Structural and Biochemical Studies of the Initiation Steps in Three Natural Product Biosynthetic Pathways

by

Andrew P. Sikkema

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biological Chemistry)
in the University of Michigan
2017

Doctoral Committee:

Professor Janet L. Smith, Chair
Associate Professor Patrick J. O'Brien
Professor David H. Sherman
Associate Professor Georgios Skiniotis
Professor John J.G. Tesmer

DEDICATION

To my Grandfather. Clear skies, smooth seas, and good fishing.

ACKNOWLEGMENTS

First, I would like to thank my mentor Professor Janet Smith for her guidance and belief in me. I am very appreciative of her creation of a great learning environment and her pursuit of interesting research questions. I would also like to thank all of the members of the Smith Lab, past and present, for being great colleagues with and for teaching me so much. I need to single out Dr. David Akey for being the structural biology rock of the Smith Lab and Meredith Skiba for taking over the AprA project and finishing a story I will not be able too.

I would like to thank my collaborators: Nathan Moss of the Professor Bill Gerwick's lab and Drs. Ashootosh Tripathi and Sung Ryeol Park of Professor David Sherman's lab. I would also like to thank the staff at the GM/CA beamline at APS, Argonne National Lab, for their support and assistance with diffraction experiments. In addition, thank you to Dr. Clay Brown and Jim Delproposto for teaching me the correct way to perform molecular cloning and protein expression experiments.

I am very appreciative of the support and guidance provided by my thesis committee: Professors Patrick O'Brien, David Sherman, Yiorgo Skiniotis, and John Tesmer. I would also like to thank Mindy Mackey and the staff of the

Biological Chemistry Department, especially Beth Goodwin, for their administrative support during my time in Ann Arbor.

Lastly, I need to thank my friends and family for their support and love. It has been a long road but everyone has been there with encouragement every step of the way. I especially need to thank my dog Red for the calming walks during the writing process and my fiancée Jessica for dealing with me when the walks were not enough.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

PKS ..... Polyketide Synthase

NRPS ..... Nonribosomal Peptide Synthetase

Ppant ..... Phosphopantetheine

ACP ..... Acyl Carrier Protein

PCP ..... Petidyl Carrier Protein

ArCP ..... Aryl Carrier Protein

FAS ..... Fatty Acid Synthase

AT ..... Acyltransferase

KS ..... Ketosynthase

KR ..... Ketoreductase

ST ..... Sulfotransferase

TE ..... Thioesterase

A ..... Adenylation

C ..... Condensation

AFE ..... Adenylate Forming Enzyme

FACL ..... Fatty Acid CoA Ligase

FAAL ..... Fatty Acid ACP Ligase

GNAT ..... GCN5-related N-acetyltransferase

MT ..... Methyltransferase

PP_i ..... Pyrophosphate

P_i ..... Phosphate

DHB ..... 2,3-Dihydroxybenzoic acid

SA ..... Salicylic Acid

3-MSA ..... 3-Methylsalicylic acid

4-MSA ..... 4-Methylsalicylic acid

5-MSA ..... 5-Methylsalicylic acid

6-MSA ..... 6-Methylsalicylic acid

C6 ..... Hexanoic Acid

C8 ..... Octanoic Acid

C10 ..... Decanoic Acid

C12 ..... Dodecanoic Acid

C14 ..... Tetradecanoic Acid

C16 ..... Hexadecanoic Acid / Palmitic Acid

C18 ..... Octadecanoic Acid / Stearic Acid

ABSTRACT

Natural products are a rich source of diverse chemical compounds, many with pharmaceutical potential. Structural and biochemical investigations into the initiation steps of three natural product biosynthetic pathways were conducted. These findings contribute to the adaptation of natural products for industrial applications or for improved pharmaceutical properties as well as the discovery of novel enzymatic chemistry.

The olefin synthase from the marine cyanobacterium *Synechococcus sp.* PCC7002 converts stearic acid to 1-nonadecene, a terminal olefin. Biologically generated length-controlled olefins have potential applications as lubricants, surfactants, chemical feedstocks, and biofuels. The fatty acid-ACP ligase (FAAL) domain of olefin synthase selects stearic acid for conversion to 1-nonadecene, dictating the length of olefin produced. Crystal structures of the SynFAAL guided the creation of a number of variants that led to the discovery of an electrostatic fatty acid selection mechanism and a better understanding of the basis of fatty acid length selectivity.

The cahuitamycin pathway of *Streptomyces gandocaensis* natively produces three cahuitamycin variants, which are all potent inhibitors of *Acinetobacter baumannii* biofilm formation. Variation originates in the starter unit selected by

aryl-AMP ligase CahJ. Crystal structures of several CahJ-substrate complexes provided a structure-based explanation of substrate selectivity in the creation of natural cahuitamycins. Together with a comprehensive set of substrate selectivity data, the structure provides a roadmap for the creation of novel cahuitamycin congeners for future investigation as biofilm inhibitors.

Apratoxin A from the marine cyanobacterium *Moorea bouillonii* contains a rare *t*-butyl group. The steps for *t*-butyl biosynthesis likely involve novel enzymatic reactions by domains within AprA, the first protein of the biosynthetic pathway. A di-domain of AprA containing a type I *S*-adenosylmethionine dependent methyltransferase (MT) domain was shown to possess both acyltransfer and methylation activity. The crystal structure of the di-domain indicated that the MT domain performs both reactions and that the other domain is vestigial.

The initiation steps are an important source of chemical diversity in natural product biosynthesis. Understanding the mechanisms of substrate recognition used by initiation enzymes could allow for the rational redesign of substrate specificity leading to biological products with improved characteristics.

CHAPTER I

Introduction

1.1 Introduction

Small-molecule natural products are complex biomolecules of stunning chemical and biological diversity. Bacteria and fungi are rich sources of these secondary metabolites, which are presumed to confer a competitive advantage to the producing organism within its environmental niche. Natural products represent an important source of pharmaceuticals due to their inherent bioactivity (1). Examples include cahuitamycin A, 1-nonadecene, and apratoxin A (Figure 1.1). Understanding the biosynthesis of natural products has multiple levels of impact ranging from basic scientific research on novel biochemical reactions to the tailoring of biosynthetic pathways to produce compounds with improved pharmaceutical or synthetic characteristics.

Given the diversity of natural products, it is not surprising that nature employs many different biosynthetic schemes in their creation. This thesis focuses on two types of biosynthetic schemes: modular polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Modular biosynthetic pathways are distinguished by their use of modular groups of domains to catalyze each step of the extension and processing of a natural product intermediate. Multiple modules

are used in series to synthesize the final compound. This leads to a one-enzyme, one-reaction paradigm, which links every enzyme of the biosynthetic pathway to a specific functional group in the completed natural product. This thesis concerns the pathways illustrated in Figures 1.2-1.4.

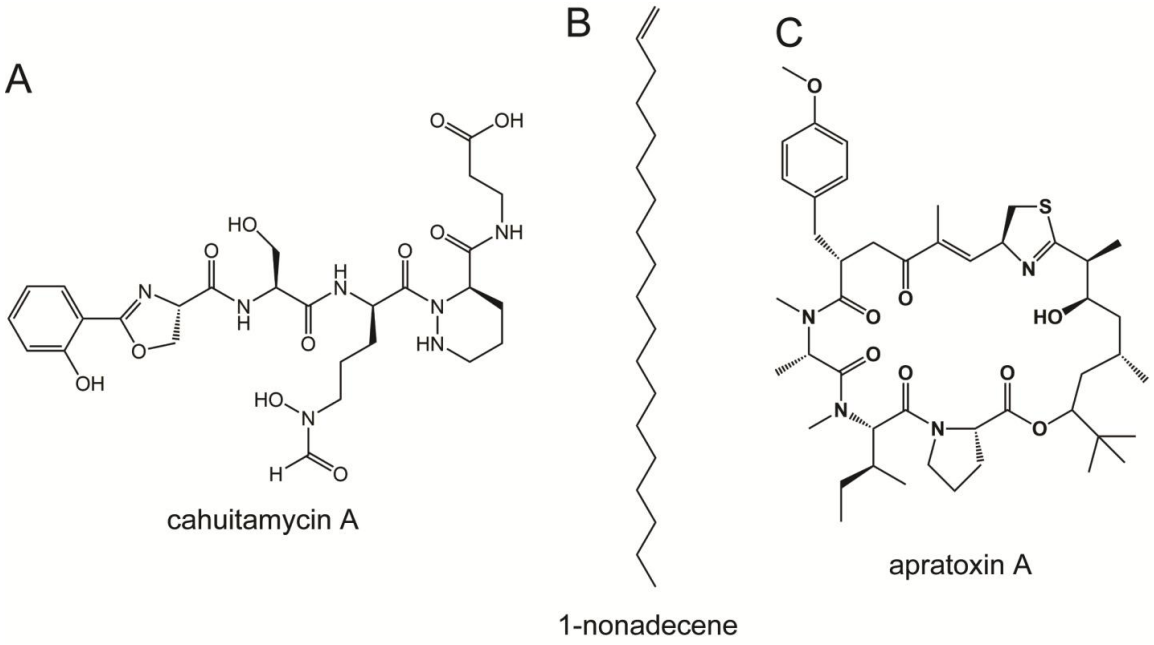

Figure 1.1 The natural product subjects of this thesis.

A) The NRPS-produced cahuitamucyin A (2) has been shown to inhibit *Acinetobacter baumanii* biofilm formation. B) The PKS-produced long-chain olefin 1-nonadecene (3) has potential industrial and biofuel applications. C) The hybrid PKS/NRPS-produced apratoxin A (4) has been shown to bind Sec61 causing cell death (5).

Modular PKS and NRPS pathways use carrier proteins to covalently tether intermediates through a thioester linkage between a carboxylate of the intermediate and the thiol of a phosphopantetheine (Ppant) prosthetic group on the carrier protein. The acyl carrier protein (ACP) for PKS modules and the peptidyl carrier protein (PCP) for NRPS modules are related to each other and to

the ACP of fatty acid biosynthesis. The carrier protein shuttles intermediates between enzyme active sites and ensures that the intermediates are not lost from the biosynthetic machinery. Carrier protein delivery is well illustrated by the movements of the ACP in the cryo-EM structures of PikAIII (6,7).

Polyketide synthase modules are functionally and evolutionarily related to type I metazoan fatty acid synthases (FAS) (8). A PKS module typically consists of between three and six domains, minimally an ACP domain, a ketosynthase (KS) domain, and an acyltransferase (AT) domain. Collectively, these domains extend the polyketide intermediate. In an important source of polyketide diversity, the AT domain selects an acyl-CoA extender unit and transfers the acyl group to the ACP (9). Most commonly, the AT selects malonyl-CoA or methylmalonyl-CoA; however, approximately a dozen acyl-CoA substrates are known (10). The ACP then shuttles the acyl group to the KS domain, where the upstream ACP has transferred its intermediate to a catalytic cysteine thiol. The KS catalyzes the decarboxylative Claisen condensation of the upstream intermediate with the acyl-ACP, thereby extending the intermediate polyketide chain by two carbon atoms. Additional diversity comes from the activities of up to three additional domains, ketoreductase (KR), dehydratase (DH) and enoylreductase (ER), which tailor the intermediate by reductive chemistry at the β-carbonyl created in the extension reaction. Modular PKS pathways are reviewed in more detail by Hill and Staunton (11) and Smith, Skiniotis and Sherman (12).

