The Pennsylvania State University
The Graduate School
Chemical Engineering Department

COMPUTATIONAL REDESIGN OF ACYL-ACP THIOESTERASES
FOR INCREASED SUBSTRATE SPECIFICITY

A Thesis in
Chemical Engineering

by

Nathanael Gifford

© 2015 Nathanael Gifford

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2015
The thesis of Nathanael Gifford was reviewed and approved* by the following:

Costas Maranas
Donald B. Broughton Professor of Chemical Engineering
Thesis Advisor

Michael Janik
Associate Professor of Chemical Engineering
John J. and Jean M. Brennan Clean Energy Early Career Professor

Howard Salis
Assistant Professor of Chemical Engineering
Assistant Professor of Agricultural and Biological Engineering

*Signatures are on file in the Graduate School
ABSTRACT

Due to limited resources and the reality of global climate change, alternatives to petroleum resources will be necessary in the near future. Biologically produced fatty acids are precursors to biodiesel, an attractive alternative to naturally occurring fossil fuels. Due to the promiscuity of thioesterases, the enzymes responsible for terminating the growth cycle of fatty acids, it is difficult to produce fatty acids with the distribution of chain lengths desired for many applications. In this work, the iterative protein redesign and optimization (IPRO) procedure is used for the redesign of Thioesterase I/Protease I/Lysophospholipase L₁ (TAP I) for increased substrate specificity for both lauric acid (12C fatty acid) and caprylic acid (8C fatty acid). IPRO uses molecular mechanics simulations alongside an MILP optimization problem formulation to generate recommended mutants for increased specificity. We also use ensemble refinement to assess the quality of mutants generated and to generate an energetic profile to approximate the specificities of the mutants for a range of fatty acid chain lengths. Mutants generated create a complex of charged residues which obstruct the binding crevice of TAP I, preventing fatty acids larger than the desired species from binding.
# TABLE OF CONTENTS

List of Figures .................................................................................................................................................. v
List of Tables ..................................................................................................................................................... vi
Acknowledgements ........................................................................................................................................... vii

Chapter 1 Background and Motivation............................................................................................................... 1
  Fatty Acid Biosynthesis Pathway ....................................................................................................................... 1
  Design Approaches .......................................................................................................................................... 3
  Iterative Protein Redesign and Optimization ................................................................................................. 4
  Thioesterase Enzyme Superfamily ................................................................................................................... 6
  Thioesterase I/Protease I/Lysophospholipase L1 (TAP I) .............................................................................. 7
  Serine-Glycine-Asparagine-Histidine (SGNH) Hydrolases .......................................................................... 8

Chapter 2 Materials and Methods .................................................................................................................... 11
  CHARMM Forcefield Parameterization ........................................................................................................... 12
  Docking the Acyl-ACP Substrate .................................................................................................................... 13
  Design Position Selection ............................................................................................................................... 14
    Leucine 11 .................................................................................................................................................. 15
    Glycine 72 .................................................................................................................................................. 16
    Leucine 76 .................................................................................................................................................. 17
    Isoleucine 107 .......................................................................................................................................... 18
    Arginine 108 ........................................................................................................................................... 19
    Alanine 111 ............................................................................................................................................. 20
    Phenylalanine 139 ..................................................................................................................................... 21
    Tyrosine 145 ............................................................................................................................................ 22
  IPRO Design Runs ........................................................................................................................................ 22

Chapter 3 Results and Conclusions ................................................................................................................ 24
  Redesign for Lauric Acid Specificity ................................................................................................................ 24
  Redesign for Caprylic Acid Specificity .......................................................................................................... 30
  Conclusions and Future Work ....................................................................................................................... 33
  References .................................................................................................................................................... 35
LIST OF FIGURES

Figure 1-1. *Escherichia coli* fatty acid biosynthesis pathway ........................................... 2

Figure 1-2. Iterative Protein Redesign and Optimization (IPRO) procedure ................. 4

Figure 1-3. Thioesterase I/Protease I/Lysophospholipase L1 (TAP) bound to octanoic acid ............................................................................................................. 7

Figure 1-4. SGNH hydrolase mechanism proposed for TAP I catalytic triad .......... 9

Figure 2-1. Leucine 11 Design Position .................................................................................. 15

Figure 2-2. Glycine 72 Design Position ........................................................................... 16

Figure 2-3. Leucine 76 Design Position ........................................................................... 17

Figure 2-4. Isoleucine 107 Design Position ........................................................................ 18

Figure 2-5. Arginine 108 Design Position ......................................................................... 19

Figure 2-6. Alanine 111 Design Position ......................................................................... 20

Figure 2-7. Phenylalanine 139 Design Position .............................................................. 21

