a second 25 µL PCR containing 0.5 µL of the universal reverse primer (100 µM), 2 µL of the digestion product, the plasmid template (~50 ng/µL), 12.5 µL Q5 ® High-Fidelity 2X Master Mix (New England Biolabs), and ~10 µL of molecular biology grade water. The PCR program was designed with the following procedure: i) denaturation at 98 °C for 30 seconds, followed by 30 cycles of ii) denaturation at 98 °C for 10 seconds, iii) annealing at 65 °C for 30 seconds, and iv) extension at 72 °C for 190 seconds (30 seconds/kb). The last step in the reaction was the final extension at 72 °C for 2 minutes. The PCR product was visualized on a gel and purified using the GeneJET PCR Purification kit (ThermoFisher Scientific). This product was transformed into competent *E. coli BL21(DE3)* cells using the cell manufacturer's heat-shock procedure; the cells were then plated on LB agar with 50 µg/mL kanamycin and incubated at 37 °C overnight.[12]

2.2.4.4 Chemical Transformation of pET28a + CFP plasmid into E. coli BL21(DE3)

The transformation of *BL21(DE3)* competent cells with the pET28a + CFP plasmid was done as described: i) add 2.5 µL of the λ-PCR product (~50 µg of plasmid) to 50 µL of competent cells and mix gently, ii) incubate on ice for 30 minutes, iii) heat shock at 42 °C

for 30 seconds without shaking, iv) immediately transfer to ice for 2 minutes, v) add 250 µL of SOC medium, then vi) incubate culture in shaker at 37 °C and 200 rpm for 1-2 hours, and lastly vii) spread 50 µL of culture on plates with LB + agar + kanamycin (50 µg/mL). The plates were incubated at 37 °C overnight.

2.2.4.5 Verifying Transformations

Positive colonies were established by colony PCR with the following reaction: template provided by the colony being checked, 0.02 µL forward primer (100 µM), 0.02 µL reverse primer (100 µM), 4.96 µL molecular biology grade water, and 5 µL One*Taq* ® Master Mix with standard buffer (New England Biolabs). The PCR program was designed with the following procedure: i) denaturation at 94 °C for 30 seconds, followed by 30 cycles of ii) denaturation at 94 °C for 30 seconds, iii) annealing at 62 °C for 60 seconds, and iv) extension at 68 °C for 60 seconds (60 seconds/kb). The last step in the reaction was the final extension at 68 °C for 5 minutes. The control for these studies was *BL21(DE3)* + pET28a without the CFP reporter. These PCR products were visualized on a 1% agarose gel with ethidium bromide.

Table 8. Primers Used in CFP Colony PCR

Primer	**Sequence**	**Annealing Temperature °C**
pART15_24_F	*aaataaacaaataggggttGCTCATGAGCCCGAAGTG*	66
pART15_AmCyanC1_R	*gatccccgggtaccatggtgTCAGAAAGGGACAACAGAG*	61

2.3 RESULTS AND DISCUSSION

2.3.1 Primer Design Results: NUPACK

2.3.1.1 Megaprimer Thermodynamic Calculations

Primer designs for λ-PCR were checked using NUPACK in two main applications: checking the secondary structure of the megaprimer, and verifying that nonspecific binding would not occur during PCR 2 of the protocol. Details on the procedure for NUPACK analysis are included in section 2.2.1.3. The table below shows an example of the NUPACK output for single-stranded thermodynamics, in this case with the megaprimer.

Table 9. NUPACK Output for Megaprimer Thermodynamics Calculations: Original CFP Forward Primer.

Column one corresponds to the nucleotide in the sequence inputted to the tool. Column two represents an unpaired nucleotide. Column three represents the probability of the nucleotide being unpaired at equilibrium.

Nucleotide	Unpaired?	Probability of Being Unpaired at Equilibrium
1	-1	0.99099
2	-1	0.75155
3	-1	0.94893
4	-1	0.94968
5	-1	0.96476
6	-1	0.97159
7	-1	0.7728
8	-1	0.98321
9	-1	0.88075
10	-1	0.89177
11	-1	0.80805
12	-1	0.848
13	-1	0.87641
14	-1	0.53844
15	-1	0.92851
16	-1	0.5519
17	-1	0.32947
18	-1	0.76025
19	-1	0.71504
20	-1	0.96968

The table above includes the first 20 nucleotides since this is the part of the forward primer sequence with homology to the plasmid. The table below, Table 10, shows the same megaprimer analysis but focuses on the reverse primer portion of the megaprimer.

Table 10. NUPACK Output for Megaprimer Thermodynamics Calculations: Original CFP Reverse Primer.

Column one corresponds to the nucleotide in the sequence inputted to the tool. Column two represents an unpaired nucleotide. Column three represents the probability of the nucleotide being unpaired at equilibrium.

Nucleotide	Unpaired?	Probability of Being Unpaired at Equilibrium
711	-1	0.83052
712	-1	0.79667
713	-1	0.80057
714	-1	0.98623
715	-1	0.97879
716	-1	0.89774
717	-1	0.85719
718	-1	0.96308
719	-1	0.9018
720	-1	0.89188
721	-1	0.97912
722	-1	0.89868
723	-1	0.8816
724	-1	0.98306
725	-1	0.92497
726	-1	0.88523
727	-1	0.98437
728	-1	0.93018
729	-1	0.87859
730	-1	0.9876
731	-1	0.97537
732	-1	0.992
733	-1	0.98541
734	-1	0.97402

Tables 9 and 10 above show the NUPACK output for the megaprimer thermodynamics. Based on these outputs, the original set of primers needed to be redesigned due to nucleotide 14, 16 and 17's low probabilities of being unpaired at equilibrium. The NUPACK output shows the entire megaprimer sequence; however, the regions with homology to the plasmid are the main areas of concern due to their involvement in the

annealing step in PCR 2. NUPACK thermodynamic calculations of the final primer design showed improvements to the design with higher probabilities overall; therefore, that design was more optimized for λ-PCR.

2.3.1.2 Checking Sequences with Homology to Plasmid

The final primer designs for λ-PCR were further checked to ensure no nonspecific binding would occur using NUPACK's ability to simulate complex formation of different sizes. Details on the procedure followed to check nonspecific binding are included in section 2.2.1.4. The table below, Table 11, shows an example of a NUPACK output for this type of analysis. For this particular output, the forward primer was run against the reverse complement of the portion of the plasmid the forward primer binds to during PCR 2.

Table 11 below shows specific base-pair interactions that are thermodynamically favorable to occur when the two strands are simulated together in NUPACK. In this example, strand 1 represents the region with homology to the plasmid within the CFP forward primer and strand 2 represents the reverse complement of the portion of pET28a in which the forward primer was designed to anneal to. Note that strand 2 does not include the entire plasmid sequence since NUPACK cannot run calculations with strands that large, so it was analyzed in multiple simulations. Higher probabilities were desired at particular locations within pET28a since the forward primer needs to have a high probability of binding where it was designed to, but should have lower probabilities elsewhere to limit nonspecific binding during PCR 2 in Figure 3. The desired probabilities for this NUPACK analysis are listed in bold in Table 11. The entire nucleotide sequence for the forward primer, strand 1 nucleotides 1-20, and nucleotides 341-360 on strand 2 had interactions with high probabilities; this was expected given the sequences input to the tool, so this region of homology used in the forward primer is acceptable based on these outputs. Note that the forward primer was run against both the entire pET28a sequence used in the design process and the reverse complement of that sequence.

Table 11. NUPACK Output for Checking for Homologous Regions Between CFP Forward Primer and pET28a.

Columns 1 and 3 list strand ID, and columns 2 and 4 list the nucleotides active in a base-pair forming between each strand. Column 5 lists the probability of the two nucleotides described in columns 1-4 being paired at equilibrium given the inputted conditions. In this case, strand 1 is the CFP forward primer and strand 2 is the reverse complement of the region of the plasmid in which the forward primer was designed for.

Strand 1	Nucleotide (Strand 1)	Strand 2	Nucleotide (Strand 2)	Probability of Two Nucleotides Being Paired at Equilibrium
1	1	1	13	0.00226
1	1	2	360	**0.3492648**
1	1	2	362	0.0024287
1	2	1	12	0.0027255
1	2	1	17	0.0018139
1	2	2	359	**0.5172099**
1	2	2	361	0.0021103
1	3	1	9	0.0090731
1	3	1	15	0.0020825
1	3	1	16	0.0035227
1	3	2	358	**0.7147922**
1	3	2	1155	0.0023376
1	4	1	12	0.0014617
1	4	1	15	0.001222
1	4	1	20	0.0016295
1	4	2	357	**0.7451135**
1	4	2	1154	0.0026491
1	5	2	356	**0.751421**
1	5	2	1153	0.0027092
1	6	1	12	0.0044705
1	6	1	17	0.0019672
1	6	1	18	0.0099127
1	6	1	19	0.0032946
1	6	1	20	0.0022467
1	6	2	355	**0.750661**
1	6	2	1152	0.0027069
1	7	1	12	0.0033116

1	7	1	16	0.0015012
1	7	1	17	0.0112721
1	7	1	18	0.003653
1	7	1	19	0.0025514
1	7	2	354	**0.7456455**
1	7	2	1151	0.0026658
1	8	1	15	0.0065712
1	8	1	16	0.0086286
1	8	2	353	**0.7238354**
1	8	2	1150	0.0020465
1	9	1	3	0.0090731
1	9	1	13	0.0018962
1	9	1	14	0.0033186
1	9	2	352	**0.6782395**
1	9	2	1149	0.0017304
1	10	1	15	0.0047228
1	10	1	16	0.0123827
1	10	2	351	**0.6846229**
1	11	1	15	0.0102828
1	11	1	16	0.0055681
1	11	2	350	**0.7312068**
1	12	1	2	0.0027255
1	12	1	4	0.0014617
1	12	1	6	0.0044705
1	12	1	7	0.0033116
1	12	2	349	**0.7410099**
1	13	1	1	0.00226
1	13	1	9	0.0018962
1	13	2	348	**0.7305228**
1	14	1	9	0.0033186
1	14	2	347	**0.7072689**
1	15	1	3	0.0020825
1	15	1	4	0.001222
1	15	1	8	0.0065712
1	15	1	10	0.0047228
1	15	1	11	0.0102828
1	15	2	346	**0.7073449**
1	16	1	3	0.0035227

1	16	1	7	0.0015012
1	16	1	8	0.0086286
1	16	1	10	0.0123827
1	16	1	11	0.0055681
1	16	2	345	**0.7227715**
1	17	1	2	0.0018139
1	17	1	6	0.0019672
1	17	1	7	0.0112721
1	17	2	344	**0.7388061**
1	18	1	6	0.0099127
1	18	1	7	0.003653
1	18	2	343	**0.7405539**
1	19	1	6	0.0032946
1	19	1	7	0.0025514
1	19	2	342	**0.7252033**
1	20	1	4	0.0016295
1	20	1	6	0.0022467
1	20	2	341	**0.6238283**

Based on the NUPACK results for the final primer sequences shown in Table 2, these designs should function correctly in λ-PCR since the regions of homology in the megaprimer generated from PCR 1 (Figure 3) are unpaired at equilibrium and the regions of homology only bind to the correct regions of pET28a with high probabilities.

2.3.2 CFP1-5 Strain Construction

The products of each PCR reaction outlined in section 2.2.2 were checked using 1% agarose gels to ensure the correct products were amplified. The sections below include graphics of each gel as well as details on product band size as well as explanations for any other bands that may be present in the band.

2.3.2.1 Gene Amplification

The first step in the λ-PCR protocol is the amplification of the target gene to generate a dsDNA product; this step adds the regions with homology to the plasmid as overhangs

that increase the length of the amplified product. Figure 5 below shows the amplification the CFP gene to be used in the construction of CFP1, CFP2, and CFP3. Figure 6 shows the amplification of the CFP gene to be used in CFP4 and CFP5 construction. The length of the CFP gene is 690 nucleotides; therefore, the product band was 750 nucleotides due to the 60 nucleotides added from the primers and RBS sequences.

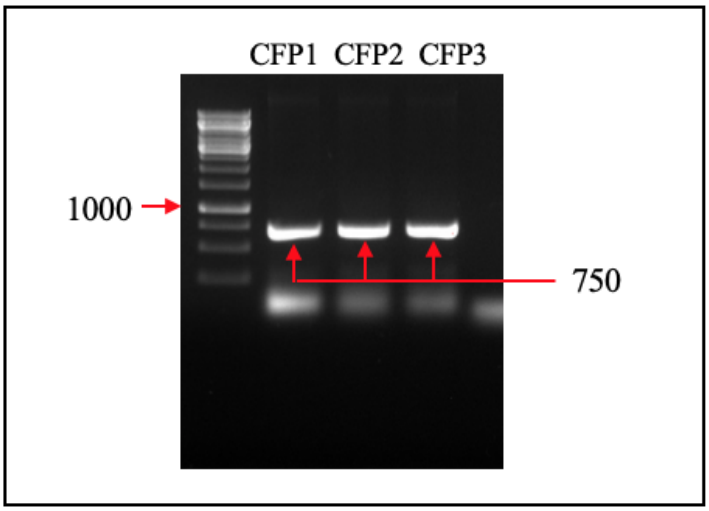


Figure 5: Gel electrophoresis of PCR 1 products in the λ-PCR protocol for CFP1-3. CFP gene length is 690 nucleotides, primer overhangs add 50-60 nucleotides depending on the RBS sequence in the strain.

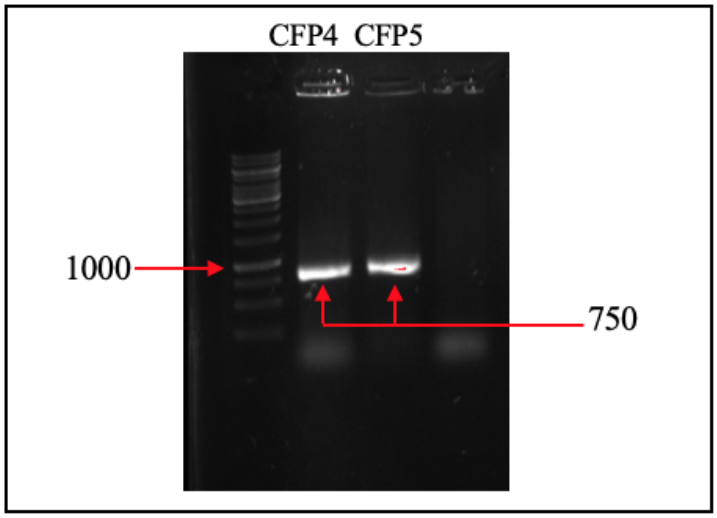


Figure 6: Gel electrophoresis of PCR 1 products in λ-PCR protocol for CFP4 and CFP5.

CFP gene length is 690 nucleotides and primer overhangs add 50-60 nucleotides depending on the RBS sequence in the strain.

2.3.2.2 Digestion and λ-PCR Products

The λ-PCR products for CFP strains 1-3 are shown in the figure below. The target product for this reaction was ~6100 nucleotides; this is the faint band appearing on the gel above the ladder band 5000. The thick bands near a band length of 3000 are potentially the linear products of nonspecific binding in the reaction or the supercoiled form of the plasmid on the gel since supercoiled products travel through the gel faster.[31] Regardless, the only product that will transform into the competent cells is the complete plasmid product at 6100 nucleotides.

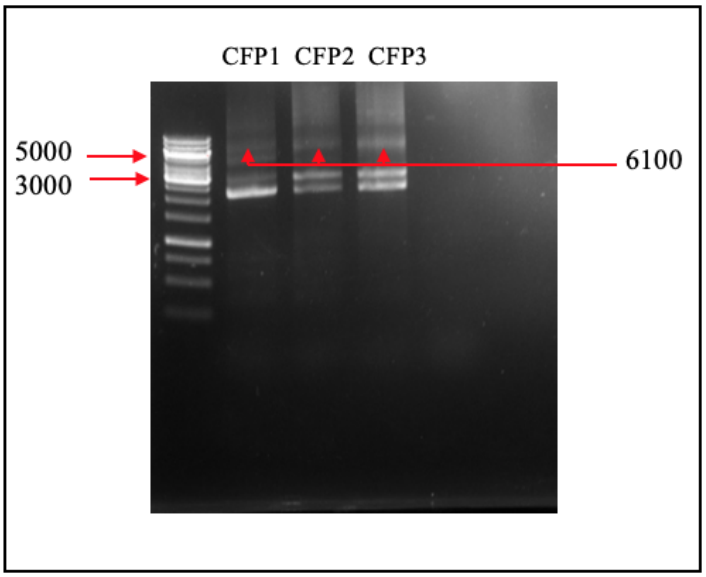


Figure 7: Gel electrophoresis of PCR 2 products in the λ-PCR protocol for CFP1-3. Thick bands near the band length 3000 are likely the supercoiled form of the plasmid or linear PCR products resulting from nonspecific binding. The target λ-PCR product is the faint band with a length of ~6100.

The second step in the protocol is the digestion of the gene amplification product overnight. Figure 8 shows the digestion products for CFP4 and CFP5 in the two lanes on the left. The appearance of digestion products on gels varies from a smear, due to the products being ssDNA, to bands smaller than the original product. This gel indicates the digestion reaction was successful. The λ-PCR products in the two lanes on the right show a product with a band size of 3000; this is likely the supercoiled form of the plasmid or a PCR product resulting from nonspecific binding of the primers since supercoiled products travel through gels faster.[31] The product was immediately transformed into *E. coli BL21(DE3)* competent cells so any products other than the plasmid at 6100 nucleotides will not successfully transform.

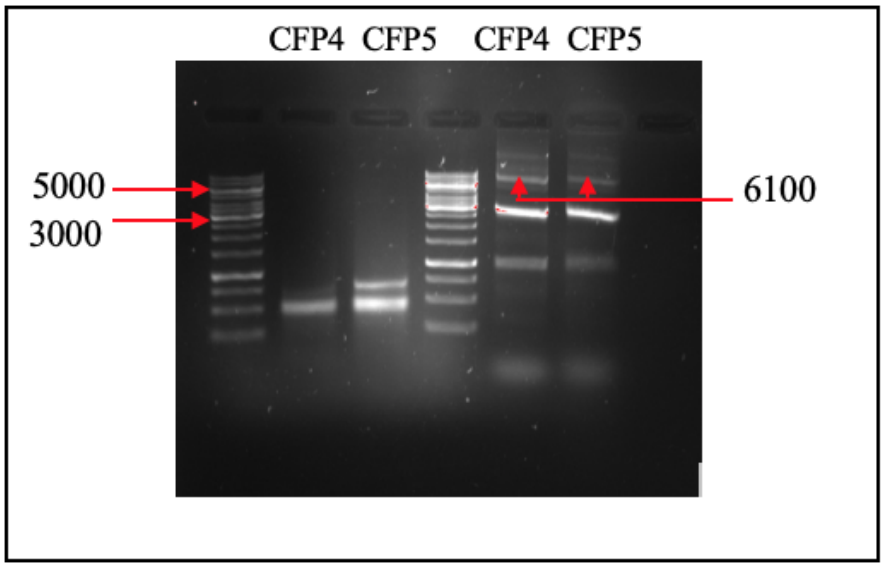


Figure 8: Gel electrophoresis of digestion and λ-PCR products for CFP4 and CFP5. Digestion products are in the first two lanes on the left and λ-PCR products in the last two lanes on the right for CFP strains 4 and 5. Thick bands near the band length 3000 are likely the supercoiled form of the plasmid or linear PCR products resulting from nonspecific binding. The target λ-PCR product is the band with a length of ~6100.

2.3.2.3 Colony PCR

Figure 9 shows gels for CFP1 and CFP2 colony PCR products. The CFP gene is 690 nucleotides in length; band length was an additional ~200 nucleotides since one primer annealed to the plasmid and one primer annealed to the 3' end of the CFP gene.

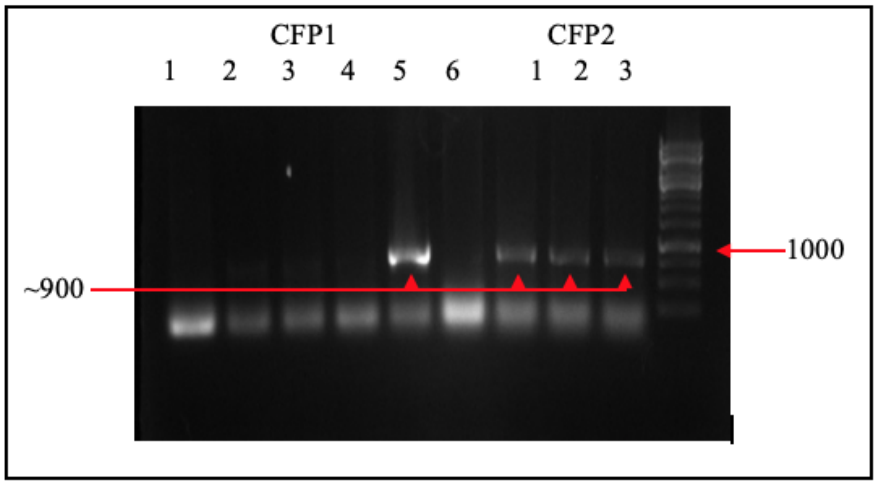


Figure 9: Gel electrophoresis of the colony PCR products for CFP1 and CFP2.

The target bands were about 900 nucleotides in length; this includes the length of the CFP gene, 690 nucleotides, plus about 200 nucleotides of the plasmid.

Figure 10 shows gels for CFP2 and CFP3 colony PCR products in a 1% agarose gel. The CFP gene is 690 nucleotides in length; band length was an additional ~200 nucleotides since one primer annealed to the plasmid and one primer annealed to the 3' end of the CFP gene.

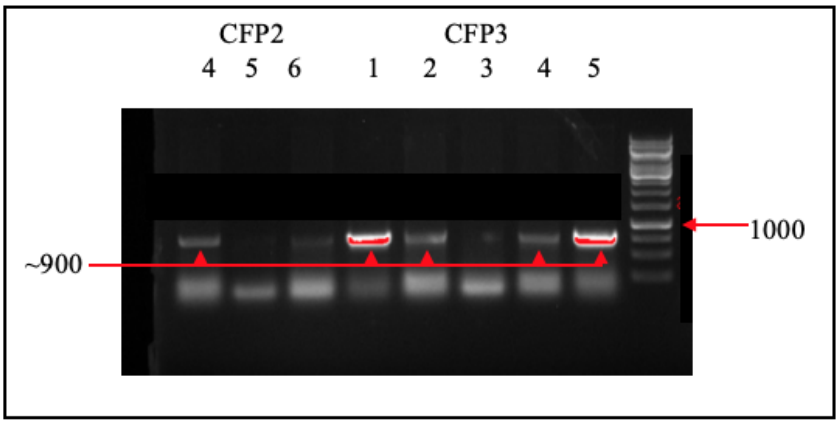


Figure 10: Gel electrophoresis of the colony PCR products for CFP2 and CFP3. The target bands were about 900 nucleotides in length; this includes the length of the CFP gene, 690 nucleotides, plus about 200 nucleotides of the plasmid.

Figure 11 shows the colony PCR products for CFP4 and CFP5 in a 1% agarose gel. The target bands were about 900 nucleotides in length, including the 690-nucleotide CFP gene with 200-nucleotide extensions due to where the forward primer anneals on the plasmid.

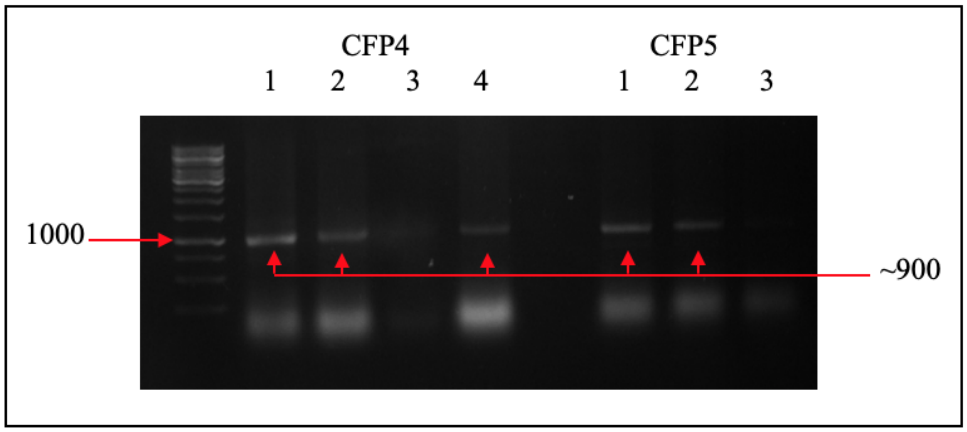


Figure 11: Gel electrophoresis of the colony PCR products for CFP4 and CFP5. The target bands were about 900 nucleotides in length; this includes the length of the CFP gene, 690 nucleotides, plus about 200 nucleotides of the plasmid.

Each of the CFP strains were constructed with primers designed using the method outlined in the previous sections of Chapter 2. The first set of primers designed for each gene and RBS proved to be successful in cloning the gene into *BL21(DE3)* on the first attempt; therefore, this shows this primer design method for λ-PCR is effective and efficient.

2.3.3 Cloning Enzymes with λ-PCR: S-adenosylmethionine (SAM) Synthase and Formaldehyde Ferredoxin Oxidoreductase

A number of projects in the Senger lab investigate heterologous protein expression for enzymes to be assembled in an *in vitro* cascade. The enzymes cloned using λ-PCR in this study are formaldehyde oxidoreductase (FOR) sourced from *Pyrococcus furiosus* and S-adenosylmethionine synthetase (SAM synthetase) sourced from *Thermatoga maritima*. The λ-PCR primer design process for each of these enzymes is outlined previously in section 2.2.2, and the cloning protocol used for each enzyme is described in detail with CFP cloning in section 2.2.4. The following sections include graphics of the products in a 1% agarose gel.

2.3.3.1 Gene Amplification

The first step in the λ-PCR protocol is to clone the target gene using chimeric primers to generate millions of copies of the double-stranded target gene with overhangs which generate the megaprimer used in the second PCR. Figure 12 shows the product bands for the first PCR for the two enzymes described previously.

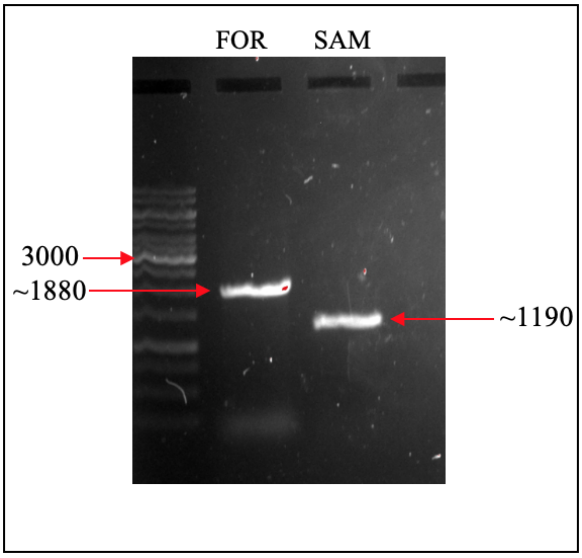


Figure 12: Gel electrophoresis of PCR 1 products in the λ-PCR protocol for SAM synthase and FOR enzymes.

The SAM synthase gene length is about 1190 nucleotides, the FOR gene length is about 1880 nucleotides, and primer overhangs add ~40 nucleotides to each of those products.

2.3.3.2 λ-PCR Products

The products from the first PCR were digested overnight; these products were used as the megaprimer in the second PCR step. The products are shown in Figure 13 below. The target product bands are highlighted in the figure with red arrows. The pET28a plasmid with the FOR gene was about 7300 nucleotides in length, and the pET28a plasmid with the SAM synthase gene was expected to be about 6600 nucleotides. The other product bands are likely the coiled forms of the plasmid or dsDNA products resulting from nonspecific binding; neither of these products will transform into the competent cells, so this reaction was considered successful.[31]

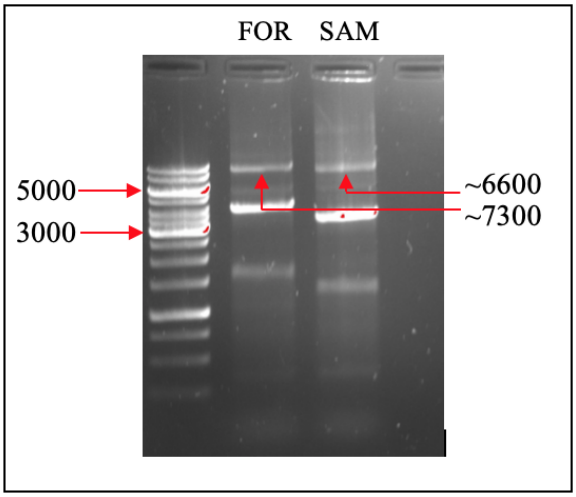


Figure 13: Gel electrophoresis of the λ-PCR products (PCR 2) for the SAM synthase and FOR enzymes.

Thick bands between the band lengths 3000 and 5000 are likely the supercoiled form of the plasmid or linear PCR products resulting from nonspecific binding. The target λ-PCR product are the bands with the following lengths: 6600 nucleotides for SAM synthase and 7300 for FOR.

2.3.3.3 Colony PCR

Figures 14 and 15 below show the colony PCR products for each enzyme after transformation into *BL21(DE3).* The target product band was slightly larger in size than the gene due to the forward primer binding to pET28a upstream of the start codon for each gene with the reverse primer annealing to the 3' end of the target gene. Both colony PCRs produced positive colonies with the original primer designs. Cloning SAM synthase and FOR cloning into *BL21(DE3)* was successful with the first set of primers designed using the method outlined previously.

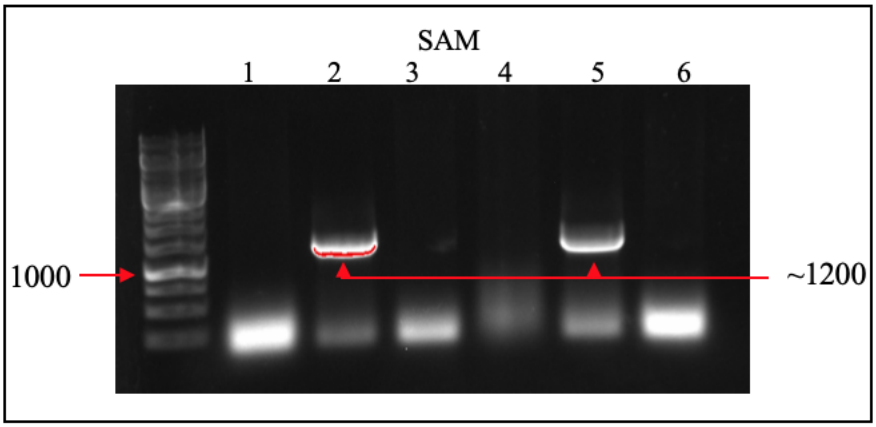


Figure 14: Gel electrophoresis of colony PCR products for SAM synthase.

The target bands shown above are about 1200 nucleotides in length, slightly larger in size than the PCR 1 product since the forward primer binds to the plasmid upstream of the gene's start codon.

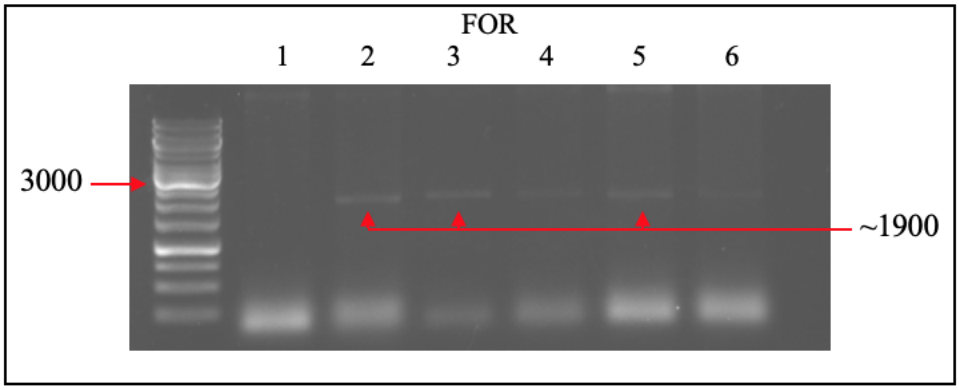


Figure 15: Gel electrophoresis of the colony PCR products for the FOR enzyme.