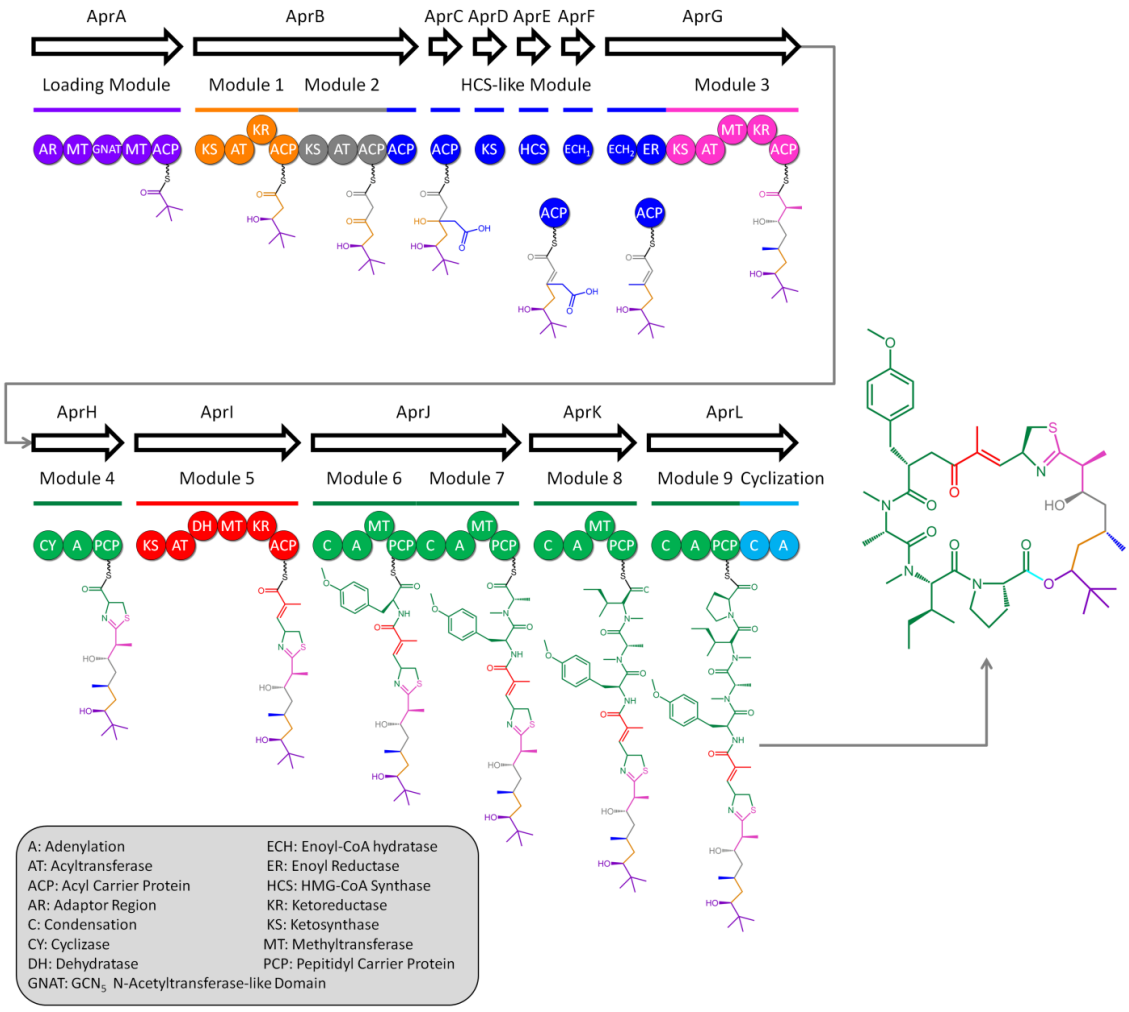

Figure 1.2 The apraxotin A biosynthetic pathway (4).

The domains and products of the pathway modules are color-coded: loading in purple, PKS extensions in orange, grey, pink, and red, β-branching in blue, NRPS extension in green, and offloading in cyan.

The minimal NPRS module consists of an adenylation (A) domain, a condensation (C) domain, and a PCP domain. The A domain selects an amino acid, or related analogue, activates its carboxylate by adenylation from ATP, and transfers the amino acid to the PCP (10,13). The PCP then transports the selected amino acid to the C domain, which catalyzes the formation of a peptide bond between the peptide intermediate from the upstream PCP and the selected

amino acid. Some NRPS modules have C domains that also catalyze a cyclization reaction to form a thiazoline or oxazoline ring, and some modules also have tailoring enzymes such as *N*-methytransferase (NMT) domains, which methylate the backbone amide. The overall chemistry of NRPS pathways was reviewed in detail by Walsh (14). In addition to purely PKS or NRPS pathways, hybrid PKS/NRPS pathways are common. This is possible because of the common thioester tethering of intermediates by carrier proteins. The apratoxin biosynthetic pathway is an excellent example of a hybrid pathway with interspersed NRPS and PKS modules (Figure 1.2) (4). The extensions and modifications described above constitute only the middle portion of PKS or NRPS pathways, which must also initiate biosynthesis and offload the final products. Product release is commonly achieved with thioesterase domains that can catalyze either thioester hydrolysis to a linear product or intramolecular cyclization to a macrolactone or cyclic peptide.

The topic of this thesis is the initiation step, among the most chemically diverse portions of modular natural product biosynthetic pathways. In many cases, the extension module machinery (*e.g*. NRPS A domains and PKS AT domains) is repurposed for the task of pathway initiation. However, the enzymes and mechanisms employed in loading an initial building block onto a carrier protein domain, reviewed from a chemical perspective by Moore and Hertweck (15), can be quite varied. Two broad categories of enzymes, the adenylate-forming

enzymes and the acyltransferases, are employed for pathway initiation (Figure 1.5).

1.2 Pathway Initiation by Adenylate-Forming Enzymes

Adenylate-forming enzymes (AFE) constitute a large superfamily grouped by common chemistry (16). AFE family members activate a substrate carboxylate with ATP to form a 5'-adenylate intermediate that releases pyrophosphate. The activated intermediate is then transferred to a nucleophilic substrate, releasing AMP. For biosynthetic initiation, the nucleophile is the Ppant thiol of a carrier protein. The AFE superfamily includes a wide variety of enzymes including firefly luciferase, fatty acid CoA ligases (FACL), NRPS A domains, and aminoacyl-tRNA synthetases. Two subclasses of AFE, NRPS A domains and fatty acid ACP ligases (FAAL), are used for natural product biosynthetic initiation.

1.3 Initiation by NRPS A Domains

The most straightforward AFE-based loading module consists of an NRPS A and a PCP. These loading modules function identically to their counterparts in extending modules. However, as the substrate selected by the loading A domain does not undergo a condensation reaction with an upstream peptide intermediate, an amino group is optional (Figure 1.5A). This allows for greater diversity in the starter unit selected by A domains, in principal limited only by the requirement for a carboxylate group that can be activated with ATP (10,17). Examples of non-amino acid starter units can be found in the cahuitamycin (2),

bacillaene (18), and rapamycin (19) pathways. These three pathways also demonstrate that initiation modules containing NRPS A domains are present at the beginning of NRPS, PKS and hybrid PKS/NRPS pathways.

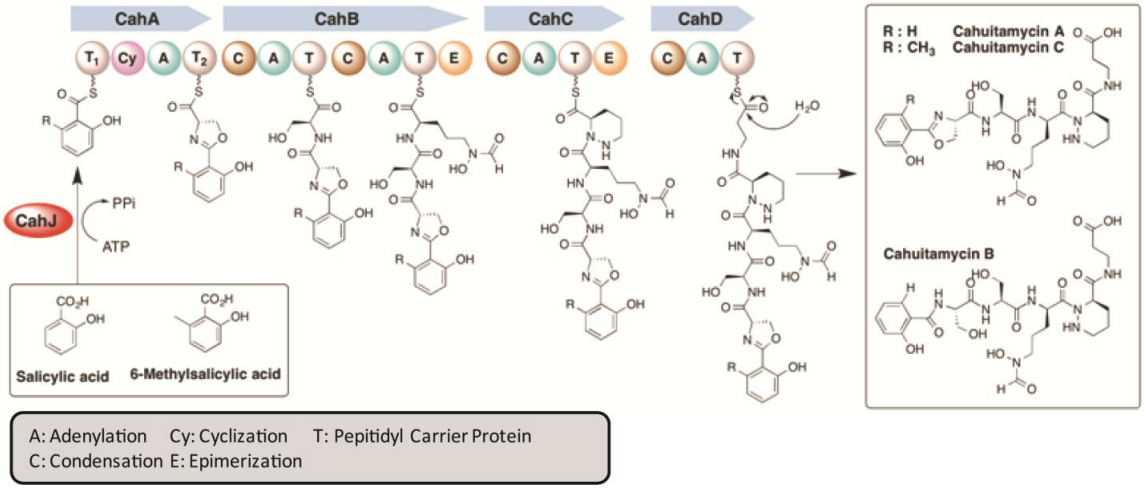

Figure 1.3 The cahuitamycins A-C biosynthetic pathway (2).

The domains are color-coded: PCP (T) domains in tan, cyclization (Cy) domain in pink, adenylation (A) domains in green, condensation (C) domains in brown, and epimerization (E) in orange. Figure courtesy of Ashootosh Tripathi.

1.3.1 Initiation of the cahuitamycin pathway

The cahuitamycins are potent inhibitors of *Acinetobacter baumanii* biofilm formation, a common nosocomial pathogen (2). The cahuitamycins are produced by a bifurcated biosynthetic pathway that has flexibility in the starter unit used to initiate biosynthesis. The starter unit significantly affects biofilm inhibition activity and is selected by a stand-alone NRPS A domain, CahJ (Figure 1.3). CahJ natively selects either salicylic acid or 6-methylsalicylic acid leading to the production of cahuitamycin A-B or cahuitamycin C, respectively. CahJ also shows rare promiscuity by acting on the non-native substrate 5-methylsalicylic acid (5MSA) leading to the production of a new cahuitamycin analog (2).

Reinforcing the importance of the starter unit, the 5MSA-derived cahuitamycin analog has approximately twofold greater biofilm inhibition activity compared to the most active natural variant.

1.4 Initiation by FAAL Domains

Despite their similar function, it is clear by sequence identity that FAAL domains are derived from primary metabolism fatty acid-CoA ligases (FACL) and not NRPS A domains (20,21). FACL and FAAL enzymes both activate fatty acids with ATP to form fatty acyl adenlyates. FACL enzymes transfer the activated fatty acid to CoA while FAAL enzymes transfer the fatty acid to a partner ACP (Figure 1.5B). A poorly conserved insert region that prevents FAAL enzymes from interacting with CoA distinguishes FAALs from FACLs (21-23). A number of FAAL enzymes have been identified at the beginning of natural product pathways including the columbamide PKS pathway (24), the jamaicamide hybrid PKS/NRPS pathway (25), and pyoverdine NRPS pathway (26). The range of fatty acid lengths in these natural product pathways is hexanoic acid (C6) to stearic acid (C18). FAALs are also employed in the biosynthesis of extremely long-chain fatty acids in *Mycobacterium tuberculosis* (27,28).

1.4.1 Pathway initiation of the Synechococcus sp. PCC7002 olefin synthase

Olefin synthases are small, polyketide-based biosynthetic pathways that convert fatty acids to olefins (29,30). Olefins have potential applications as biofuels, surfactants, lubricants, and chemical feedstocks (31). Olefin production is

initiated by an FAAL enzyme that selects the fatty acid that will be converted to an olefin. The fatty acid then undergoes a canonical PKS extension and ketoreduction. The extended fatty acid is then offloaded by a sulfotransferase and thioesterase domain pair, which generates a terminal olefin that is one carbon atom shorter than the extended fatty acyl group (32-34). In the case of the *Synechococcus sp.* PCC 7002 olefin synthase, stearic acid (octadecanoic acid) is converted to 1-nonadecene (Figure 1.4) (3,30). All identified olefin synthases produce long-chain olefins (29) that are too long for current industrial applications (31). An understanding of the FAAL domain's fatty acid selection mechanism is required to engineer olefins of more useful chain length.

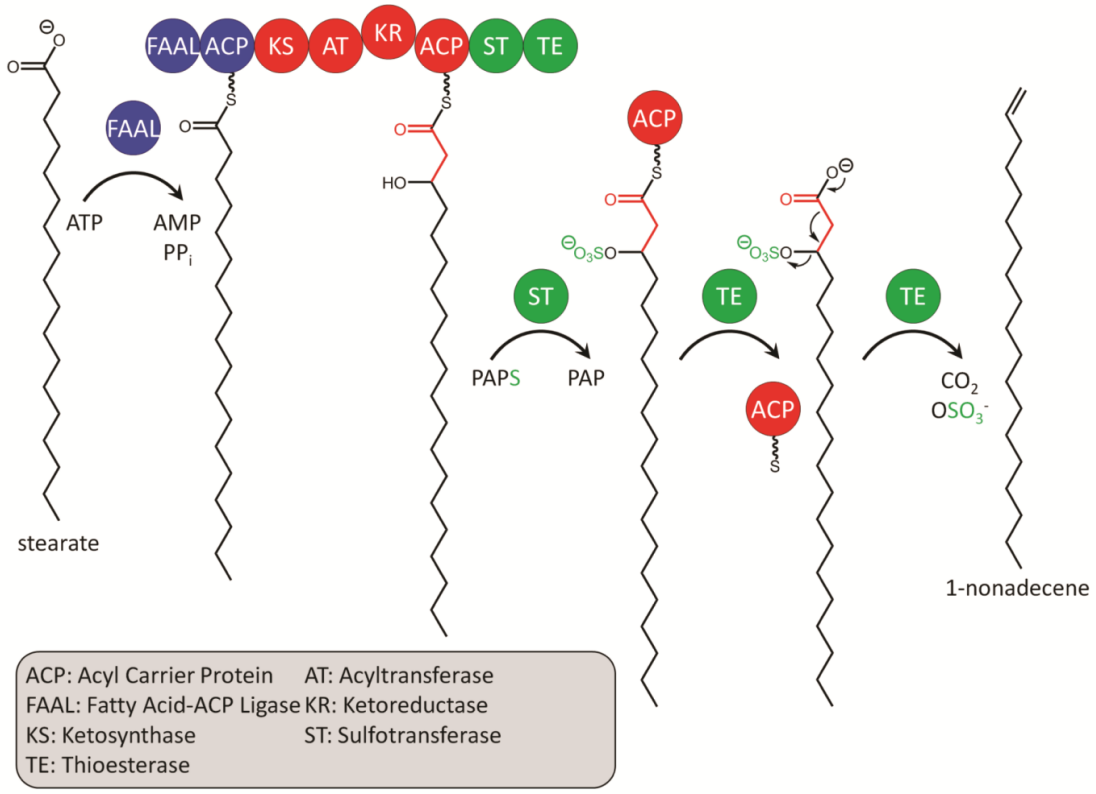

Figure 1.4 The olefin synthase pathway from *Synechococcus* sp. PCC7002 (3).