Figure 2-8. Tyrosine 145 Design Position ......................................................................... 22

Figure 3-1. Probability matrix of point mutations for lauric acid specificity .......... 26

Figure 3-2. Mutant 8 complex of charged residues ......................................................... 27
LIST OF TABLES

Table 3-1. IPRO recommended mutants for lauric acid specificity in TAP I..............24
Table 3-2. Lauric acid specific TAP I mutant ensemble refined energies ...............28
Table 3-3. Lauric acid specific TAP I mutant ensemble refined energies ...............29
Table 3-4. IPRO recommended mutants for caprylic acid specificity in TAP I.........30
Table 3-5. Caprylic acid specific TAP I mutants ensemble refined energies............32
ACKNOWLEDGEMENTS

Thank you to Dr. Mike Janik and Dr. Costas Maranas for their invaluable guidance and support in my studies.

Thank you to Matt Grisewood for the collaboration and help in troubleshooting as well as countless useful discussions.

Thank you to Bob Pantazes for development work on IPRO, as well as the guidance and mentorship as I began my research.

Thank you to the Janik and Maranas lab groups. It has been a pleasure to work with such dedicated, helpful, and friendly labmates.

Thank you to the members of the Penn State Christian Grads group for support and encouragement throughout my studies.

Thank you to Dr. Howard Salis for serving on this committee.
Chapter 1

Background and Motivation

One of today’s eminent challenges is the replacement of fossil fuels as the primary source of energy in the transportation sector. Fossil fuels are limited in supply, with peak production predicted to be reached in 2030 at the latest [1]. Further, combustion of fossil fuels is considered to be a primary cause of climate change, motivating more carbon neutral alternatives [2]. A plethora of alternatives have been proposed, including electric vehicles, hydrogen fuel cells, ethanol, and biodiesel. Over the past eight years, biodiesel has grown in use and demand rapidly [3], presenting itself as an attractive target for biological production.

Microbial production of fossil fuel alternatives is an attractive alternative for synthesis of long chain alcohols and free fatty acids [4, 5]. Medium and long chain fatty acids are common microbial metabolites that are precursors for fatty acid esters (FAEs) which are diesel alternatives.

Fatty Acid Biosynthesis Pathway

Microbial production of free fatty acids occurs through cyclical elongation of a fatty acid bound to an acyl carrier protein (ACP) (Figure 1-1). The cycle is initiated when acetoacetyl-ACP is reduced by β-ketoacyl-ACP reductase to the β-hydroxyacyl-ACP. The hydroxyl group is then eliminated by β-hydroxyacyl-ACP dehydratase, resulting in
the dehydrated enoyl-ACP. This unsaturation is then reduced by enoyl-ACP reductase, yielding the acyl-ACP product. The cycle continues by the condensation of malonyl-ACP with the acyl-ACP by β-ketoacyl-ACP synthase. β-ketoacyl-ACP synthase decarboxylates the malonyl-ACP, generating a nucleophilic carbon. The nucleophilic carbon then attacks the thioester group, thereby generating a new ketoacyl-ACP that is lengthened by two carbons. This cycle will continue elongating the acyl-ACP two carbons at a time until the acyl group is hydrolyzed from the ACP by acyl-ACP Thioesterase, generating the fatty acid product.

Figure 1-1. *Escherichia coli* fatty acid biosynthesis pathway

Fatty acids produced can then be processed to generate a number of biofuel alternatives. Biodiesel is produced by the esterification of fatty acids with methanol or
ethanol, generating fatty acid esters (FAEs), i.e., biodiesel. Properties of the resulting biodiesel including heating value, oxidative stability, and cold flow properties are directly related to the chain length of the fatty acid[6]. It would therefore be desirable to design enzymes that change and increase specificity for the production of specific chain length fatty acids. Acyl-ACP thioesterase catalyzes the terminating step of the fatty acid elongation cycle, presenting itself as a clear target for design. Increased specificity of the thioesterase for desired chain lengths will result in greater production of the desired species.

**Design Approaches**

Enzymes have been redesigned both experimentally and computationally as well as *de novo*. Attempts at *de novo* protein design have met with moderate success, achieving active enzymes, though not on the order of wild type activity [7]. Lack of strongly active *de novo* designs motivates attempts at redesign of natural enzymes. Experimental redesign methods have seen good results [8]. However, the complexity and scope of designing enzymes makes this a time consuming endeavor. Enzyme design is therefore an ideal application for the less time consuming and comparatively inexpensive resources of computational methods. Much work has been focused on the development of computational enzyme design methodologies, using tools such as: Quantum Mechanics (QM) [9-11], Mixed Quantum Mechanics/Molecular Mechanics (QM/MM) [9, 12], Molecular Mechanics (MM) [9, 13], and Molecular Dynamics (MD) [9, 13]. However,
the vast majority of this work has focused on applications of computational methods through rational design rather than by a computational algorithm.