The target bands shown above are about 1900 nucleotides in length, slightly larger in size than the PCR 1 product since the forward primer binds to the plasmid upstream of the gene's start codon.

3 SYNTHETIC RBS STUDIES

3.1 INTRODUCTION

A number of methods can be used to improve heterologous protein expression, including codon optimization to circumvent codon bias and modifying the 5' and 3' ends of mRNA in a variety of ways to improve stability.[21,32,33] These approaches have been largely successful; however, drawbacks have been identified with these methods to maximize protein expression. Codon optimization has been shown to impact protein folding and function due to key amino acid substitutions.[32,34] Stabilizers modified in sequences upstream of the start codon or added to the 3' end of mRNAs become ineffective in certain environments. In one example, modifying the sequence upstream of the target gene improved mRNA stability in *E. coli*, but only in T4-infected cells.[33,35] Synthetic RBS sequences improve protein translation by modifying the sequence upstream of the target gene to increase translation initiation, the rate-limiting step of protein translation.[15] This method of improving protein yields can be applied in any prokaryotic species, assuming the ribosomal RNA sequence is known, and does not modify the final protein since all changes occur in the untranslated region of the mRNA.[13, 15]

Designing a synthetic RBS upstream of a protein coding sequence allows for greater control over protein expression by directly impacting translation initiation, but it does not impact protein structure nor function; as a result, it is a desirable method to use for heterologous protein production with enzymatic cascades.[13] The initial goal of this project was to evaluate the accuracy of a tool used for predicting and designing RBS sequences in *E. coli BL21(DE3),* the Salis Lab RBS Calculator, and to use the results to optimize protein production for a cascade designed to convert citrate into fumarate. Fluorescent proteins, in particular the cyan fluorescent protein, was used to evaluate the tool since fluorescence is proportional to the TIR once the culture reaches steady state growing conditions.[19] The initial assay results combined with changes to the RBS Calculator drove new questions regarding the 16S rRNA sequences that may be present in our particular strain of *E. coli.* The Salis Lab removed the *BL21(DE3)* strain with 16S rRNA sequence *TAACCGTAG* from the RBS Calculator; this sequence was used to

design RBS sequences 1-3 in this project. Consequently, the designs were run in the Prediction tool with the *BL21(DE3)* strain with 16S rRNA sequence *ACCTCCTTA* to show how the strains should behave, assuming the latter 16S rRNA sequence is the only sequence present in the strain. The tool's outputs revealed predicted TIRs less than 50 au for each strain, indicating each RBS should produce little protein. Given the high fluorescence levels received with CFP1, containing an RBS with a predicted TIR of 12 au, further experiments were designed and executed in an attempt to pinpoint what may be occurring in our strain of *BL21(DE3)*.

A number of methods were investigated to observe which RBS characteristics, primarily focusing on the Gibbs free energy of a number of interactions, have the greatest impact on protein expression. Additionally, a number of factors were tested experimentally to determine the optimum way to maximize protein expression in *BL21(DE3)*.

3.2 MATERIALS AND METHODS

3.2.1 Design Procedure for Strains CFP1-3

The process employed to design the primers used to construct strains CFP1-3 with λ-PCR is outlined in section 2.2.1. The same pair of primers, shown in Table 2, was used for each CFP strain, including CFP4 and CFP5; the only difference between the CFP strains is the synthetic RBS sequence added as a linker sequence in the forward primer. The synthetic RBS sequences were designed to assess protein expression at different levels using the cyan fluorescent reporter.

3.2.1.1 Salis Lab RBS Calculator

The RBS Calculator is an algorithm developed by the Salis lab at Pennsylvania State University that assesses heterologous protein expression in bacteria and designs new RBS sequences if protein expression is poor. For the course of this project, the threshold for adequate protein translation was chosen to be 1000 au. This section includes the use of the RBS Calculator's Predict and Design tools in λ-PCR primer design for cyan fluorescent protein expression in *E. coli*; details regarding the algorithm's free energy calculations in the RBS Calculator is in section 1.1.

The performance of gene expression in a new host organism can be assessed using the Prediction tool in the RBS Calculator by inputting at least the first 35 nucleotides of a gene's mRNA sequence and selecting the expression organism.[13] For this project, the following 80-nucleotide sequence was used to predict CFP translation in *E. coli BL21(DE3)*. This sequence is the first 80 nucleotides of the CFP gene:
5'- *ATGGCTCTTTCAAACAAGTTTATCGGAGATGACATGAAAATGACCTACCATATGG ATGGCTGTGTCAATGGGCATTACTT* – 3'

After submitting the mRNA sequence and selecting the organism in the algorithm, the RBS Calculator outputs predicted Translation Initiation Rates (TIR) for every start codon found in the mRNA sequence. If the predicted TIR for the correct start codon is greater than 1000 au, then designing a synthetic RBS is unnecessary; if the predicted TIR falls

under this threshold, then a synthetic RBS must be added to the forward primer for this design to ensure adequate protein expression. The mRNA sequence for the CFP gene had a predicted TIR below 1000 au, so synthetic RBS sequences were designed and added to the primer designs.

Synthetic RBS sequences can be designed downstream of the promoter and upstream of a gene's start codon to control protein expression. The design of these synthetic RBS sequences is complex and is completed with the use of the RBS Calculator's Design tool. The first 80 nucleotides of the mRNA sequence and the 40 nucleotides directly upstream of the insertion site on the plasmid were inputted to the tool as the Protein Coding Sequence and Pre-Sequence, respectively. Additionally, the tool requires a chosen target TIR for any chosen level of protein expression. The tool then outputs a synthetic RBS sequence that meets the criteria as best as possible. The output includes the position of the start codon with the synthetic RBS, TIR, ΔG_{total}, $\Delta G_{mRNA\text{-}rRNA}$, $\Delta G_{spacing}$, $\Delta G_{standby}$, ΔG_{start}, ΔG_{mRNA}, and a link to the initial secondary structure of the first ~60 nucleotides of the mRNA at equilibrium at 37 °C; descriptions of each Gibbs free energy term are included in the section 1.1. Accounting for strong secondary structure formation is a key step for accurately predicting protein translation since any stable structures occurring in the RBS region can prohibit the 30S subunit from binding and initiating translation.[36] If there are additional constraints that must be included in the design inputs, the "Design with Constraints" tool should be used. This interface is set up similarly to the Design tool; however, it contains an additional input that can be used to specify the length of the RBS sequence in addition to any nucleotides that must be included in the design.

Three synthetic RBS sequences with different target TIRs were designed for the CFP using the Salis Lab RBS Calculator to assess protein expression based on the rates produced by the tool. The inputs to the Design tool included: the first 80 nucleotides of the CFP gene, listed previously; three targeted translation rates evenly spaced apart; and 40 nucleotides directly upstream of the insertion site, 5'-*ATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCAGCA*-3'. The original

synthetic RBS sequences and their corresponding predicted translation rates are listed in the table below.

Table 12. Initial CFP RBS Designs and Predicted TIR

RBS	**Sequence**	**Predicted TIR (au)**
CFP1	*TTAACAATTCCCCTGGTTATTTTT*	1029
CFP2	*GCATCCTGCGGCCTAAAT*	3311
CFP3	*CCCAGACCACCTACATCTTTTTTA*	2002

3.2.1.2 NUPACK: Checking Megaprimer Thermodynamics

After creating a synthetic RBS sequence, the RBS was added to the forward primer in the New England Biolabs NEBuilder Assembly Tool®; the RBS is added to the forward primer using the "Spacer Sequence" option built into the Assembly tool. This did not change the annealing temperature, but added the synthetic RBS sequence to the forward primer at the 5' end of the target gene primer and the 3' end of the region with homology to the plasmid. After the primer sequences have been updated, NUPACK was used to check the single-stranded megaprimer's secondary structure during the second PCR of the λ-PCR protocol to verify the addition of a synthetic RBS did not prevent the megaprimer from annealing correctly during PCR 2 in Figure 3. The synthetic RBS sequences proved to generate unwanted secondary structure in the megaprimers, so these sequences were redesigned to ensure proper plasmid construction. The final RBS sequences for CFP1-3 and their predicted translation rates are displayed below in Table 13.

Table 13. Final RBS Designs for CFP1-3

The 16S rRNA sequence used during this design process with the RBS Calculator was *TAACCGTAG.*

RBS	**Sequence**	**Predicted TIR (au)**
CFP1	*CTTTTGGTTTCAG*	500
CFP2	*TGCAACTTTTCGGTTTCAGAC*	1500
CFP3	*CAACTTTTCGGTTTCAGAC*	1000

Details regarding the primer design process for these three strains are included in Chapter 2, section 2.2.1. The final primer designs, including the synthetic RBS in the forward primer, are shown in Table 7 in section 2.2.4.1.

3.2.2 Design Procedure for Strains CFP4 and CFP5

3.2.2.1 Salis Lab RBS Calculator Modifications

One of the key factors to verify when assessing protein translation is to check the binding free energies associated with the mRNA-rRNA complex because translation initiation is the limiting factor.[19] As a result, it is important to know the 16S rRNA sequence in the system. Some organisms listed in the RBS Calculator contain multiple selections for the same organism with the only difference between the two selections being the rRNA sequence listed in parentheses next to the species name. This is due to some organisms potentially having multiple 16S rRNA sequences present in their genomes.

The Salis Lab RBS Calculator designs sequences by finding the minimum free energy (MFE) with the thermodynamic relationship outlined in section 1.1 for each iteration of a proposed RBS sequence while satisfying the RBS Calculator inputs as best as possible. The thermodynamic model relies heavily on the accuracy of the 16S rRNA sequence used to model the MFE for each potential RBS.[16] Consequently, any inaccuracies in the

16S rRNA sequences would have detrimental effects on the RBS Calculator's predictions.

Recent changes to annotations in the NCBI genome for *E. coli BL21(DE3)* cascaded changes to the RBS Calculator, affecting the RBS Calculator's TIR predictions for designs CFP1-3. As a result, running the three designs in the Prediction tool with the sequence currently suspected to be the only 16S rRNA sequence present in *BL21(DE3)* showed each strain producing very little of the CFP reporter, directly contradicting experimental results for strains CFP1 and CFP3.

These changes to the RBS Calculator along with the initial results of the plate reader assays for CFP strains 1, 2, and 3, described in section 3.3, compared to the outputs from the Salis lab RBS Calculator drove a new question: are there multiple sequences for the 3' end of the 16S rRNA in our strain of *E. coli BL21(DE3)*? Moving forward, two new strains were designed so each would produce high levels of protein by binding to only one of the 16S rRNA sequences suspected to be present in our strain of *E. coli BL21(DE3).* The following section entails how these strains were designed with respect to the first three CFP strains.

3.2.2.2 CFP4 and CFP5 RBS Design

The extensity of the λ-PCR primer design process ensures that the primers will be successful at cloning new plasmids using a megaprimer, constructed from the gene template, and a template plasmid. The success achieved during cloning for the first three strains resulted in similar primers being used for constructing strains 4 and 5 with only one difference: the RBS sequences (spacer sequence in the forward primer) were changed to fit the goals of the assays for strains 4 and 5. As a result, the same chimeric primers were chosen at the start of the design process so the same insertion site was targeted for the new CFP strains; changes to the NEBuilder Assembly only affected the "Spacer" sequence in the forward primer. There were a few changes to the design process for the

primers for strains 4 and 5 since these RBS sequences were designed using the RBS Calculator with some manual changes due to the updates to the tool.

The Design tool in the RBS Calculator generates a number of outputs: $\Delta G_{mRNA\text{-}rRNA}$, ΔG_{start}, $\Delta G_{standby}$, $\Delta G_{spacing}$, and ΔG_{mRNA}. Following a similar structure, designs for RBS sequences 4 and 5 were broken down into the following segments: standby sequence, binding sequence, and spacer sequence; Figure 16 shows this setup. More details regarding these RBS Calculator outputs are included in the section 1.1. RBS sequences 4 and 5 were designed to have the same free energy associated with the design components with the only difference between the two sequences being the binding region for the mRNA and rRNA, noted as $\Delta G_{mRNA\text{-}rRNA}$. Therefore, their outputs were designed to have the same free energies associated with the standby, spacer, and mRNA sequences with different values for the mRNA-rRNA binding interaction. The spacing and standby sequences, represented by five nucleotides on the 5' and 3' ends of the RBS sequence, were optimized in each design so the energetics of these sequences between the two strains were as similar as possible. It was previously reported that a spacer sequence length of five nucleotides is optimal for protein expression, so the spacer sequences in each RBS contained five nucleotides.[18] Additionally, they were designed so they had similar free energies as predicted by the RBS Calculator in order to prevent unwanted sources for differences in performance between the two designs. The binding sequence for each RBS was created using the reverse complement of the 3' end of the two 16S rRNA sequences assessed with these strains. Additionally, ΔG_{mRNA} was designed to be similar between the two strains since any differences in the mRNA transcripts' secondary structures near the RBS can severely decrease protein expression, thus adding another variable to the system. [19, 36] The table below displays the outputs in the RBS Calculator for the design of RBS sequences 4 and 5.

Table 14. Salis Lab Outputs for CFP4 and CFP5

RBS	rRNA	TIR	ΔG_{total}	$\Delta G_{mRNA-rRNA}$	$\Delta G_{spacing}$	$\Delta G_{standby}$	ΔG_{start}	ΔG_{mRNA}
4	original	425	2.37	-17.85	1.52	4.06	-2.76	-17.91
5	original	27929	-6.93	-23.7	0.0	3.68	-2.76	-16.03
4	new	40959	-7.78	-27.45	0.0	5.56	-2.76	-17.91
5	new	413	2.43	-13.68	0.01	2.97	-2.76	-16.03

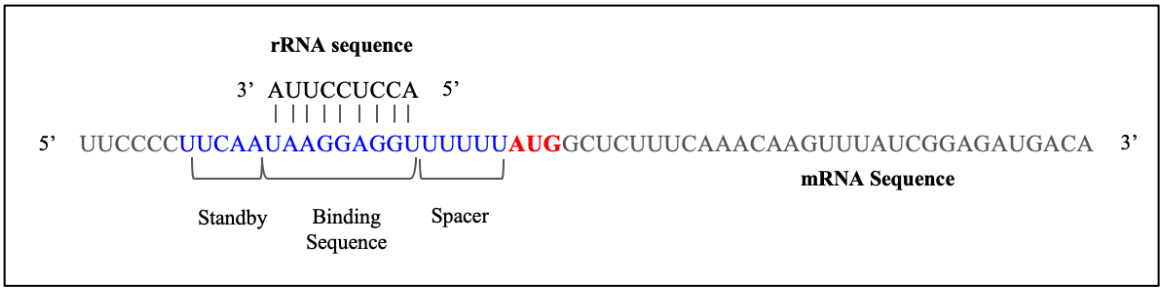


Figure 16: RBS 4 and 5 Breakdown of Design Components [16]

3.2.2.3 NUPACK

After constructing the primer sequences, including the synthetic RBS sequences, the megaprimer thermodynamics were evaluated at the primer annealing temperature using NUPACK to ensure the synthetic RBS did not cause undesirable folding to occur during the cloning reaction. The regions with homology to the plasmid (~20 nucleotides on the 5' and 3' ends) were checked to verify these nucleotides were unpaired at equilibrium. These results showed the megaprimer designs were suitable for the cloning reaction. The region of the primers homologous to the plasmid were previously checked during the design of strains CFP1-3 to ensure no other binding occurred between the megaprimer and the plasmid, so this portion of the design process in NUPACK was omitted for strains CFP4 and CFP5.

3.3.3 Further Analysis: Modeling Translation Initiation

3.3.3.1 NUPACK Thermodynamic Analysis of mRNA-rRNA Complex

During translation initiation, the 3' end of the 16S rRNA strand binds to the mRNA sequence at the RBS.[37] The strength of the interactions in this complex is correlated to the strength of the synthetic RBS; however, these interactions must occur in precise locations on each nucleic acid for translation to initiate correctly.[13]

Experimental results for strains CFP1-3 outlined in section 3.3.1 consistently revealed a relationship different from what was predicted with the Salis Lab RBS Calculator; strain CFP2 produced a similar amount of protein as the control, a blank pET28a plasmid, and CFP1 produced the most protein. As a result, simulations of the mRNA-rRNA complex and the mRNA transcript alone were performed to see if secondary structures in the mRNA prohibited the rRNA from binding properly, or if the rRNA was binding somewhere upstream or downstream of the RBS.

The mRNA sequence began where transcription occurs under the T7 promoter, which for this design included 29 nucleotides upstream the 5' end of the RBS.[38] Variations of the mRNA were simulated in NUPACK in an effort to observe what may be happening at the

molecular level. These variations included simulations involving strictly the RBS; simulations of the RBS with up to 29 nucleotides upstream; simulations of the RBS, up to 29 nucleotides upstream of the RBS, and the first 30 nucleotides of the CFP gene; and simulations of the complete mRNA transcript (up to 29 nucleotides upstream, the RBS, and the entire CFP gene). The rRNA sequences in each simulation were entered into NUPACK as they appeared on the Salis Lab RBS Calculator; simulations were performed for both rRNA sequences suspected to be present in our strain of *BL21(DE3)*, the Shine-Dalgarno sequence and the non-canonical rRNA sequence. For each simulation, the mRNA and rRNA nucleotide strands were inputted into NUPACK as separate strands, and the maximum complex size was chosen to be "2." These simulations were done at a range of temperatures, and appropriate concentrations for each molecule were chosen in NUPACK. NUPACK results show the bonds most thermodynamically favorable to form between the nucleotide strands and the probability of each bond occurring at the selected temperature at equilibrium. The probability of the bonds forming at equilibrium and the free energy of the complex reflect the strength of the RBS sequence with regard to the rRNA used in the simulations.

3.3.3.2 Salis Lab RBS Calculator

The RBS Calculator underwent modifications during the period it was used to design RBS sequences for CFP expression; these changes are described in detail in section 3.2.2.1. As a result, all CFP designs were run in the RBS Calculator against both potential 16S rRNA sequences by typing in the rRNA sequence to the tool rather than selecting it from the list of organisms provided. This generated predicted TIRs for each CFP design with both 16S rRNA sequences potentially present in our lab's strain of *BL21(DE3)*. These outputs were compared to the experimental data for each design as well as the NUPACK mRNA-rRNA analyses to see which predictions most closely resembled the experimental data for our system.

Using the Prediction tool, the CFP sequences were input to the tool as follows: 35 nucleotides directly upstream of the insertion site on pET28a, the synthetic RBS

sequence, and the first 80 nucleotides of the CFP gene. The nucleotides upstream of the RBS were included in the Prediction tool since this tool finds all start codons in an mRNA transcript no matter how far downstream in the mRNA transcript they are located. Additionally, the 35 nucleotides upstream of the RBS were included since this region can impact translation initiation, and the RBS has been shown to include at least 35 nucleotides upstream of the start codon.[13] The start codon beginning at the 5' end of the CFP gene was used to compare predicted TIR for each CFP strain.

3.3.4 Buffer and Media Preparation

Liquid Lennox Broth (LB) media was used in the plate reader assays. A 50 mM stock solution of Isopropyl β-D-1-thiogalactopyranoside (IPTG) was prepared previously and diluted for experiments.

3.3.5 *In vivo* Assays

3.3.5.1 Microplate Reader Assays: Constant IPTG Concentration

All *in vivo* assays were performed for a minimum of 24 hours in 96 well microplates with a Synergy H4 microplate reader (BioTek; Winooski, VT); this microplate reader kept temperature and shaking speed consistent throughout the assays. Culture growth was measured with Optical Density (OD) measured at 600 nm, and AmCyan fluorescence was measured with excitation and emission wavelengths of 458/489 nm. Each CFP strain and the control were grown overnight (~16 hours) in liquid LB media with 50 µg/mL of kanamycin at 37 °C and 200 rpm prior to each plate reader assay. Each *in vivo* assay reinoculated a small volume (~1 µL/1 mL fresh LB + 50 µg/mL kanamycin) in fresh media. The newly inoculated culture was vortexed to ensure the bacteria spread evenly throughout the fresh media. The first *in vivo* assay contained newly inoculated culture with 50 µM IPTG to induce CFP expression; each replicate was induced at the start of the assay. The assay was performed at 30 °C and 150 rpm. The second *in vivo* assay included the newly inoculated culture and employed the same conditions; in addition, half of the 200 µL wells were induced at the start of the assay and the other half were induced 8 hours into the assay. The third *in vivo* assay included the same design and conditions as the second assay apart from an increased IPTG concentration to 3 mM.

3.3.5.2 Microplate Reader Assays: Variable IPTG Concentration

The remaining *in vivo* assays contained the same experimental setup as the previous *in vivo* assays, however variable IPTG concentrations were used in an attempt to find the optimal IPTG concentration for heterologous protein expression. IPTG concentrations used in these assays include: 1 µM, 10 µM, 50 µM, 100 µM, 1 mM, and 3 mM. OD and

CFP fluorescence were measured for a minimum of 24 hours. The fourth *in vivo* experiment was performed at 30 °C and 150 rpm while the fifth occurred at 37 °C and 150 rpm.

3.3.5.3 Evaluating Fluorescence Levels

Relative fluorescence was calculated for each replicate in each *in vivo* plate reader using Equation 1 shown below.

$$Relative\ Fluorescence = \frac{Total\ Fluorescence}{Optical\ Density} \quad \textbf{(1)}$$

Measurements for cultures induced at the start of the assay were taken once the culture reached the stationary phase; if the culture was directly compared to cultures induced after the assay began, then fluorescence and OD measurements at 18 hours were used. Measurements for cultures induced after the assay began were taken during the stationary phase, 18 hours after induction.

3.4 RESULTS AND DISCUSSION

3.4.1 Factors Impacting Cyan Fluorescent Protein Expression

Tables 15, 16, 17, and 18 below show the relative fluorescence in each well in the *in vivo* experiments. Strain, induction time, IPTG concentration, and temperature are the four main factors that varied throughout the experiments. The following sections include statistical analysis on the results, showing which factors were statistically significant and contrasts for variables within each factor. Full graphs displaying OD and fluorescence throughout each experiment are available in Appendix C.

Table 15. Relative fluorescence for each CFP strain, In Vivo Experiments 1-3

		CTRL	CFP1	CFP2	CFP3
***In Vivo* trial 1, 50 µM IPTG**	**induced @ 0 hrs**	72.38	737.96	43.69	225.02
		65.47	853.17	43.16	176.94
		26.75	940.88	77.13	274.59
		28.97	721.94	24.16	231.00
	Average	**48.39**	**813.49**	**47.03**	**226.89**
***In Vivo* trial 2, 50 µM IPTG**	**induced @ 0 hrs**	73.80	1125.87	47.05	976.98
		8.37	959.09	32.93	541.36
		45.30	1140.24	54.26	745.47
		50.86	1177.85	34.51	925.32
	Average	**44.58**	**1100.76**	**42.19**	**797.28**
	induced @ 8 hrs	38.67	451.45	32.62	131.38
		15.38	196.13	16.03	223.71
		40.72	527.68	33.44	257.72
		54.87	614.62	29.66	299.29
	Average	**37.41**	**447.47**	**27.94**	**228.02**
***In Vivo* trial 3, 3 mM IPTG**	**induced @ 0 hrs**	13.45	710.85	10.14	6.59
		22.89	890.86	39.11	25.50
		32.67	1695.92	2.60	23.12
		43.60	1797.09	25.97	0.95
	Average	**28.15**	**1273.68**	**19.45**	**14.04**
	induced @ 8 hrs	11.14	477.94	27.66	29.62
		13.70	478.30	37.67	17.80
		47.74	475.00	30.00	30.52
		43.82	568.73	28.38	31.75
	Average	**29.10**	**499.99**	**30.93**	**27.42**

Table 16. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 4 & 5

	induced @ 0 hrs	**1 µM**	**10 µM**	**50 µM**	**100 µM**	**1 mM**	**3 mM**
***In Vivo* 4 (30 °C)**	**CTRL**	15.20	33.42	50.16	60.27	35.53	46.84
	CFP1	118.90	410.83	1298.32	1490.11	1655.71	1480.26
	CFP2	13.95	16.04	37.00	66.95	58.76	27.49
	CFP3	509.39	352.69	609.39	679.30	779.41	521.14
***In Vivo* 5 (37 °C)**	**induced @ 0 hrs**	**1 µM**	**10 µM**	**50 µM**	**100 µM**	**1 mM**	**3 mM**
	CTRL	38.02	44.44	55.49	24.10	20.81	29.88
	CFP1	122.07	213.56	1001.22	1129.57	162.90	138.65
	CFP2	44.70	14.42	17.97	29.58	40.60	1.95
	CFP3	779.16	471.77	750.41	757.28	890.32	895.33

Table 17. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 6 & 7

	Induced @ 0 hrs	**CTRL**	**CFP4**	**CFP5**
***In Vivo* 6 (30 ℃)**		41.29	34.67	34.57
		63.99	65.74	66.74
		87.35	21.32	42.19
	50 µM	80.68	58.46	59.85
	Average	**68.33**	**45.05**	**50.84**
		49.96	25.04	51.22
		45.42	30.00	29.24
		32.65	59.08	46.51
	500 µM	45.52	62.50	25.20
	Average	**43.39**	**44.16**	**38.04**
	induced @ 0 hrs	**CTRL**	**CFP4**	**CFP5**
***In Vivo* 7 (37 ℃)**		20.63	18.75	7.96
	50 µM	42.29	16.88	46.17
		61.10	43.13	60.61
	500 µM	38.82	4.52	51.63

Table 18. Relative fluorescence for each CFP strain with varying IPTG concentrations, In Vivo Experiments 6 & 7 with varying IPTG Concentrations

***In Vivo* 6 (30 °C)**	**induced @ 0 hrs**	**1 µM**	**50 µM**	**100 µM**	**500 µM**	**1 mM**	**3 mM**
	CTRL	58.23	28.28	121.81	194.99	44.68	44.95
	CFP4	5.80	66.38	68.86	42.70	12.76	53.19
	CFP5	9.13	71.27	51.17	36.51	31.20	42.63
***In Vivo* 7 (37 °C)**	**induced @ 0 hrs**	**1 µM**	**50 µM**	**100 µM**	**500 µM**	**1 mM**	**3 mM**
	CTRL	9.63	16.70	47.12	41.51	14.44	16.65
	CFP4	12.88	31.80	9.76	31.96	94.99	36.36
	CFP5	41.43	41.71	32.16	23.05	63.75	44.27

3.4.1.1 Evaluating RBS Calculator Predictions with Plate Reader Assays

Statistical analysis was first performed to evaluate protein translation for each strain compared to the outputs from the Salis lab RBS Calculator. Multiple linear regression analysis, or ANOVA if more appropriate, and Tukey's Honest Significant Difference (HSD) tests were performed for each experiment to determine if each strain is performing as predicted compared to the other strains. For experiment *In Vivo* 1, each strain was expressed with the same concentration of IPTG and induced at the beginning of the plate reader experiment. The p-value associated with each strain performing equally was significant ($p < 5\ e^{-9}$), indicating that protein expression for each strain is statistically significantly different. Tukey's HSD on this dataset produced the probabilities of each pair of strains being equal. The following strains are statistically significantly different based on their means ($p < 0.005$): CTRL-CFP1, CTRL-CFP3, CFP2-CFP1, CFP1-CFP3, and CFP3-CFP2. These results disagree with the RBS Calculator predictions since plotting the 95% confidence intervals for each strain, shown below in Figure 17, showed CFP2 produced the least protein among the three strains and showed no significant difference in protein expression compared to the control. Additionally, CFP1 was significantly different from each strain and the control and produced the most protein; therefore, CFP1 significantly outperformed each of the strains although the predicted TIR was smaller than CFP2 and CFP3.

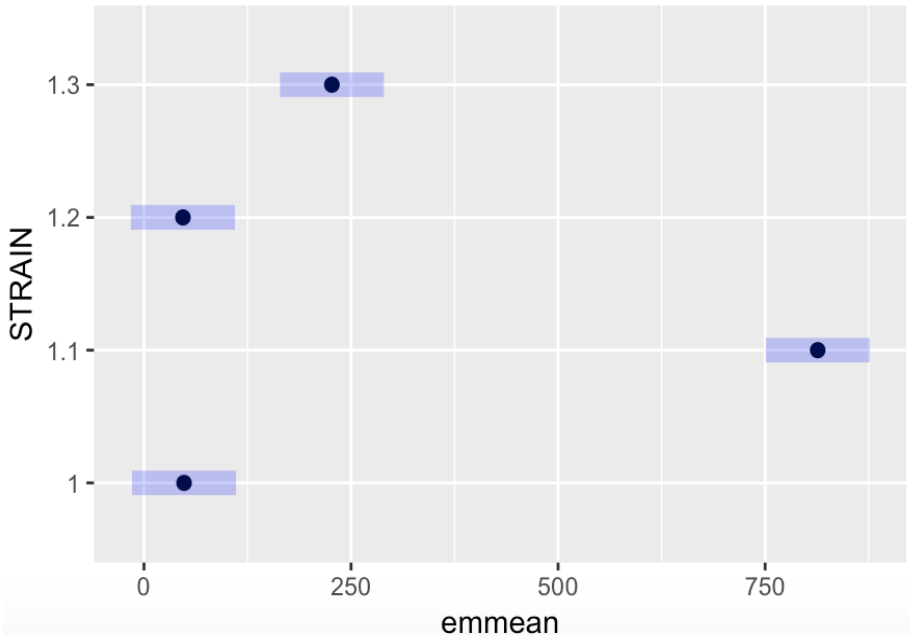


Figure 17. 95% Confidence Intervals for Average Fluorescence for CFP1-3. Confidence intervals were generated by RStudio. Strain 1 is the control, strain 1.1 is CFP1, strain 1.2 is CFP2, and strain 1.3 is CFP3. The x-axis displays fluorescence means for each strain with a 95% confidence interval.

In Vivo 2 was designed so each strain's performance and time of induction were tested to see which combination of factors yields the most protein. A two-factor ANOVA was performed for *In Vivo* 2 was performed with the two factors being "Induction Time" and "Strain" since half of the cultures were induced at the start of the plate reader assays and the remaining cultures were induced 8 hours after the plate reader assay began. ANOVA for the two factors generated significant p-values for the "Strain" factor ($p < 1\ e^{-12}$), the "Induction Time" factor ($1\ e^{-7}$), and the interaction between the "Induction Time" and "Strain" factors was significant as well ($p < 5\ e^{-6}$). Tukey's HSD for this dataset proved the following strain contrasts to be significant: CTRL-CFP1, CTRL-CFP3, CFP1-CFP3, CFP1-CFP2, and CFP2-CFP3 ($p < .0005$). Plots of the 95% confidence interval showed CFP1 is significantly higher than the other strains used in this study, similar to the results for *In Vivo 1*; therefore, CFP1 produced significantly more protein than any other strain in this experiment.

ANOVA for *In Vivo* 3 produced similar results as the statistical analysis for *In Vivo* 2 with significant p-values ($p < 0.05$) for "Strain," "Induction Time," and the interaction between these two factors. Inducing the culture at the beginning of the assay significantly produced the most protein compared to inducing 8 hours into the culture. The following contrasts showed each strain producing significantly different amounts of protein ($p < 0.0001$): CTRL-CFP1, CFP1-CFP3, and CFP1-CFP2. This analysis is in agreement with the analysis for *In Vivo* 1 and is different from the results with *In Vivo* 2 since CFP1 produced statistically significantly higher amounts of protein than CFP3. Additionally, the effect of the interaction between the two variables, strain and induction time, on relative fluorescence was found to be significant ($p < 0.001$).

Analyzing *In Vivo* 2 and *In Vivo* 3 together required three factors: "Strain," "Induction Time," and "IPTG Concentration." Multiple linear regression analysis revealed "Strain" to be a significant factor impacting relative fluorescence levels with this dataset. Additionally, the following interactions were shown to be significant ($p < 5\ e^{-5}$): Strain:IPTG Concentration and Strain:Induction Time. Tukey's HSD analysis on each strain showed the following strains to be significantly different with regard to relative fluorescence ($p < 0.005$): CTRL-CFP1, CTRL-CFP3, CFP1-CFP2, CFP1-CFP3, and CFP2-CFP3.