The domains and products of the pathway modules are color-coded: loading in blue, PKS extension in red, and offloading in green.

1.5 Pathway Initiation by Acyltransferase Enzymes

Natural product acyltransferase enzymes transfer an acyl group from a specific acyl-CoA to an ACP Ppant thiol. The majority of pathways that use an acyl-transfer initiation strategy employ a standard PKS AT domain; however, some pathways use a GCN5-related N-acetyltransferase (GNAT)-related domain (35). The transfer of an acyl group from a CoA thioester to a Ppant thiol is isoenergetic (36) and should require only proper active site geometry, *i.e*. no catalytic residues are required. Thus it is possible that other, as-yet-unidentified classes of enzymes may catalyze acyltransfer reactions.

1.6 Initiation by PKS AT Domains

Considered the canonical form of PKS initiation, loading modules containing AT domains function similarly to their counterparts in extension modules and consist of an AT domain and an ACP (37). Loading module ATs act identically to other AT domains by selecting the proper acyl-CoA and transferring the acyl group to the ACP, via an acyl-enzyme intermediate at a catalytic Ser. The majority of pathways are initiated with either acetyl- or propionyl-ACP, which can be generated by direct transfer from acetyl- or propionyl-CoA or by the decarboxylation of malonyl- or methylmalonyl-ACP through the action of an additional decarboxylating KS enzyme (Figure 1.5C) (38,39). Although most pathways are initiated with acetyl- or propionyl-ACP, pathways utilizing isobutryl-CoA (40), benzoyl-CoA (41), and 2-methylbutryl-CoA (42) as a starter unit are

common. As a more extreme example, the borrelidin pathway initiates with cyclopentanedicarboxylate (43), a significant departure from the common acetyl and propionyl starter units and an example of the remarkable adaptability of AT domains.

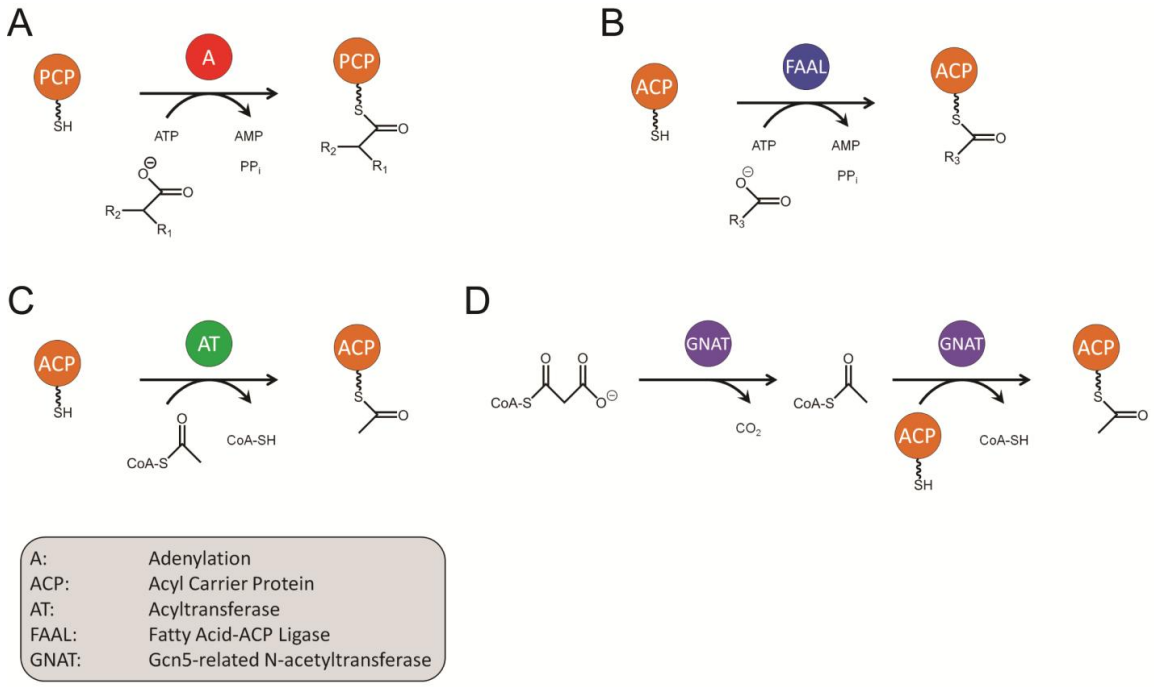

Figure 1.5 The four major routes to initiation of natural product biosynthesis.

A) Initiation by NRPS A domains utilizes ATP to activate amino acids or related molecules, and load them onto PCP domains. For amino acids, R_1 is NH_3 and R_2 is an amino acid side chain. For non-amino acids molecules, R_1 and R_2 can be any moiety. B) FAAL domains initiate biosynthesis similarly to A domains but utilize fatty acids as substrates instead of amino acids. R_3 is $(CH_2)_nCH_3$ were n is typically between 4 and 16. C) PKS AT domains initiate biosynthesis by transferring acyl groups from CoA to an ACP domain. D) GNAT-based loading modules decarboxylate malonyl-CoA to acetyl-CoA and then transfer the acetyl group to an ACP domain.

1.7 Initiation by GNAT-like Domains

First discovered in a natural product pathway in the pederin biosynthetic pathway (44) and first characterized in the curacin biosynthetic pathway (35), initiation by

GNAT domains has been identified in at least twelve pathways. As a superfamily, GNAT domains catalyze the transfer of acyl groups from acyl-CoA to a diverse group of substrates and are found across all domains of life (45). The family was first defined by the yeast general control non-repressible 5 (GCN5)-related histone *N*-acetyltransferase (46). In the context of natural product pathways, GNAT enzymes act similarly to AT domains to transfer of an acyl group from CoA to an ACP, but without use of an acyl-enzyme intermediate (Figure 1.5D). Investigations of the curacin loading module (CurA) GNAT demonstrated that the enzyme utilizes malonyl-CoA as a substrate, which it decarboxylates prior to acyl transfer (35). Examples of loading modules containing GNAT domains are know from gene cluster assignments across diverse organisms such as sponge symbionts (44,47,48), myxobacteria (49), and cyanobacteria (4,50,51); however, only the CurA GNAT has been characterized biochemically or structurally. A subset of GNAT-based loading modules also contain putative methyltransferase (MT) domains, which are proposed to methylate the acetyl-ACP generated by the GNAT domain (49,50).

1.7.1 Initiation of the apratoxin pathway

Apratoxin A has been shown to bind to the translocation protein Sec61, inhibiting cotranslational translocation and causing cell death (5,52,53). The loading module of the apratoxin pathway (AprA) contains two putative MT domains, a feature that has not yet been observed elsewhere (Figure 1.1) (4). Additionally, the AprA GNAT domain aligns poorly with the other loading module GNATs and

appears to lack the amino acids responsible for the decarboxylation activity of the CurA GNAT. Together this may indicate that AprA utilizes a unique mechanism of biosynthetic initiation despite the presence of a GNAT domain. The product of the AprA loading module is a *t*-butyl containing pivalyl-ACP, both of which are rarely seen in natural product pathways. The putative MT domains are proposed to create the pivalyl-ACP by performing three SAM-dependent methylations on the GNAT generated acetyl-ACP.

1.8 Thesis Overview

The studies described in the following chapters seek to broaden our understanding of the initiation steps of natural product biosynthesis though structural and biochemical investigations into three biosynthetic pathways: the olefin synthase of the marine cyanobacterium *Synechococcus sp.* PCC 7002, the cahuitamycin pathway of *Streptomyces gandocaensis*, and the apratoxin pathway of the marine cyanobacterium *Moorea bouillonii*. For the *Synechococcus sp.* PCC 7002 olefin synthase, crystal structures of the FAAL domain guided the creation of a number of variants that led to the characterization of an electrostatic fatty acid selection mechanism and a better understanding of fatty acid length selectivity. For the cahuitamycin pathway, crystal structures of several CahJ-substrate complexes provided a structure-based explanation of substrate utilization probed with a large, chemically diverse set of substrates. Lastly, for the apratoxin pathway structural and biochemical data is presented showing that the loading module GNAT is significantly different

from the GNAT of the curacin pathway in both structure and function. Additionally, structural data are presented showing that the second MT domain is similar to the recently published structure of the CurJ methyltransferase (54).

CHAPTER II

An Electrostatic Fatty Acid Selection Mechanism by the Olefin Synthase FAAL domain from *Synechococcus sp.* PCC7002

2.1 Notes

This chapter is being prepared for publication in a manuscript titled " An Electrostatic Fatty Acid Selection Mechanism by the Olefin Synthase FAAL domain from Synechococcus sp. PCC 7002" with the authors: Sikkema AP, Sturgis R, Gerwick WH, Sherman DH, Smith, JL.

Author Contributions:
A.P.S. and J.L.S. designed experiments. A.P.S and R.S performed experiments. A.P.S. wrote the manuscript. A.P.S. and J.L.S. edited the manuscript.

2.2 Abstract

Long and medium chain α-olefins have applications as lubricants, surfactants, chemical feedstocks, and biofuels. The marine cyanobacterium *Synechococcus sp.* PCC 7002 contains an olefin synthase that converts stearic acid to 1-nonadecene. Stearic acid is selected and activated by the fatty acid-ACP ligase (FAAL) domain at the N-terminus of the olefin synthase prior to canonical reactions of polyketide synthase (PKS) extension and β-ketoreduction. The extended fatty acid is then sulfonated at the β-hydroxy position and offloaded via concerted desulfation and decarboxylation to generate an α-olefin product that is one carbon longer than the fatty acid substrate. The length of the selected fatty acid selected by the FAAL domain dictates the length of the final α-olefin.

Understanding the FAAL mechanism of fatty acid chain length selection would enable the engineering of olefin synthases to produce specific α-olefins. Here crystal structures of SynFAAL in complex with AMP and acyl adenylates is report. The structures guided the creation of a number of SynFAAL variants that led to the characterization of a mechanism for chain-length selectivity of fatty acid substrates which is supported with biochemical data.

2.3 Introduction

Microbial production of hydrocarbons has drawn increasing interest as a potential alternative to petroleum-based fuels and chemicals (31,55,56). Cyanobacteria have two major pathways for the conversion of fatty acids to hydrocarbons (29). The majority of cyanobacteria contain a “FAAR/ADO” system, which converts fatty acids to alkanes via the dual action of a fatty acyl acyl-carrier-protein (ACP) reductase (FAAR) and an aldehyde deformylating oxygenase (ADO). Alternatively, some cyanobacteria contain olefin synthases, small biosynthetic pathway composed of one or two polypetides that convert fatty acids to α-olefins. Although α-olefins have applications as energy sources, they have potentially higher value as lubricants, surfactants and chemical feedstocks (31).