**Iterative Protein Redesign and Optimization**

The Iterative Protein Redesign & Optimization (IPRO)[14, 15] procedure seeks to generate favorable mutants through a fully automated the process. IPRO optimizes proteins for substrate binding and binding specificity [16, 17] by using the molecular mechanics force field CHARMM [18] to evaluate energies of randomly selected mutants.

![Figure 1-2. Iterative Protein Redesign and Optimization (IPRO) procedure](image)
IPRO begins with (A) a random local backbone perturbation surrounding the chosen design residues. This step allows for a larger design space by increasing and/or changing the feasible rotamers to select at and around the design positions. This prevents the procedure from repeatedly predicting the same designs, allowing for exploration of other options. Following this step IPRO (B) generates a table of rotamer-rotamer and rotamer-backbone energy interactions. This provides the energetic data necessary for selection of the optimal rotamers. Rotamer selection proceeds via (C) an MILP problem formulation which is solved using the General Algebraic Modeling System (GAMS). This formulation selects of the optimal rotamer for each design position from all rotamers of allowed residues at a design position. Optimal rotamers for the surrounding sidechain windows are also selected from the rotamers of the wild type residue at each position. After rotamer selection, the backbone is (D) relaxed by a CHARMM energy minimization. Following this relaxation, the target molecule is redocked by small perturbations of the rigid target molecule position, selecting the most energetically favorable position as evaluated by CHARMM. If the energy of the complex at this point is less favorable than the current best mutant, the mutant is rejected, and a new iteration is begun. If the mutation is the most favorable so far, the mutant then undergoes ensemble refinement. Ensemble refinement is performed by generating several unique ensembles through minor structural perturbations. These perturbed structures then undergo several IPRO iterations with rotamers only selected from those of the mutant residues selected at each position. The energies of all iterations from all ensembles are averaged to generate an ensemble refined energy. The mutations are then accepted or rejected on the basis of
the metropolis criterion applied to the ensemble refined energy. This procedure is repeated iteratively until the prescribed iteration limit is met.

This mode of redesign is called protein engineering, the redesign of protein sequences for favorable protein characteristics. The microorganism *E. coli* has been the target of engineering for the overproduction of fatty acids [5, 19]. There are two primary methods changing the chain length specificity, heterologous expression of alternative thioesterases [5, 19-22] and protein engineering [23, 24]. IPRO was used to engineer a thioesterase for increased substrate specificity.

**Thioesterase Enzyme Superfamily**

Thioesterases (TEs) are a superfamily of enzymes that hydrolyze the bond between a sulfur atom and a carbonyl group. Sequence analysis of these enzymes has divided them into 23 distinct families whose sequences are almost entirely unrelated [25]. Phylogenetic characterization has shown that evolutionary algorithms can successfully differentiate these families into subfamilies of enzymes. Each enzyme in a subfamily is more closely related to the sequences of those within the subfamily than to those in other subfamilies[26]. Sequence alignments within these subfamilies have been used to successfully identify clustering within the subfamily, consistent with the clustering of substrate specificities [26]. This suggests that phylogenetic characterization is an effective method of selection of important sequence positions for the determination of substrate specificity.
**Thioesterase I/Protease I/Lysophospholipase L1 (TAP I)**

*E. coli* Thioesterase I/Protease I/Lysophospholipase L1 (TAP I) is a multifunctional SGNH hydrolase classified in the TE3 family [25, 27]. Leaderless expression of TAP has been demonstrated to increase free fatty acid production. However, due to the promiscuity of TAP, a range of fatty acids (C8-C18) is produced [5]. It would therefore be desirable to redesign TAP for higher substrate specificity for medium chain length acyl-ACPs in order to produce useful biofuel precursors.

![Figure 1-3. Thioesterase I/Protease I/Lysophospholipase L1 (TAP) bound to octanoic acid](image-url)

Structural studies of TAP have solved a crystal structure of free TAP, as well as TAP bound to caprylic acid (8:0) (PDBs: 1IVN, 1U8U [27, 28]). The catalytic triad has been identified as Ser10-Asp154-His157, and the oxyanion hole has been identified as
Ser10-Gly44-Asn73[27]. It has been shown that long chain substrates initiate a switch loop movement in TAP, likely via rearrangements initiated by hydrophobic contacts. Further, abolishment of this movement via mutation L109P resulted in a 10-fold decrease in catalytic efficiency for lauric (12:0) and palmitic (16:0) acids [28]. Additional NMR studies have shown several loops in TAP I surrounding the active site to have significant flexibility, suggesting binding site plasticity as a significant contributor to the lack of substrate specificity exhibited[29].