Two-factor ANOVA was performed for *In Vivo* 4 with "Strain" and "IPTG Concentration" as factors. The resulting ANOVA showed strain to be a statistically significant factor ($p < 1\ e^{-4}$). Contrasts between each strain showed the following strains to produce significantly different protein ($p < 0.05$): CTRL-CFP1, CTRL-CFP3, CFP1-CFP2, and CFP2-CFP3. CFP1 and CFP3 produced the most protein, but were not significantly different when directly contrasted. Strain proved to be a significant factor after performing ANOVA for experiment *In Vivo* 5 ($p < 1\ e^{-3}$). Contrasts showed the following strains to be statistically significant ($p < 0.05$): CTRL-CFP1, CTRL-CFP3, CFP1-CFP2, and CFP2-CFP3. Plotting these 95% confidence intervals for mean strain relative fluorescence showed CFP3 producing more protein than any other strain in this dataset, contrary to previous analyses showing CFP1 producing the most protein;

however, this difference was not statistically significant ($p > 0.05$). Combining datasets *In Vivo* 4 and *In Vivo* 5 showed "Strain" to be a significant factor in protein expression ($p < 1\ e^{-6}$). Additionally, the interaction term for the variables "Strain" and "Temperature" was significant ($p < 0.005$), indicating an effect from these two variables interacting together on protein expression. Contrasts of each strain showed the following pairs to be significant ($p < 0.0001$): CTRL-CFP1, CTRL-CFP3, CFP1-CFP2, and CFP2-CFP3.

Two-factor ANOVA was performed for experiment *In Vivo* 6 with "Strain" and "IPTG Concentration" as the factors impacting relative fluorescence. "Strain" was not a significant factor with a p-value of 0.0535. Both strains, CFP4 and CFP5, were not significantly different from the control ($p > 0.05$) and contrasts revealed no significant difference between any of the strains. *In Vivo* 7 had the same experimental design as experiment *In Vivo* 6, but the temperature increased from 30 ℃ to 37 ℃. ANOVA for experiment *In Vivo* 7 revealed "Strain" was not significant ($p > 0.05$), and contrasts between the strains showed no significance.

3.4.1.2 Determining Optimum IPTG Concentration and Time of Induction for Expression under T7 Promoter

Experiment *In Vivo* 2 evaluated each strain's fluorescence at 50 μM IPTG with half the cultures induced at the start of the experiment and half induced 8 hours into the experiment. ANOVA for *In Vivo* 2 for time of induction showed a significant difference in relative fluorescence when inducing at different times ($p < 1\ e^{-7}$). Inducing at the start of the plate reader experiment produces significantly higher fluorescence 18 hours post-induction than inducing 8 hours into the experiment. ANOVA for experiment *In Vivo* 3 showed significant differences in fluorescence levels at different times of induction ($p < 0.05$). Similarly to *In Vivo* 2, inducing at the start of the experiment produced the most protein.

Experiments *In Vivo* 2 and *In Vivo* 3 implemented the same designs; however, they tested different concentrations of IPTG to assess the concentration of the inducer on protein

expression. Analysis of the two IPTG concentrations used in experiments *In Vivo* 2 and *In Vivo* 3 showed "IPTG concentration" and "Induction Time" to be significant factors affecting protein expression; additionally, the interaction terms "Strain: Induction Time" and "Strain: IPTG" were significant with p-values less than 5 e^{-5}. Plotting the 95% confidence intervals revealed 50 μM IPTG producing significantly more protein than 3 mM IPTG at the 0.05 significance level. Additionally, inducing at the beginning of the experiment instead of 8 hours after the start of the experiment produced significantly higher protein at the 0.0001 significance level.

Experiments *In Vivo* 4 and *In Vivo* 5 investigated protein expression with the same experimental design but utilized different temperatures. *In Vivo* 4 and *In Vivo 5* tested protein expression at varying IPTG concentrations; however, all cultures were induced at the beginning of the assay, eliminating induction time as a factor. ANOVA was performed for *In Vivo* 4 with "Strain" and "IPTG Concentration" as factors. The resulting ANOVA showed "IPTG Concentration" was not significant with regard to protein expression ($p > 0.05$). Given that "IPTG Concentration" was not a significant factor in protein expression, optimum IPTG concentration could not be determined for this dataset. ANOVA for the dataset *In Vivo* 5 produced the same conclusion: "IPTG" was not a significant factor for each culture's relative fluorescence ($p > 0.05$). Analyzing both datasets together with multiple linear regression analysis, *In Vivo* 4 and *In Vivo* 5, with three factors, "Strain," "IPTG Concentration," and "Temperature," showed "IPTG Concentration" to be a significant factor impacting protein expression at the 0.05 significance level with a p-value of 0.039. No direct contrasts between IPTG concentrations tested in *In Vivo* 4 and *In Vivo* 5 proved to be significant with regard to protein expression ($p > 0.05$).

ANOVA was performed for experiment *In Vivo* 6 with "Strain" and "IPTG Concentration" as the factors impacting relative fluorescence. Both factors were not significant with "IPTG Concentration" having a p-value of 0.2132. Contrasts between each IPTG concentration revealed no significant difference in relative fluorescence ($p > 0.05$). *In Vivo* 7 had the same experimental design as experiment *In Vivo* 6, but the

temperature increased from 30 ℃ to 37 ℃. ANOVA for experiment *In Vivo* 7 revealed both factors were not significant ($p > 0.05$), and contrasts among the IPTG concentrations tested showed no significance.

3.4.1.3 Temperature Effects on Protein Expression

Preliminary NUPACK analysis of each strain's mRNA (the CFP gene with the RBS) was performed at 30 °C following the first four experiments in an attempt to investigate why the designs were not behaving as designed. These graphics shown below in Figure 18 predicted secondary structures impacting the ability of the 16S rRNA to correctly bind to the RBS2 sequence in the mRNA strand for CFP2. Simulating this complex formation at 37 °C showed disappearance of the primary hairpin with fewer nucleotides bound with secondary structures in the RBS. As a result, experiment *In Vivo* 5 was performed at 37 °C to see any changes in protein expression for CFP2. Experimental results showed no clear differences in protein expression at 30 °C and 37 °C for CFP1 and CFP2; however, CFP3 had higher levels of expression at the higher temperature. NUPACK calculations for the mRNA for CFP3 did not reveal any key changes between 30 °C and 37 °C in the secondary structure near the RBS.

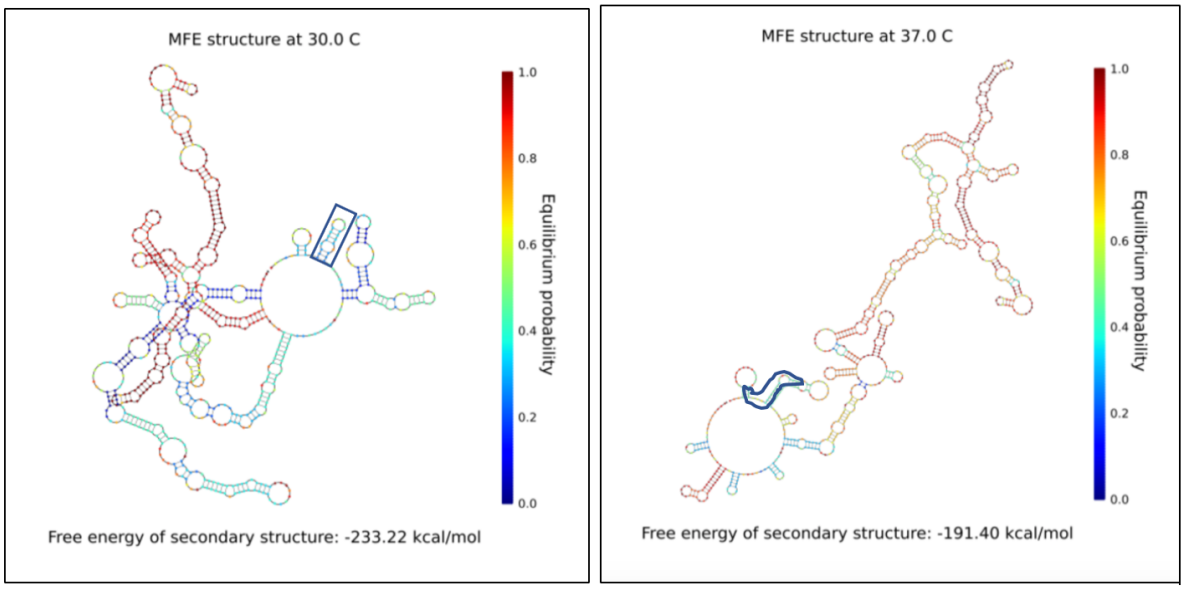


Figure 18: NUPACK graphics of the secondary structure of the mRNA for CFP2 at 30 and 37 °C

Multiple linear regression for the dataset showing the difference in protein expression between *In Vivo* 4 and *In Vivo* 5 showed no significant difference in relative fluorescence levels at the two temperatures when factoring in each strain and each IPTG concentration tested. The p-value was found to be 0.1102. A one-sample t-test was done to illustrate any potential difference in protein expression at 30 °C and 37 °C for the strain CFP3 in particular due to the observed difference in fluorescence in the plate reader assay compared to the other strains. This t-test showed the difference in relative fluorescence at the two temperatures for CFP3 is significant ($p < 0.01$) indicating a statistically significant difference in expression levels. As a result, further NUPACK analysis was done in an effort to discover what is happening during mRNA-rRNA complex formation at a molecular level. Multiple linear regression analysis on experiments *In Vivo* 6 and *In Vivo* 7 together reveal temperature to be a significant factor ($p < 0.05$) on relative fluorescence when taking all variables into account.

3.4.2 NUPACK: Complex Formation Results

Plate reader experiments for CFP1-3 revealed a relationship contradicting what the Salis Lab RBS Calculator predicted. In the assays CFP2 fluorescence levels were very similar to the control, a blank pET28a, and CFP1 consistently outperformed both CFP2 and CFP3. Given that CFP1 was designed to produce the least protein yet produced fluorescence levels as high as 3000, it was hypothesized that the complex may not be forming as anticipated due to secondary structures in the mRNA not predicted by the RBS Calculator. Thermodynamic calculations for the mRNA-rRNA complex with a range of temperatures were completed in an attempt to ensure the bonds were forming in the correct locations at equilibrium at the temperatures tested. Experimental results showed CFP1 producing high protein concentrations at 30 °C and 37 °C, CFP2 not producing significant protein at 30 °C or 37 °C, and CFP3 producing protein at 30 °C and 37 °C, but significantly more protein at 37 °C. As a result, variations of the mRNA sequence were analyzed in NUPACK in order to determine which conditions produced results matching those of the plate reader experiments. The following three figures show the mRNA-rRNA complex forming at equilibrium at the specified temperature such that protein expression would yield results matching those of the plate reader experiments. The figures below show the formation of the mRNA-rRNA complex, with the mRNA composed of the RBS, the first 30 nucleotides of the CFP gene, and 29 nucleotides upstream of the RBS in the designs for this project to encompass where transcription begins with T7 RNA polymerase.[39]

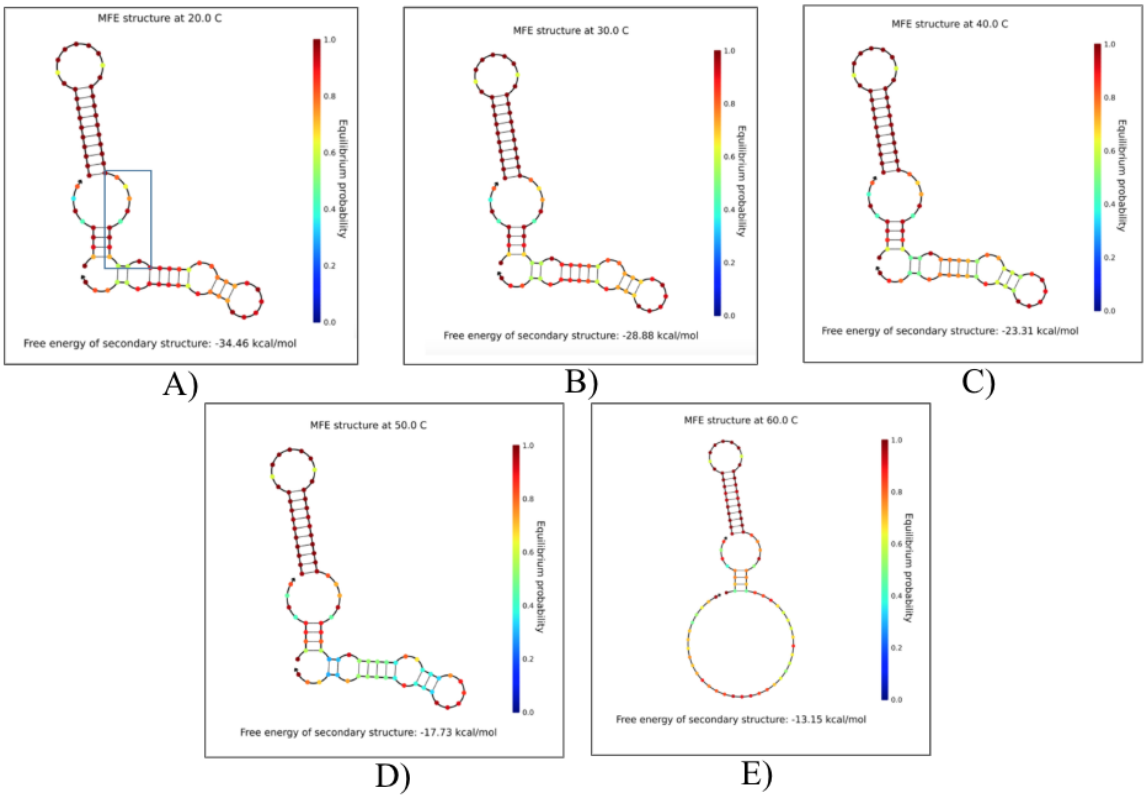


Figure 19: CFP1 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation

Each calculation was performed with both 16S rRNA sequences hypothesized to be participating in protein expression in our lab's strain of *BL21(DE3).* Figure 19 above shows the mRNA, consisting of RBS 1 (CFP1), 29 nucleotides upstream of the RBS, and the first 30 nucleotides of the CFP gene, interacting with the 16S rRNA sequence used in the RBS Calculator to design the first three RBS sequences, the non-canonical sequence. Each graphic shows the mRNA-rRNA complex's structure at 20 °C, 30 °C, 40 °C, 50 °C and 60 °C. The rRNA binds to the RBS, outlined in the 20 °C graphic, as expected at 30 °C and 40 °C; as a result, protein expression resulting from this mRNA-rRNA complex would be successful at both temperatures, similar to CFP1 performance in the plate reader experiments at 30 °C and 37 °C.

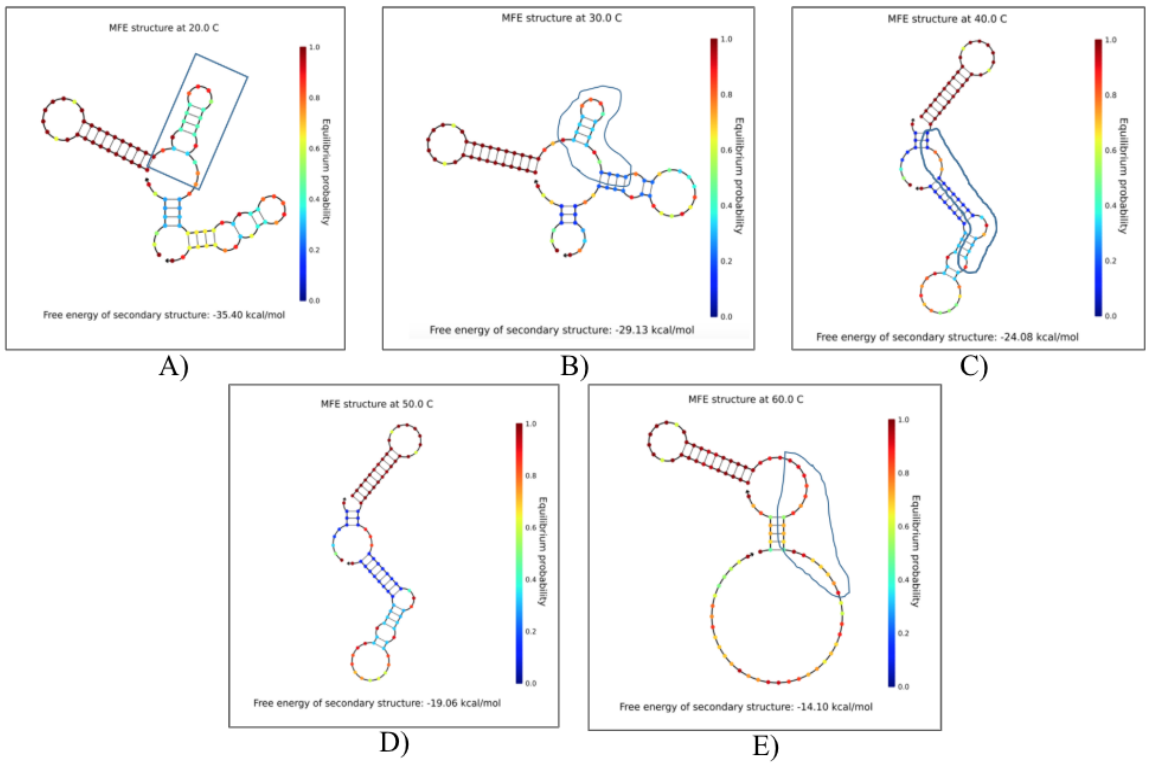


Figure 20: CFP2 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation

Figure 20 shows the CFP2 mRNA-rRNA complex forming at temperatures 20 °C to 60 °C with the RBS outlined in graphics that show changes to the complex's structure. The mRNA sequence consists of the RBS, 29 nucleotides upstream of the RBS, and the first 30 nucleotides of the CFP gene. The rRNA sequence used in this simulation is the sequence used in the initial design process for CFP1-3. CFP2 did not produce protein significantly different from the control at 30 °C and 37 °C; consequently, plausible simulations of the mRNA-rRNA complex forming would show structures preventing the 16S rRNA from binding to the RBS correctly, thus preventing translation initiation at 30 °C and 37 °C. The figure above shows this relationship due to the hairpin forming in the RBS at lower temperatures as well as the RBS favorably binding to nucleotides in the CFP gene rather than the rRNA.

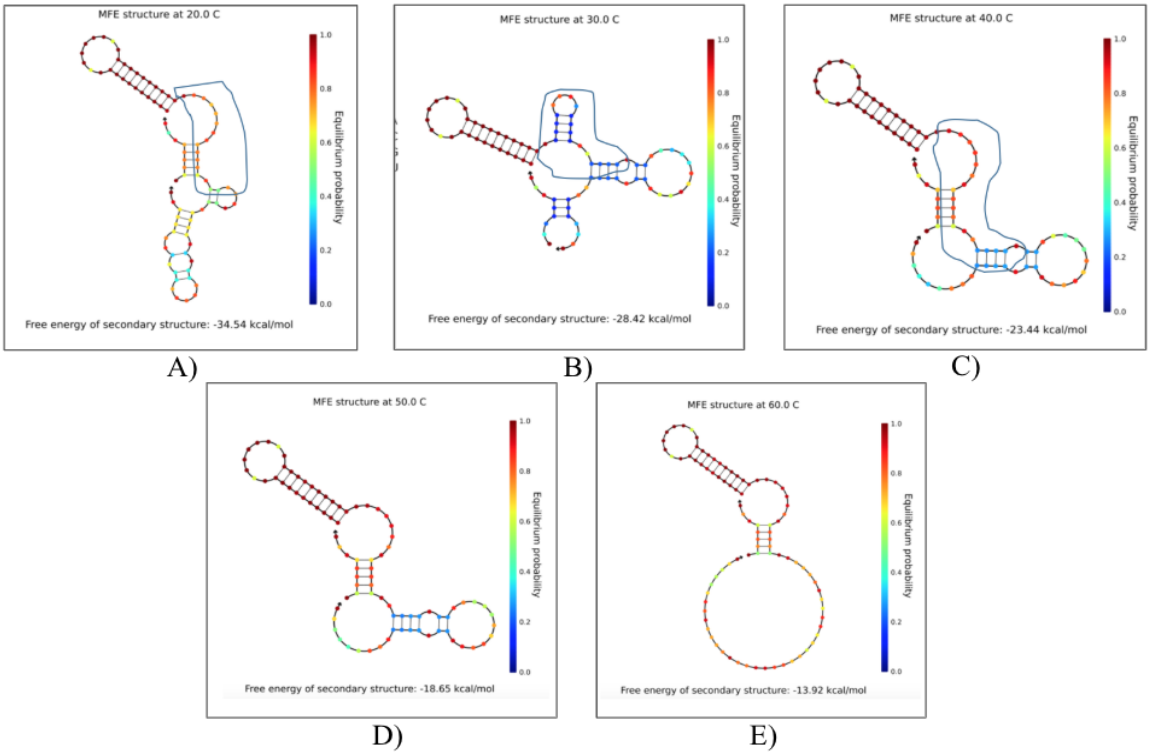


Figure 21: CFP3 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) -*TAACCGTAG* Free Energy of Complex Formation

Figure 21 shows thermodynamic calculations in NUPACK and corresponding graphics of mRNA-rRNA complex formation for CFP3 with the RBS outlined. Experimental results showed CFP3 produced protein at 30 °C and 37 °C but produced significantly larger amounts of protein at 37 °C. As a result, accurate displays of the mRNA-rRNA complex should show secondary structures preventing the RBS binding correctly to the rRNA at 30 °C; additionally these secondary structures should disappear at 37 °C, allowing the RBS to bind to the rRNA correctly. The graphics above show CFP3 forming a hairpin within the RBS at 30 °C, preventing the rRNA from binding. The simulation at 40 °C shows this structure unlikely to exist at equilibrium, and the rRNA binds to the exposed RBS.

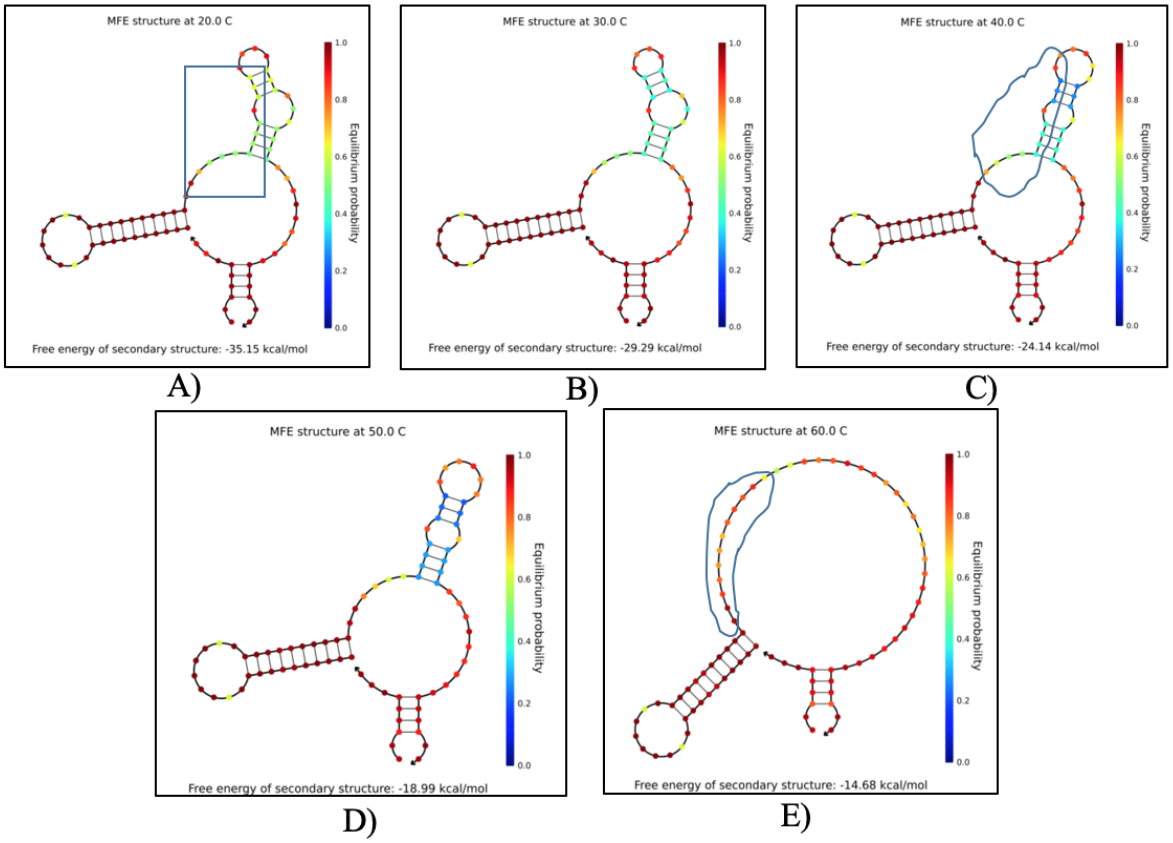


Figure 22: CFP1 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation

Figure 22 shows thermodynamic calculations with the same conditions as Figure 19; however, the 16S rRNA sequence used in the calculations is the canonical rRNA sequence suspected to be the only 16S rRNA present in *BL21(DE3).* The graphics of the complex do not match the experimental results, CFP1 producing high amounts of protein at 30 °C and 37 °C, due to hairpins forming in the RBS, thus preventing the rRNA from binding correctly at temperatures 50 °C and below. Figure 22 below shows the mRNA-rRNA complex forming between CFP2 and the 16S rRNA sequence *ACCTCCTTA*; additionally, relatively stable hairpins are present in RBS 2 at temperatures below 50 °C. The RBS and rRNA do not show the complex forming correctly at any temperature due to the hairpin structures in the RBS and more favorable binding between nucleotides in the CFP gene and the rRNA. This relationship matches the experimental data because CFP2 did not produce protein amounts different from the control at 30 °C nor 37 °C.

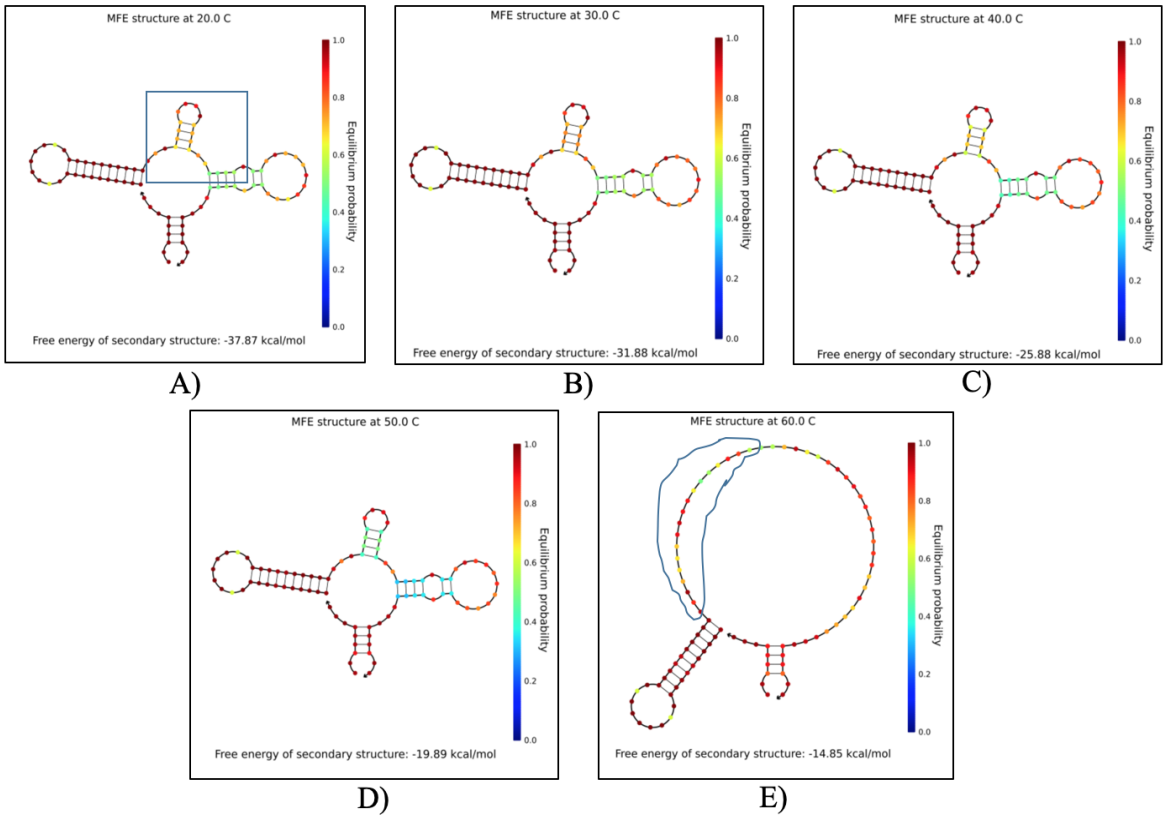


Figure 23: CFP2 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation

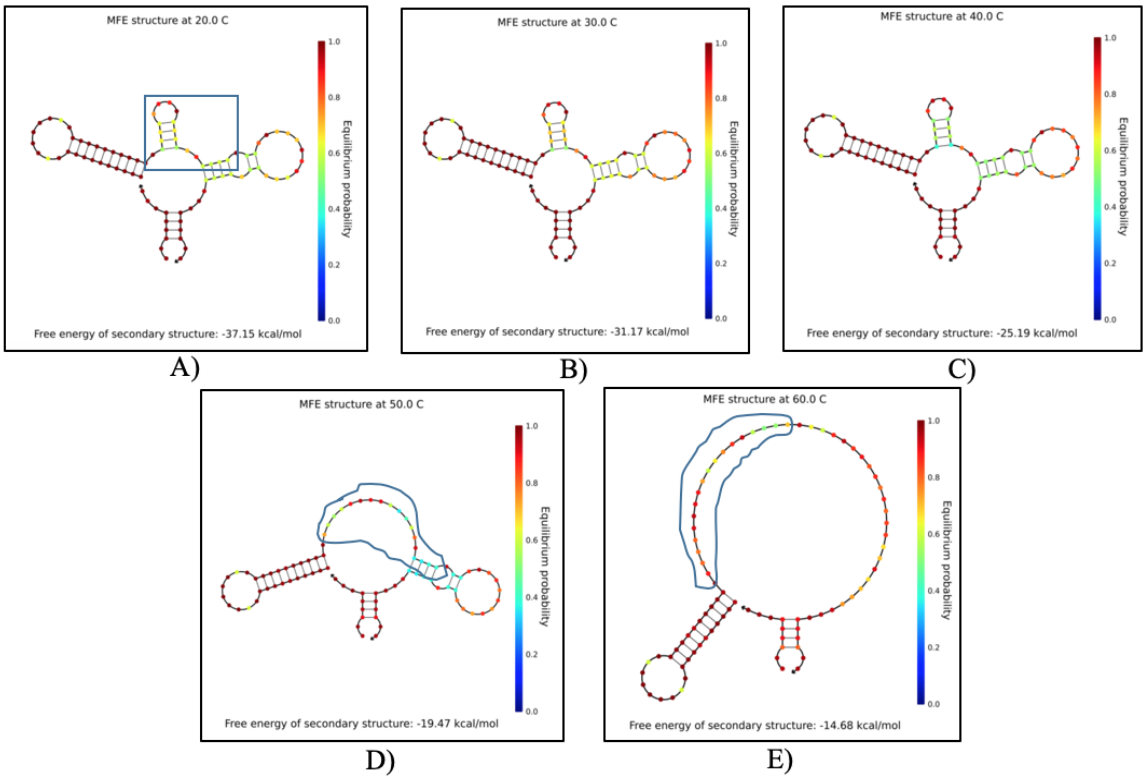


Figure 24: CFP3 (mRNA beginning at 3' end of T7 promoter, including 30 nucleotides of CFP gene) - *ACCTCCTTA* Free Energy of Complex Formation

Figure 24 shows the free energy of the mRNA-rRNA complex and binding patterns between CFP3 and the rRNA sequence not used during the design process, *ACCTCCTTA*. These calculations show the rRNA binding to the CFP gene, downstream of the RBS, at temperatures 60 °C and below. Additionally, an unstable hairpin forms in the RBS, preventing the rRNA from binding correctly and initiating protein translation.