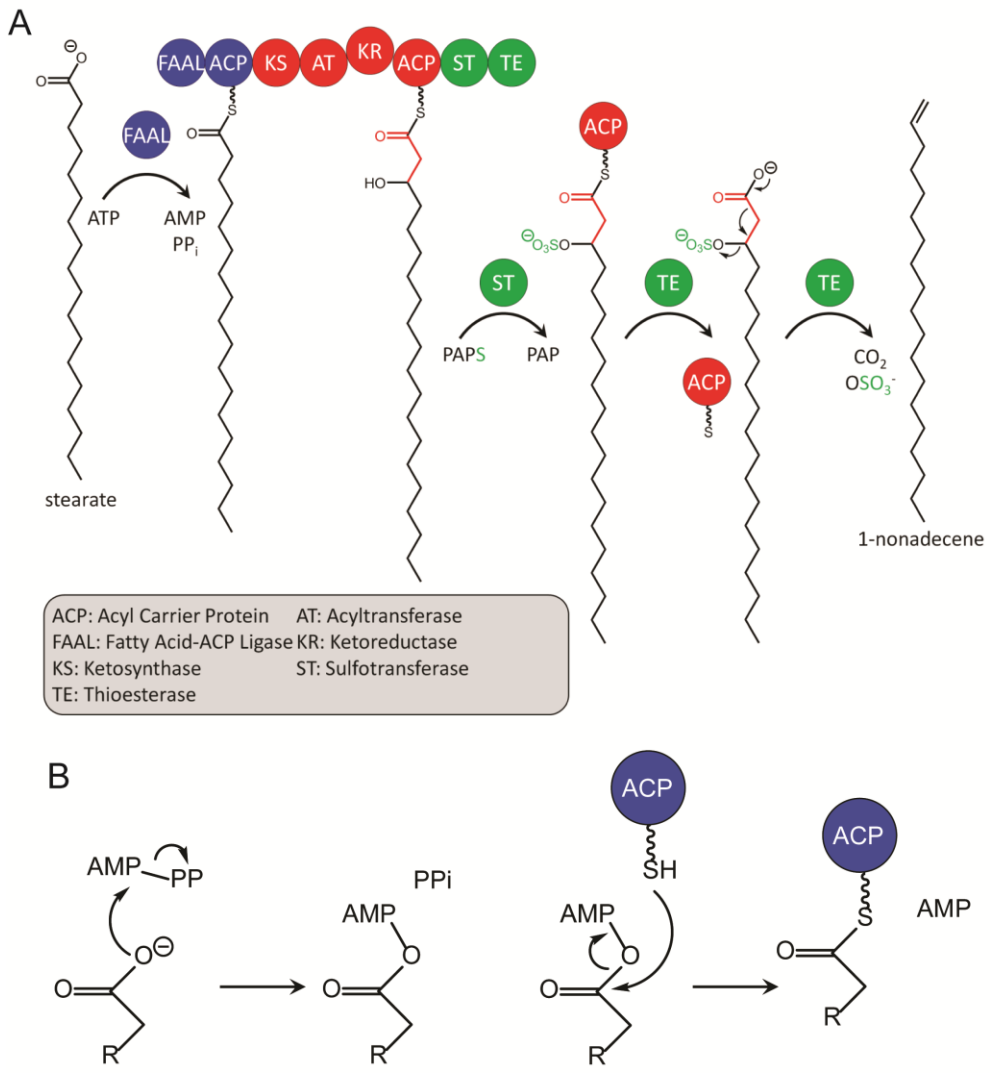

Figure 2.1 Olefin biosynthesis and FAAL activity.

A) The olefin synthase pathway from *Synechococcus* sp. PCC 7002. B) The two-step FAAL reaction; 5'-adenylation and acyl transfer to ACP.

Olefin synthases consist of six enzymatic domains: fatty acid ACP ligase (FAAL), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR), sulfotransferase (ST), and thioesterase (TE), and two ACP domains (ACP1 and ACP2) (Figure 2.1A). The FAAL initiates biosynthesis by selecting a fatty acid, activating it by adenylation from ATP, and transferring it to the phosphopantetheine cofactor of ACP1 (Figure 2.1B). The fatty acyl chain then undergoes canonical reactions of

polyketide synthase (PKS) extension and β-ketoreduction, catalyzed by the AT, KS and KR (12). The two-carbon extension reaction transfers the fatty acyl chain to ACP2. α-olefin production is completed by the unusual ST and TE offloading reactions in which β-hydroxyacyl-ACP sulfonation is followed by chain release via concerted desulfation and decarboxylation to generate an α-olefin product that is one carbon longer than the fatty acid substrate (32-34).

The *Synechococcus sp.* PCC 7002 olefin synthase has been characterized *in vivo* to produce 1-nonadecene. However, feeding odd-number chain-length fatty acids to *Synechococcus sp.* PCC 7002 resulted in even-number chain-length olefins, for example feeding heptadecanoic acid resulted in 1-octadecene (3). *In vivo*, olefin production is required for survival at lowered growth temperatures, indicating a biological function in the regulation of membrane fluidity (30). The chain length of α-olefins affects important chemical and physical properties such as energy density and melting temperature, which are critical to their use as biofuels and in industrial applications (31). Thus, the ability to control olefin chain length though protein engineering could be a major advantage of producing olefins in bacterial systems. In order to achieve such control the mechanism of FAAL selectivity of fatty acid chain-length must be understood.

FAALs are members of the adenylate-forming enzyme superfamily, which includes nonribosomal peptide synthetase (NRPS) adenylation domains, fatty acid-CoA ligases (FACLs), and firefly luciferase (16). The FAALs are

distinguished from FACLs by an ~20-amino acid insertion, which is proposed to prevent formation of a CoA binding pocket in the N-terminal domain where it is located. Removal of the insertion restores the ability of an FAALs to transfer fatty acids to CoA (21,22).

FAALs are found in number of biological contexts. Those most closely related to the olefin synthase FAALs are within the biosynthetic pathways of natural products, including jamaicamide (25), hectochlorin (57), and columbamide (24). The natural-product FAALs act on shorter fatty acids than olefin synthase FAALs (hexanoic acid for jamaicamide and hectochlorin and dodecanoic acid for columbamide), indicating a variable selectivity in the substrate chain-length. JamA, the FAAL from the jamaicamide pathway, can act on fatty acids between 5 and 7 carbons long in a variety of saturation states (58,59). *Mycobacterium tuberculosis* also contains a number of FAALs (27) such as those linked to the production of the mycolic acids (23). Crystal structures have been reported for *M. tuberculosis* FadD32 FAAL, a potential drug target (60) and for a few putative FAALs of unknown function from *E. coli* and *L. pneumophila* (20).

To improve our understanding of the substrate selectivity of FAALs, we examined the *Synechococcus sp.* PCC 7002 olefin synthase FAAL domain (SynFAAL). Here we report the crystal structure of SynFAAL in complex with AMP and acyl adenylates together with biochemical data on the chain-length selectivity for acyl-ACP1 (SynACP1) substrates. We also demonstrate that the chain-length

selectivity can be manipulated by mutagenesis of the FAAL fatty acid binding tunnel.

2.4 Results

2.4.1 Biochemical Characterization of SynFAAL

The fatty acid ACP ligase activity of SynFAAL was assayed by HPLC revealing a robust conversion of holo-ACP to stearoyl-ACP (Figure 2.2A). Next, we evaluated SynFAAL selectivity for fatty acid chain length via a high-throughput malachite green colorimetric assay (61). In this coupled assay, the pyrophosphate produced by the fatty acid adenylation step is converted by inorganic pyrophosphatase to orthophosphate (Pi), which is detected with malachite green. We tested SynFAAL with a panel of fatty acids from four to twenty carbons in length. Surprisingly, SynFAAL had a bimodal substrate length profile. Holo-ACP1 was acylated with the natural substrate, stearic acid (octadecanoic acid - C18) at an apparent rate of 0.23 ± 0.03 $\mu M/min^{-1}$ whereas palmitoylation (hexadecanoic acid - C16) was threefold slower (0.085 ± 0.009 $\mu M/min^{-1}$). In contrast, acylation with tetradecanoic (C14), dodecanoic (C12), and decanoic (C10) acids was up to twofold faster than from the native substrate (0.32 ± 0.04 $\mu M/min^{-1}$, 0.49 ± 0.06 $\mu M/min^{-1}$, 0.53 ± 0.04 $\mu M/min^{-1}$, respectively), and acylation with substrates shorter than C10 or longer than C18 was 3-6 times slower than the C18 natural substrate (Figure 2.2B). It is likely that SynFAAL distinguishes stearic acid (C18) from palmitic acid (C16) but not from C10 - C14 fatty acids, which are expected to have very low concentrations *in vivo*. Hence

we conclude that the greater activity observed with C10 - C14 substrates reflects a lack of selectivity against them.

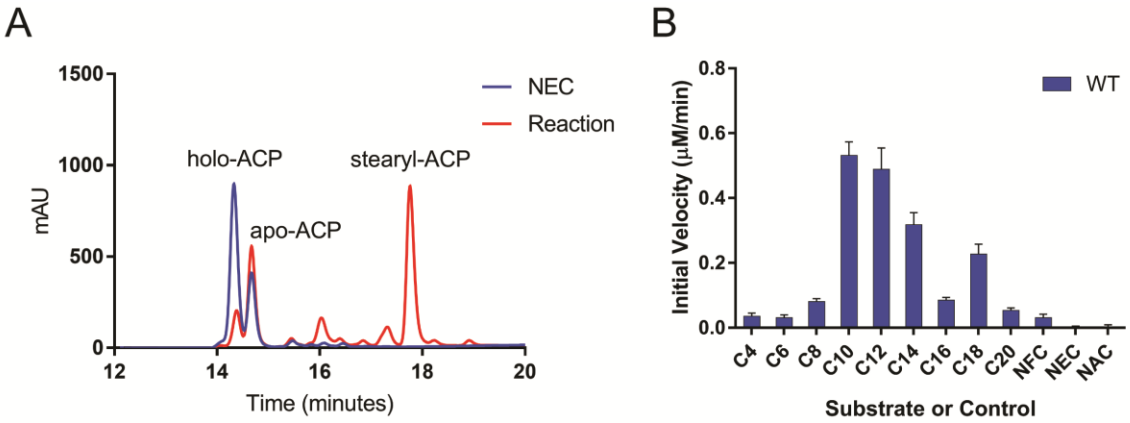

Figure 2.2 SynFAAL fatty acids ligase activity with ACP.

A) HPLC elution of SynACP1 species in a reaction mix with SynFAAL, stearic acid and ATP (red trace) compared to a negative control without SynFAAL (blue trace). Samples were incubated overnight at 20°C. Some apo-ACP1 contaminates both samples as the ACP1 was incompletely loaded with the phosphopantetheine cofactor. The identity of stearoyl-ACP1 at 17.9 min was confirmed by mass spec analysis. B) Activity of SynFAAL on a panel of even-number fatty acids of chain length between 4 (C4) and 20 (C20) carbons measured by a malachite-green-based pyrophosphate detection assay. Negative controls include NFC: no fatty acid, NEC: no enzyme, NAC: no ACP. Error bars indicate one standard deviation of 3 replicates in triplicate.

2.4.2 Overall Structure of SynFAAL

To aid our understanding of substrate selectivity, we determined the crystal structure of SynFAAL at 2.0Å. Screening yielded SynFAAL crystals in the presence of ATP (Table 2.1). The crystals grew, with poor reproducibility, as clusters of plates that required raster scanning with a 10-μm diameter X-ray beam to identify sites where the beam intercepted a single crystal. Size exclusion chromatography indicated that SynFAAL is a monomer in solution, and crystals have two monomers in the asymmetric unit. Both monomers are in the same

conformation. SynFAAL has a similar structure to other members of the FAAL subfamily of adenylate-forming enzymes (20,23,62), consisting of an N-terminal domain (amino acids 2-449) and a smaller C-terminal domain (450-570). The closest homologue of known structure is the *L. pneumophila* fatty acyl-adenylate ligase (20) (43% identity, RMSD 0.88Å, PDB code: 3LNV), which was used to solve the structure of SynFAAL by molecular replacement. The fatty acid binding pocket is contained within the N-terminal domain and the ATP binding pocket is formed at the interface of the two domains (Figure 2.3). Co-crystallization with ATP resulted in a complex with AMP and PP_i due to slow hydrolysis during crystallization.