**Serine-Glycine-Asparagine-Histidine (SGNH) Hydrolases**

SGNH hydrolases are a superfamily of enzymes which hydrolyze bonds through a catalytic network composed of a Serine, Glycine, Asparagine, and Histidine residue. These are also typically accompanied by an acidic residue such as a glutamic acid to complete the hydrogen shuffling pathway. SGNH hydrolases have a wide range of native substrates including esters, thioesters, lipids, phospholipids, and proteins. Investigations of a variety of SGNH hydrolases found them to be fairly promiscuous, having activity on a variety of esters as well as some non-native functionalities [30]. SGNH hydrolases have overall low sequence identity across the family, however they have exhibited structural homology for most members for which an NMR or crystal structure exist [27, 31, 32].
SGNH hydrolases catalytic machinery is conserved as the four residues described each placed in different regions of the sequence, while sharing structural homology. They have been studied using site directed mutagenesis in TAP I, indicating these and the conserved sequence blocks surrounding them are important for catalytic activity. The importance of these residues suggests a catalytic mechanism analogous to that of common serine hydrolases, such as chymotrypsin, as seen in figure 1-4. In this mechanism, the Serine residue is activated by a proton shuffling network containing the Histidine and Aspartate residues. This nucleophilic Serine then attacks the ester carbonyl of the substrate forming a tetrahedral intermediate structure with a highly negative
oxygen. This oxygen is stabilized by the oxyanion hole of the enzyme. The carbonyl is then reformed, and the ACP thiolate group leaves forming the covalent adduct species. This species is then attacked by water activated by general base, again forming a tetrahedral intermediate. The reformation of the carbonyl then causes the serine to leave, allowing the covalent adduct to leave as a free fatty acid.

In this work, IPRO is used for the generation of mutants that may improve selectivity for lauric acid (12:0), while decreasing selectivity for longer chain fatty acids. Further mutants were generated to compare mutants recommended to improve specificity for caprylic acid (8:0) for comparison with recommendations for longer chain length specificity. These mutants were then analyzed to generate a specificity profile for all saturated fatty acids from eight to eighteen carbons in length.

Specificity is increased by lowering $K_M$ for the desired substrate while raising $K_M$ for the competing substrates in the Michaelis-Menten kinetic model. $K_M$ is related to the interaction energy (IE) of the enzyme with the substrate. Therefore, this will be achieved by lowering the IE with the desired substrate or raising it with the undesired substrate.
Chapter 2
Materials and Methods

All work was performed using computational resources provided through the Research Computing and Cyberinfrastructure Lion-Xf cluster run by the high performance computing group. All IPRO runs were performed using 20 Intel Xeon X5675 Six-Core 3.06 GHz processor cores. All energy minimizations were performed using the CHARMM forcefield with Generalized Born solvation. Energy calculations included bond, angle, electrostatic, Urey-Bradley, improper dihedral, and van der Waal contributions.

Interaction Energies (IE) were calculated by evaluating the energy of the unbound enzyme and substrate as well as the minimized energy of the Enzyme-Substrate Complex, as seen in the below equation.

\[
IE = G_{ES} - G_E - G_S
\]

Specificity was enforced by improving the binding with the desired substrate while eliminating binding to the undesired substrate in IPRO. To improve binding with the desired substrate, the criteria for acceptance of an IPRO iteration (as seen in figure 1-2) is that for each accepted iteration the interaction energy of the desired substrate with the enzyme is improved, enforcing stricter requirements for improvement with each accepted iteration. Similarly, for the elimination of binding to the undesired substrate the criteria is that the interaction energy of the enzyme with the undesired substrate must be worse than the previous best iteration, enforcing successively stricter criteria for the
binding of the undesired substrate. By this methodology IPRO selects mutants with the highest affinity for the desired substrate and the lowest for the undesired substrate.

This methodology was applied to improve the specificity toward lower chain length substrates for the TAP I-Acyl-ACP complex. The desired effect is the elimination of activity towards the longer chain substrate and all longer chains. An assumption implicit in these calculations is that the elimination of binding of a substrate thereby eliminates binding of all longer chain substrates. This assumption is predicated on all resulting mutants operating under principles of steric hindrance and binding crevice obstruction. If in fact the binding crevice is obstructed preventing an alkyl chain from binding, presumably all longer chains will be obstructed from binding.