These graphics reveal the interactions occurring at a molecular level between the two strands to form a complex. It was expected that each RBS would bind to the rRNA in a similar fashion with differences in the probability of bond formation due to different RBS strength; however, these graphics revealed that the rRNA sequence, *ACCTCCTTA*, did not bind correctly to the RBS in the majority of the simulations. Additionally, *TAACCGTAG*, the rRNA sequence removed from the RBS Calculator, had a higher probability of binding to each RBS as predicted, and resembled the experimental data

more closely. Given the discrepancies between the RBS Calculator and the experimental results following the modifications made to the RBS Calculator, it was unclear which rRNA sequence may be involved with protein expression in these CFP strains.

3.4.2.1 Modeling RBS-rRNA Interactions in NUPACK

Modeling the mRNA and rRNA interactions in NUPACK was done in a number of fashions, simulating solely the RBS, the RBS with portions of the sequence directly upstream of the 5' end of the RBS, the RBS with upstream sequences and the first 30 nucleotides of the CFP gene, and the entire CFP mRNA transcript (region upstream, RBS, CFP gene). Each version of the mRNA was simulated with both potential 16S rRNA sequences to see which relationships matched those observed during the course of the plate reader experiments. The strongest relationship occurred when simulating the non-canonical 16S rRNA sequence, *TAACCGTAG,* with the RBS sequences for each design tested. This data included the CFP designs described previously, another CFP design with the pUC19 plasmid transformed in *E. coli* 10β, SAM synthase cloned into pET28a and transformed into *BL21(DE3),* and a diaphorase enzyme cloned into *BL21(DE3).* SDS-page gels displaying expression of each enzyme are shown below in Figures 25 and 26.

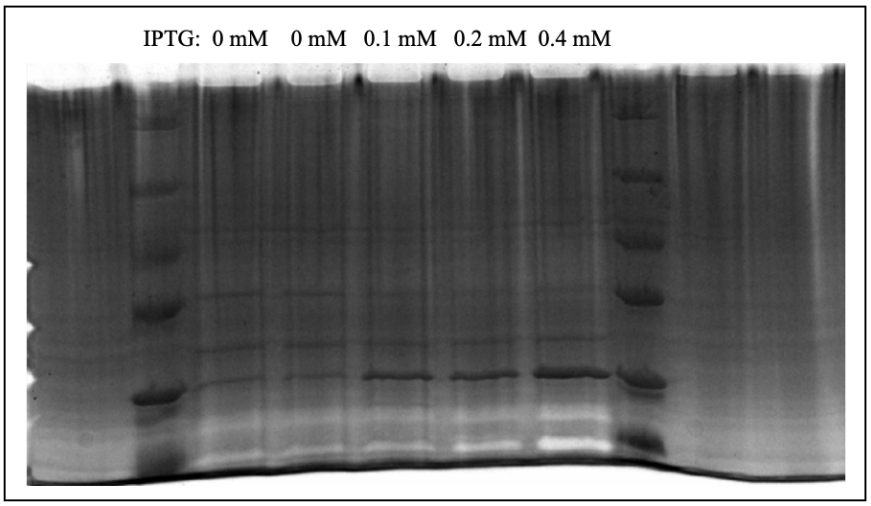


Figure 25. Diaphorase SDS-page gel.

Diaphorase expression in *BL21(DE3)* with 0, 0.1, 0.2, and 0.4 mM IPTG. The ladder is Unstained Protein Molecular Weight Marker from Thermo Scientific. The numbering above each lane corresponds to the concentration of IPTG used for induction.

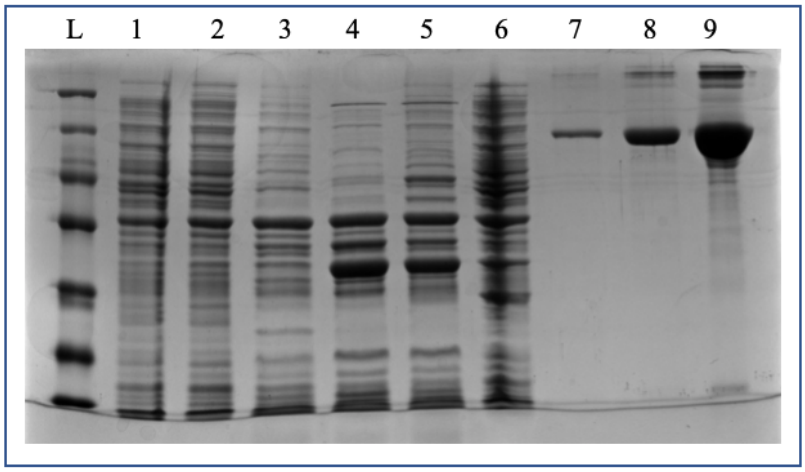


Figure 26. SAM Synthase SDS-page gel.

SDS-page gel for SAM synthase expression in *BL21(DE3)*. The ladder is Unstained Protein Molecular Weight Marker from Thermo Scientific. Lane 1 contains uninduced pET28a without heat treatment and diluted 10-fold. Lane 2 contains induced pET28a without heat treatment and diluted. Lane 3 contains pET28a induced and heat-purified. Lane 4 contains the SAM synthase strain uninduced and heat treated. Lane 5 contains the SAM synthase strain both induced and heat-treated. Lane 6 contains the induced SAM synthase strain with no heat-treatment and diluted. Lanes 7-9 contains BSA at the following concentrations: 0.08 g/L, 0.4 g/L, and 2 g/L.

Table 19 below shows each strain included in this analysis and the RBS sequences used to improve expression of each protein. The Gibbs free energy of the RBS-rRNA complex with the non-canonical 16S rRNA sequence is included in this table. The relationship displayed in Figure 27 shows the difference in the Gibbs free energy of the complex between the sum of the final state components and the initial state, represented with Equation 3 below. Note that the free energy of each 16S rRNA sequence is 0 kcal/mol at both 30 °C and 37 °C.

(3) $$\Delta G_{RBS-rRNA} - (\Delta G_{RBS} + \Delta G_{rRNA}) = \Delta G_{system}$$

Table 19. RBS sequence simulated in NUPACK listed by protein expressed.

Gibbs free energy of the system, expression level, and temperature in which expression and NUPACK thermodynamic calculations occurred are included for each RBS.

Protein/Strain	**RBS Sequence**	**ΔG_{system} (kcal/mol)**	**Expression**	**Temperature (°C)**
CFP1	*TTAACAATTCCCCTGGTTATTTTT*	-7.31	1700	30
CFP2	*GCATCCTGCGGCCTAAAT*	-4	67	30
CFP3	*CCCAGACCACCTACATCTTTTTTA*	-5.73	976	30
CFP4	*TTCAATAAGGAGGTTTTTT*	-5.99	62	30
CFP5	*CCCCCCTACGGTTAAAAAA*	-13.77	71	30
CFP_pUC19	*AGGAGGAA*	-5.32	1270	30
Diaphorase	*AAGAAGGAGATATACCAT*	-5.32	See Figure 25	30
SAM Synthase	*TTAAGAAGGAGATATACC*	-5.32	See Figure 26	30
CFP1	*TTAACAATTCCCCTGGTTATTTTT*	-6.58	1130	37
CFP2	*GCATCCTGCGGCCTAAAT*	-4.38	45	37
CFP3	*CCCAGACCACCTACATCTTTTTTA*	-5.78	895	37
CFP4	*TTCAATAAGGAGGTTTTTT*	-5.78	95	37
CFP5	*CCCCCCTACGGTTAAAAAA*	-12.88	64	37

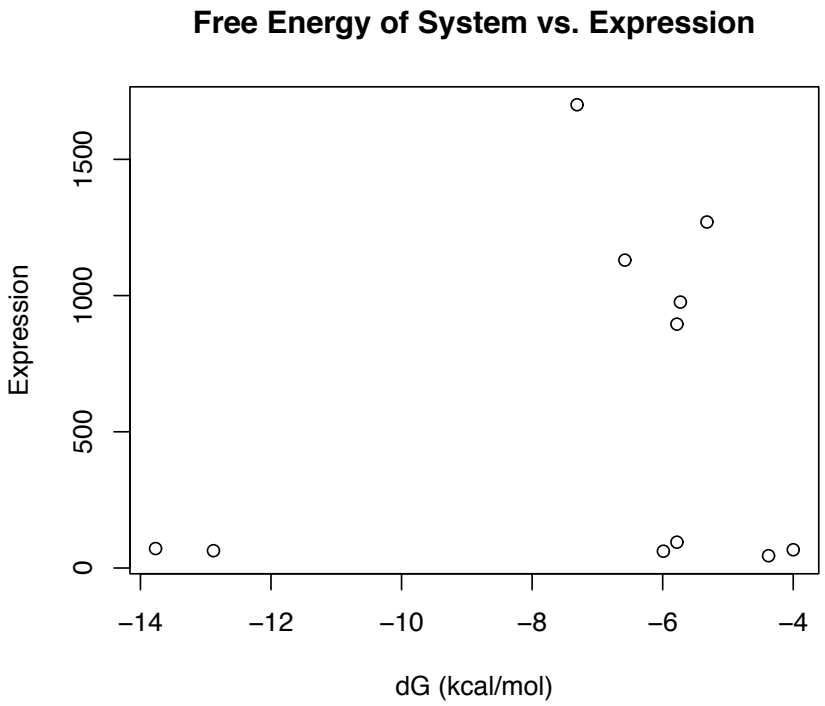


Figure 27. Change in Gibbs free energy vs. expression

Gibbs free binding energies associated with each RBS and rRNA, *TAACCGTAG,* with their corresponding expression at the temperatures in which NUPACK calculations were performed.

Given the levels of expression received with each RBS, it appears there is a maximum binding energy associated with the RBS-rRNA complex in which adequate expression can occur. Deviations from this relationship are likely due to secondary structures forming near the RBS in the mRNA, thus preventing the 30S ribosome unit and 16S rRNA from docking and binding at the RBS correctly. For example, CFP3 and CFP4 have very similar binding energies associated with the formation of the RBS-rRNA complex; however, their expression of CFP varies greatly. It is important to take the initial structure of each nucleic acid into account when modeling the formation of the complex since the free energy of the complex alone includes all secondary structures present; therefore, finding the energy of the RBS-rRNA complex requires subtracting the free energy of the secondary structure present prior to the formation of the complex. For

example, CFP2 has a high free energy of complex formation because of the secondary structures present at equilibrium in the RBS. As a result, this high energy of formation for the complex largely comes from the secondary structures present at equilibrium in the RBS, so including the structures present in the nucleic acid alone yield inaccurate binding strengths for the complexes. Figure 28 below shows the relationship present when modeling the free energy of the complex without accounting for initial secondary structures and expression.

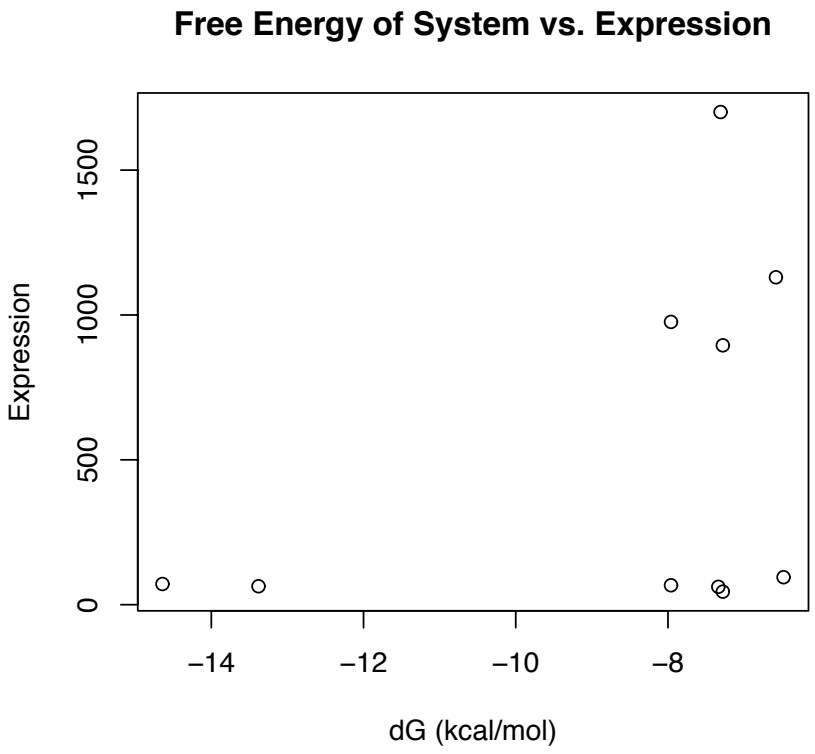


Figure 28. Change in Gibbs free energy vs. expression, without initial structures

Gibbs free binding energies associated with each RBS and rRNA, *TAACCGTAG,* with their corresponding expression at the temperatures in which NUPACK calculations were performed. Initial secondary structures in each molecule were not taken into account with the Gibbs free binding energy calculations.

The secondary structures for each mRNA show hairpins of different strengths forming near the RBS; the graphic of CFP3 shows a hairpin with weaker interactions while CFP4 has much stronger bonds located in the hairpin with the RBS. This is shown in Figures 28 and 29.

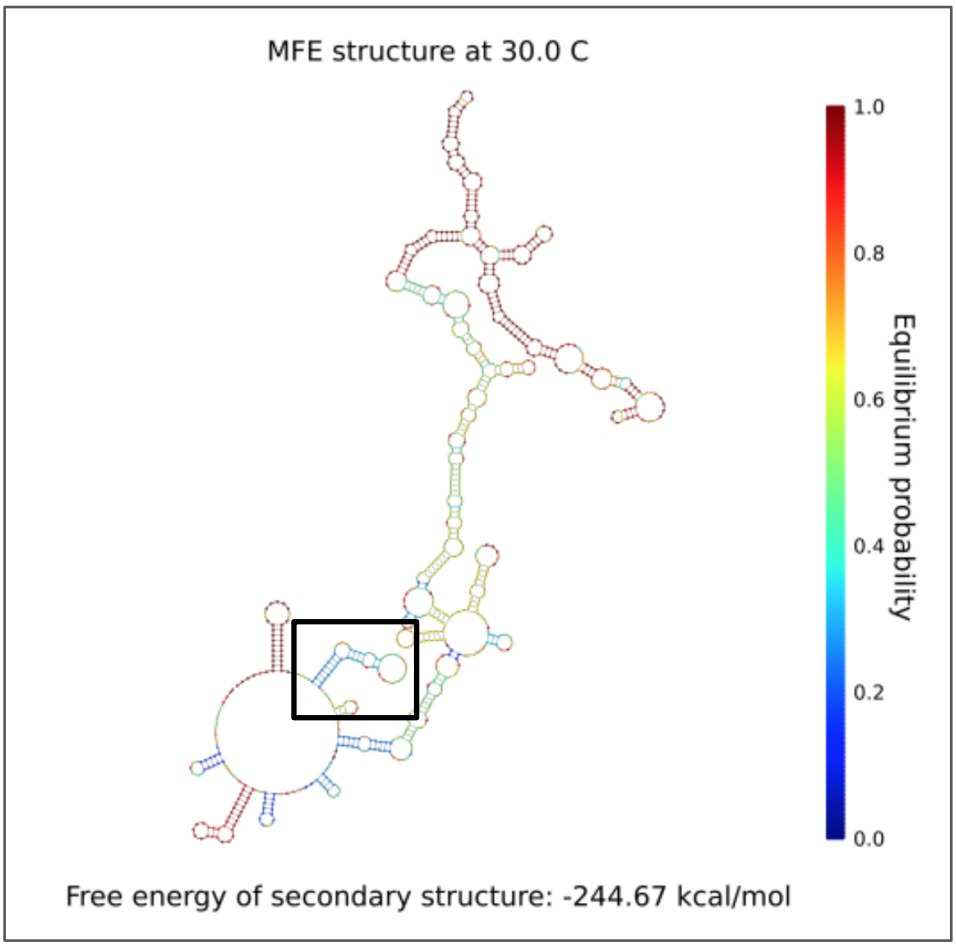


Figure 29. Full mRNA transcript of CFP3 with the hairpin containing the RBS sequence outlined in the box.

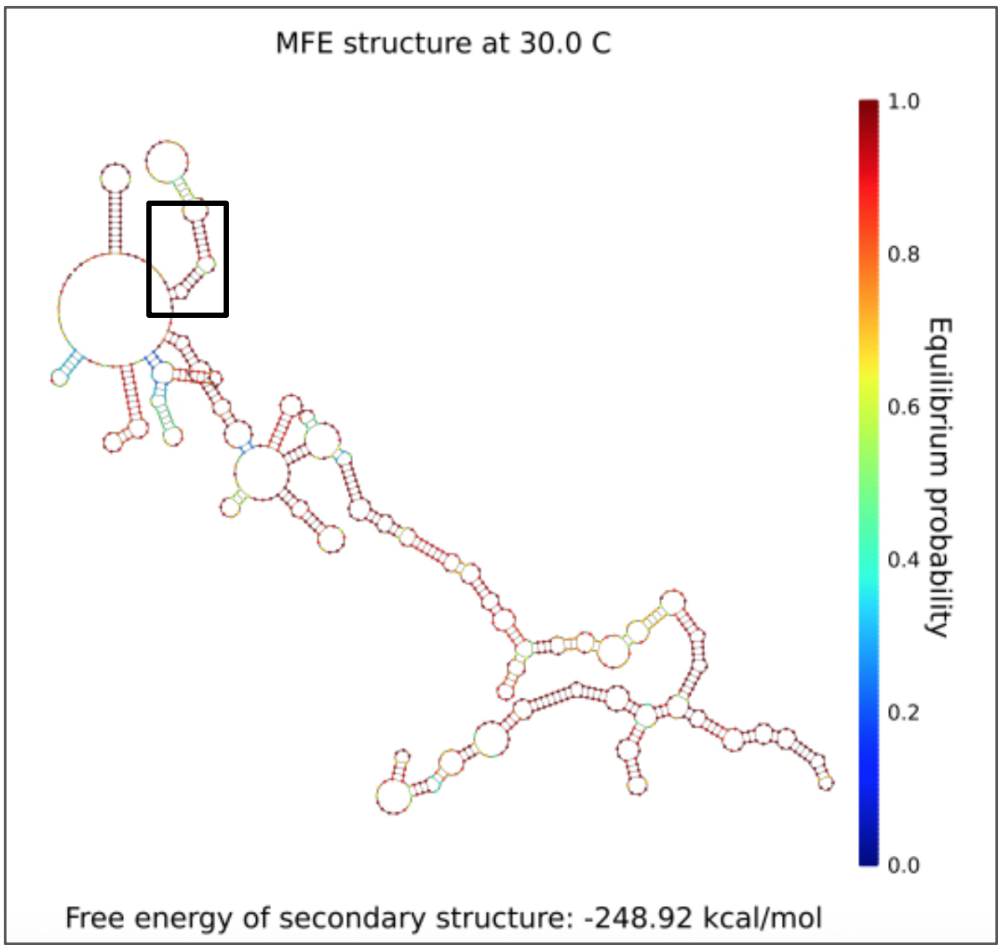


Figure 30. Full mRNA transcript of CFP4 with the hairpin containing the RBS sequence outlined in the box.

After seeing deviations in expression dependent on the secondary structures present around the RBS, this impact of secondary structure on expression was investigated further. Using the simulations generated in NUPACK previously, a figure showing the relationship between RBS accessibility and expression was generated. The figure below shows the relationship between the average probability of a nucleotide in the RBS sequence being unpaired at equilibrium at 30 °C or 37 °C; this value was computed by obtaining the probabilities of each nucleotide in the entire mRNA transcript being unpaired at equilibrium and computing the average probability of each nucleotide being unpaired for only the region containing the RBS. As Figure 30 shows, expression increases as accessibility of the RBS, the untranslated part of the mRNA transcript

upstream of the start codon, increases. Deviations from this linear relationship are likely due to the model shown in Figure 27 in which there is likely a maximum binding energy for the RBS-rRNA complex associated with adequate protein expression.

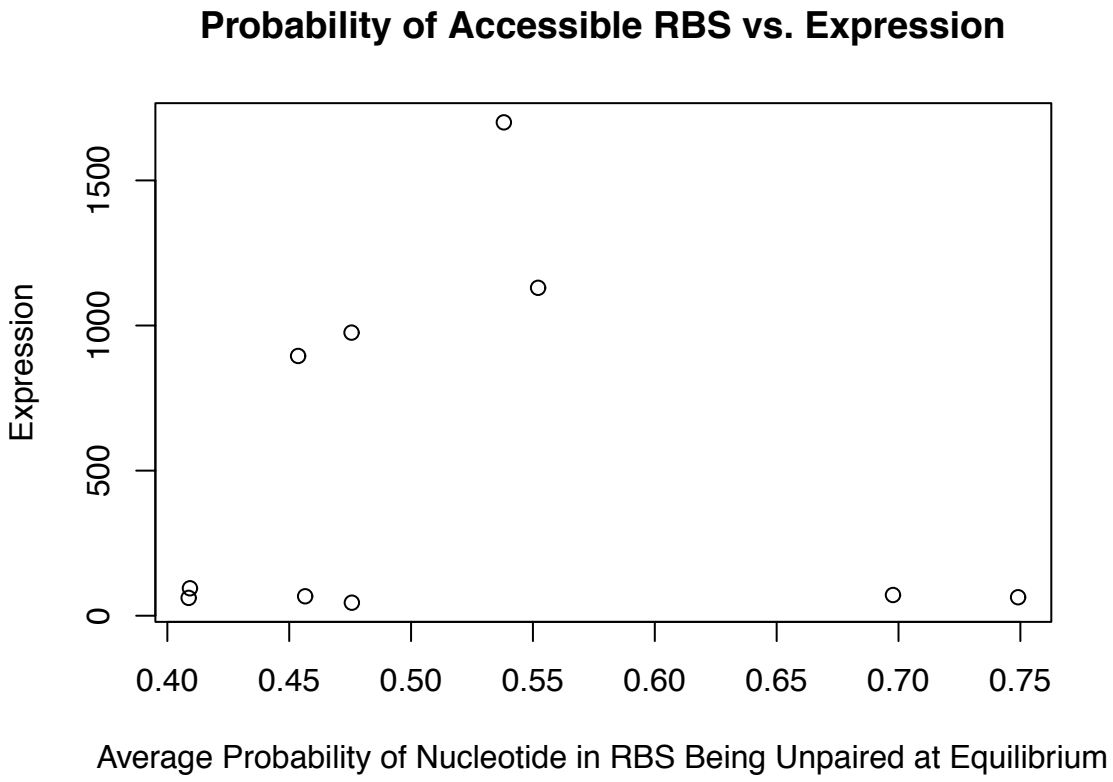


Figure 31. Accessibility of RBS in an mRNA transcript vs. expression.
"Average probability of Nucleotide in RBS Being Unpaired at Equilibrium" was calculated by finding the average probability of each nucleotide in the RBS being unpaired at equilibrium in the entire mRNA transcript.

3.4.3 RBS Calculator Outputs for Each CFP Strain with Canonical and Non-Canonical 16S rRNA Sequences

Table 20 below displays the RBS Calculator outputs for each CFP strain simulated with both suspected 16S rRNA sequences. Based on the canonical 16S rRNA sequence, *ACCTCCTTA,* CFP4 should be producing the most protein and highest fluorescence among all five strains. CFP2, CFP3, and CFP5 have similar TIRs, so these strains should be producing relatively similar protein, and consequently, fluorescence. CFP1 should be producing the least protein compared to the other four strains. The simulations with the non-canonical 16S rRNA sequence, *TAACCGTAG,* predict CFP5 to produce the most protein and CFP1 producing the least protein. Additionally, the RBS Calculator predicted CFP2 and CFP3 to produce similar amounts of protein due to their similar TIRs.

The RBS Calculator outputs for each 16S rRNA sequence were compared to experimental results in order to determine which rRNA sequence is likely present and active in our strain of *BL21(DE3).* The RBS Calculator results do not match CFP fluorescence measured through plate reader experiments since simulations with both rRNA sequences predicted CFP1 to produce the least protein; however, CFP1 statistically significantly produced the most protein compared to any other strain. Additionally, CFP4 and CFP5 did not produce protein levels significantly different from the control, pET28a. CFP4 and CFP5 were each designed to perform optimally with one of the suspected 16S rRNA sequences; however, both strains did not successfully express proteins. Details regarding data analysis of protein expression are included in section 3.3.1.

Table 20. Salis Lab RBS Calculator Outputs for Each CFP Design Simulated with Both Potential 16S rRNA Sequences

RBS	rRNA Sequence	Start Position	TIR	ΔG_{total}	$\Delta G_{mRNA-rRNA}$	$\Delta G_{spacing}$	$\Delta G_{standby}$	ΔG_{start}	ΔG_{mRNA}
1	*ACCTCCTTA*	48	28.13	8.4	-12.57	0.0	6.27	-2.76	-17.76
2	*ACCTCCTTA*	56	275.16	3.33	-15.38	0.01	3.31	-2.76	-18.19
3	*ACCTCCTTA*	54	321.59	2.99	-13.5	0.01	2.97	-2.76	-16.31
4	*ACCTCCTTA*	54	40958	-7.78	-27.45	0.0	5.56	-2.76	-17.91
5	*ACCTCCTTA*	54	413.94	2.43	-13.68	0.01	2.97	-2.76	-16.03
1	*TAACCGTAG*	48	52.79	7	-16.71	1.52	7.18	-2.76	-17.76
2	*TAACCGTAG*	56	1158.4	0.14	-20.87	0.0	5.56	-2.76	-18.19
3	*TAACCGTAG*	54	1282.88	-0.09	-18.99	0.0	5.33	-2.76	-16.31
4	*TAACCGTAG*	54	425.07	2.37	-17.85	1.52	4.06	-2.76	-17.91
5	*TAACCGTAG*	54	27929.04	-6.93	-23.7	0.0	3.68	-2.76	-16.03

4 CONCLUSION AND FUTURE DIRECTIONS

4.1 CONCLUSION

A procedure for λ-PCR primer design was created to ensure accurate cloning products when using this technique in the laboratory. This procedure was used to design primers for each CFP strain; all primer sequences were identical aside from the RBS which was added as the linker sequence in the forward primer. These RBS sequences were designed using the RBS Calculator designed by the Salis lab. CFP expression for each strain was evaluated in plate reader assays with both whole-cell and cell-free systems, and a number of variables were changed to optimize conditions for protein expression, including: induction time, IPTG concentration, temperature, and in the cell-free experiments, media. ANOVA and multiple linear regression revealed induction time, IPTG concentration, and media were statistically significant factors impacting protein expression. Additionally, each strain did not perform as predicted by the RBS Calculator since overall CFP1 produced the most protein compared to the other strains, and CFP2 did not produce protein concentrations significantly different from the control, the pET28a plasmid transformed into *BL21(DE3).* Whole-cell experiments produced significantly higher fluorescence, and consequently more protein than cell-free experiments. NUPACK was used to further investigate binding between the mRNA and potential 16S rRNA sequences active in the strain of *BL21(DE3)* used in this project. As a result, a new method for designing RBS sequences for optimum protein expression was developed based on relationships between expression data generated in plate reader experiments and NUPACK thermodynamic calculations for the RBS-rRNA complex. Additionally, CFP expression was modeled with secondary structures present within and surrounding the RBS to visualize the relationship between RBS accessibility and expression.

4.2 FUTURE DIRECTIONS

A new method for designing RBS sequences for protein expression was developed based on the data received with the plate reader experiments and simulations in NUPACK. In the future new RBS sequences should be developed with the CFP reporter to further test the relationship described in this thesis.

Appendix A: CELL-FREE STUDIES

A.1 INTRODUCTION

While metabolic engineering practices have significantly improved yields and helped synthesize new bioproducts in bacteria and yeast, such as sugar alcohols, acids, and biopharmaceutical proteins, there are a number of limitations associated with whole cell metabolic engineering.[40,41,42] Some pharmaceutical products and their intermediates have been shown to be toxic to cells, restricting their production using biological methods.[43,44] Cells also have transport limitations since any materials entering and exiting the cell must pass through the semipermeable membrane.[45] Additionally, genetic engineering of whole-cells can result in unintended effects, thus causing further delays in reaching the goal of the modifications; additionally, a significant portion of resources are dedicated to cell growth, putting a limit on bioreactor productivity.[46] *In vitro* metabolic engineering practices would allow for improved control over reaction environments compared to *in vivo* metabolic engineering since there are no membranes that can limit chemical transport and disruptive cellular machinery can be removed from the system.

A.1.1 Cell-Free Metabolic Engineering

In vitro metabolic engineering, or cell-free metabolic engineering (CFME), is circumventing many of the limitations associated with *in vivo* metabolic engineering, and commonly uses purified enzymes or cell lysate to produce valuable products with biological inputs.[44] A variety of methods have been investigated for high protein expression, including overproducing the protein *in vivo* and purifying the proteins from cellular machinery as well as cell-free protein expression. Overexpression of proteins not involved in cellular metabolism has been shown to harm growth in *E. coli* and damage cellular machinery, causing *in vitro* protein expression to provide more opportunity for improving heterologous protein expression.[47] Cell-free protein expression was tested in this project in an effort to assess its potential to yield high quantities of protein for enzymatic cascades, but did not yield consistent results.

A.2 MATERIALS AND METHODS

A.2.1 Buffer and Media Preparation

Liquid Lennox Broth (LB) media was used in the plate reader assays. The Lysis Buffer used in the *in vitro* studies was prepared with the following recipe: 20 mM Na_2HPO_4, 50 mM NaCl, 10 mM imidazole, and 2% triton by volume.

A.2.2 Strain Construction

The strains, plasmids, and genes used in the cell-free studies were identical to the strains used in the *in vivo* studies. The primer design process and strain construction procedures are outlined in Chapter 2 with the RBS design process included in Chapter 3.

A.2.3 Microplate Reader Assays: Constant IPTG Concentration

All *in vitro* assays were performed for a minimum of 96 hours in 96-well microplates with a Synergy H4 microplate reader (BioTek; Winooski, VT); this microplate reader kept temperature and shaking speed consistent throughout the assays at 30 °C and 150 rpm. Culture growth was monitored with Optical Density (OD), measured at 600 nm, and CFP fluorescence was measured with excitation and emission wavelengths of 458/489 nm. The section below details the sonication procedures for *in vitro* protein expression. Cells were resuspended with LB containing 50 µg/mL of kanamycin or fresh lysis buffer following sonication and assayed in the plate reader with 50 µM IPTG. Total fluorescence was measured at 90 hours for ANOVA and multiple linear regression analysis.

A.2.4 Microplate Reader Assays: Variable IPTG Concentration

The same IPTG concentrations tested with the *in vivo* protocol were tested in the *in vitro* experiments: 1 µM, 10 µM, 50 µM, 100 µM, 1 mM, and 3 mM. *In vitro* plate reader

assays were done with cells resuspended in either LB with 50 μg/mL kanamycin or lysis buffer.

A.2.5 Sonication Procedure

Each CFP strain and the control were grown overnight in 50-75 mL of LB media with 50 μg/mL kanamycin. Each culture was centrifuged at 10,000 rpm in 2-minute intervals and the supernatant was discarded. The 50-mL tubes for each strain were weighed before and after centrifuging to receive the final weight of the cells. The lysis buffer used in these cell-free studies was composed of 20 mM Na_2HPO_4, 50 mM NaCl, 10 mM imidazole, and 2% triton by volume. Lysis buffer was added to each strain using Equation 2:

(2)

$$\textit{Lysis Buffer volume (in mL)} = \textit{weight of cells (in grams)} * 10$$

After the lysis buffer was added, the cells were vortexed until all cell debris was thoroughly broken up, and the cells were immediately placed on ice. The sonication probe was wiped with ethanol prior to use, and each strain was sonicated for a total of 7 minutes per gram of cell material; any cells weighing less than a gram were sonicated for 7 minutes. The temperature setting was off, and the amplitude was set at 30% with 7 seconds on and 5 seconds off. Following sonication, the lysed cells were centrifuged at 10,000 rpm for 2 minutes and resuspended in LB media with 50 μg/mL kanamycin or lysis buffer for the plate reader assays.