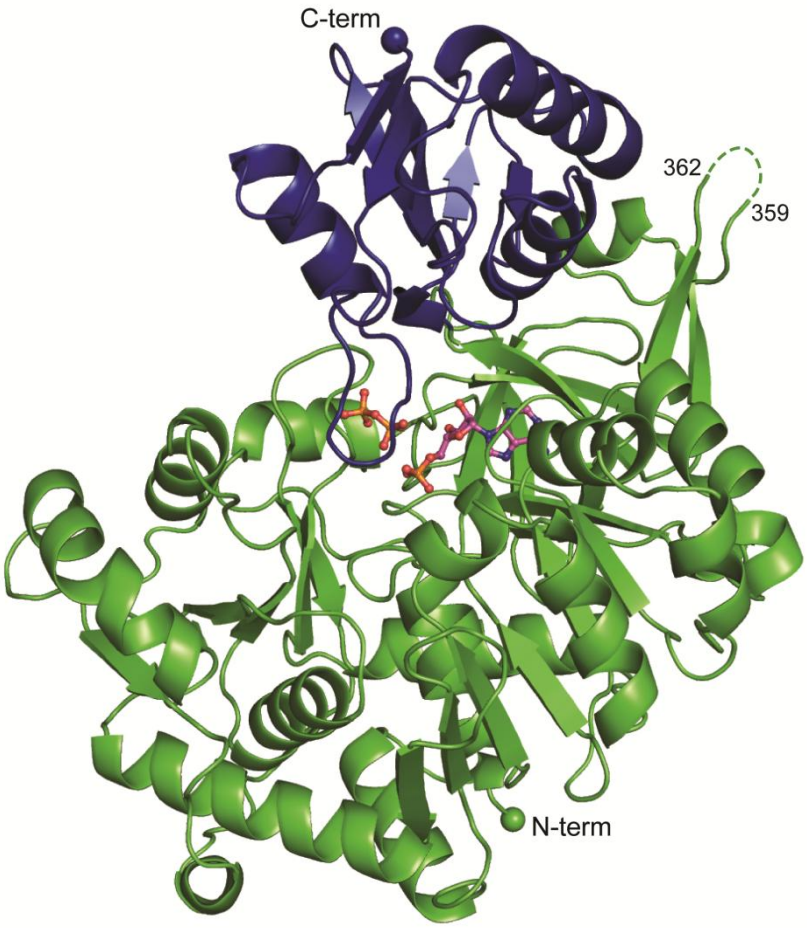

Figure 2.3 Structure of SynFAAL at 2.0Å.

AMP and PP_i bind between the N-terminal domain (green) and the C-terminal domain (blue). Ligands are shown as ball-and-stick in atomic coloring (magenta C, red O, blue N, orange P). The N- and C-termini are shown as spheres.

Table 2.1 X-ray Data Collection and Refinement Statistics

Data Collection	SynFAAL	SynFAAL/R336A + Stearyl Adenylate	SynFAAL/R336A/A229I + Palmityl Adenylate
Space group	*C*2	*C*2	*C*2
Cell dimensions a/b,c (Å)	201.3, 60.9, 95.9	199.1, 60.6, 95.1	201.2, 60.9, 95.8
α,β,γ (°)	90, 101.2, 90	90, 101.1, 90	90, 101.1, 90
X-ray source	APS 23ID-D	APS 23ID-D	APS 23ID-B
Wavelength (Å)	1.033	1.033	1.033
d_{min} (Å)	1.99 (2.11-1.99)[1]	2.13 (2.26-2.13)	2.25 (2.38-2.25)
R_{merge}	0.094 (0.647)	0.126 (0.753)	0.144 (0.663)
Wilson B factor	34.2	34.4	36.8
Avg I/σ(I)	8.96 (1.49)	8.20 (1.79)	6.58 (1.61)
Completeness (%)	97.0 (86.9)	99.3 (97.1)	99.2 (97.4)
Multiplicity	3.08 (2.63)	4.05 (3.94)	3.41 (3.45)
Total observations	233,105 (28,483)	253,195 (38,712)	185,070 (29,379)
$CC_{1/2}$	0.995 (0.708)	0.992 (0.684)	0.990 (89.6)
CC*	0.999 (0.911)	0.998 (0.901)	0.997 (0.972)
Refinement			
Data range (Å)	29.82-1.99	50.50-2.13	29.4-2.25
Reflections used in refinement	75,765	62,444	54,183
R_{work}/R_{free}	18.7/23.9	19.9/25.1	20.4/27.5
Number of non-hydrogen atoms	9,427	9,167	9,039

protein	8,654	8,665	8,652
ligands	112	112	68
water	661	378	319
Amino acid residues	1,134	1,136	1,135
Deviation from ideality			
bond lengths (Å)	0.009	0.010	0.009
bond angles (°)	1.05	1.14	1.23
Average B-factor	36.1	36.8	42.5
macromolecules	35.6	36.8	42.5
ligands	40.6	39.1	47.7
solvent	41.0	36.6	39.7
Ramachandran plot			
favored (%)	98.0	96.5	96.6
allowed (%)	1.9	3.3	3.0
outliers (%)	0.1	0.2	0.4

[1]Values in parentheses pertain to outermost shell of data.

A fatty acid binding pocket is apparent as a long, largely hydrophobic tunnel extending into the N-terminal domain. Sparse density in the tunnel may be a co-purifying fatty acid at low occupancy or poorly localized waters. I attempted to crystallize a SynFAAL stearoyl adenylate complex, however the addition of fatty acid to the crystallization solutions prevented crystal growth and I was unable to soak fatty acids into previously grown crystals.

2.4.3 Crystal Contact Engineering

The challenge of producing single crystals of SynFAAL and a lack of success in crystallizing a complex with the stearoyl adenylate intermediate led me to investigate crystal lattice contacts that might be improved by mutagenesis. I identified an unfavorable contact between Arg336 side chains in two FAAL molecules at a crystallographic twofold axis (Figure 2.4). An alanine substitution at this position improved the reproducibility and speed of crystallization and increased the frequency of single crystal growth. SynFAAL/R336A had activity and substrate selectively almost identical to the wild type and was used for all subsequent crystal structures (Table 2.2).

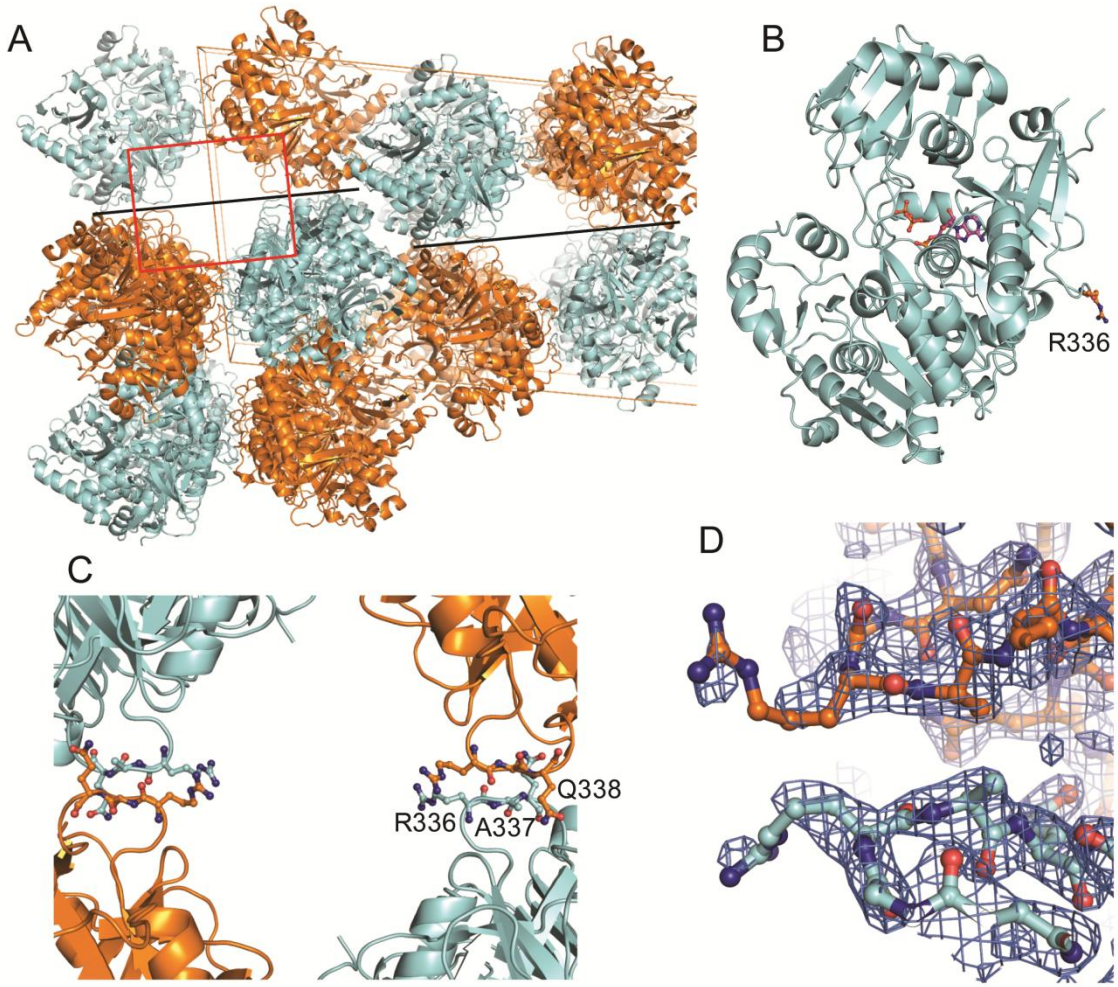

Figure 2.4 Crystal Packing of SynFAAL.

A) Packing of wild type SynFAAL in crystals of space group *C*2, viewed along the crystallographic *b* axis. The two independent SynFAAL molecules are rendered in orange and cyan. The poor lattice contact is indicated with a black line. The red box indicates the portion of the lattice shown in part C. B) The position of Arg336 in SynFAAL is more than 25 Å from the site of acyl transfer. Bound AMP and PP_i are rendered in magenta ball-and-stick to show the location of the active site. R336 shown in orange ball-and-stick. C) Close up view of the weak contact around the crystallographic twofold axis with contacting amino acids shown in ball-and-stick. D) Electron density at the lattice contact (2Fo-Fc contoured at 1σ) with protein atoms shown in ball-and-stick. In neighboring molecules, the Arg336 Cα atoms are only 6 Å apart and their side chains are poorly ordered beyond the Cγ atom due to electrostatic repulsion. Substitution of Ala at position 336 resulted in more reproducible crystallization and higher quality crystals.

2.4.4 Fatty Acid Binding Tunnel of SynFAAL

The improved crystallization behavior of SynFAAL/R336A allowed us to capture a stearoyl adenylate intermediate (Table 2.1, Figure 2.5). The C18 stearoyl chain binds in the long, hydrophobic binding tunnel as expected, and occupies the full length of the tunnel (Figure 2.5C and D). The tunnel lies between the curved β-sheet and two helices of the core domain. It is narrow for much of its length, but has a triangular bulge containing three ordered water molecules near the C-12 position of the stearoyl chain (Figure 2.5C). Surprisingly, the fatty acid does not form a hydrogen bond with SynFAAL; the carbonyl oxygen of the thioester is pointed toward the adenine ring of AMP instead of the protein (Figure 2.5A).

2.4.5 Electrostatic Lock for Fatty Acid Selection

Although the fatty acid binding tunnel has a predominantly nonpolar surface (Figure 2.6A), it has a positive surface potential at the top and a negatively charged surface at the bottom (Figure 2.6B and C). As expected, the AMP and PPi sites are positively charged. Conversely, a strong negative patch near the bottom of the binding tunnel between the C15 and C17 positions is due to the Asp199 side chain (Figure 2.7A). I speculated that the negative charge at the C15-C17 position in the tunnel may be an “electrostatic lock” that effectively reduces selectivity for palmitic acid (C16). To explore this possibility, we characterized the activity of SynFAAL/D199L and SynFAAL/D199A with the panel of fatty acids (Figure 2.7B). Importantly, both nonpolar substitutions

resulted in an approximately two-fold increase in activity with palmitic acid (C-16) as well as slight increases with tetradecanoic acid (C-14) and stearic acid (C-18).