**CHARMM Forcefield Parameterization**

The CHARMM forcefield must be parameterized for all atoms in the model in order to represent the behavior and interactions of the enzyme-substrate complex. The CHARMM forcefield utilizes topology and parameter files in order to fully describe these interactions. The topology file contains the connectivity, partial charges and atom types for all atoms in all molecules in the model. The parameter file contains all of the force constants for all bonds, angles, and dihedral angles between all atom types that are necessary for the model.

All twenty standard amino acids have complete existing parameters which are packages with the CHARMM forcefield. However, the fatty acids to be studied, as well
as the phosphopantetheine linker and serine residue bound to the linker are not parameterized in the packaged CHARMM files.

In order to model the interaction energy of the acyl-ACP substrates, the fatty acid, the phosphopantetheine linker, and the covalently modified residue must be parameterized for the CHARMM force field. Parameterization was accomplished using the CHARMM General Force Field (CGenFF) [33, 34] alongside manual parameterization by analogy. CGenFF is known to parameterize poorly for general biomacromolecules due to its generalized training set [33, 34] and therefore could not be used for the parameterization of the covalently modified residue. This parameterization was achieved by analogy to existing parameterizations of serine residues and a variety of phosphate containing lipids. CGenFF was used for the parameterization of the phosphopantetheine linker and the fatty acid. The boundary between the two parameterization methods was parameterized by comparison and combination of the two, with relative qualities of parameters assessed using an internal CGenFF penalty score calculated according to relative similarity to the training set.

Docking the Acyl-ACP Substrate

Structural studies of E. coli acyl carrier protein have shown that it holds the acyl group in a highly plastic internal binding cavity[35]. The existing crystal structure (PDB: 2FAE [35]) therefore cannot be docked to TAP without obtaining a structure which holds the fatty acid externally. To this end, the fatty acid and phosphopantetheine linker were
extracted from the interior of the ACP by the rotation of a single dihedral angle in order to minimize disruption of the structure.

Once the physically relevant acyl-ACP structure was obtained, docking was performed by alignment of the Acyl group of the Acyl-ACP to the caprylic acid bound in the TAP crystal structure via Root Mean Square Deviation (RMSD) minimization. Important contacts maintained include the Ser10-Asp154-His157 catalytic triad to the sulfur of the thioester bond, as well as the Ser10-Gly44-Asn73 oxyanion hole to the carboxylic acid of the thioester bond. The system was then minimized using the CHARMM force field.

**Design Position Selection**

Design positions within TAP were selected on the basis of proximity to the substrate binding crevice, particularly near the terminal end of bound lauric acid (12:0). Further, they were chosen in an attempt to minimize interaction with the switch loop (residues 75-80) that maintains the dynamic behavior necessary for activity on medium and long chain fatty acids[27]. Additionally, positions known to have mutants resulting in reduced activity were avoided[28].

Three primary modes for changing of specificity were considered in the selection of design positions. First, mutation of residues lining the binding crevice could result in a less linear binding crevice. This shape, and concurrent crevice length, could result in the terminal end of the bound fatty acid being in a more easily obstructed portion of the binding crevice. Second, mutation of large residues which run parallel to the terminal end
of the binding crevice could result in changes in angles and rotamers resulting in more sterically crowded packing. Third, mutation of small residues near the terminal end of the fatty acid could result in significant changes in steric and therefore specificity. The design positions selected are summarized below. Each design position is highlighted in yellow for clarity.

**Leucine 11**

![Figure 2-1. Leucine 11 Design Position](image)

Leu11 lines the underside of the binding crevice as shown here and will affect the packing surrounding the fatty acid. Changes which require a longer path in the binding crevice could allow for greater steric control at the terminal end of the fatty acid. Leu 11 has been targeted as an attractive target for site directed mutagenesis by Lo et al.[27]
Glycine 72 lines the side of the binding crevice. Its small size provides ample opportunity for introduction of residues to change the crevice shape. Gly72 has also been targeted as an attractive target for site directed mutagenesis by Lo et al.[27]
Leucine 76

Figure 2-3. Leucine 76 Design Position

Leu76 lines the upper side of the binding crevice, and blocks a significant portion of the fatty acid from solvent access. A longer sidechain or different rotamer angle could cause significant changes in shape of the binding crevice.
Isoleucine 107

Figure 2-4. Isoleucine 107 Design Position

Ile107 lines the side of the binding crevice. A change in rotamer angle could cause the sidechain to change the shape of the binding crevice.
Arginine 108

Arg108 runs parallel to the binding crevice essentially along the length of the portion of the crevice binding the terminal end of the 14C fatty acid. Any change to packing here could drastically affect specificity.
Alanine 111

Ala111 lies close to the terminal end of the fatty acid, but is insufficiently large to hinder binding. A change to a larger residue could result in significant change to specificity.