A.3 RESULTS AND DISCUSSION

The following sections include ANOVA and multiple linear regression analysis for a number of variables tested in the plate reader experiments in addition to Tukey's HSD contrasts. Graphics for each experiment are available with the supplementary material in section A.5.

A.3.1 Strain Performance

Tables 21, 22 and 23 below shows the fluorescence values for strains CFP1-3, experiments *in vitro* 1, 2, and 3, at 90 hours into the plate reader assay. The following factors were taken into account for the multiple linear regression statistical analysis: "Strain," "IPTG Concentration," and "Block," with "Block" representing each separate experiment; this accounts for any deviations in measurements due to the independence of the experiments since the factors tested are consistent among experiments. Analysis of experiments *in vitro* 1, 2, and 3 together revealed "Strain" and "Block" to be statistically significant factors ($p < 5\ e^{-5}$), and the following interaction terms were also significant ($p < 0.0005$): Strain:IPTG, Strain:Media, and Strain:Block. The significance of the interaction terms in the model show a relationship between each of the factors, showing expression by each strain was impacted by another mechanism involving IPTG concentration, media, and block. Tukey's HSD contrasts showed the following strains to be statistically significantly different ($p < 0.005$): CTRL-CFP1, CTRL-CFP3, CFP1-CFP2, and CFP2-CFP3.

Table 21. Experiment *In Vitro* 1 Results

Strain	Fluorescence	Media	IPTG Concentration (µM)
CTRL	11.56	Lysis Buffer	50
	16.48	Lysis Buffer	50
	31.84	Lysis Buffer	50
	37.12	Lysis Buffer	50
CFP1	82.36	Lysis Buffer	50
	49.44	Lysis Buffer	50
	45.28	Lysis Buffer	50
	57.16	Lysis Buffer	50
CFP2	81.08	Lysis Buffer	50
	35.76	Lysis Buffer	50
	47.08	Lysis Buffer	50
	41.24	Lysis Buffer	50
CFP3	67.56	Lysis Buffer	50
	61.4	Lysis Buffer	50
	32	Lysis Buffer	50
	41.2	Lysis Buffer	50
CTRL	37.6	LB	50
	43.12	LB	50
	38.64	LB	50
	47.4	LB	50
CFP1	19.6	LB	50
	39.92	LB	50
	55.6	LB	50
	19.8	LB	50
CFP2	73.75	LB	50
	32.75	LB	50
	47.5	LB	50
	46.75	LB	50

CFP3	507	LB	50
	61	LB	50
	50.75	LB	50
	36.75	LB	50

Table 22. Experiment *In Vitro 2* Results

Strain	Fluorescence	Media	IPTG Concentration (µM)
CTRL	68.65	LB	1
	3.70	LB	10
	76.30	LB	50
	56.65	LB	100
	78.61	LB	1000
	87.00	LB	3000
CFP1	105.09	LB	1
	365.61	LB	10
	376.22	LB	50
	464.04	LB	100
	509.96	LB	1000
	535.30	LB	3000
CFP2	52.43	LB	1
	52.35	LB	10
	61.70	LB	50
	67.04	LB	100
	81.17	LB	1000
	78.65	LB	3000
CFP3	102.87	LB	1
	443.74	LB	10
	366.17	LB	50
	374.35	LB	100
	410.13	LB	1000
	380.17	LB	3000
CTRL	12.39	Lysis Buffer	1
	32.04	Lysis Buffer	10
	42.30	Lysis Buffer	50
	60.35	Lysis Buffer	100

	57.26	Lysis Buffer	1000
	54.04	Lysis Buffer	3000
CFP1	35.17	Lysis Buffer	1
	140.52	Lysis Buffer	10
	186.52	Lysis Buffer	50
	183.52	Lysis Buffer	100
	199.35	Lysis Buffer	1000
	181.78	Lysis Buffer	3000
CFP2	36.00	Lysis Buffer	1
	59.83	Lysis Buffer	10
	37.39	Lysis Buffer	50
	19.39	Lysis Buffer	100
	41.65	Lysis Buffer	1000
	38.04	Lysis Buffer	3000
CFP3	42.30	Lysis Buffer	1
	73.17	Lysis Buffer	10
	114.74	Lysis Buffer	50
	95.83	Lysis Buffer	100
	103.83	Lysis Buffer	1000
	113.00	Lysis Buffer	3000

Table 23. Experiment *In Vitro* 3 Results

Strain	Fluorescence	Media	IPTG Concentration (µM)
CTRL	31	Lysis Buffer	50
	5.75	Lysis Buffer	50
	32.25	Lysis Buffer	500
	3.25	Lysis Buffer	500
	61.75	LB	50
	64	LB	50
	67	LB	500
	55.5	LB	500
CFP1	968.5	Lysis Buffer	50
	1059.25	Lysis Buffer	50
	798.75	Lysis Buffer	500
	558.5	Lysis Buffer	500
	1925	LB	50
	1778.75	LB	50
	1587.75	LB	500
	1578.25	LB	500
CFP2	41.5	Lysis Buffer	50
	18.25	Lysis Buffer	50
	21.5	Lysis Buffer	500
	22.5	Lysis Buffer	500
	46.5	LB	50
	65.25	LB	50
	50.5	LB	500
	69.5	LB	500
CFP3	942	Lysis Buffer	50
	598.75	Lysis Buffer	50
	1044.25	Lysis Buffer	500
	514.75	Lysis Buffer	500

	709.25	LB	50
	519	LB	50
	706	LB	500
	576	LB	500

Tables 24 and 25 show the fluorescence values for strains CFP4 & CFP5, experiments *In Vitro* 4 and 5, at 90 hours into each of the two plate reader assays. “Strain,” “IPTG Concentration,” “Block,” and “Media” were the factors taken into account for this multiple linear regression analysis of the two experiments. Additionally, interaction terms between each factor were included in the model. “Strain” was not a significant factor impacting protein expression ($p > 0.05$), and none of the interaction terms involving “Strain” were significant. “Block” was significant ($p < 0.005$) meaning there were significant differences in expression between experiments. Tukey’s HSD contrasts between each strain showed statistically insignificant differences between each strain’s performance.

Table 24. Experiment *In Vitro* 4 Results

Strain	Fluorescence	Media	IPTG Concentration (µM)
	32.25	Lysis Buffer	50
	16	Lysis Buffer	50
	40.25	Lysis Buffer	500
	28.5	Lysis Buffer	500
	34	LB	50
	9.25	LB	50
	63.25	LB	500
CTRL	73.25	LB	500
	13.75	Lysis Buffer	50
	21.25	Lysis Buffer	50
	14.5	Lysis Buffer	500
	17.5	Lysis Buffer	500
	46.75	LB	50
	44.25	LB	50
	35.25	LB	500
CFP4	70	LB	500
	33.75	Lysis Buffer	50
	15	Lysis Buffer	50
	29	Lysis Buffer	500
	29.25	Lysis Buffer	500
	34.5	LB	50
	59.5	LB	50
	73.5	LB	500
CFP5	32	LB	500

Table 25. Experiment *In Vitro* 4 Results

Strain	Fluorescence	Media	IPTG Concentration (µM)
	45.5	Lysis Buffer	50
	56	Lysis Buffer	50
	33	Lysis Buffer	500
	32.25	Lysis Buffer	500
	94	LB	50
	81.75	LB	50
	76.75	LB	500
CTRL	36.5	LB	500
	10.25	Lysis Buffer	50
	4.5	Lysis Buffer	50
	17.25	Lysis Buffer	500
	28.5	Lysis Buffer	500
	92.5	LB	50
	97.5	LB	50
	105	LB	500
CFP4	62.5	LB	500
	40.5	Lysis Buffer	50
	42	Lysis Buffer	50
	24	Lysis Buffer	500
	21.5	Lysis Buffer	500
	97.25	LB	50
	73.75	LB	50
	78	LB	500
CFP5	39.75	LB	500

A.3.2 Optimum media for protein expression

Multiple linear regression analysis showed "Media" to be a statistically significant factor impacting cell-free protein expression for strains CFP1-3 ($p < 5\ e^{-5}$). Resuspending the lysed cells in LB media consistently produced higher protein concentrations than resuspending the lysed cells in lysis buffer. Additionally, there was significant interaction between the "Strain" and "Media" factors ($p < 0.0005$).

Multiple linear regression analysis of experiments *In Vitro* 4 and 5 revealed "Media" to be a statistically significant factor impacting protein expression ($p < 1\ e^{-8}$). In agreement with previous analysis, LB media produced higher protein concentrations than resuspending cells in lysis buffer. Additionally, the interaction terms between "Media" and "Block" and "Media" and "Strain" generated significant impacts on expression due to mechanisms acting between each factor ($p < 0.05$).

A.3.3 Optimum IPTG Concentration for Maximum Protein Expression

Statistical multiple linear regression analysis of strains CFP1-3 in a cell-free system showed "IPTG Concentration" to be a significant factor when modeling protein expression ($p < 5\ e^{-5}$). Additionally, the interaction term involving "Strain" and "IPTG Concentration" was significant, showing a relationship between these two factors impacting expression. For experiments *In Vitro* 4 & 5, "IPTG Concentration" was not a significant factor affecting protein expression ($p > 0.05$); however, the interaction between "IPTG Concentration" and "Block" was significant ($p < 0.01$).

A.3.4 Cell-Free vs. *In Vivo* System

This thesis investigates a variety of factors impacting protein expression using both cell-free and whole-cell methods. Multiple linear regression analysis was performed for experiments *In Vivo* 4 and *In Vitro* 2 due to the employment of the exact same conditions in both experiments with the only difference being the system. The only difference between these experiments is the method of measuring fluorescence: fluorescence in the *In Vivo* system was measured with relative fluorescence once the system reached the

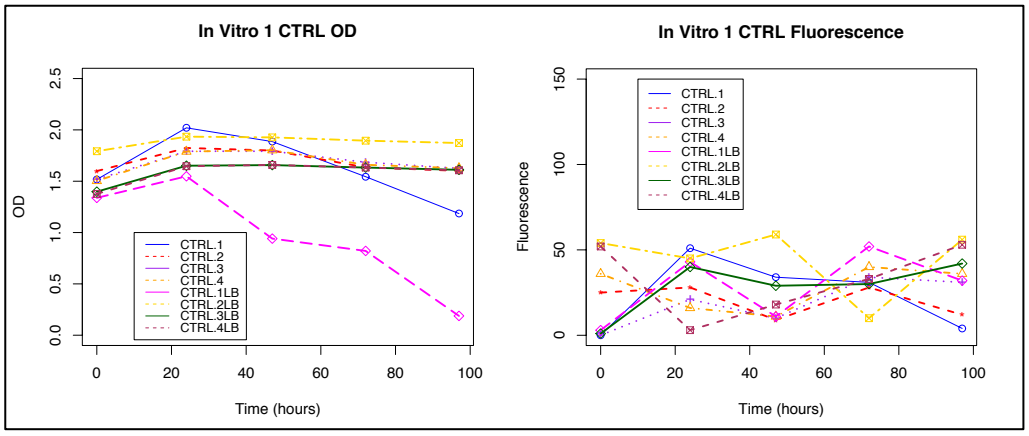
stationary phase and at least 18 hours post-induction while the cell-free system fluorescence was measured using total fluorescence at 90 hours. This analysis with "Strain," "IPTG Concentration," and "System" as the three main factors revealed "System' to be a statistically significant factor when considering protein expression with the *in vivo* system producing significantly greater amounts of protein ($p < 0.005$). While the *in vivo* system yielded higher fluorescence, this may be due to the proteins being contained inside cell membranes, indicating unequal distribution of proteins in the system with some areas (cells) containing very high fluorescence while areas containing media would have lower fluorescence. Consequently, if the laser measuring protein hits a cell with high density of fluorescent proteins, this would yield a much higher fluorescence that may not accurately reflect the entire system. Cell-free systems have greater dispersion of proteins since these proteins are not contained in membranes. Therefore, distribution of these proteins about the reactor is relatively equal; consequently, these systems would produce lower fluorescence due to the equal dispersion of proteins. In the future more research should be done to normalize these systems so system performance, or protein expression, can be directly compared.

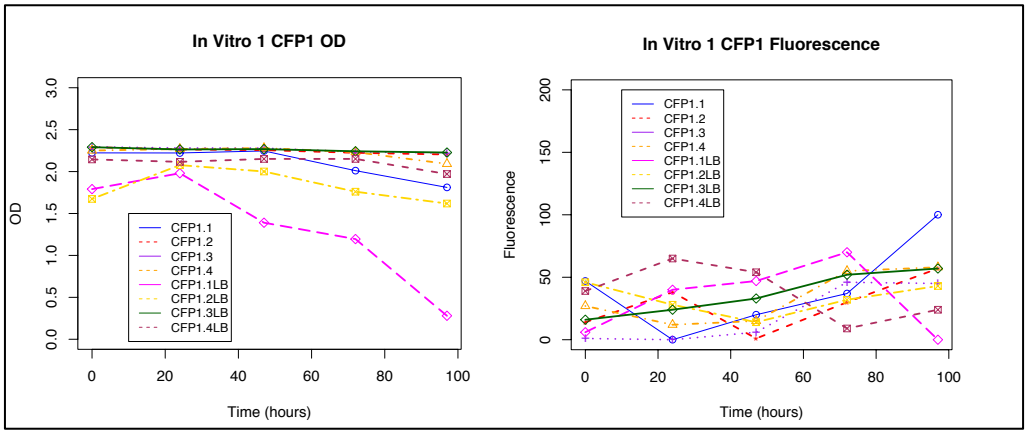
A.4: CONCLUSION

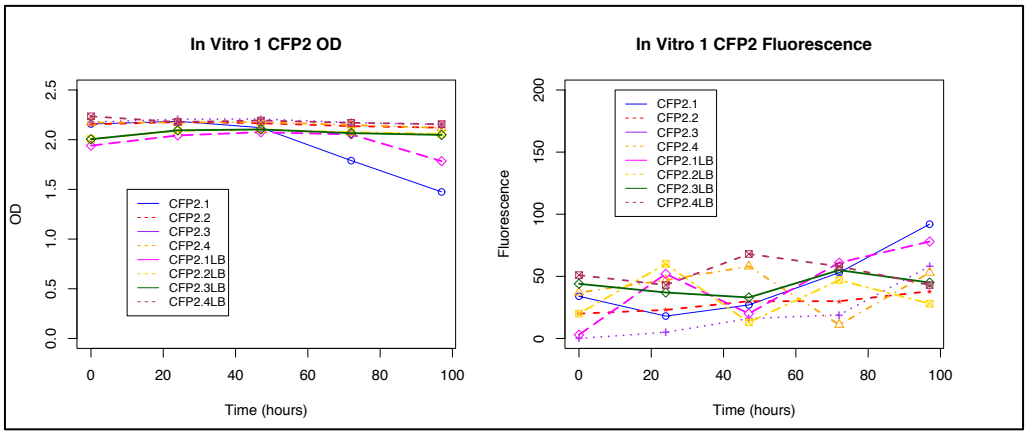
Cell-free systems offer a number of benefits compared to *in vivo* systems including eliminating membrane transport limitations, allow for more direct genetic modifications, and cell viability is no longer a concern when synthesizing toxic products.[44, 45, 46] However, in this study the cell-free system did not produce greater amounts of protein; this may be due to the differences in protein dispersion between the *in vivo* and *in vitro* systems. In fact, the *in vivo* system produced significantly greater amounts of protein in less than half the time of the cell-free system. The factors altered throughout the studies showed similar impacts on the system for both whole-cell and cell-free experiments with IPTG concentration impacting protein expression; however, optimum IPTG concentration could not be determined for any experiment due to high variability in the results. Strain performance was slightly different from performance in a whole-cell system, but still did not accurately depict predictions from the RBS Calculator. CFP1

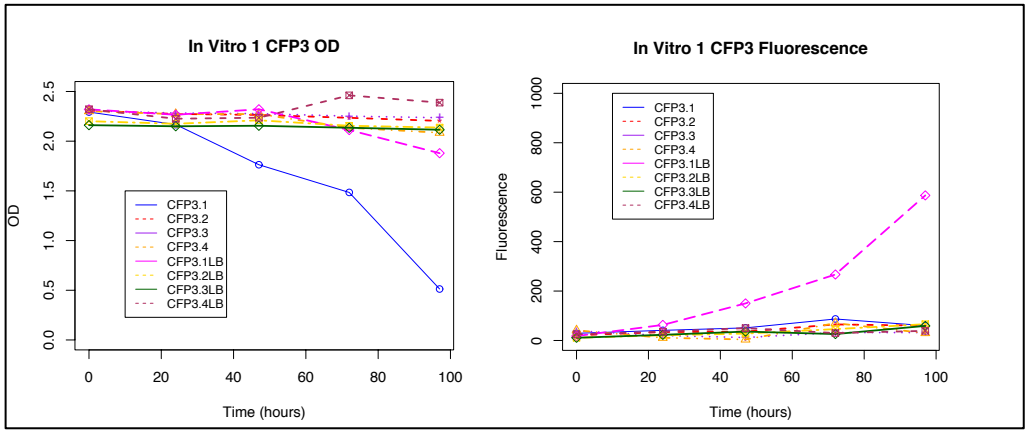
produced the most protein among the strains, similarly to the whole-cell experiments. Additionally, it was shown that cell-free systems functioned much more efficiently when resuspended in LB media compared to the lysis buffer.

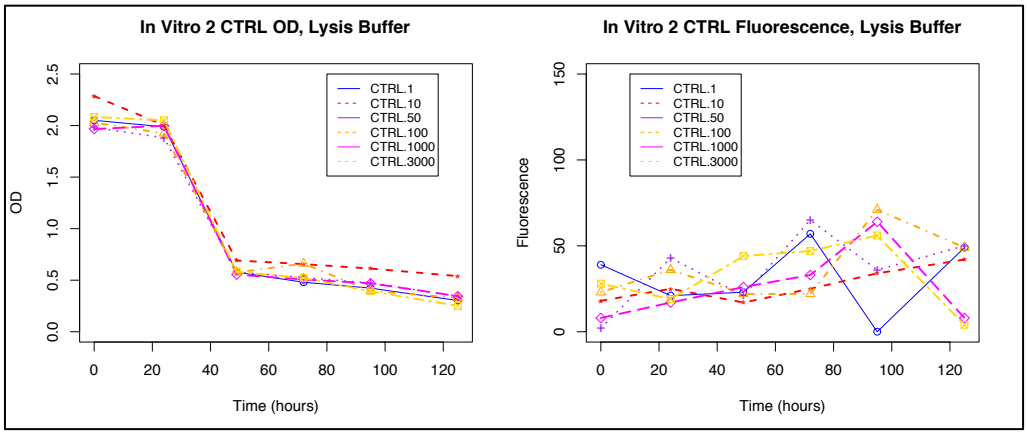
A.5: SUPPLEMENTARY DATA

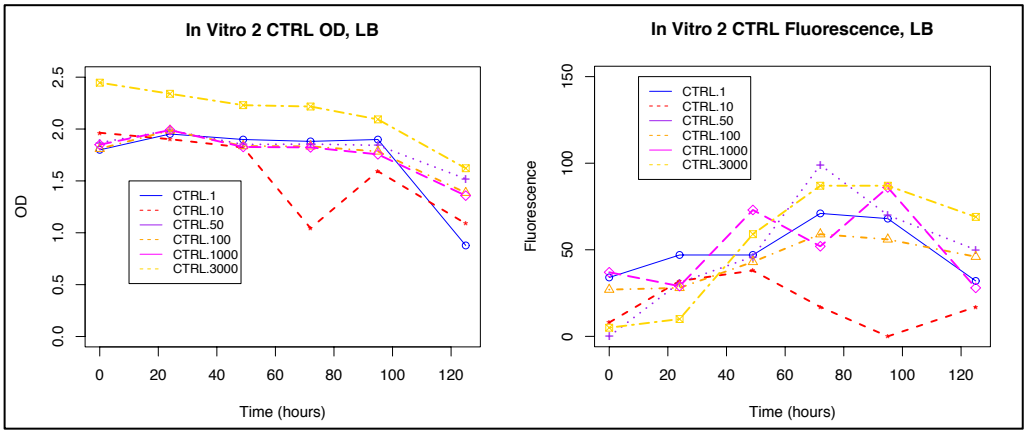


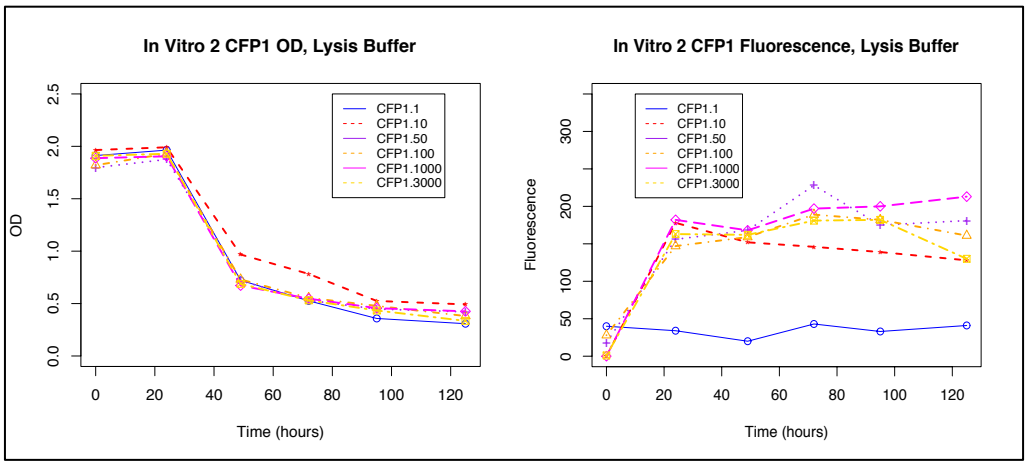


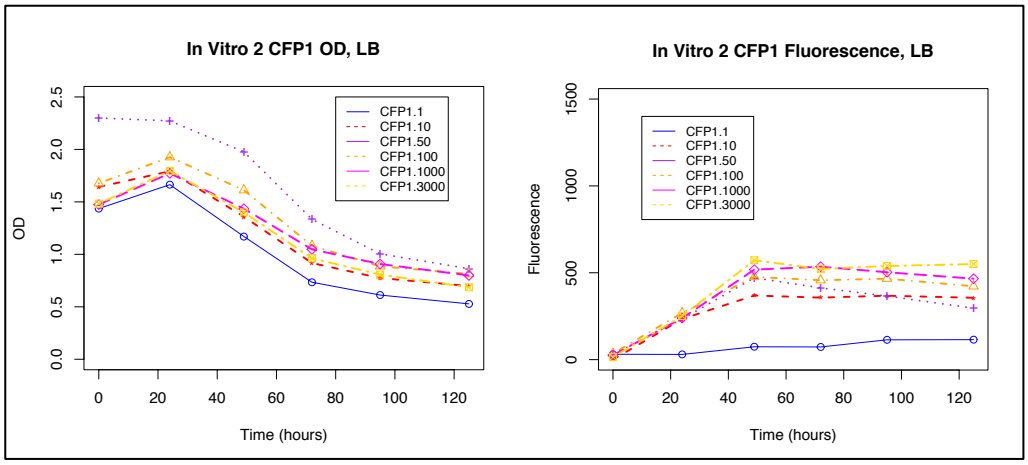


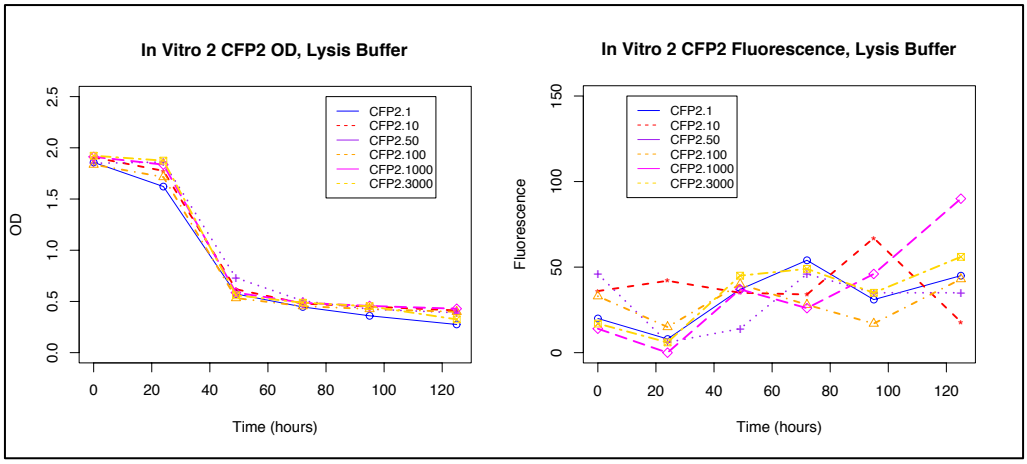

In Vitro 1 CFP3 OD
OD
Time (hours)
CFP3.1
CFP3.2
CFP3.3
CFP3.4
CFP3.1LB
CFP3.2LB
CFP3.3LB
CFP3.4LB
In Vitro 1 CFP3 Fluorescence
Fluorescence
Time (hours)

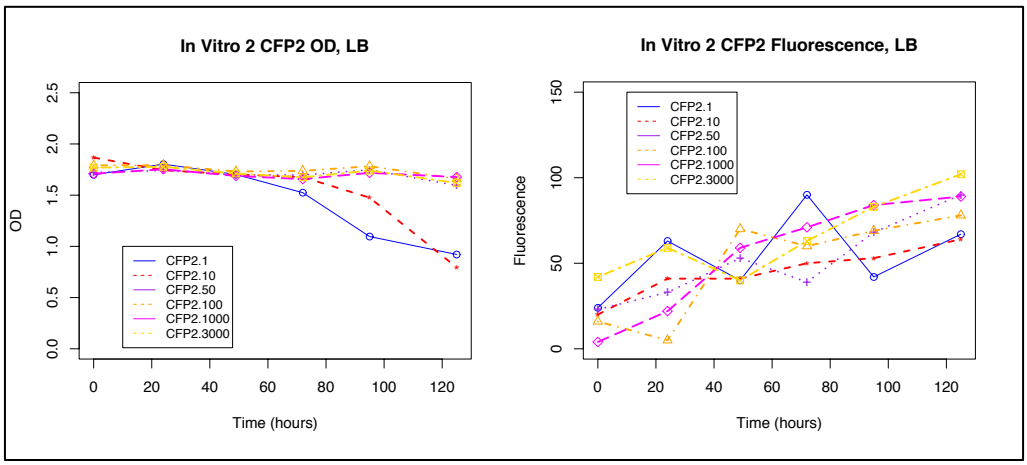

In Vitro 2 CTRL OD, Lysis Buffer
OD
Time (hours)
CTRL.1
CTRL.10
CTRL.50
CTRL.100
CTRL.1000
CTRL.3000
In Vitro 2 CTRL Fluorescence, Lysis Buffer
Fluorescence
Time (hours)

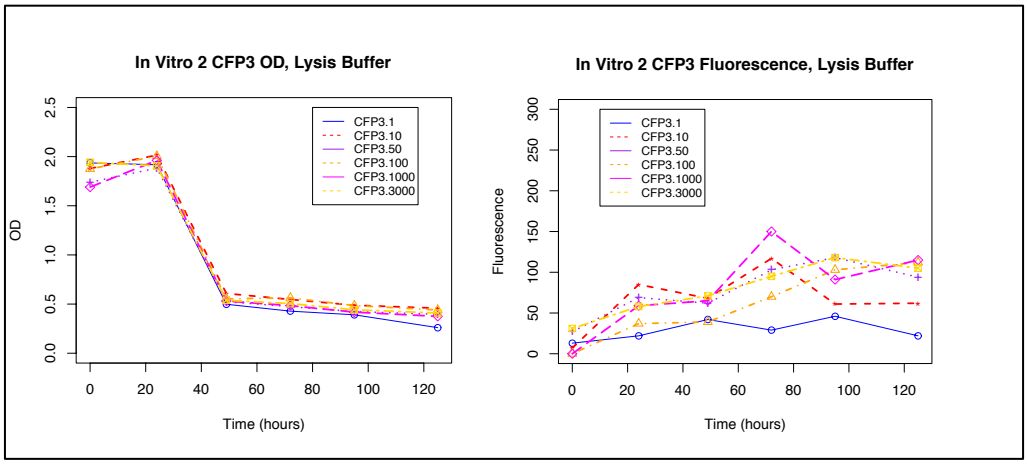

In Vitro 2 CTRL OD, LB
OD
Time (hours)
CTRL.1
CTRL.10
CTRL.50
CTRL.100
CTRL.1000
CTRL.3000
In Vitro 2 CTRL Fluorescence, LB
Fluorescence
Time (hours)

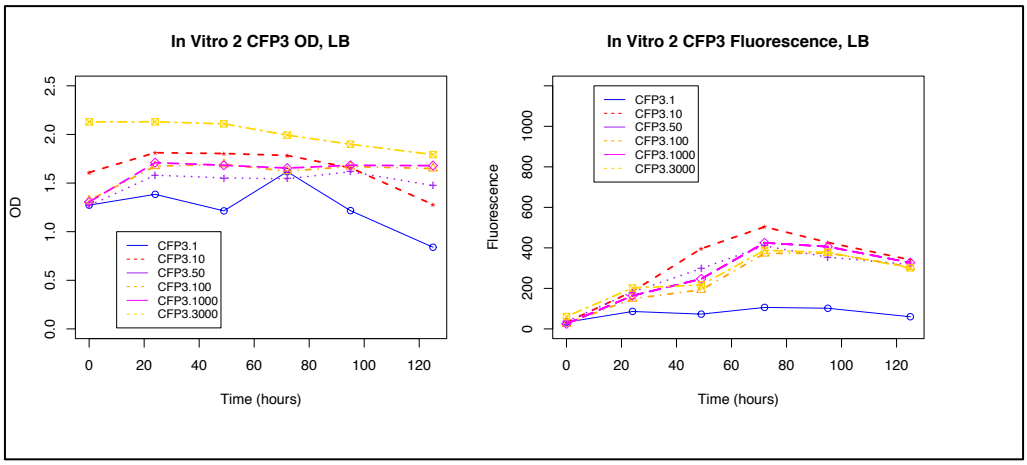

In Vitro 2 CFP1 OD, Lysis Buffer
OD
Time (hours)
CFP1.1
CFP1.10
CFP1.50
CFP1.100
CFP1.1000
CFP1.3000
In Vitro 2 CFP1 Fluorescence, Lysis Buffer
Fluorescence
Time (hours)

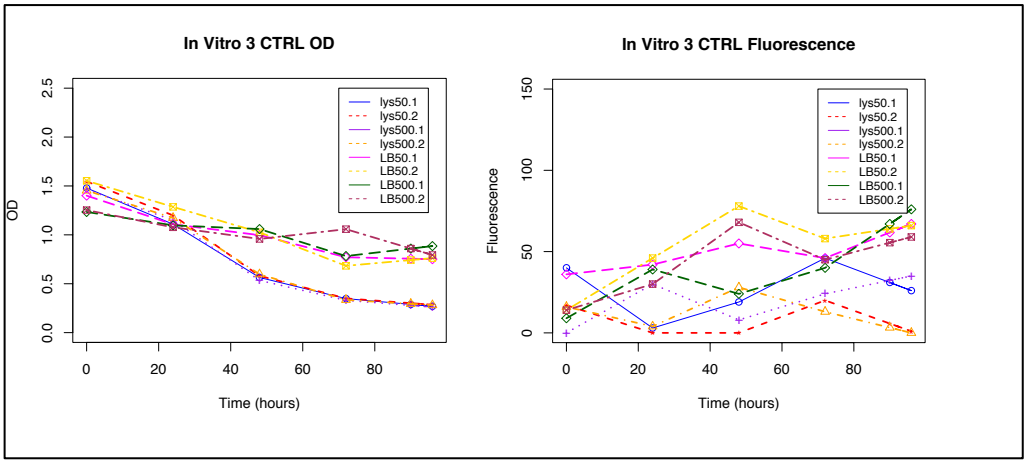

In Vitro 2 CFP1 OD, LB
OD
Time (hours)
CFP1.1
CFP1.10
CFP1.50
CFP1.100
CFP1.1000
CFP1.3000
In Vitro 2 CFP1 Fluorescence, LB
Fluorescence
Time (hours)