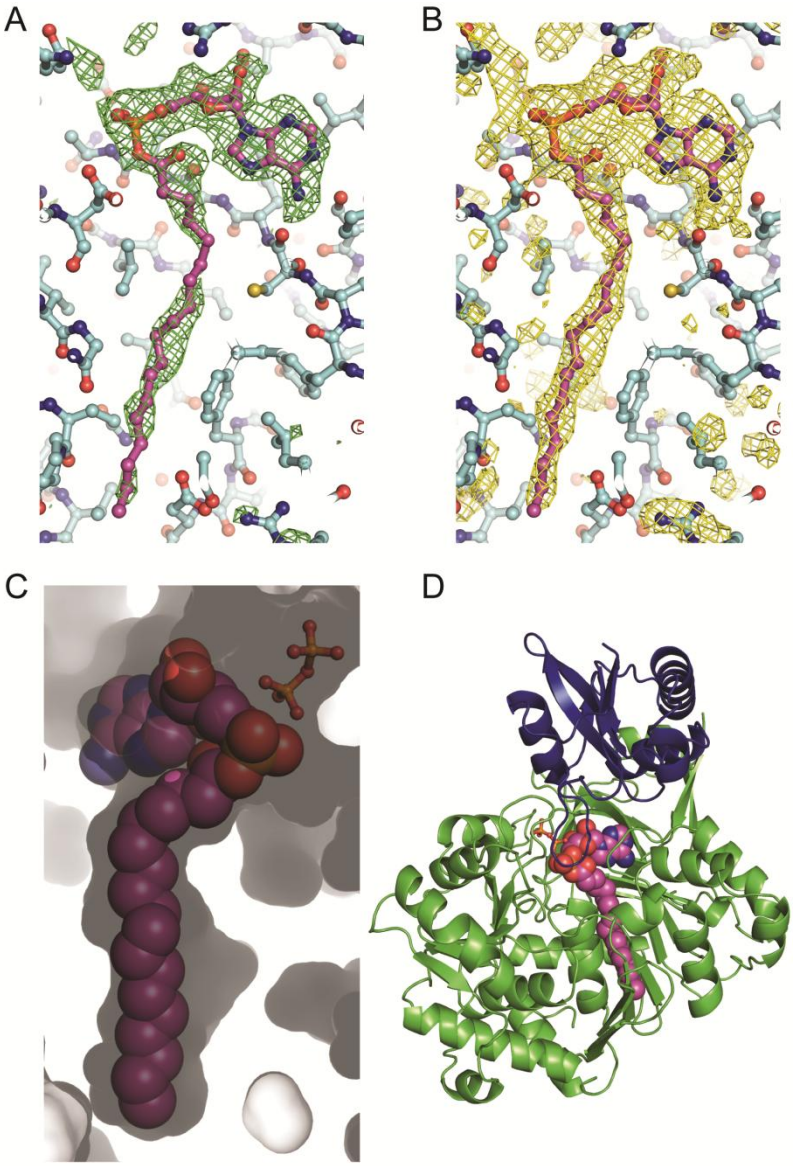

Figure 2.5 Binding of fatty acyl chains in the SynFAAL substrate tunnel.

Electron density for stearoyl adenylate in the substrate tunnel. Fo-Fc simulated-annealing omit density at 3σ (A) (green mesh) and 2σ (B) (yellow mesh). Atoms are shown in ball-and-stick with cyan C for the protein and magenta C for stearoyl adenylate. C) Length and shape of the substrate binding tunnel. The tunnel surface is rendered in gray, stearoyl adenylate in spheres and PP_i in ball-and-stick. D) Position of stearoyl adenylate in the overall SynFAAL structure. SynFAAL is rendered as in Fig. 2 with ligands shown as in C.

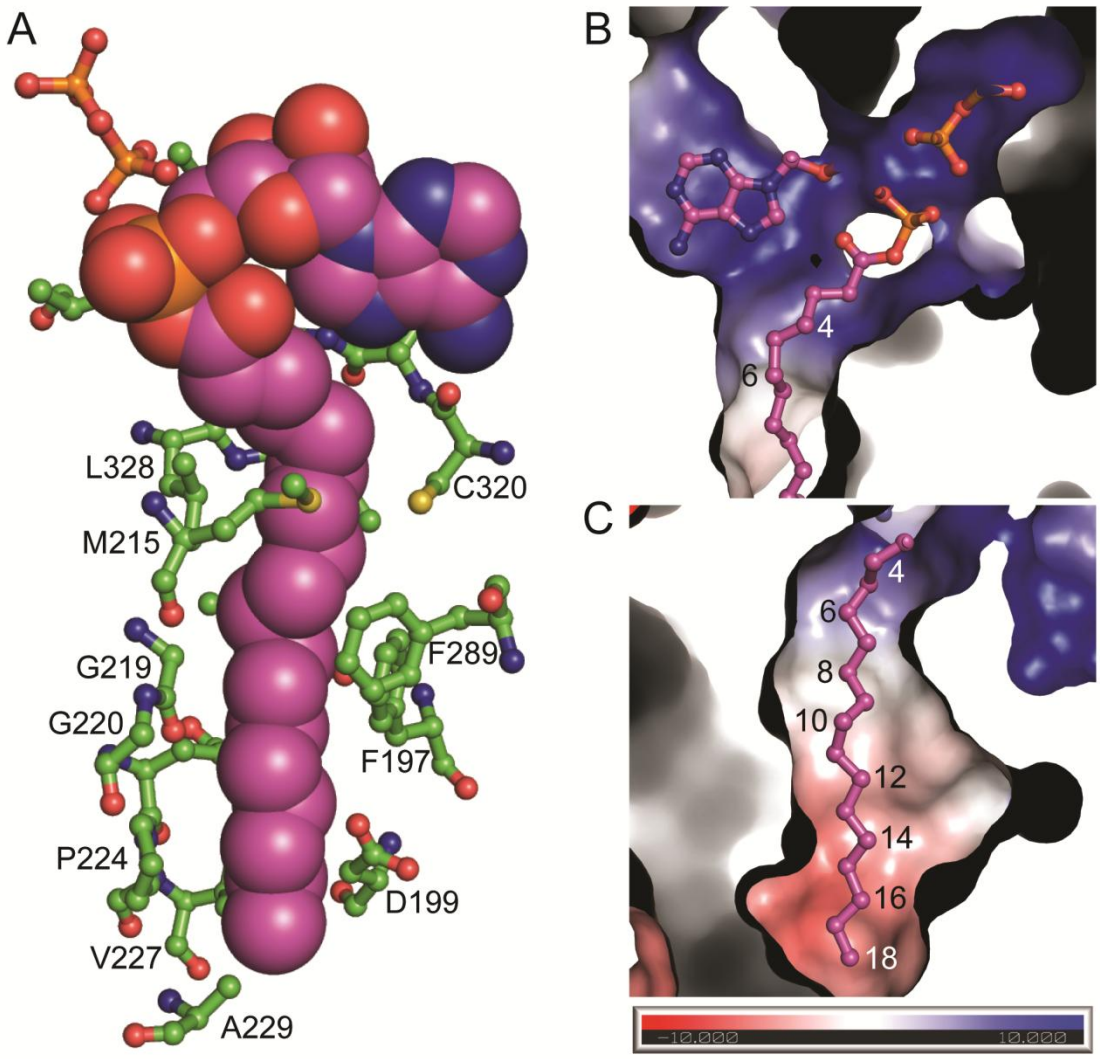

Figure 2.6 Surface features of the SynFAAL substrate tunnel.

A) Stearoyl adenylate (spheres) and surrounding amino acids (ball-and-stick with green C). B) and C) Electrostatic surface potential of the substrate binding tunnel, colored from electronegative to electropositive (red to blue, -10 to +10 kT). The top of the tunnel is highly electropositive for nucleotide binding (B), whereas the middle is largely hydrophobic for fatty acid binding (C). The bottom of the tunnel is electronegative.

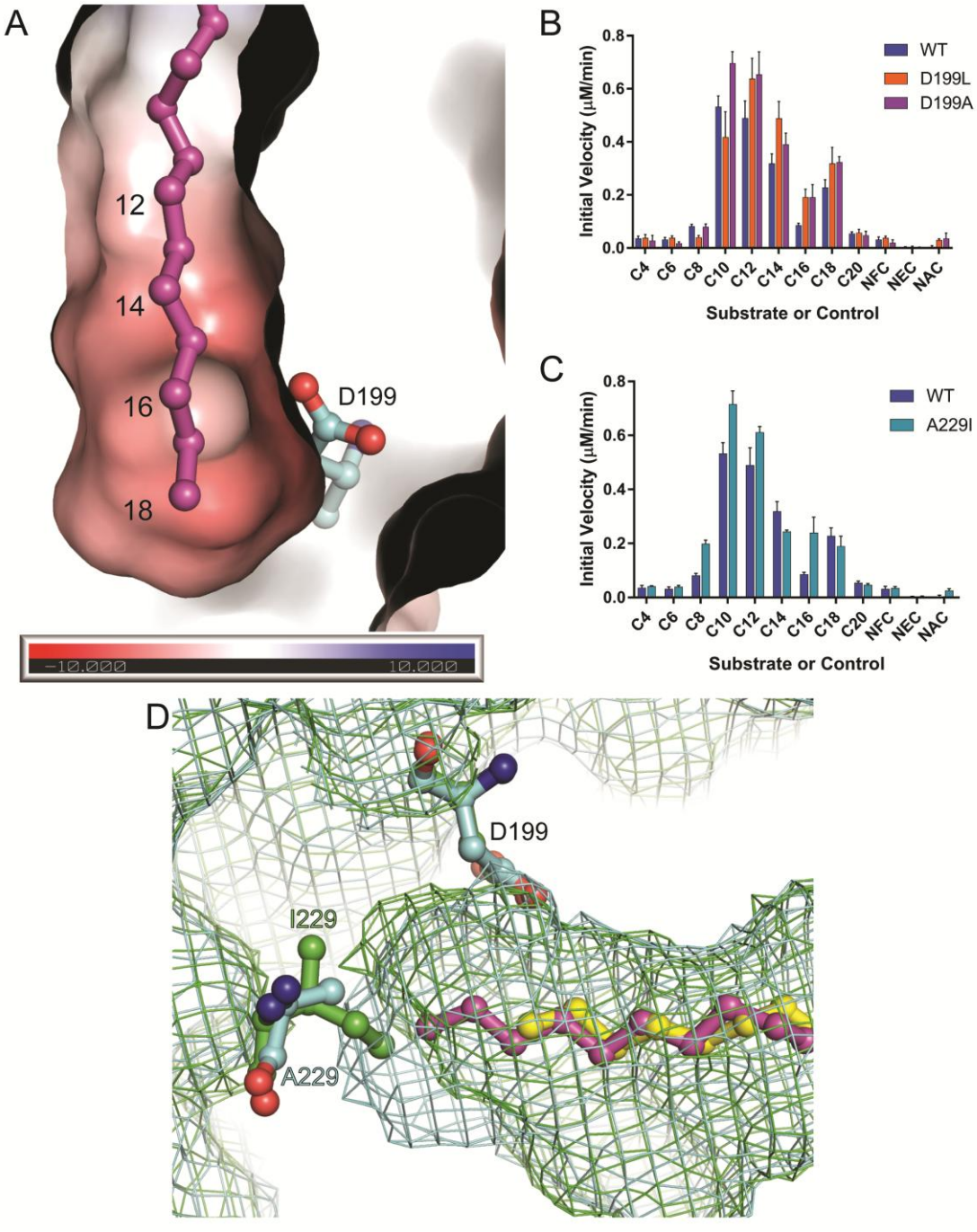
A
12
14
16
18
D199
−10.000
10.000
B
Initial Velocity (μM/min)
0.8
0.6
0.4
0.2
0.0
WT
D199L
D199A
C4 C6 C8 C10 C12 C14 C16 C18 C20 NFC NEC NAC
Substrate or Control
C
Initial Velocity (μM/min)
0.8
0.6
0.4
0.2
0.0
WT
A229I
C4 C6 C8 C10 C12 C14 C16 C18 C20 NFC NEC NAC
Substrate or Control
D
D199
I229
A229

Figure 2.7 Electrostatic lock for fatty acid selection.

A) Electrostatic surface potential of the bottom of the substrate binding tunnel. Asp199 accounts for the negative surface potential in this region. The tunnel surface, bound stearoyl adenylate and Asp199 are rendered as in Fig. 6. B) and C) Activity of SynFAAL variants, D199L and D199A (B) and A229I (C) on a panel of fatty acids of chain length between 4 (C4) and 20 (C20) carbons. Negative controls are as in Fig. 2. Error bars indicate one standard deviation of at least two replicates in triplicate. D) Overlay of wild type SynFAAL stearoyl adenylate complex and the SynFAAL/R336A/A22I palmitoyl adenylate complex (RMSD 0.3 Å). WT is shown in cyan and A229I variant is shown in green with amino acids shown in ball-and-stick and protein surfaces as meshes. Stearoyl adenylate (magenta) and palmitoyl adenylate (yellow) shown in ball-and-stick. The protein surfaces show the shift of the bottom of the tunnel in the upward direction with A229I.

To further explore how the charge of the binding tunnel affects substrate selection we shortened the substrate tunnel by substituting Ile for Ala at position 229 near the bottom of the tunnel. SynFAAL/A229I (Figure 2.6A) had an almost 3-fold increase in activity with palmitic acid relative to the wild type while not significantly altering activity with other fatty acids (Figure 2.7C). Additionally, in the crystal structure of the SynFAAL/R336A/A229I complex with palmitoyl adenylate the bottom of the binding pocket expands slightly outward to accommodate the larger isoleucine which prevents the binding tunnel from decreasing significantly in size (Figure 2.7D). The isoleucine substitution increased the overall hydrophobicity of the bottom of the tunnel. The bound palmitoyl adenylate adopts a very similar conformation to stearoyl adenylate in the wild type enzyme.