Figure 2-6. Alanine 111 Design Position
Phenylalanine 139

Figure 2-7. Phenylalanine 139 Design Position

Phe139 contacts the binding crevice near the terminal end of the fatty acid. Any change will likely result in change to packing near the terminal end of the fatty acid. A change to a more linear large amino acid could obstruct the binding crevice.
Tyrosine 145

Figure 2-8. Tyrosine 145 Design Position

Tyr145 lines the underside of the binding crevice near the terminal end of the fatty acid. Any change could result in a change in packing.

**IPRO Design Runs**

These design positions were then used in IPRO to obtain a library of recommended mutants to increase specificity.

IPRO was run for 1000 iterations using no ensemble refinement with two binding assemblies. The first to eliminate binding for the 14 carbon Acyl-ACP structure, while the second improved binding for the 12 carbon Acyl-ACP structure. These mutants were then evaluated for binding each Acyl-ACP structure from 8 to 18 carbons in length. Each
of these binding assemblies underwent ensemble refinement to obtain a specificity profile for each recommended mutant.

IPRO was then run for 1000 iterations with no ensemble refinement to eliminate binding for the 10 carbon acyl-ACP structure, while improving binding to the 8 carbon acyl-ACP structure. The recommended mutants were again evaluated with ensemble refinement for the binding of each acyl ACP structure from 8 to 18 carbons.
Chapter 3

Results and Conclusions

Redesign for Lauric Acid Specificity

Design runs for the 12C fatty acid specific thioesterase generated 9 unique mutants which are summarized in table 3-1. Design position 72 was omitted from the table as it resulted in no mutations, remaining the wild type Glycine in all mutants.

Table 3-1. IPRO recommended mutants for lauric acid specificity in TAP I

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Design Position</th>
<th>Energies</th>
<th>Best IE (C12)</th>
<th>Best IE (C14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>L L I R A F Y</td>
<td>183.6  267.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>L L I R A F K</td>
<td>188.3  267.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>G L I R A F Y</td>
<td>-25.5  370.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>G K K A F Y</td>
<td>-143.1 381.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>G K K A F Y</td>
<td>-90.2  415.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>G K K A E Y</td>
<td>-81.0  494.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>G K K A E R</td>
<td>155.4  479.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>G K K A E K</td>
<td>127.0  517.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>G K K E E R</td>
<td>-279.6 628.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>G K K E E K</td>
<td>-257.9 628.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A few patterns immediately evidence themselves, most notably a preference for charged residues. It appears that uncharged leucine 76 and isoleucine 107 are both mutated to lysine in all mutants 4-9. Additionally the large aromatic phenylalanine 139, rather than mutating to a more linear amino acid as predicted mutates preferentially to the glutamic acid residue. Tyrosine 145 also mutates in several instances to charged residues including lysine (mutants 1, 7,9) and arginine (mutants 6,8). Additionally there appears to
be no strong preference for either positive or negatively charged residues, indicating that these residues are not stabilizing binding of the desired substrate, but rather destabilizing the undesired. This is reflected in the large positive change in interaction energy for the 14 carbon species observed in the majority of mutants.

Comparison of the frequency and correlation of point mutations could reveal information about the source of the differences in interaction energy between the species. The mutants were analyzed by studying the recurrence of point mutations to identify correlated and anticorrelated mutants predicted by IPRO. This was represented by a probability matrix (figure 2), which shows the percentage of times each point mutation reoccurred for each point mutation. Red squares represent 100% correlation, blue 0% correlation. Note that positions are grouped and separated by black lines.
Figure 3-1. Probability matrix of point mutations for lauric acid specificity

This representation allows the identification of the relative importance of mutant interactions. Note that mutant and I107K and R108K are completely correlated, indicating that the introduction of one or the other introduces no favorable change in specificity, while their simultaneous introduction appears to have a positive impact on lauric acid specificity. Further analysis reveals a similar correlation between mutations I107K, R108K, A111E, F139E, and Y145K.
It was noted that Mutant 8, possessing nearly all of these mutations, resulted in the largest gap in binding assembly interaction energies, suggesting greatest improved specificity. The structure of mutant 8 can be seen in figure 3-2.

Figure 3-2. Mutant 8 complex of charged residues

As seen in figure 3-2, these mutants draw together through a network of hydrogen bonds to form a complex of charged residues that obstruct the binding crevice. We see instances of both positive and negatively charged residues providing an electrostatic driving force for the clustering of these point mutations. The resulting steric clash introduced begins to hinder binding of the acyl-ACP at approximately the fourteenth
carbon, allowing the smaller species to bind, while disallowing binding for the larger species.

The top few mutants were then examined using ensemble refinement, generating energetic information for 8 to 18 carbon acyl-ACPs. Additional mutants generated by combination or exclusion of point mutations from other mutants were also examined to generate a specificity profile for the modeled mutants. The ensemble refined interaction energies are shown in Table 3-2.