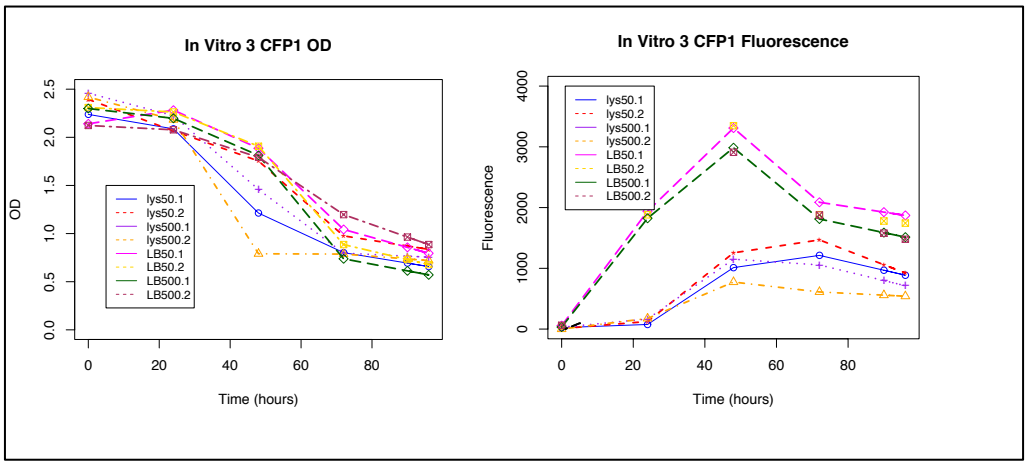

In Vitro 2 CFP2 OD, Lysis Buffer
OD
Time (hours)
CFP2.1
CFP2.10
CFP2.50
CFP2.100
CFP2.1000
CFP2.3000
In Vitro 2 CFP2 Fluorescence, Lysis Buffer
Fluorescence
Time (hours)

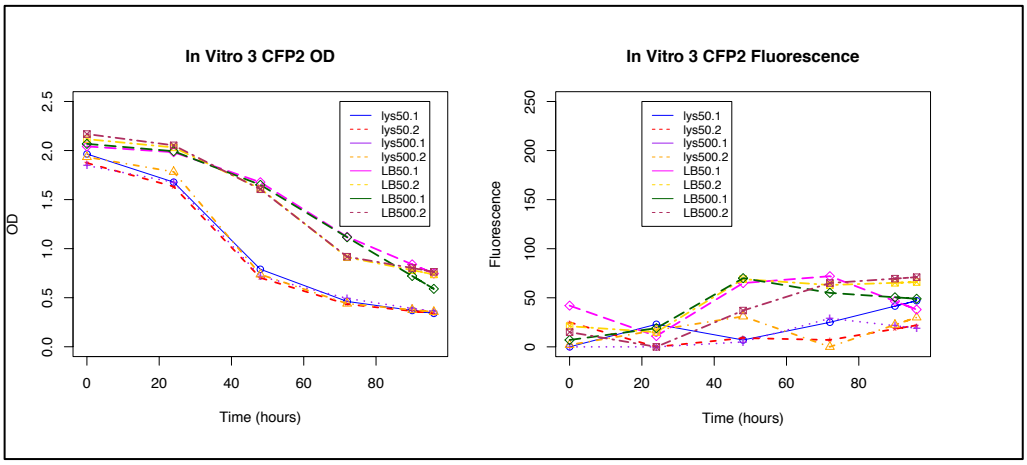

In Vitro 2 CFP2 OD, LB
OD
Time (hours)
CFP2.1
CFP2.10
CFP2.50
CFP2.100
CFP2.1000
CFP2.3000
In Vitro 2 CFP2 Fluorescence, LB
Fluorescence
Time (hours)

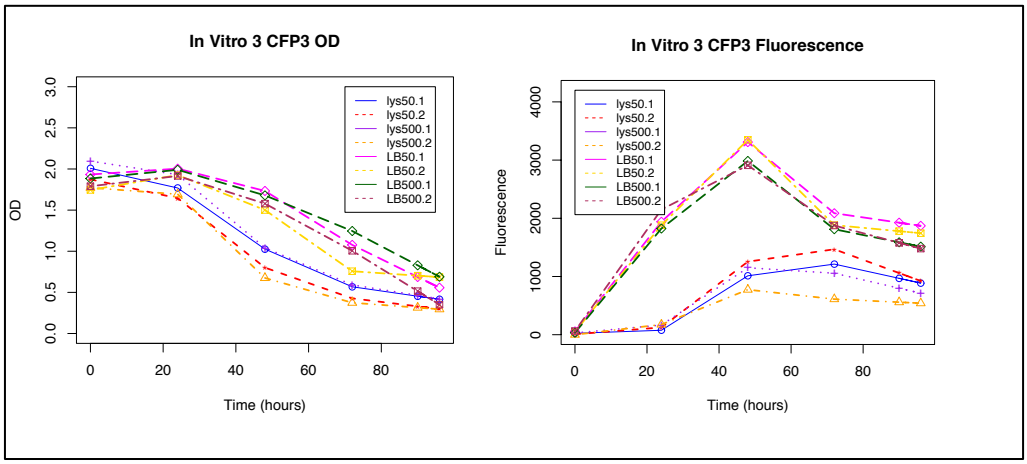

In Vitro 2 CFP3 OD, Lysis Buffer
OD
Time (hours)
CFP3.1
CFP3.10
CFP3.50
CFP3.100
CFP3.1000
CFP3.3000
In Vitro 2 CFP3 Fluorescence, Lysis Buffer
Fluorescence
Time (hours)

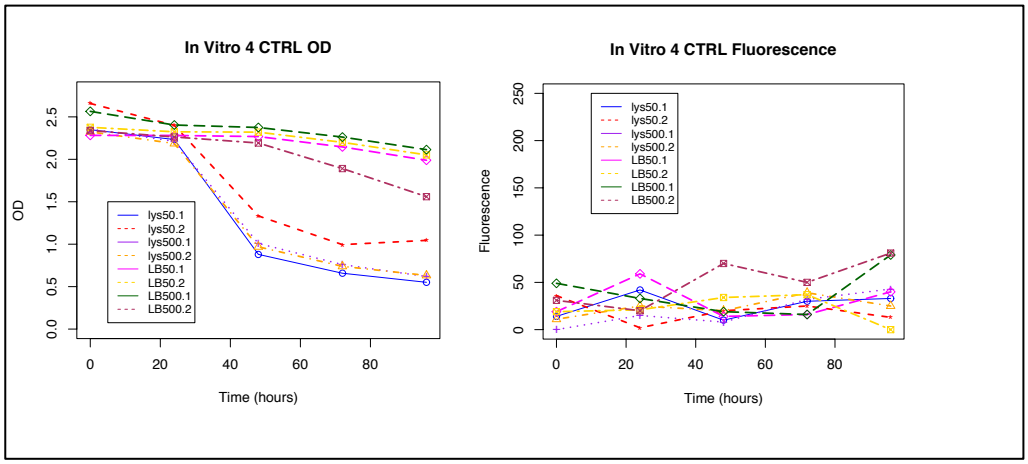

In Vitro 2 CFP3 OD, LB
OD
Time (hours)
CFP3.1
CFP3.10
CFP3.50
CFP3.100
CFP3.1000
CFP3.3000
In Vitro 2 CFP3 Fluorescence, LB
Fluorescence
Time (hours)

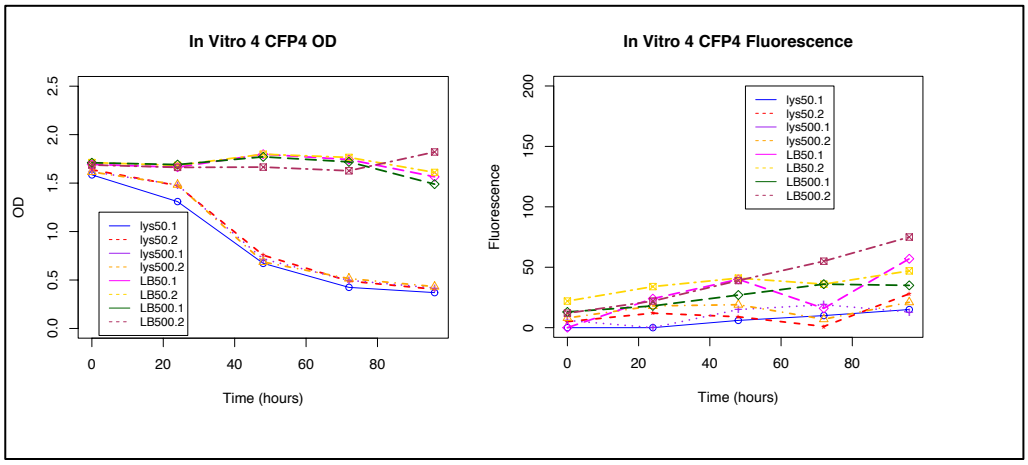

In Vitro 3 CTRL OD
OD
Time (hours)
In Vitro 3 CTRL Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2

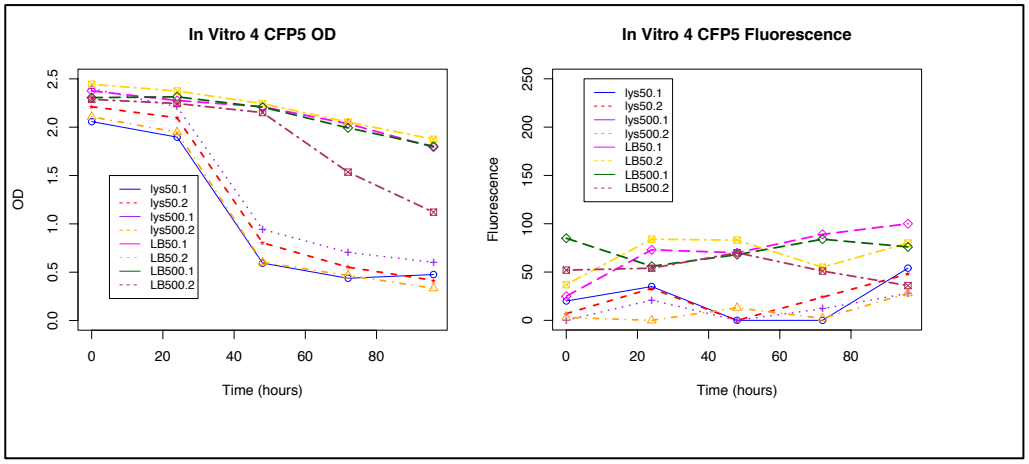

In Vitro 3 CFP1 OD
OD
Time (hours)
In Vitro 3 CFP1 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2

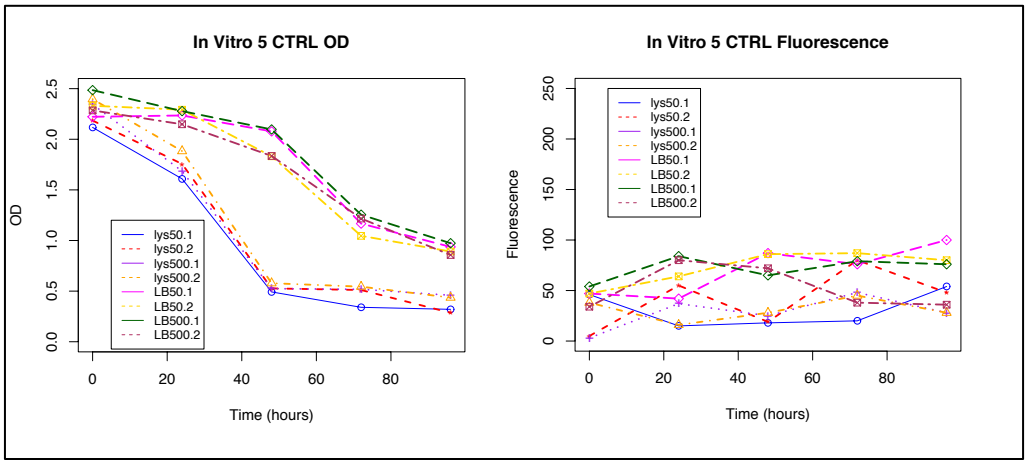

In Vitro 3 CFP2 OD
OD
Time (hours)
In Vitro 3 CFP2 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2

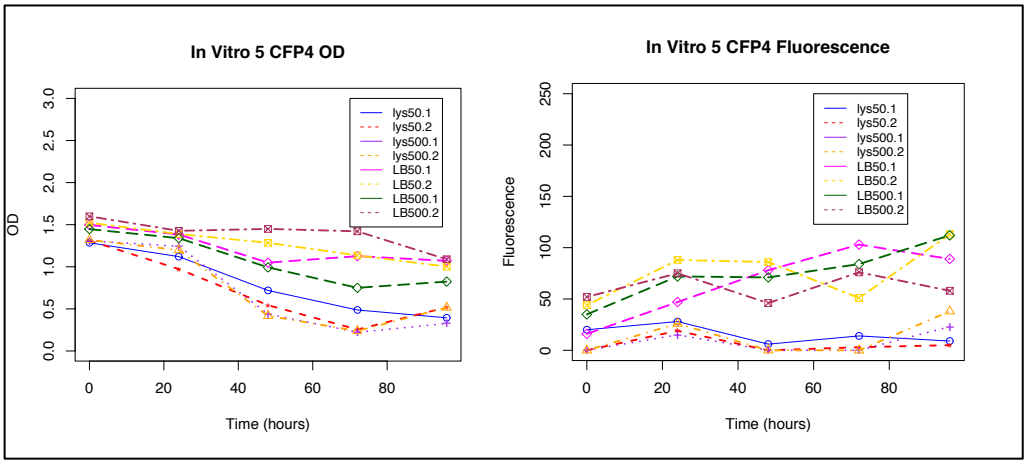

In Vitro 3 CFP3 OD
OD
Time (hours)
In Vitro 3 CFP3 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2

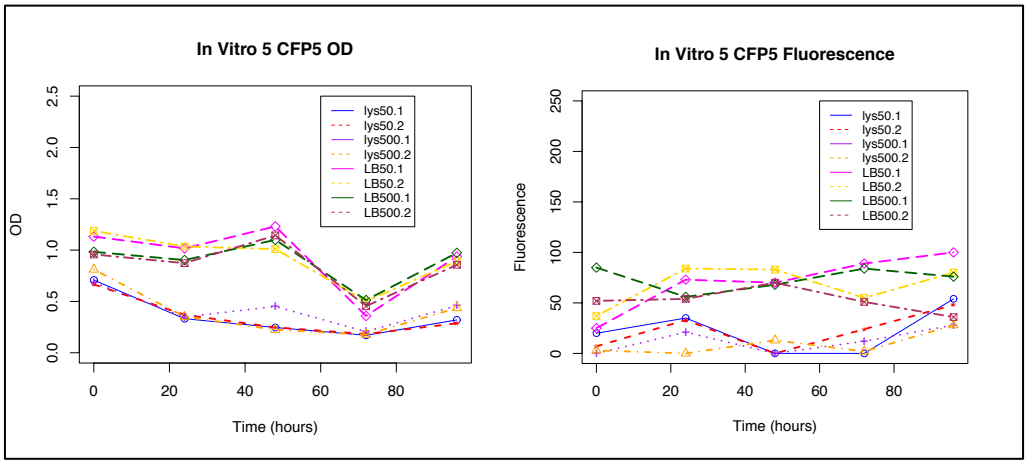

In Vitro 4 CTRL OD
OD
Time (hours)
In Vitro 4 CTRL Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2


In Vitro 4 CFP4 OD
OD
Time (hours)
In Vitro 4 CFP4 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2


In Vitro 4 CFP5 OD
OD
Time (hours)
In Vitro 4 CFP5 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2


In Vitro 5 CTRL OD
OD
Time (hours)
In Vitro 5 CTRL Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2


In Vitro 5 CFP4 OD
OD
Time (hours)
In Vitro 5 CFP4 Fluorescence
Fluorescence
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2


In Vitro 5 CFP5 OD
OD
Time (hours)
lys50.1
lys50.2
lys500.1
lys500.2
LB50.1
LB50.2
LB500.1
LB500.2
In Vitro 5 CFP5 Fluorescence
Fluorescence
Time (hours)

Appendix B: R-STUDIO CODE FOR DATA ANALYSIS

```
#in vivo 1

vivo1 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
1_STAT.csv", header=T)
vivo1

vivo1$STRAIN <- as.factor(vivo1$STRAIN)
fit <- lm(REL.FLUOR ~ STRAIN, data=vivo1)
summary(fit)
anova(fit)

qqnorm(fit$residuals)  # checking residuals
qqline(fit$residuals)

library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

#in vivo 2
vivo <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
2_STAT.csv", header=T)
vivo

vivo$STRAIN <- as.factor(vivo$STRAIN)
vivo$IND.TIME <- as.factor(vivo$IND.TIME)
fit <- lm(REL.FLUOR ~ STRAIN + IND.TIME + STRAIN:IND.TIME, data=vivo)
summary(fit)
anova(fit)
```

```
library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IND.TIME")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

# in vivo 3

vivo3 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/IN VIVO
3_STAT.csv", header=T)
vivo3

vivo3$STRAIN <- as.factor(vivo3$STRAIN)
vivo3$IND.TIME <- as.factor(vivo3$IND.TIME)
fit3 <- lm(REL.FLUOR ~ STRAIN + IND.TIME + STRAIN:IND.TIME, data=vivo3)
summary(fit3)
anova(fit3)

library(emmeans)
fitA_e <- emmeans(fit3, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")
```

```
library(emmeans)
fitA_e <- emmeans(fit3, "IND.TIME")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

## in vivo 2 & 3

vivo23 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
2&3 50_3000.csv", header=T)
vivo23

vivo23$STRAIN <- as.factor(vivo23$STRAIN)
vivo23$IPTG <- as.factor(vivo23$IPTG)
vivo23$IND.TIME <- as.factor(vivo23$IND.TIME)
vivo23$BLOCK <- as.factor(vivo23$BLOCK)
fit <- lm(REL.FLUOR ~ STRAIN + IPTG + IND.TIME + STRAIN:IPTG +
IPTG:IND.TIME + STRAIN:IND.TIME, data=vivo23)
summary(fit)
anova(fit)

library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IPTG")
pairs(fitA_e)
plot(fitA_e)
```

```
cld(fitA_e,adjust="Tukey")

# in vivo 4

vivo4 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
4_STAT.csv", header=T)
vivo4

vivo4$STRAIN <- as.factor(vivo4$STRAIN)
vivo4$IPTG <- as.factor(vivo4$IPTG)
fit <- lm(REL.FLUOR ~ STRAIN + IPTG, data=vivo4)
summary(fit)
anova(fit)

library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

# in vivo 5

vivo5 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/IN VIVO 5
37 STAT.csv", header=T)
vivo5
```

```
vivo5$STRAIN <- as.factor(vivo5$STRAIN)
vivo5$IPTG <- as.factor(vivo5$IPTG)
fit15 <- lm(REL.FLUOR ~ STRAIN + IPTG, data=vivo5)
summary(fit15)
anova(fit15)

library(emmeans)
fitA_e <- emmeans(fit15, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit15, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

##in vivo 4&5

vivo45 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
45_STAT.csv", header=T)
vivo45

vivo45$STRAIN <- as.factor(vivo45$STRAIN)
vivo45$IPTG <- as.factor(vivo45$IPTG)
vivo45$TEMP <- as.factor(vivo45$TEMP)
fit <- lm(REL.FLUOR ~ STRAIN + IPTG + TEMP + STRAIN:IPTG + IPTG:TEMP +
STRAIN:TEMP, data=vivo45)
summary(fit)
```

```
anova(fit)

library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "TEMP")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

vivotempstrain3 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in
vivo/In Vivo 5 Temp_strain3.csv",header=T)
vivotemp3
DIFF <- vivotempstrain3$DIFF
t.test(vivotempstrain3$DIFF, data=vivotemp, alternative="greater")

# in vivo 6

vivo6 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo 6
STAT.csv", header=T)
vivo6
```

```
vivo6$STRAIN <- as.factor(vivo6$STRAIN)
vivo6$IPTG <- as.factor(vivo6$IPTG)
fit16 <- lm(REL.FLUOR ~ STRAIN + IPTG, data=vivo6)
summary(fit16)
anova(fit16)

library(emmeans)
fitA_e <- emmeans(fit16, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit16, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

# in vivo 7

vivo7 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo 7
STAT.csv", header=T)
vivo7

vivo7$STRAIN <- as.factor(vivo7$STRAIN)
vivo7$IPTG <- as.factor(vivo7$IPTG)
fit17 <- lm(REL.FLUOR ~ STRAIN + IPTG, data=vivo7)
summary(fit17)
anova(fit17)
```

```
library(emmeans)
fitA_e <- emmeans(fit17, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit17, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

# in vivo 67

vivo67 <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/in vivo/In Vivo
67 STAT.csv", header=T)
vivo67

vivo67$STRAIN <- as.factor(vivo67$STRAIN)
vivo67$IPTG <- as.factor(vivo67$IPTG)
vivo67$TEMP <- as.factor(vivo67$TEMP)
fit167 <- lm(REL.FLUOR ~ STRAIN + IPTG + TEMP + STRAIN:TEMP +
IPTG:TEMP + STRAIN:IPTG, data=vivo67)
summary(fit167)
anova(fit167)

library(emmeans)
fitA_e <- emmeans(fit167, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")
```

```
library(emmeans)
fitA_e <- emmeans(fit167, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit167, "TEMP")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")
#In vitro 1-3

vitro <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/In
Vitro13_STAT.csv", header=T)
vitro

vitro$STRAIN <- as.factor(vitro$STRAIN)
vitro$IPTG <- as.factor(vitro$IPTG)
vitro$MEDIA <- as.factor(vitro$MEDIA)
vitro$BLOCK <- as.factor(vitro$BLOCK)
fit <- lm(FLUOR ~ STRAIN + IPTG + MEDIA + BLOCK + STRAIN:IPTG +
STRAIN:MEDIA + STRAIN:BLOCK + IPTG:BLOCK + IPTG:MEDIA +
MEDIA:BLOCK, data=vitro)
summary(fit)
anova(fit)


library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
```

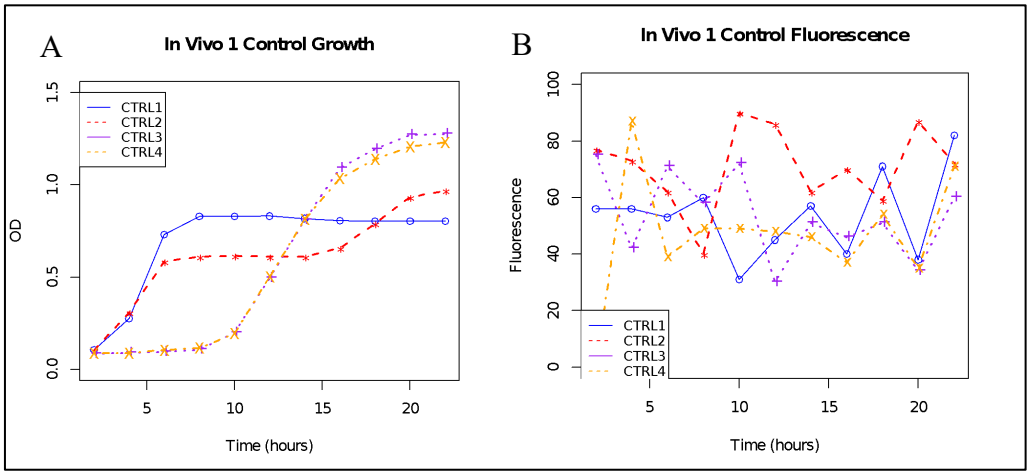
```
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "MEDIA")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

#in vitro 4&5
vitro <- read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/In
Vitro_45_STAT.csv", header=T)
vitro

vitro$STRAIN <- as.factor(vitro$STRAIN)
vitro$IPTG <- as.factor(vitro$IPTG)
vitro$MEDIA <- as.factor(vitro$MEDIA)
vitro$BLOCK <- as.factor(vitro$BLOCK)
fit <- lm(FLUOR ~ STRAIN + IPTG + MEDIA + BLOCK + STRAIN:IPTG +
STRAIN:MEDIA + STRAIN:BLOCK + IPTG:BLOCK + IPTG:MEDIA +
MEDIA:BLOCK, data=vitro)
summary(fit)
anova(fit)
```

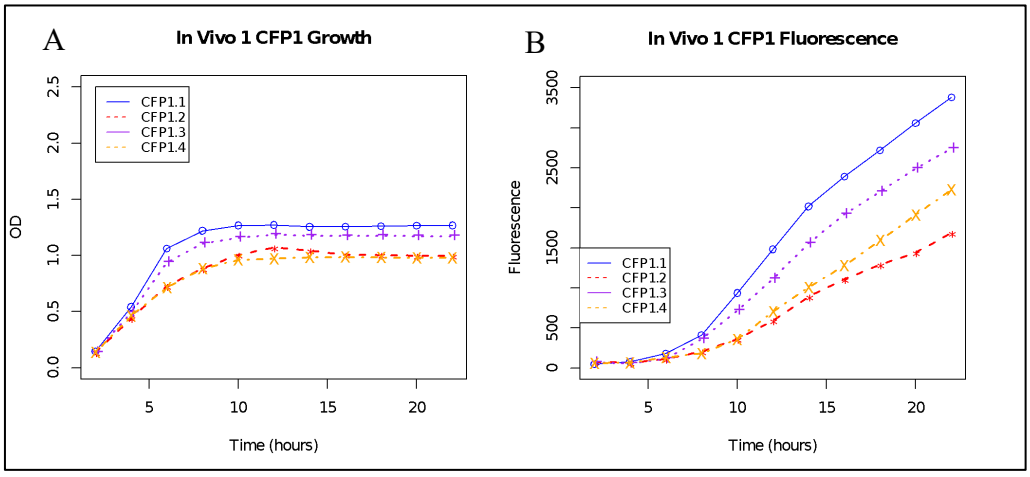
```
library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "IPTG")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "MEDIA")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

#in vivo 4 vs in vitro 2
vivovitro <-
read.csv("/Users/EmilyBerg/Desktop/Research/Experiments/InVivoVsVitro_STAT.csv",
header=T)
vivovitro

vivovitro$STRAIN <- as.factor(vivovitro$STRAIN)
vivovitro$IPTG <- as.factor(vivovitro$IPTG)
vivovitro$SYSTEM <- as.factor(vivovitro$SYSTEM)
fit <- lm(FLUOR ~ STRAIN + IPTG + SYSTEM + IPTG:SYSTEM + IPTG:STRAIN +
STRAIN:SYSTEM, data=vivovitro)
summary(fit)
```

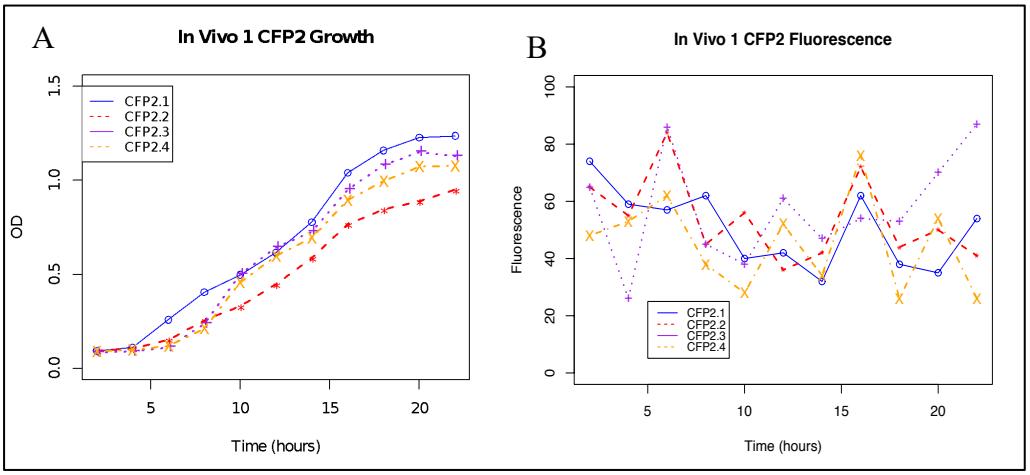
```
anova(fit)


library(emmeans)
fitA_e <- emmeans(fit, "STRAIN")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "SYSTEM")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")

library(emmeans)
fitA_e <- emmeans(fit, "MEDIA")
pairs(fitA_e)
plot(fitA_e)
cld(fitA_e,adjust="Tukey")
```

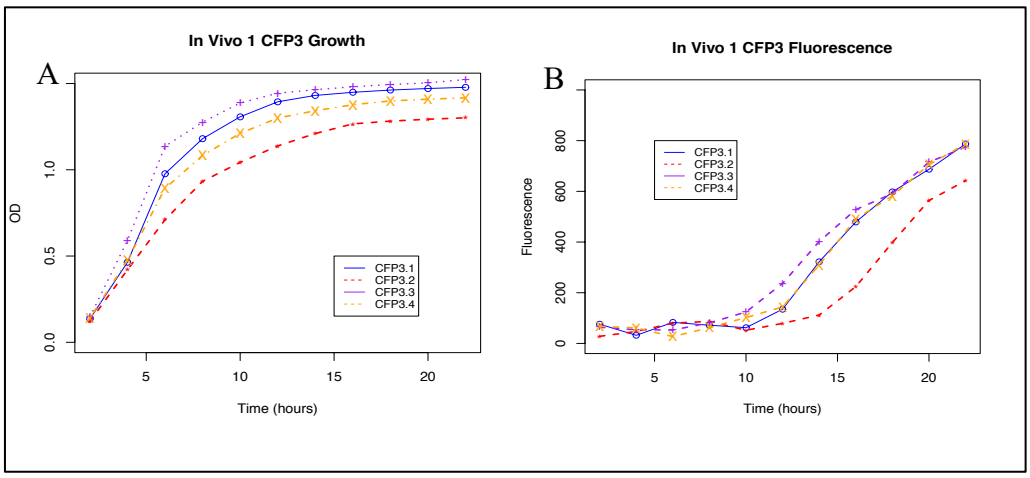
Appendix C: OD AND FLUORESCENCE READING- DATA BY EXPERIMENT