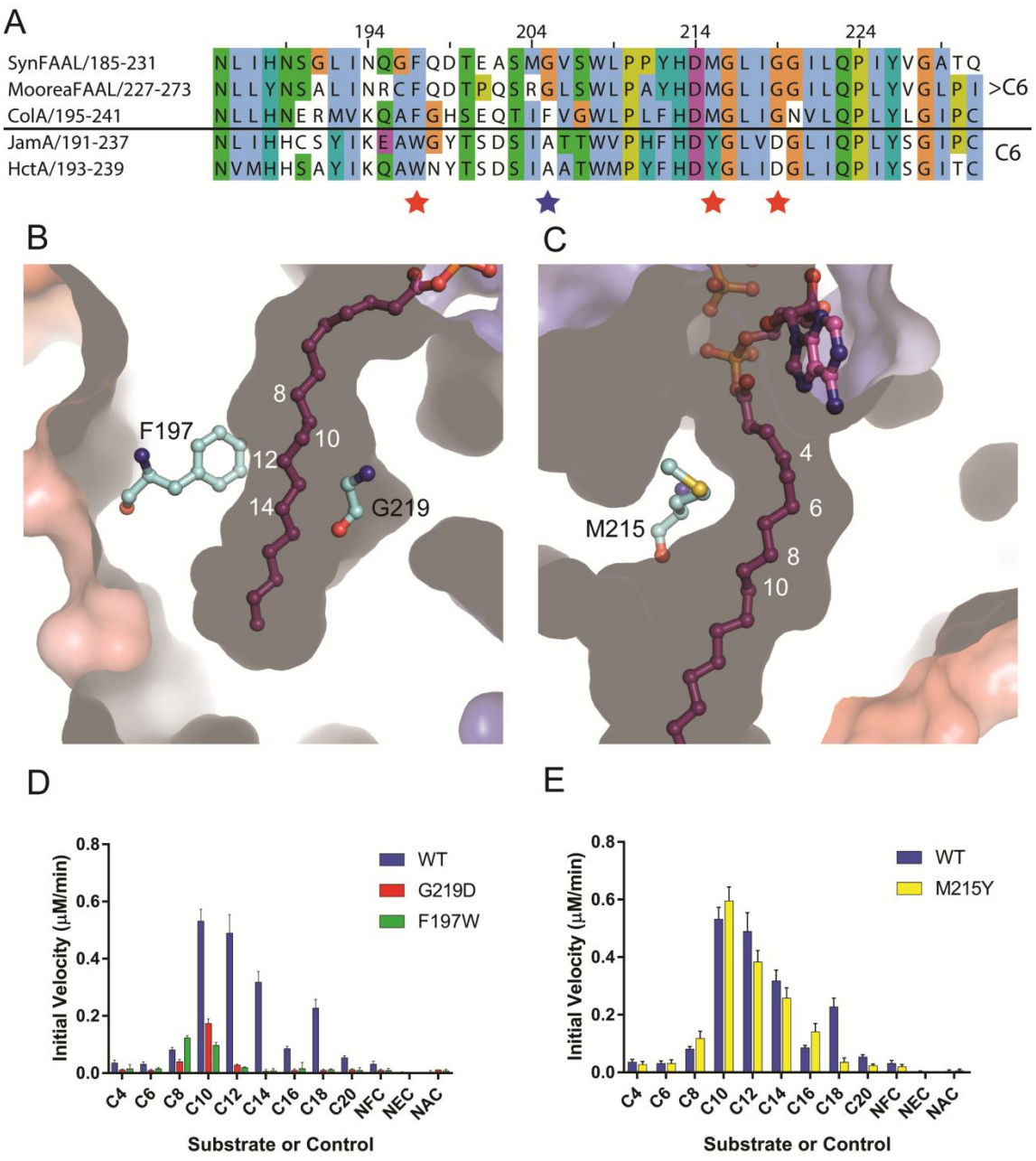
A
194
204
214
224
SynFAAL/185-231
MooreaFAAL/227-273
ColA/195-241
JamA/191-237
HctA/193-239
>C6
C6
B
C
F197
G219
M215
8
10
12
14
4
6
D
E
Initial Velocity (μM/min)
WT
G219D
F197W
M215Y
C4
C6
C8
C10
C12
C14
C16
C18
C20
NFC
NEC
NAC
Substrate or Control

Figure 2.8 Amino acids affecting fatty acid selectivity.

A) Sequence alignment of six cyanobacterial FAALs in the region that lines the substrate tunnel. Red stars indicate positions that are differentially conserved among sequences of FAALs with natural substrates longer than six carbons (above the horizontal line), and those that act on hexanoic acid (below the line). Residue 205 (blue star) appears to be a selectivity filter for substrate length in ColA (C12, Phe205) compared to SynFAAL and MooreaFAAL (C18, Gly205). B) and C) Positions of differentially conserved amino acids relative to the fatty acid binding tunnel. The tunnel surface is rendered as in Fig. 5, and highlighted amino acids and stearoyl adenylate as in Fig. 6. D) and E) Activity of SynFAAL variants, F197W and G219D (D) and M215Y (E) relative to the wild type on a panel of fatty acids with chain length between 4 (C4) and 20 (C20) carbons. Negative controls are as in Fig. 2. Error bars indicate one standard deviation for at least 2 replicates in triplicate. Accession numbers for sequences: SynFAAL: CP000951.1, MooreaFAAL: KF550302.1, ColA: KP715425.1, JamA: AY522504.1, HctA: AY974560.1.

2.4.6 Fatty Acid Selectivity in the Binding Tunnel

To probe SynFAAL chain-length selectivity, we made a number of substitutions to residues lining the fatty acid binding tunnel. These efforts were guided by a sequence alignment of cyanobacterial FAALs with known substrates of differing length: C6 (JamA and HctA), and C12 or longer (SynFAAL, MooreaFAAL and ColA) (Figure 2.8A). We identified three differentially conserved positions in the two groups: Phe197, Met215, and Gly219 (SynFAAL numbering) for the long-substrate FAALs and Trp197, Tyr215 and Asp219 for the C6-selective FAALs (Figure 2.8B and C). We also identified one position, Gly205, that may discriminate between C18 (SynFAAL) and C16 (MooreaFAAL) (29) substrates and C12 (ColA FAAL) (24) substrates. Each of these positions in SynFAAL was mutagenized to its counterpart in the shorter-chain FAALs, along with a F197W/M215Y double substitution, and their activities with the panel of fatty acids were compared to the wild type.

Phe197 and Gly219 are located halfway along the binding tunnel near the stearoyl C12 position. Relative to the wild type, SynFAAL/F197W and SynFAAL/G219D had significantly decreased activity for nearly all substrates and no detectable activity with fatty acids longer than 12 carbons (Figure 2.8D). Despite the lower activity, SynFAAL/G219D had a strong preference for decanoic acid whereas SynFAAL/F197W had similar activity with octanoic and decanoic acids with levels approximately 50% that of SynFAAL with stearic acid. Met215 is located at a bend near the top of the substrate tunnel between the C3 and C7 positions (Figure 2.8C). Surprisingly, given its location, the M215Y variant had little impact on activity with fatty acids shorter then palmitic acid (C16). However, activity with palmitic acid was increased relative to wild type, and activity with stearic acid was nearly abolished (Figure 2.8E). SynFAAL/G205F and the doubly substituted SynFAAL/F197W/M215Y had no detectable activity with any of the tested substrates (Table 2.2).

2.5 Discussion

Microbial production of α-olefins for biofuels and industrial applications (surfactants, lubricants, feedstocks) shows promise. For olefin synthases to be a suitable means of production, the length of the produced olefin must be controllable. Understanding the FAAL mechanism of fatty acid chain length selection would enable the engineering of olefin synthases to produce specific α-olefins. Towards this end I determined AMP and acyl adenylates complex

structures of synFAAL by X-ray crystallography. I also performed biochemical assays investigating the chain-length selectivity for fatty acid substrates. SynFAAL represents the first FAAL domain from an olefin synthase to be purified and characterized and the first FAAL from cyanobacteria to be crystallized. The structures guided the creation of a number of SynFAAL variants that led to the characterization of an electrostatic fatty acid selection mechanism and an understanding of some aspects of the fatty acid length selectivity.

Determining the structure of SynFAAL was challenging. An initial crystal condition was difficult to obtain and took significant optimization. After optimization crystals still had high mosaicity. Dealing with these issues required extensive screening of crystals and the use of small beam sizes combined with rastering to find areas of the crystals with suitable diffraction. During attempts to obtain a SynFAAL structure in complex with stearyl adenylate it was discovered that the addition of fatty acids negatively impacted crystal formation and diffraction quality crystals could not be obtained. Attempts to soak fatty acids into already formed crystal were unsuccessful as well. Therefore, crystal contact engineering was used to generate more robust crystals that could tolerate the addition of fatty acids in co-crystallization. The crystal contact containing Arg336 was identified as the best target for engineering and replacement of the arginine with alanine resulted in larger and more robust crystals (Figure 2.4). Improved crystals were more tolerant of fatty acids and resulted in the strearyl adenylate structure.

The bimodal substrate utilization profile of SynFAAL is unprecedented among the studied FAAL enzymes (Figure 2.2B) (26,63). In previous cases the FAAL enzyme displayed high selectivity for one fatty acid chain-length with a rapid drop off in activity with changes in length. A clearer picture of the biological reason for the bimodal distribution can be gained by looking at the substrates available *in vivo*. Previous work has shown that *Synechococcus sp.* PCC 7002 contains almost 20-fold more palmitic acid than stearic acid, with approximately three-fold less tetradecanoic acid than stearic acid. No shorter fatty acids were reported (29). It has also been shown on multiple occasions that only 19 carbon long olefins are produced (3,29,64), indicating the olefin synthase strongly selects against the more abundant palmitic acid. Fatty acids shorter than palmitic acid are either not present or in low abundance so there is little pressure to select against them. This is consistent with the results presented here showing that Asp199 creates a negatively charged patch near the C16 position of fatty acids to select against palmitic acid (Figure 2.7A). The presence of the charged patch may also explain previous work showing that feeding pentadecanoic acid to *Synechococcus sp.* PCC 7002 did not result in 1-hexadecene production (3).

The SynFAAL/A229I variant was originally created with the expectation that it would fill the bottom of the binding tunnel and prevent stearic acid from binding, but not affect the usage of other fatty acids. Because of the slight movement of loops making up the bottom of the binding tunnel, the added bulk of the

isoleucine did not significantly decrease the size of the tunnel (Figure 2.7D). The mutation did increase the hydrophobicity of the tunnel which resulted in increased in palmitic acid usage, further supporting the use of an electrostatic selection mechanism. The competing factors of increased hydrophobicity and tunnel constriction apparently offset resulting in stearic acid activity similar to WT.

Phe197 and Gly219 are both located close to the C12 position of fatty acids and it was expected that replacing either with bulkier side chains would block the usage of fatty acids longer than 12 carbons. Both SynFAAL/F197W and SynFAAL/G219D variants behaved as expected (Figure 2.8). The SynFAAL/F197W variant also caused a slight increase in octanoic acid activity. The ideal length of olefins for biofuels applications is 11-15 carbons long; this length provides a suitable balance of energy density and melting temperature (31). The SynFAAL/G219D variant is highly selective for decanoic acid, which would produce 1-undecene. This demonstrates the feasibility of engineering the SynFAAL enzyme for the production of specific α-olefins. With additional experiments it may be possible to produce SynFAAL variants with high specificity for a variety of fatty acid chain-lengths. However, for the production of short α-olefins it could be more efficient to exchange the entire FAAL domain for one that natively works on shorter fatty acids, such as JamA (25).

Met215 is near the C6 position of fatty acids and its replacement with Tyr was expected to occlude the binding tunnel. However, experiments showed that only

the utilization of long chain fatty acids was affected (Figure 2.8E). My current hypothesis is that the tyrosine does not position itself in the binding tunnel, but rotates 180° about the Cα-Cβ bond to occupy space to the left of the methionine position seen in Figure 2.8C. Consequently, the fatty acid chain would not need to adopt the curved confirmation around Met215 and would take a more direct path to the bottom of the binding tunnel. The more direct path would cause stearic acid to clash with the bottom of the tunnel, leading to the sharp decrease in activity. Attempts to crystallized a SynFAAL/R336A/M215Y double variant to test this hypothesis did not produce crystals of sufficient diffraction quality. The SynFAAL/M215Y/F197W double variant had no detectable activity, and we speculate that the additional Trp caused the tyrosine to adopt a tunnel occluding conformation.

It was found during *in vivo* characterization that the *Synechococcus sp.* PCC 7002 olefin synthase is involved in the production of a second olefin product, 1,14-nonadecadiene (3). Subsequent studies indicate that the olefin synthase is likely utilizing the desaturated fatty acid C18:1 (Δ13) as the starting fatty acid (30). The triangular bulge near the C-12 position of the stearoyl chain appear to be able to accommodate a cis-double bonded fatty acid chain (Figure 2.5C).