Table 3-2. Lauric acid specific TAP I mutant ensemble refined energies

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Design Position</th>
<th>Ensemble Refined Interaction Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>8C</td>
</tr>
<tr>
<td>WT</td>
<td>L G L I R A F Y</td>
<td>-121.0</td>
</tr>
<tr>
<td>7</td>
<td>G K K K A E K</td>
<td>-83.0</td>
</tr>
<tr>
<td>8</td>
<td>G K K K E E R</td>
<td>-69.0</td>
</tr>
<tr>
<td>9</td>
<td>G K K K E E K</td>
<td>-69.7</td>
</tr>
<tr>
<td>10</td>
<td>L G L K K E H K</td>
<td>-85.8</td>
</tr>
<tr>
<td>8*</td>
<td>L G L K K E E R</td>
<td>-64.9</td>
</tr>
<tr>
<td>9*</td>
<td>L G L K K E E K</td>
<td>-79.0</td>
</tr>
</tbody>
</table>

Ensemble refinement gave energetics that show favorable interaction energies for both the 12C and 14C species, with unfavorable binding of the 16C species. It is worth noting that the binding of the 12C species was shown to be more favorable than the 14C species, reflecting some specificity for the desired species, however a similar result is found for the wild type enzyme. In contrast, the specificity change for the 16C species appears to be present in mutants 7-9, while absent in the wild type species. In order to more easily view the specificity profile, a table of the differences in interaction energy between each bound species was created (Table 3-3).
Table 3-3. Lauric acid specific TAP I mutant ensemble refined energies

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Design Position</th>
<th>Chain length Energy Difference (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8C</td>
<td>12C</td>
</tr>
<tr>
<td></td>
<td>10C</td>
<td>14C</td>
</tr>
<tr>
<td></td>
<td>16C</td>
<td>18C</td>
</tr>
<tr>
<td>WT</td>
<td>L G L I R A F Y</td>
<td>-99.2</td>
</tr>
<tr>
<td>7</td>
<td>G G K K A E K</td>
<td>140.5</td>
</tr>
<tr>
<td>8</td>
<td>G G K K E E R</td>
<td>259.6</td>
</tr>
<tr>
<td>9</td>
<td>G G K K E E K</td>
<td>-299.3</td>
</tr>
<tr>
<td>10</td>
<td>L G L K E H K</td>
<td>-226.6</td>
</tr>
<tr>
<td>8*</td>
<td>L G L K E E R</td>
<td>-326.2</td>
</tr>
<tr>
<td>9*</td>
<td>L G L K E E K</td>
<td>-226.6</td>
</tr>
</tbody>
</table>

In this table we notice that as expected, as the size of the substrate increases from 8 carbons to 12 carbons the interaction energy between the enzyme and substrate increase. This is anticipated as the increased size increases the number of possible positive interactions. We then see a distinct increase in interaction from 12 to 14 carbons for all species, and another increase in interaction energy from 14 to 16 carbons for mutants 7-9. Further, notice again that the 18C species again binds more favorably than the 16C species. This is likely an artifact of the simulation, as each of the chain lengths are docked to the wild type enzyme, which is then mutated and put through ensemble refinement. It is likely that the 18C species occupies the space taken up by the charged residue complex, and is sufficiently favorable to prevent the displacement which would result in steric hindrance of the large substrate binding. This is not reflective of the nature of binding.
Redesign for Caprylic Acid Specificity

Designs were also generated to select mutants for specificity between C8 and C10, generating 5 unique mutants (Table 3-4). Note that design positions 11 and 72 are omitted from this table as no mutations were generated at these positions.

Table 3-4. IPRO recommended mutants for caprylic acid specificity in TAP I

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Design Position</th>
<th>Energies</th>
<th>Best IE (C8)</th>
<th>Best IE (C10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT</td>
<td>L I R A F Y</td>
<td>82.24</td>
<td>247.63</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>L I R A F K</td>
<td>-73.79</td>
<td>252.38</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>L I R E F K</td>
<td>-77.84</td>
<td>256.6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>L K K E F K</td>
<td>-139.03</td>
<td>296.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>L K K E E K</td>
<td>-165.04</td>
<td>299.88</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>K K K E E K</td>
<td>-281.78</td>
<td>610.71</td>
<td></td>
</tr>
</tbody>
</table>

Notice that again, IPRO selected mutants with increased electrostatic interactions suggesting that specificity is similarly increased by obstructing the binding site with a complex of charged residues. Further, these designs seem remarkably similar to those generated for C12 specificity, motivating the comparison of the designs.