C.1: Experiment in vivo 1

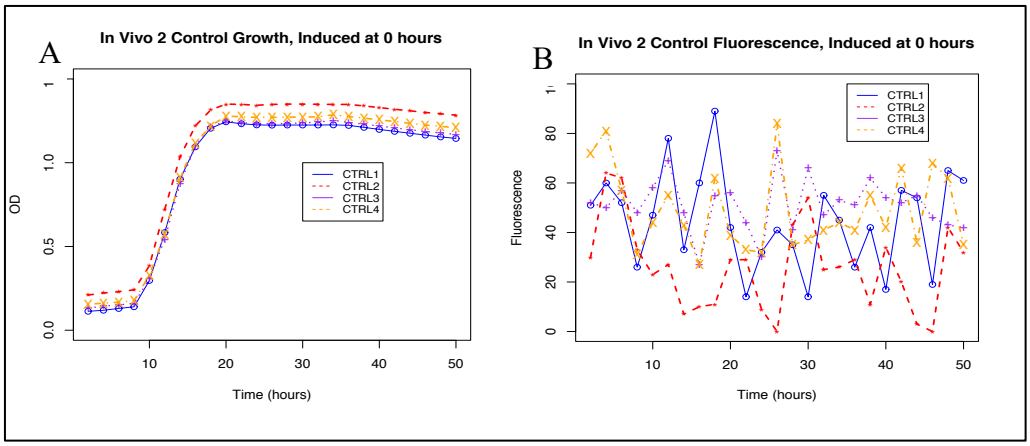


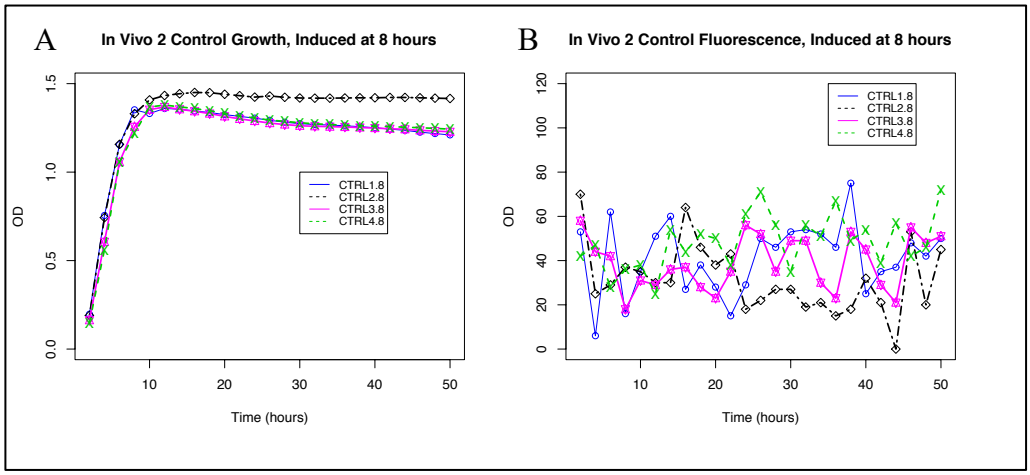


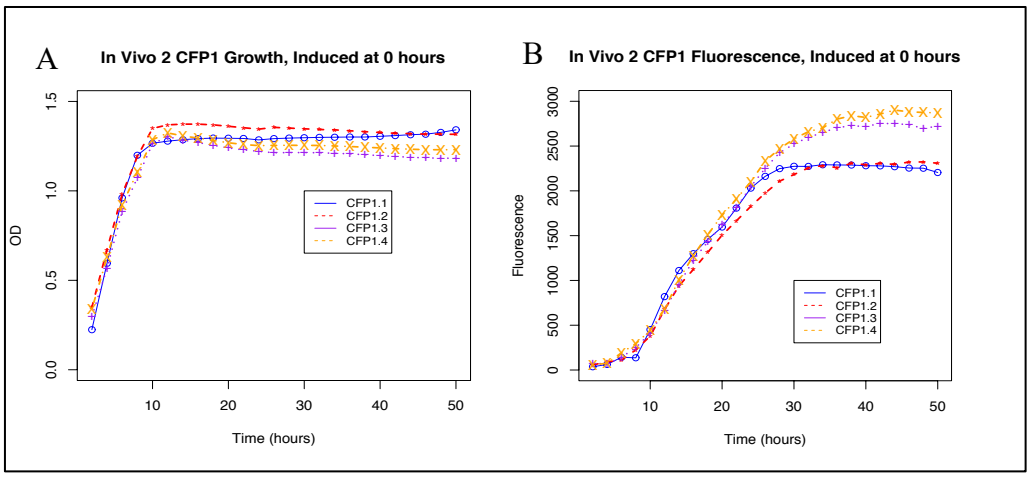

A
In Vivo 1 CFP2 Growth
OD
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4
B
In Vivo 1 CFP2 Fluorescence
Fluorescence
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4

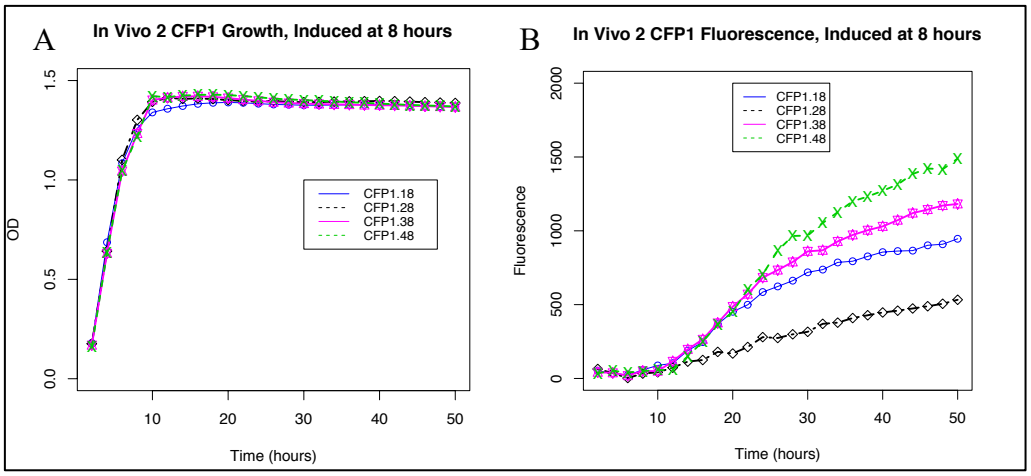

In Vivo 1 CFP3 Growth
A
OD
Time (hours)
CFP3.1
CFP3.2
CFP3.3
CFP3.4
In Vivo 1 CFP3 Fluorescence
B
Fluorescence
Time (hours)
CFP3.1
CFP3.2
CFP3.3
CFP3.4

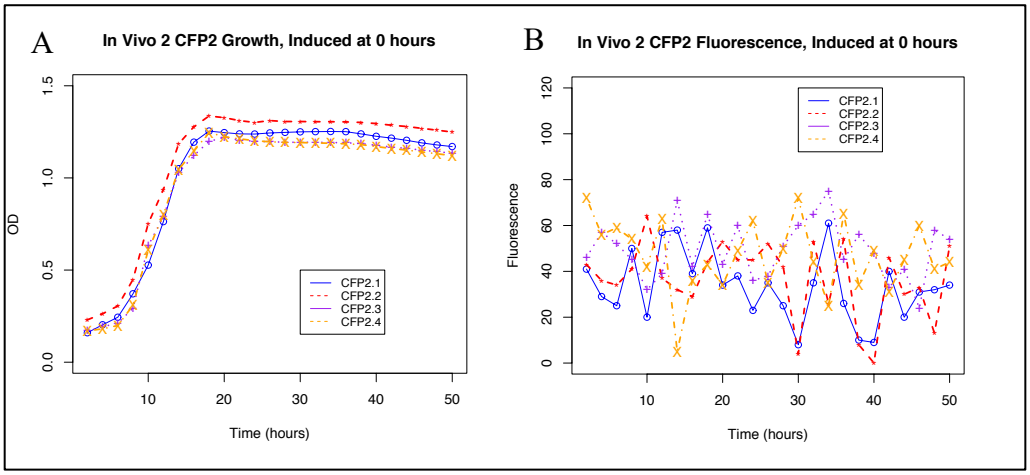
C.2: Experiment in vivo 2

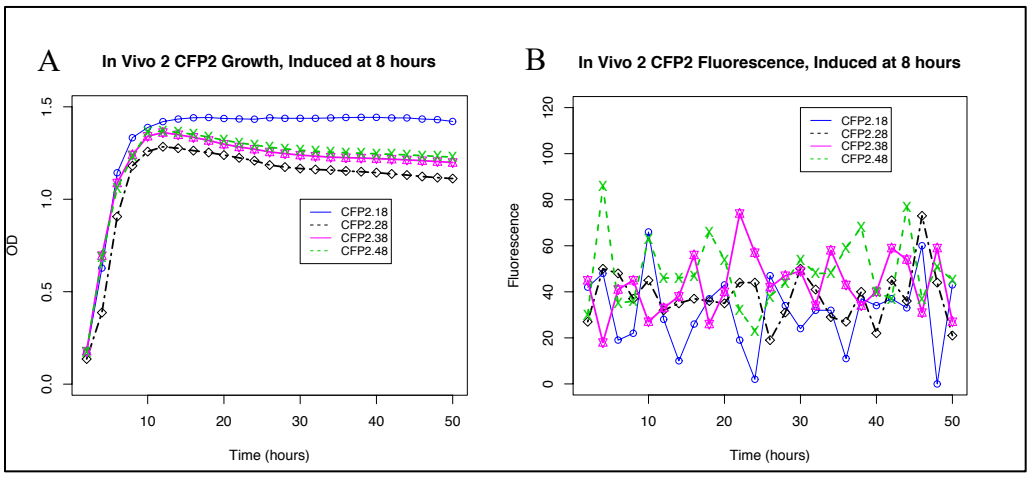


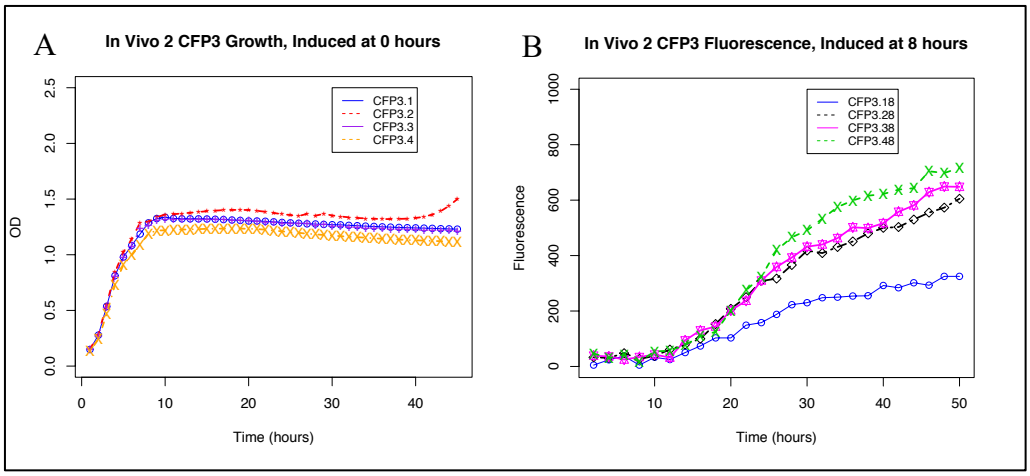


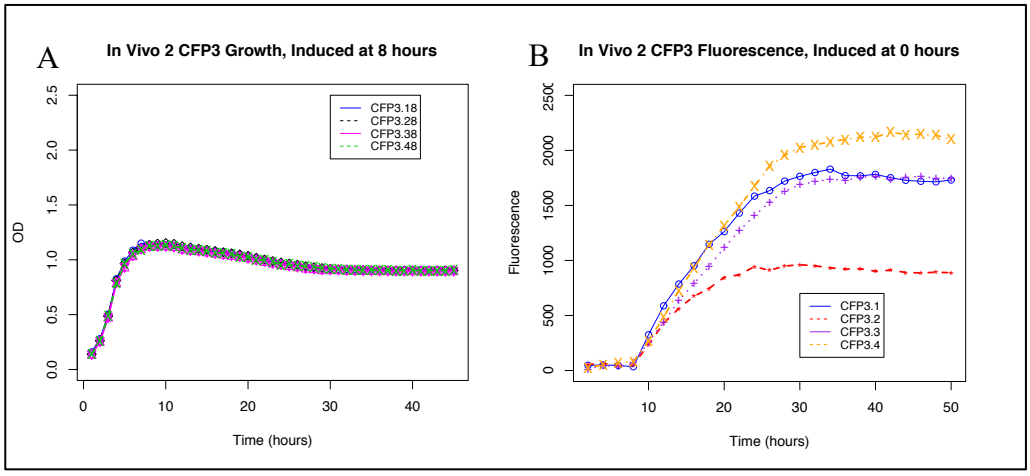



A
In Vivo 2 CFP1 Growth, Induced at 8 hours
OD
Time (hours)
CFP1.18
CFP1.28
CFP1.38
CFP1.48
B
In Vivo 2 CFP1 Fluorescence, Induced at 8 hours
Fluorescence
Time (hours)
CFP1.18
CFP1.28
CFP1.38
CFP1.48

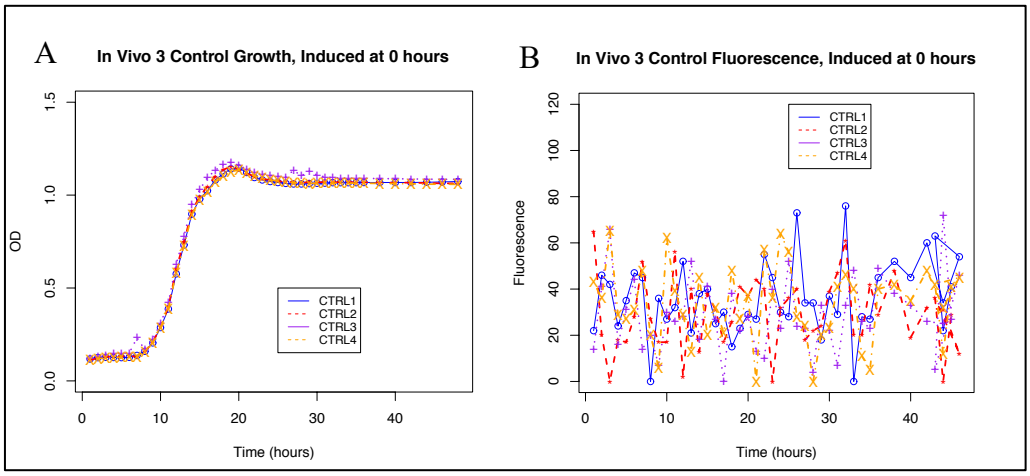

A
In Vivo 2 CFP2 Growth, Induced at 0 hours
OD
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4
B
In Vivo 2 CFP2 Fluorescence, Induced at 0 hours
Fluorescence
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4

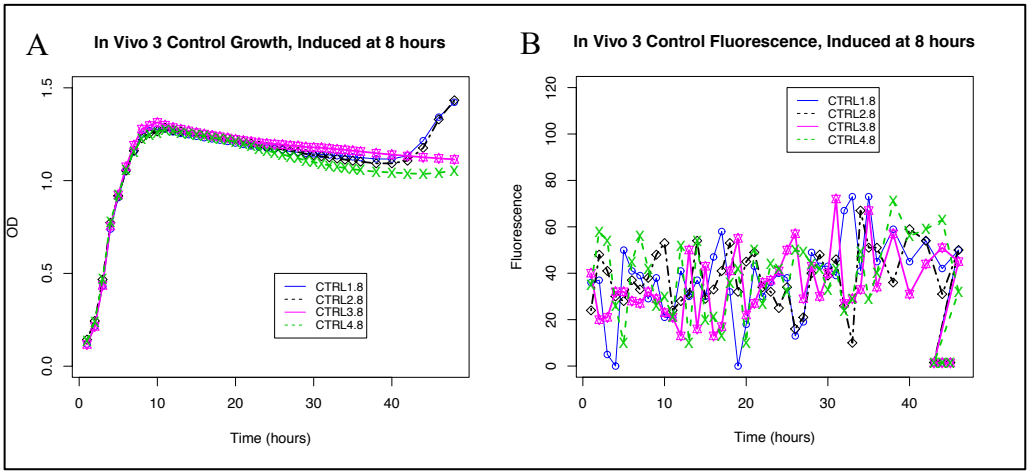

A
In Vivo 2 CFP2 Growth, Induced at 8 hours
OD
Time (hours)
CFP2.18
CFP2.28
CFP2.38
CFP2.48
B
In Vivo 2 CFP2 Fluorescence, Induced at 8 hours
Fluorescence
Time (hours)
CFP2.18
CFP2.28
CFP2.38
CFP2.48

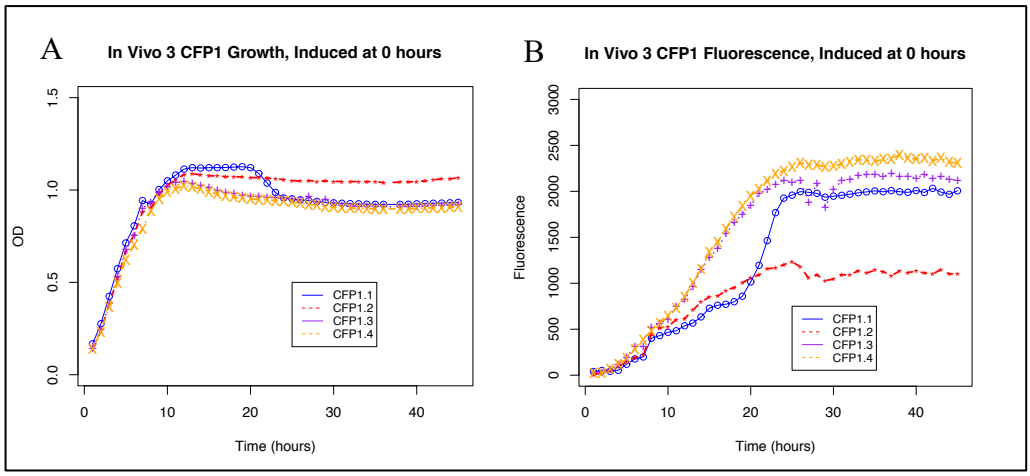

A
In Vivo 2 CFP3 Growth, Induced at 0 hours
OD
Time (hours)
CFP3.1
CFP3.2
CFP3.3
CFP3.4
B
In Vivo 2 CFP3 Fluorescence, Induced at 8 hours
Fluorescence
Time (hours)
CFP3.18
CFP3.28
CFP3.38
CFP3.48

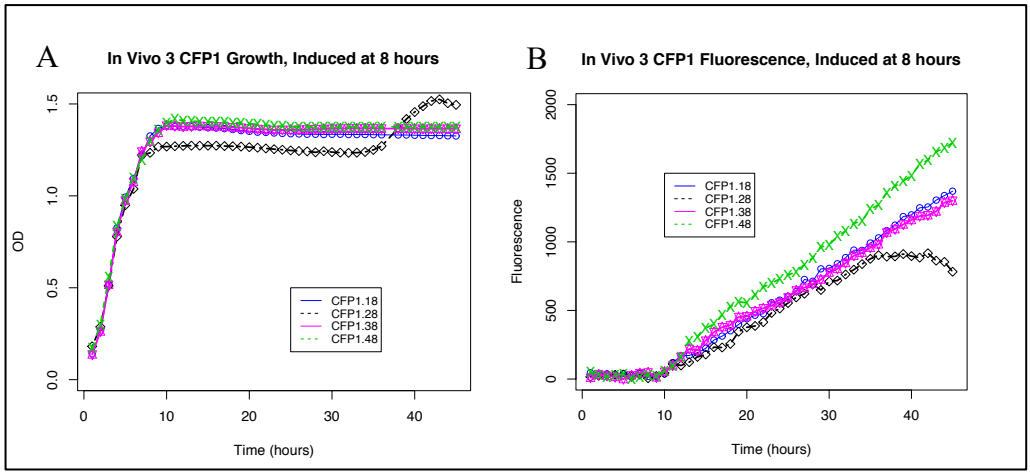

A
In Vivo 2 CFP3 Growth, Induced at 8 hours
OD
Time (hours)
CFP3.18
CFP3.28
CFP3.38
CFP3.48
B
In Vivo 2 CFP3 Fluorescence, Induced at 0 hours
Fluorescence
Time (hours)
CFP3.1
CFP3.2
CFP3.3
CFP3.4

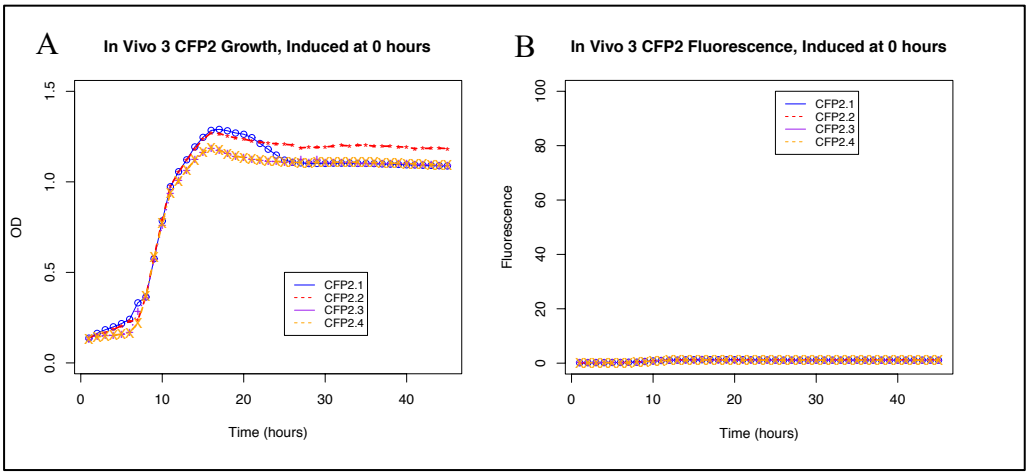
C.3: Experiment in vivo 3

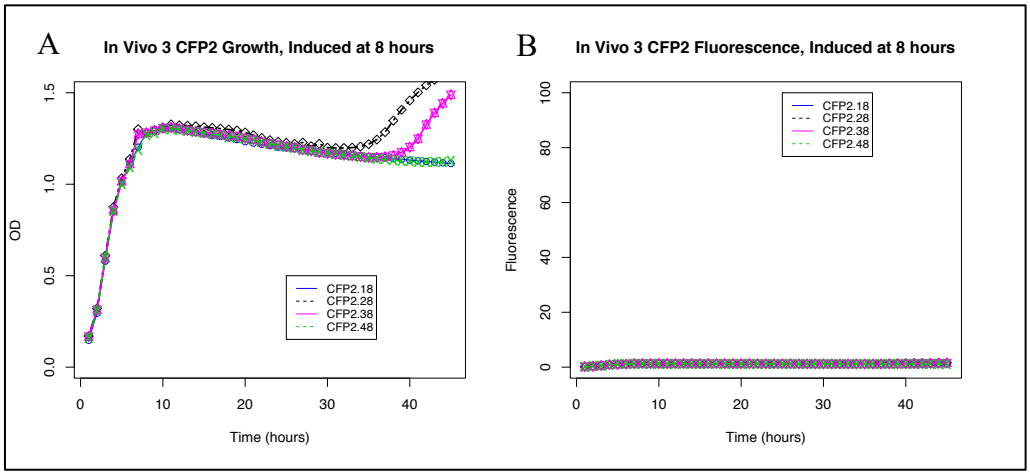




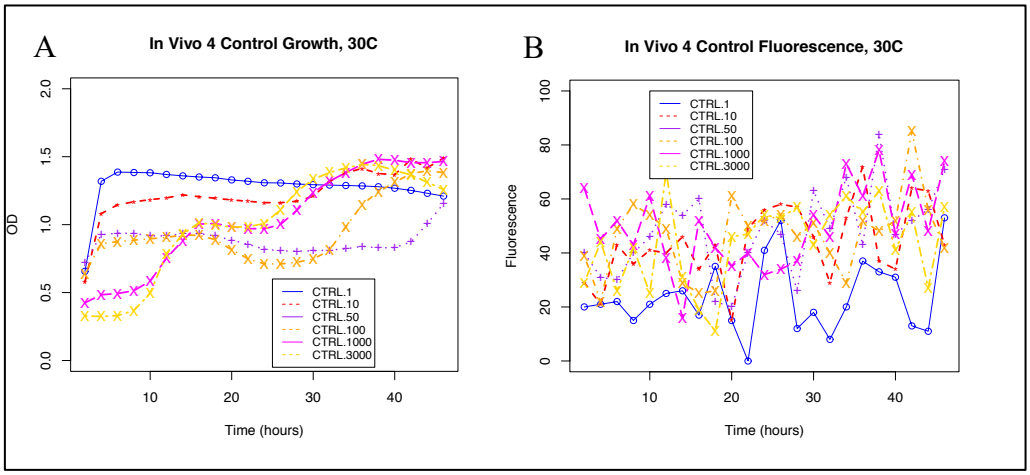

A
In Vivo 3 CFP1 Growth, Induced at 0 hours
OD
Time (hours)
CFP1.1
CFP1.2
CFP1.3
CFP1.4
B
In Vivo 3 CFP1 Fluorescence, Induced at 0 hours
Fluorescence
Time (hours)
CFP1.1
CFP1.2
CFP1.3
CFP1.4

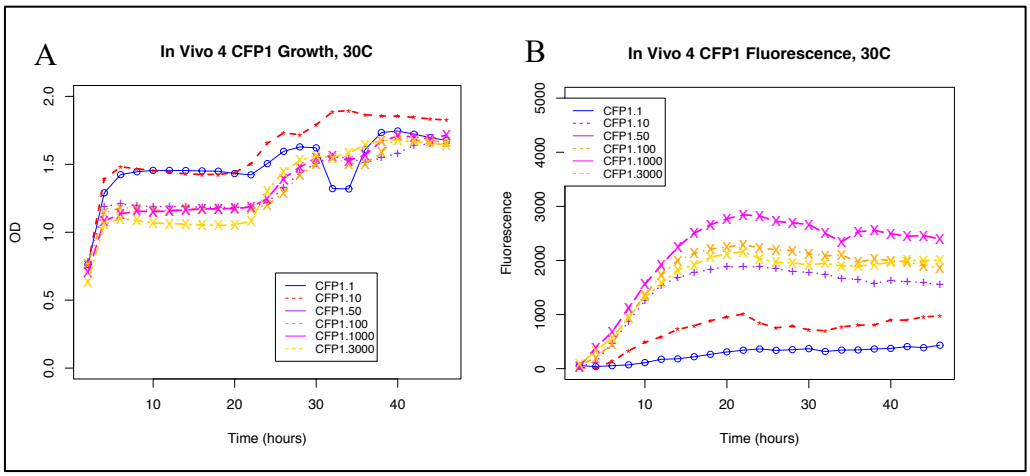

A
In Vivo 3 CFP1 Growth, Induced at 8 hours
OD
Time (hours)
CFP1.18
CFP1.28
CFP1.38
CFP1.48
B
In Vivo 3 CFP1 Fluorescence, Induced at 8 hours
Fluorescence
Time (hours)
CFP1.18
CFP1.28
CFP1.38
CFP1.48

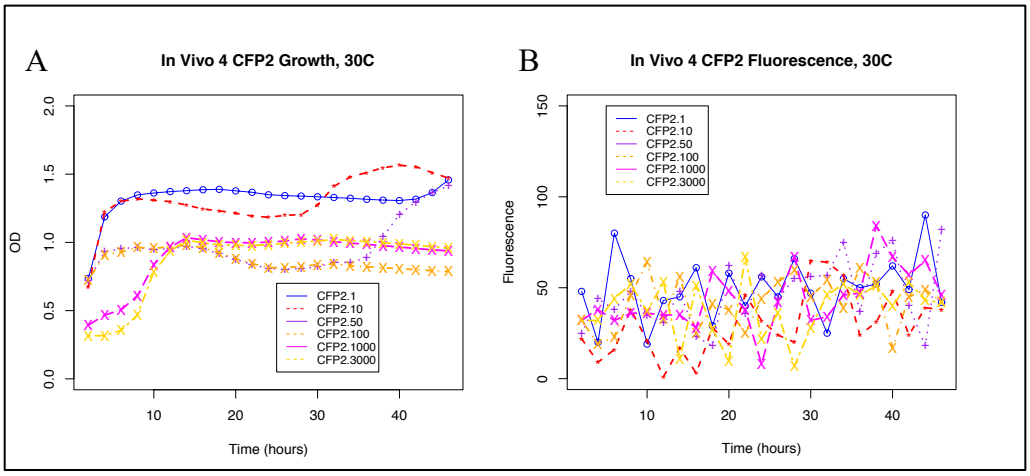

A
In Vivo 3 CFP2 Growth, Induced at 0 hours
OD
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4
B
In Vivo 3 CFP2 Fluorescence, Induced at 0 hours
Fluorescence
Time (hours)
CFP2.1
CFP2.2
CFP2.3
CFP2.4

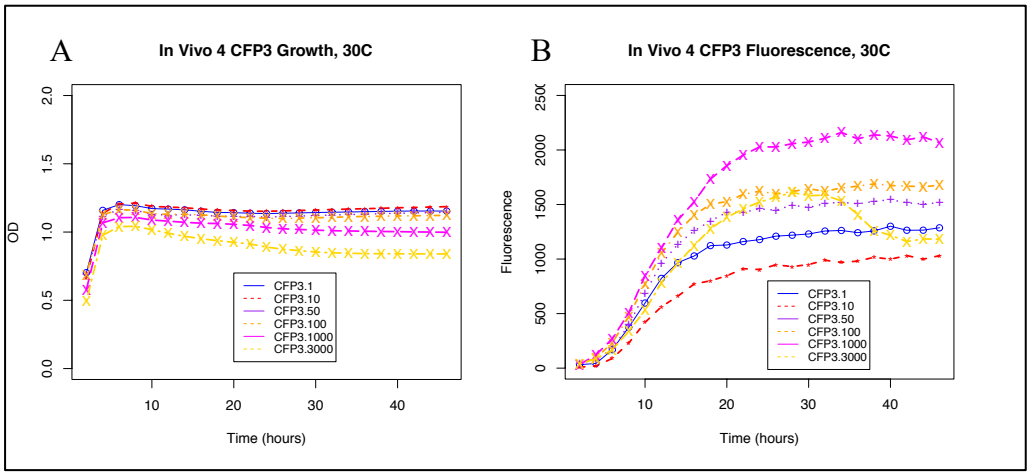

A
In Vivo 3 CFP2 Growth, Induced at 8 hours
OD
Time (hours)
CFP2.18
CFP2.28
CFP2.38
CFP2.48
B
In Vivo 3 CFP2 Fluorescence, Induced at 8 hours
Fluorescence
Time (hours)
CFP2.18
CFP2.28
CFP2.38
CFP2.48

C.4: Experiment in vivo 4

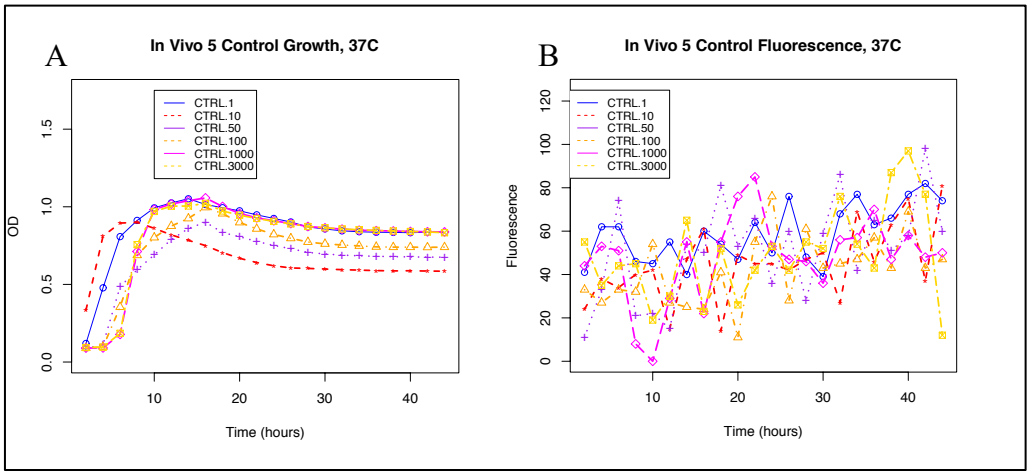


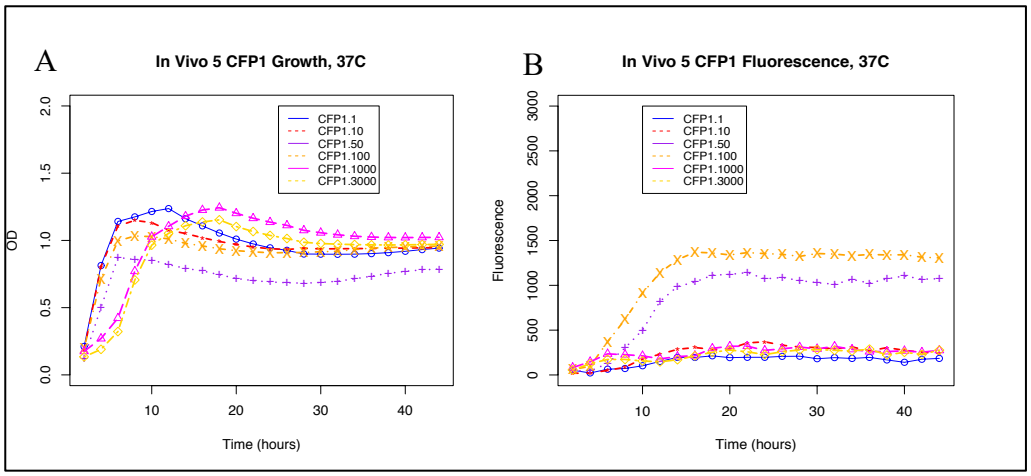


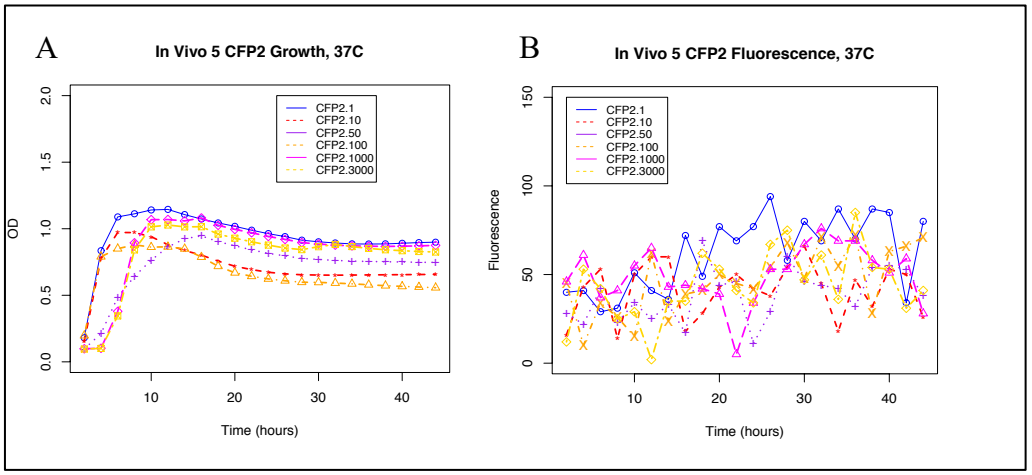

A
In Vivo 4 CFP2 Growth, 30C
OD
Time (hours)
CFP2.1
CFP2.10
CFP2.50
CFP2.100
CFP2.1000
CFP2.3000
B
In Vivo 4 CFP2 Fluorescence, 30C
Fluorescence
Time (hours)

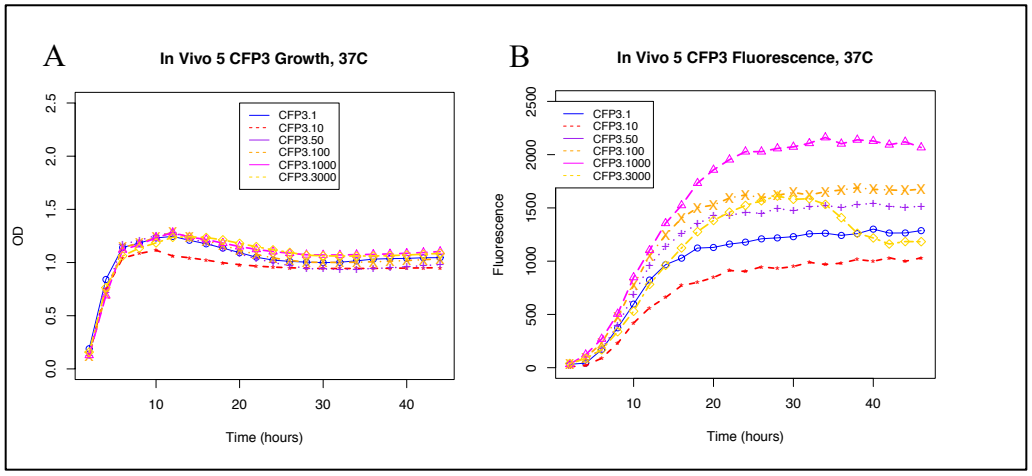

A
In Vivo 4 CFP3 Growth, 30C
OD
Time (hours)
CFP3.1
CFP3.10
CFP3.50
CFP3.100
CFP3.1000
CFP3.3000
B
In Vivo 4 CFP3 Fluorescence, 30C
Fluorescence
Time (hours)

C.5: Experiment in vivo 5

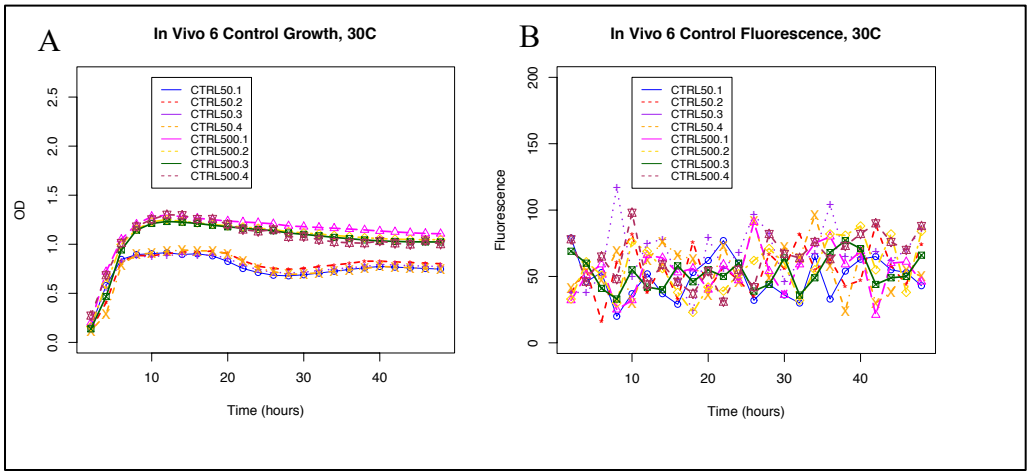


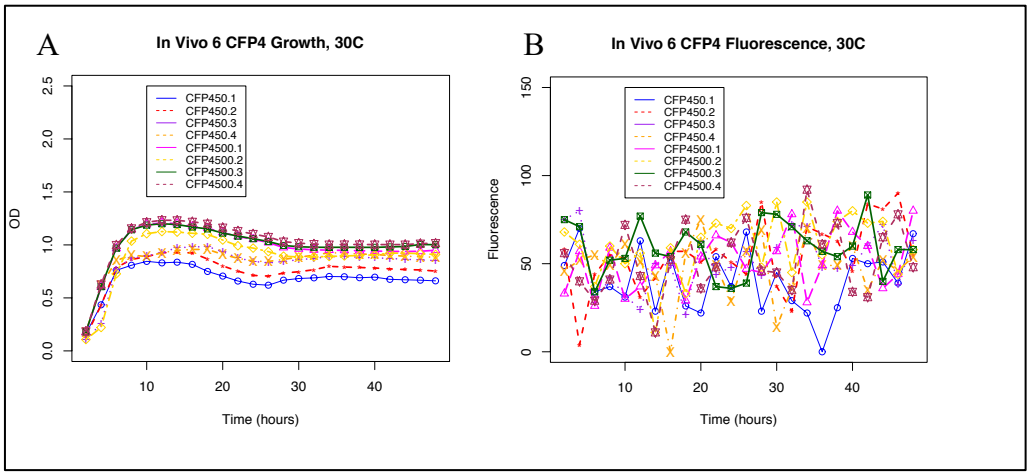


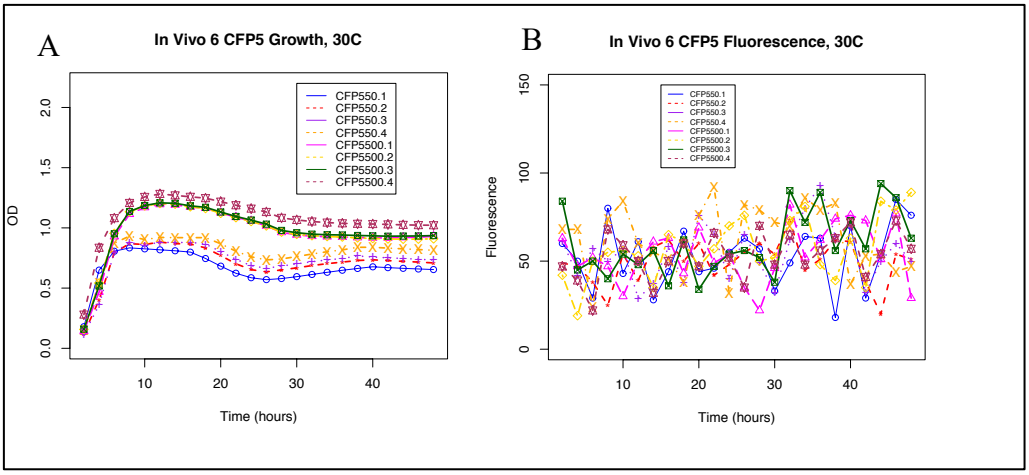


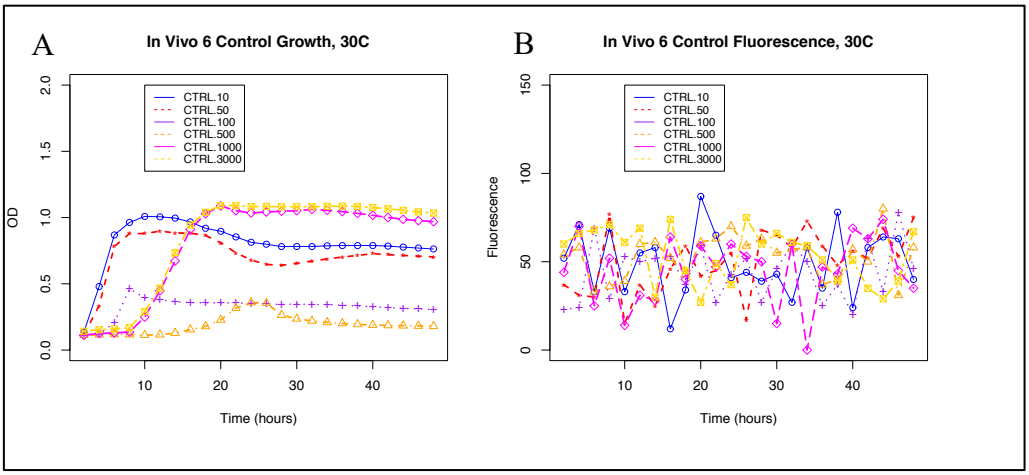


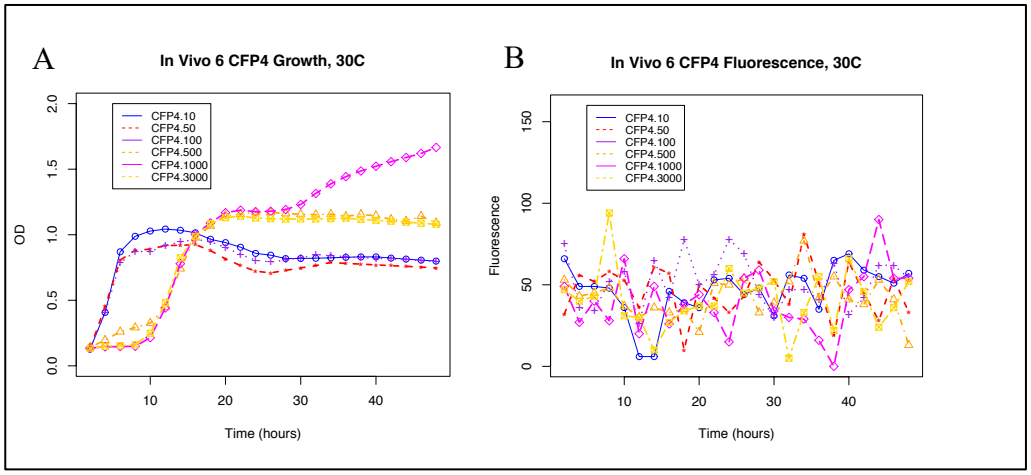
C.6: Experiment in vivo 6

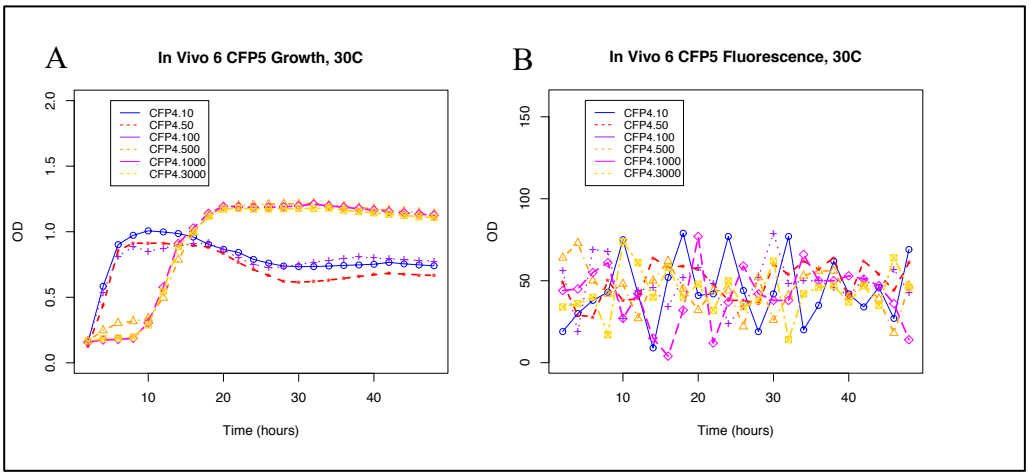




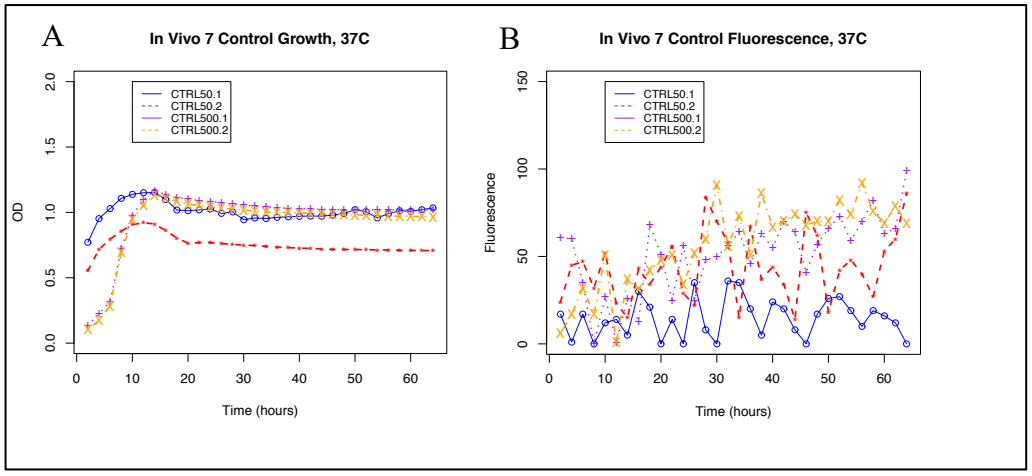

A
In Vivo 6 CFP5 Growth, 30C
OD
Time (hours)
CFP550.1
CFP550.2
CFP550.3
CFP550.4
CFP5500.1
CFP5500.2
CFP5500.3
CFP5500.4
B
In Vivo 6 CFP5 Fluorescence, 30C
Fluorescence
Time (hours)

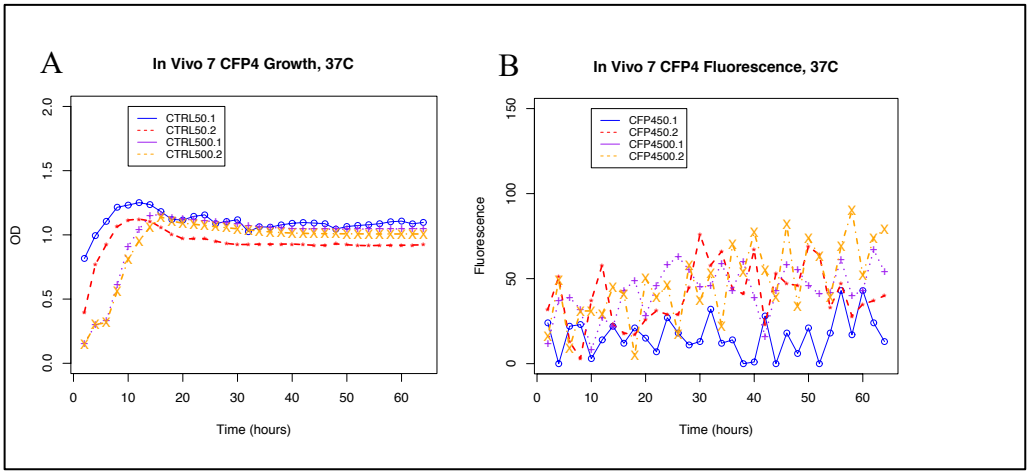

A
In Vivo 6 Control Growth, 30C
OD
Time (hours)
CTRL.10
CTRL.50
CTRL.100
CTRL.500
CTRL.1000
CTRL.3000
B
In Vivo 6 Control Fluorescence, 30C
Fluorescence
Time (hours)

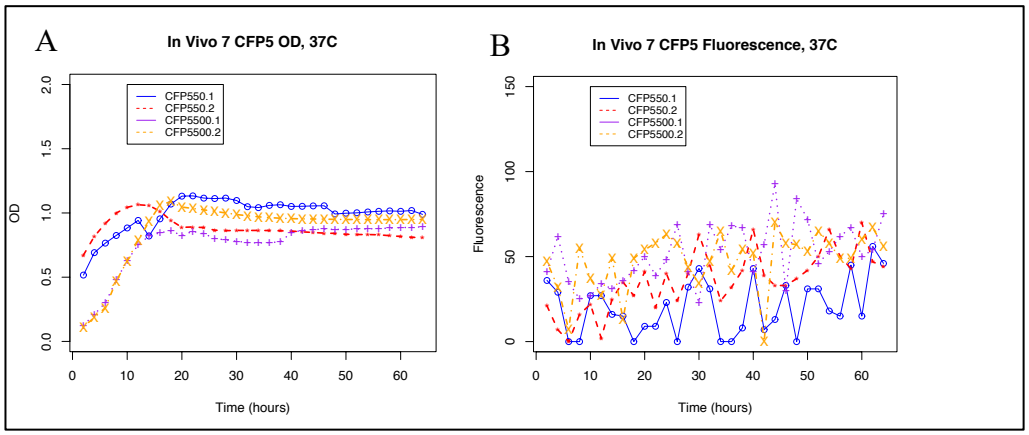

A
In Vivo 6 CFP4 Growth, 30C
OD
Time (hours)
CFP4.10
CFP4.50
CFP4.100
CFP4.500
CFP4.1000
CFP4.3000
B
In Vivo 6 CFP4 Fluorescence, 30C
Fluorescence
Time (hours)

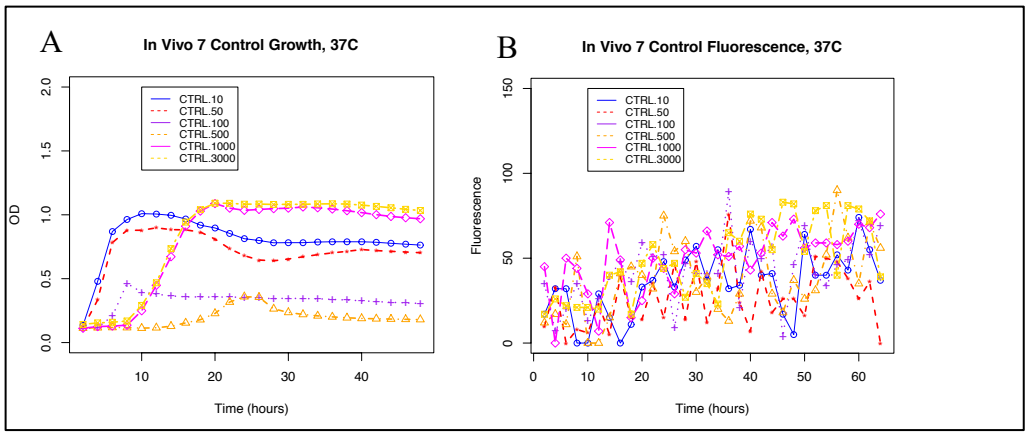

A
In Vivo 6 CFP5 Growth, 30C
CFP4.10
CFP4.50
CFP4.100
CFP4.500
CFP4.1000
CFP4.3000
OD
0.0
0.5
1.0
1.5
2.0
10
20
30
40
Time (hours)
B
In Vivo 6 CFP5 Fluorescence, 30C
CFP4.10
CFP4.50
CFP4.100
CFP4.500
CFP4.1000
CFP4.3000
OD
0
50
100
150
10
20
30
40
Time (hours)

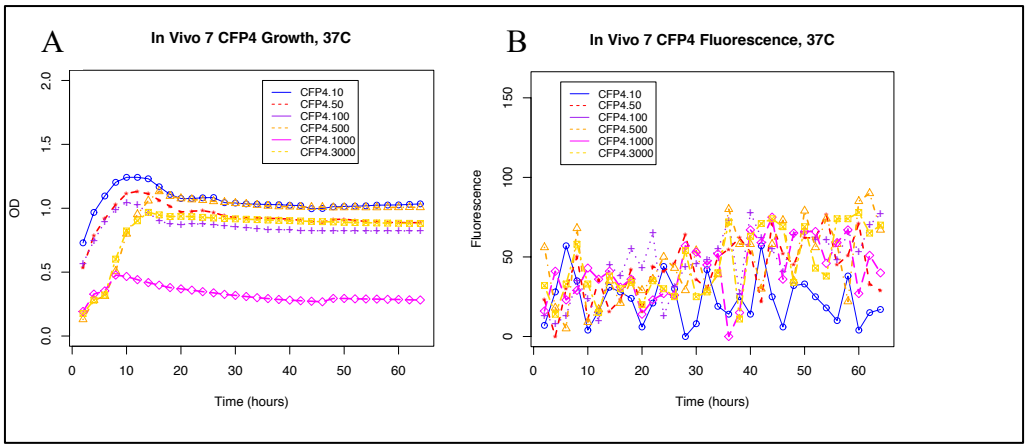
C.7: Experiment in vivo 7

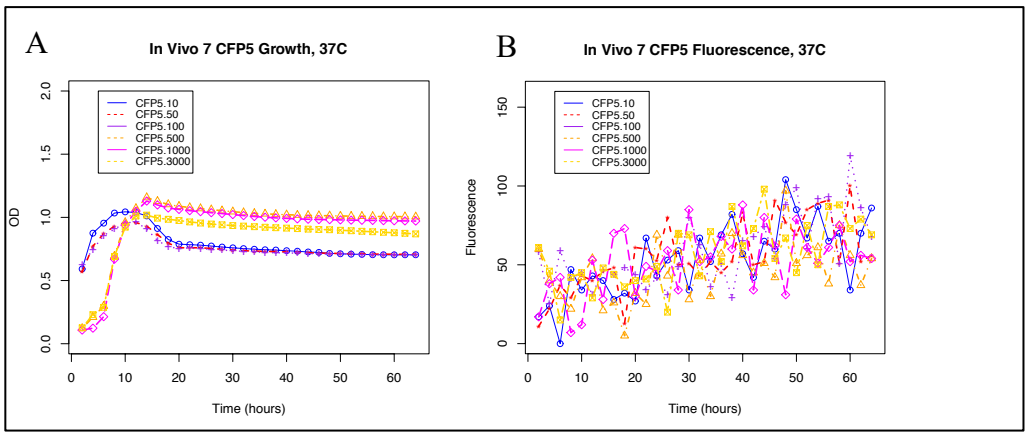







A
In Vivo 7 Control Growth, 37C
CTRL.10
CTRL.50
CTRL.100
CTRL.500
CTRL.1000
CTRL.3000
OD
Time (hours)
B
In Vivo 7 Control Fluorescence, 37C
Fluorescence
Time (hours)


A
In Vivo 7 CFP4 Growth, 37C
CFP4.10
CFP4.50
CFP4.100
CFP4.500
CFP4.1000
CFP4.3000
OD
Time (hours)
B
In Vivo 7 CFP4 Fluorescence, 37C
Fluorescence
Time (hours)


A
In Vivo 7 CFP5 Growth, 37C
CFP5.10
CFP5.50
CFP5.100
CFP5.500
CFP5.1000
CFP5.3000
OD
Time (hours)
B
In Vivo 7 CFP5 Fluorescence, 37C
Fluorescence
Time (hours)

REFERENCES

(1) Khan, S., Ullah, M. W., Siddique, R., Nabi, G., Manan, S., Yousaf, M., and Hou, H. (2016) Role of Recombinant DNA Technology to Improve Life. *Int J Genomics 2016*.

(2) Keasling, J. D., and Bang, S.-W. (1998) Recombinant DNA techniques for bioremediation and environmentally-friendly synthesis. *Current Opinion in Biotechnology 9*, 135–140.

(3) Lin, B., and Tao, Y. (2017) Whole-cell biocatalysts by design. *Microb Cell Fact 16*.

(4) Wang, J., Shen, X., Jain, R., Wang, J., Yuan, Q., and Yan, Y. (2017) Establishing a novel biosynthetic pathway for the production of 3,4-dihydroxybutyric acid from xylose in Escherichia coli. *Metab. Eng. 41*, 39–45.

(5) Simeonidis, E., and Price, N. D. (2015) Genome-scale modeling for metabolic engineering. *J. Ind. Microbiol. Biotechnol. 42*, 327–338.

(6) Yadav, V. G., De Mey, M., Lim, C. G., Ajikumar, P. K., and Stephanopoulos, G. (2012) The Future of Metabolic Engineering and Synthetic Biology: Towards a Systematic Practice. *Metab Eng 14*, 233–241.

(7) Rosano, G. L., and Ceccarelli, E. A. (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol 5*.

(8) Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison Iii, C. A., and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods 6*, 343–345.

(9) Hillson, N. J. (2011) DNA Assembly Method Standardization for Synthetic Biomolecular Circuits and Systems, in *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology* (Koeppl, H., Setti, G., di Bernardo, M., and Densmore, D., Eds.), pp 295–314. Springer New York, New York, NY.

(10) Guo, B., and Bi, Y. (2002) Cloning PCR Products, in *PCR Cloning Protocols* (Chen, B.-Y., and Janes, H. W., Eds.), pp 111–119. Humana Press, Totowa, NJ.

(11) Silva, G. B. da, and Ivo, P. (2018) Discovery of a Novel Microalgal Strain Scenedesmus Sp. A6 and Exploration of Its Potential as a Microbial Cell Factory.

(12) Tanniche, I., Fisher, A. K., Gillam, F., Collakova, E., Zhang, C., Bevan, D. R., and Senger, R. S. (2019) λ-PCR for precise DNA assembly and modification.

(13) Salis, H. M. (2011) Chapter two - The Ribosome Binding Site Calculator, in *Methods in Enzymology* (Voigt, C., Ed.), pp 19–42. Academic Press.

(14) Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011) NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry 32*, 170–173.

(15) Reeve, B., Hargest, T., Gilbert, C., and Ellis, T. (2014) Predicting Translation Initiation Rates for Designing Synthetic Biology. *Front Bioeng Biotechnol 2*.

(16) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946–950.

(17) Roberts, T. M., Bikel, I., Yocum, R. R., Livingston, D. M., and Ptashne, M. (1979) Synthesis of simian virus 40 t antigen in Escherichia coli. *PNAS 76*, 5596–5600.

(18) Chen, H., Bjerknes, M., Kumar, R., and Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res 22*, 4953–4957.

(19) Espah Borujeni, A., Channarasappa, A. S., and Salis, H. M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res. 42*, 2646–2659.

(20) Shaham, G., and Tuller, T. (2018) Genome scale analysis of Escherichia coli with a comprehensive prokaryotic sequence-based biophysical model of translation initiation and elongation. *DNA Res 25*, 195–205.

(21) Hannig, G., and Makrides, S. C. (1998) Strategies for optimizing heterologous protein expression in Escherichia coli. *Trends in Biotechnology 16*, 54–60.

(22) Mathews, D. H. (2006) Revolutions in RNA Secondary Structure Prediction. *Journal of Molecular Biology 359*, 526–532.

(23) Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure11Edited by I. Tinoco. *Journal of Molecular Biology 288*, 911–940.

(24) Jones, C. P., and Ferré-D'Amaré, A. R. (2015) RNA quaternary structure and global symmetry. *Trends in Biochemical Sciences 40*, 211–220.

(25) Tinoco, I., and Bustamante, C. (1999) How RNA folds. *Journal of Molecular Biology 293*, 271–281.

(26) Null, A. P., Hannis, J. C., and Muddiman, D. C. (2000) Preparation of single-stranded PCR products for electrospray ionization mass spectrometry using the DNA repair enzyme lambda exonuclease. *Analyst 125*, 619–626.

(27) Frenkel, F. E., and Korotkov, E. V. (2009) Using Triplet Periodicity of Nucleotide Sequences for Finding Potential Reading Frame Shifts in Genes. *DNA Res 16*, 105–114.

(28) Gopal, G. J., and Kumar, A. (2013) Strategies for the Production of Recombinant Protein in Escherichia coli. *Protein J 32*, 419–425.

(29) Briand, L., Marcion, G., Kriznik, A., Heydel, J. M., Artur, Y., Garrido, C., Seigneuric, R., and Neiers, F. (2016) A self-inducible heterologous protein expression system in *Escherichia coli*. *Scientific Reports 6*, 33037.

(30) Tanniche, I. (2013, February 8) Correlating antisense RNA performance with thermodynamic calculations. Thesis, Virginia Tech.

(31) Ishido, T., Ishikawa, M., and Hirano, K. (2010) Analysis of supercoiled DNA by agarose gel electrophoresis using low-conducting sodium threonine medium. *Anal. Biochem. 400*, 148–150.

(32) Parret, A. H., Besir, H., and Meijers, R. (2016) Critical reflections on synthetic gene design for recombinant protein expression. *Current Opinion in Structural Biology 38*, 155–162.

(33) Makrides, S. C. (1996) Strategies for achieving high-level expression of genes in Escherichia coli. *Microbiol Rev 60*, 512–538.

(34) Zucchelli, E., Pema, M., Stornaiuolo, A., Piovan, C., Scavullo, C., Giuliani, E., Bossi, S., Corna, S., Asperti, C., Bordignon, C., Rizzardi, G.-P., and Bovolenta, C. (2017) Codon Optimization Leads to Functional Impairment of RD114-TR Envelope Glycoprotein. *Molecular Therapy - Methods & Clinical Development 4*, 102–114.

(35) Gorski, K., Roch, J. M., Prentki, P., and Krisch, H. M. (1985) The stability of bacteriophage T4 gene 32 mRNA: a 5’ leader sequence that can stabilize mRNA transcripts. *Cell 43*, 461–469.

(36) Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene 361*, 13–37.

(37) Gualerzi, C. O., and Pon, C. L. (2015) Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell Mol Life Sci 72*, 4341–4367.

(38) Ikeda, R. A., Lin, A. C., and Clarke, J. (1992) Initiation of transcription by T7 RNA polymerase as its natural promoters. *J. Biol. Chem. 267*, 2640–2649.

(39) Imburgio, D., Rong, M., Ma, K., and McAllister, W. T. (2000) Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry 39*, 10419–10430.

(40) Akinterinwa, O., Khankal, R., and Cirino, P. C. (2008) Metabolic engineering for bioproduction of sugar alcohols. *Current Opinion in Biotechnology 19*, 461–467.

(41) Curran, K. A., Leavitt, J. M., Karim, A. S., and Alper, H. S. (2013) Metabolic engineering of muconic acid production in Saccharomyces cerevisiae. *Metab. Eng. 15*, 55–66.

(42) Nielsen, J. (2013) Production of biopharmaceutical proteins by yeast. *Bioengineered 4*, 207–211.

(43) Hong, S. H., Kwon, Y.-C., and Jewett, M. C. (2014) Non-standard amino acid incorporation into proteins using Escherichia coli cell-free protein synthesis. *Front Chem 2*, 34.

(44) Dudley, Q. M., Anderson, K. C., and Jewett, M. C. (2016) Cell-Free Mixing of Escherichia coli Crude Extracts to Prototype and Rationally Engineer High-Titer Mevalonate Synthesis. *ACS Synth Biol 5*, 1578–1588.

(45) Harris, D. C., and Jewett, M. C. (2012) Cell-free biology: exploiting the interface between synthetic biology and synthetic chemistry. *Current Opinion in Biotechnology 23*, 672–678.

(46) Guo, W., Sheng, J., and Feng, X. (2017) Mini-review: In vitro Metabolic Engineering for Biomanufacturing of High-value Products. *Comput Struct Biotechnol J 15*, 161–167.

(47) Dong, H., Nilsson, L., and Kurland, C. G. (1995) Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. *Journal of Bacteriology 177*, 1497–1504.