A comparison of the designs for C8 specificity and C12 specificity reveals that there are no point mutations unique to the C8 designs, rather they are recapitulations of mutations from the C12 designs containing some unique combinations of mutations. Comparison of C8 mutant 5 to C12 mutant 9 reveals that mutations L76K, I107K, R108K, A111E, F139E, and Y145K are all recapitulated, the only distinguishing mutation is the lack of the L11G mutation in the mutant 5 design. The primary distinction
between the best C12 designs and the C8 designs is the preference for the Y145K mutation over the Y145R mutation.

Figure 3-3. Mutant 5 complex of charged residues binding caprylic acid

Visualization of the charged complex reveals the importance of the choice of the Y145K mutation over the Y145R mutation. Note that in contrast to the prior conformation of the complex, the R108K and Y145K mutants, rather than maintaining the A111E contact, make a contact with wild type residue E142. This contact forces the loop of Y145K to curve inward towards the binding crevice, causing the delta carbon of the lysine to intrude into the portion of the crevice occupied by carbons 9 and 10. This
conformation change to the binding crevice do to the change in relative strength of the E142 and A111E contacts provides strong rationale for the more common recurrence of this mutation.

All mutants generated for increased C8 specificity were evaluated for binding of all acyl ACP chain lengths from 8C to 18C using ensemble refinement (Table 3-5).

Table 3-5. Caprylic acid specific TAP I mutants ensemble refined energies

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Design Position</th>
<th>Ensemble Refined Interaction Energy (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11 72 76 107 108 111 139 145</td>
<td>8C 10C 12C 14C 16C 18C</td>
</tr>
<tr>
<td>WT</td>
<td>L G L I R A F Y</td>
<td>37.5 -2.9 16.7 -109.9 0.7 15.5</td>
</tr>
<tr>
<td>1</td>
<td>L G L I R A F K</td>
<td>2.3 -53.2 -69.2 -70.9 -92.0 56.2</td>
</tr>
<tr>
<td>2</td>
<td>L G L I R E F K</td>
<td>11.4 -10.3 6.2 -32.3 -200.4 71.9</td>
</tr>
<tr>
<td>3</td>
<td>L G L K K E F K</td>
<td>24.5 -8.8 13.5 -17.1 -233.2 12.1</td>
</tr>
<tr>
<td>4</td>
<td>L G L K K E E K</td>
<td>22.4 17.8 13.5 -22.5 -101.7 82.2</td>
</tr>
<tr>
<td>5</td>
<td>L G K K K K E E K</td>
<td>54.2 70.6 30.2 41.4 -127.3 114.6</td>
</tr>
</tbody>
</table>

Ensemble refined energies show an undesired specificity profile. This is likely due to the importance of the E142 contact in maintaining a favorable position for the charged residue complex. With the multitude of possible favorable electrostatic contacts available, it is apparent that a change in specificity dependent upon the positioning of a single electrostatic contact will be much more sensitive to ensemble averaging than a complex that will block the binding crevice regardless of a minor change in electrostatic contacts.
Conclusions and Future Work

In this thesis I have described efforts to redesign *E. coli* Thioesterase I/Protease I/Lysophospholipase L1 (TAP I) for increased substrate specificity using the Iterative Protein Redesign and Optimization procedure. Designs were generated for increased specificity towards both hydrolysis of lauric acid (12C) and caprylic acid (8C) from *E. coli* Acyl Carrier Protein. These designs were further evaluated by the calculation of ensemble refined interaction energies for each Acyl-ACP from the 8C bound species to the 18C bound species. This provided a more rigorous computational evaluation of the quality of the recommended designs as well as a more complete description of the computationally predicted specificity profile of these designs.

The designs generated exhibited clear rationale for the predicted increase in specificity, with interesting and unique methods for generating sterically guided changes in specificity. The designs generated new information about possible avenues to changing substrate specificity, as well as identifying residues lining the binding crevice which are not strong targets for redesign.

Experimental efforts have been unable to recapitulate the computational results thus far. Future work could include improving the quality of computationally predicted designs. The most likely cause of inaccuracy is the treatment of the enzyme as a static structure. The importance of dynamic features attributed to the activity and promiscuity of TAP I provide opportunity for improvement. Molecular dynamics simulations could be
used to identify the flexible loops which allow changes to the binding crevice shape. These loops could be targeted to decrease active site flexibility. Additionally, the contacts formed by C8 designs provided other possible design positions to target for increased specificity. TAP I is known to have activity on both ACP bound acyl groups as well as CoA bound acyl groups, participating in the beta-oxidation pathway. Other future work could include building specificity against the CoA bound species, resulting in greater yields of the desired fatty acids in *E. coli.*
References