Moving forward it will be important to examine the variants which modulate substrate selectivity in the context of the entire olefin synthase pathway. It has been shown that the sulfotransferase is functional on shorter substrates (32), but

it is unknown how well shorter substrate will be tolerated by the remainder of the olefin synthase. It is tempting to speculate that the remaining olefin synthase enzymes do not have selection mechanisms for substrate length and engineering would be straightforward; however, future efforts will need to be made to examine this possibility. Currently the production of α-olefins for biofuels and industrial applications by biological means shows potential.

2.6 Methods and Materials

2.6.1 Cloning and Mutagenesis

Synthetic DNAs encoding the SynFAAL (amino acids 2-570) and SynACP1 (amino acids 600-677) regions of *Synechococcus sp.* PCC 7002 olefin synthase were purchased from IDT. Expression plasmids pMCSG9-*SynFAAL* and pMCSG7-*SynACP1* were created by insertion into the pMCSG9 and pMCSG7 vectors, respectively, and verified by sequencing. *SynFAAL* variants were generated using the QuikChange Lighting kit (Agilent) according to the vendor's directions. Primers were designed using the Agilent QuikChange Primer Design tool.

2.6.2 Bacterial strains, media and culture conditions

Cultures were grown in Terrific Broth (TB) medium with 4% glycerol and ampicillin (100 µg/mL). *E. coli* strain BAP1 (65) was transformed with pMCSG7-*synACP*, and strain BL21(DE3) was transformed with pMCSG9-*synFAAL*. Cultures were grown at 37°C in 500 mL TB to an OD_{600} of 1.5, cooled to 20°C for

at least 1 hr, induced with IPTG (final concentration 0.2 mM), and grown for an additional 14-18 hr before harvesting by centrifugation. Cell pellets were stored at -20°C until used for protein purification.

2.6.3 Purification of wild type SynFAAL and SynFAAL Variants

Cell pellets were resuspended in approximately 5 mL Buffer A (100 mM Tris pH 7.5, 300 mM NaCl, 5% glycerol) containing 15 mM imidazole per 1 g cell paste. All purification steps were performed at 4°C. Resuspended cells were incubated with DNase (4 mg), lysozyme (10 mg), and $MgCl_2$ (4 mM) for 30 min, then lysed by three passes though an Avestin EmulsiFlex-C3 homogenizer and clarified by centrifugation at 30,000 x g. The soluble fraction was loaded onto a 5-mL HisTrap Ni NTA column (GE Healthcare) in Buffer A and eluted with an imidazole gradient from 30-300 mM. The eluted protein was treated overnight at 4°C with tobacco etch virus (TEV) protease (1:30 ratio TEV:SynFAAL plus 2 mM DTT) to remove the purification (polyhistidine) and solubility (maltose binding protein) tags, dialyzed into Buffer B (50 mM Tris pH 7.5, 150 mM NaCl, 2.5% glycerol), and applied to a 5-mL HisTrap Ni NTA column. The flow-through fraction was further purified by size exclusion chromatography with a HiLoad 16/600 Superdex 200 column (GE Healthcare) in Buffer B. The FAAL eluted as a monomer of apparent molecular weight 51 kDa, and yielded 100 mg of purified protein from 500 mL *E. coli* culture. The protein was concentrated to 20 mg/mL, flash cooled with liquid N_2, and stored at -80°C.

2.6.4 Purification of SynACP1

SynACP1 was purified identically to SynFAAL except the buffer for purification by metal affinity chromatography was 100 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 5 mM TCEP and the polyhistidine tag was not cleaved. In order to ensure full phosphopantetheinylation, SynACP1 was incubated for 16 hr at 30°C with the nonspecific phosphopantetheinyl transferase Sfp (66) at a 1:40 ratio in a reaction mix also containing 16 mM ATP, 4 mM pantethine, 4 mM DDT, and 0.2 mg/mL each of CoaA, CoaD, and CoaE (67). Holo-SynACP1 was purified from the reaction mix using a HiLoad 16/600 Superdex 75 column (GE Healthcare) equilibrated with 100 mM Tris pH 7.5, 250 mM NaCl, 5% glycerol, 5 mM TCEP. Fractions were pooled, concentrated, flash cooled with liquid N_2, and stored at -80°C.

2.6.5 Crystallization

Prior to crystallization, 5 mg/mL wild type SynFAAL was treated with 10 mM ATP and 0.4 mM $MgCl_2$; SynFAAL/R336A and SynFAAL/R336A/A229I with 0.4 mM each of ATP, $MgCl_2$, and fatty acid (dissolved in ethanol). Protein solutions were clarified by centrifugation at 20,000 x g for 30 min at 4°C. Crystals grew in 3 days to two weeks at 20°C by vapor diffusion in 1.8 µL sitting drops containing 0.6 µL protein stock and 1.2 µL reservoir solution (0.17-0.21 M NDSB-256, 26-33% PEG 3350, 0.19-0.21 M ammonium acetate, 0.1M sodium acetate pH 5.5). Crystals were harvested in nylon loops and flash cooled in liquid N_2.

2.6.6 Diffraction Data Collection and Structure Determination

Diffraction data were collected at 100 K on GM/CA beamline 23ID-D or 23ID-B at the Advanced Photon Source (APS) at Argonne National Laboratory (Argonne, IL). For each crystal, 360º of data were collected using 0.2º rotation and 0.2 sec exposure per image. Data were processed using XDS (68) (Table 2.1). The structure of SynFAAL with bound AMP was solved by molecular replacement (MR) with Phaser (69) in the PHENIX software suite (70) from the structure of the *L. pneumophila* fatty acyl-adenylate ligase (20) (43% identity, PDB code: 3LNV) with the C-terminal domain removed. The initial MR model was expanded using AutoBuild (71) to generate a 77% complete model of SynFAAL, and the remaining portions of the structure were completed manually using Coot (72). Refinement was performed using phenix.refine (73). The structure of SynFAAL/R336A/A229I was solved by difference Fourier methods as the crystals were isomorphous. SynFAAL/R336A with stearoyl adenylate bound was not isomorphous and was solved by molecular replacement using the SynFAAL-AMP complex as a search model. Restraints for fatty acyl adenylates were generated with eLBOW (74). The structure (Table 1) is complete except for residues 359-362. Buried surface area was analyzed with PISA (75). Sequence alignments were created with Jalview (76,77) using Clustal O (78). Electrostatic surface potentials were calculated with the PDB2PQR server (79,80). Structures were validated with MolProbity (81). Images and figures were prepared using PyMOL (82).

2.6.7 Malachite Green Assay

Catalytic activity of wild type and variant SynFAAL was determined using an end-point colorimetric assay coupled to pyrophosphatase (61). All assays were performed in 48-well PCR plates. A master mix consisting of reaction buffer (50 mM Tris pH 7.8, 100 mM NaCl, 10 mM $MgCl_2$), 200 µM ATP, 0.1 U/mL inorganic pyrophosphatase (New England Biolab), 20 µM SynACP1, and 1 µM SynFAAL was prepared on ice and kept as cool as possible during subsequent steps. 39.2 µL of master mix was added to wells. 0.8 µL fatty acid dissolved in ethanol was then added to the reaction (final concentrations 20 µM fatty acid, 2% ethanol). Plates were placed in a thermocycler and incubated at 30°C for 30 min and then cooled to 4°C. 33 µL reaction mixture was then transferred to a 96 well reaction plate containing 66 µL reaction buffer. Reactions were developed by the addition of 20 µL molybdate/malachite green reagent. The released P_i concentration was determined at 620 nm on a Molecular Devices SpectraMax M5 384 variable wavelength spectrometer. End-point data was used to calculated the initial rate of reaction. Reactions were performed in triplicate and repeated at least twice. At least one replicate was performed at a different concentration with data normalized to 1 µM SynFAAL. For variants with reaction rates similar to WT, replicates were performed at a concentration of 0.5 µM SynFAAL and run for 30 min. For variants with slow reaction rates, replicates were performed at a concentration of 2 µM SynFAAL and run for 60 min. Reactions without SynFAAL, SynACP1, or fatty acid added were used as negative controls. Absorbance measurements were converted to P_i concentrations using a P_i standard curve

from 0-25 μM. The 30 min end-point was chosen based on previous experiments establishing that the reaction was linear under those conditions (Figure 2.9). Briefly, WT SynFAAL was incubated with stearic acid under the same reactions conditions used for end-point assays. Time points were taken at 0, 5, 10, 15, 20, 25, 30, 35, 40, 50, and 60 min. For each time point, 33 μL reaction mixture was quenched with 66 μL reaction buffer plus 10 μL formic acid. Once all time points were collected, reactions were developed and quantified the same as end-point assays.

2.6.8 Detection of FAAL products by HPLC/MS

30 μL reaction mixtures consisted of 50 mM Tris pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 5 mM TCEP, 10 mM ATP, 10 μM SynFAAL, and 50 μM holo-SynACP1. Reactions were initiated by the addition of 500 μM fatty acid substrate in ethanol (final concentration 0.5% ethanol) and incubated at 20°C for 18 hr. A reaction mixture without SynFAAL was used as a negative control. 15 μL of reaction mixture was injected onto an Agilent 1290 HPLC system after incubation. Samples were analyzed by RP-HPLC (Jupiter 5μ C4 column 300Å, 250 X 2.00 mm) with a 20 min gradient of 30-90% acetonitrile with 0.2% trifluoroacetic acid at a flow rate of 0.2 mL/min, peak fractions were collected for LC/MS analysis. Peak fractions analyzed by intact protein and phosphopantetheine ejection (83) on an Agilent 6520 Q-TOF mass spectrometer equipped with an Agilent 1290 HPLC system. Samples were separated for MS analysis by RP-HPLC (Aeris widepore C4 column 3.6 μm 50 x 2.10 mm) with a 10 min gradient of 10-95%

acetonitrile with 0.1% formic acid at a flow rate of 0.5 mL/min. Intact protein and phosphopantetheine data were analyzed using the Agilent Mass Hunter Qualitative Analysis software with the maximum entropy deconvolution algorithm.

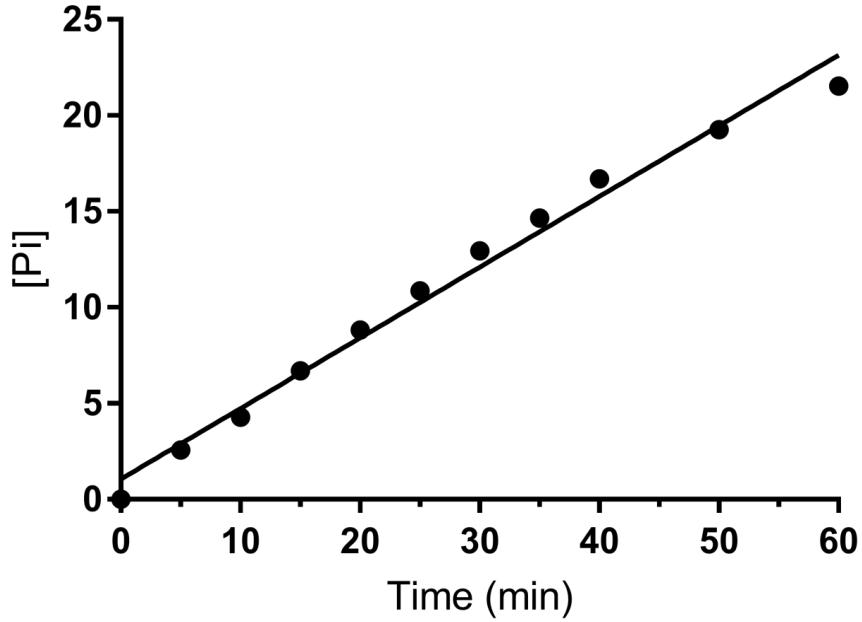

Figure 2.9 Linearity of SynFAAL reactions.

WT SynFAAL was incubated with SynACP and stearic acid at the same concentrations used for end-point malachite green assays. 11 time points taken over 1 hr and quantified using the malachite green reagent. Time course was fit using linear regression (R^2 = 0.986). The linearity indicated that the reaction is well within the linear range at 30 min.

Table 2.2 End-point initial velocities (μM/min) from malachite green assay.

	Wild Type	**F197W**	**D199A**	**D199L**	**G205F**
C4	0.036 ± 0.009	0.02 ± 0.01	0.03 ± 0.02	0.04 ± 0.01	0.01 ± 0.01
C6	0.031 ± 0.009	0.015 ± 0.004	0.017 ± 0.007	0.038 ± 0.008	0.005 ± 0.004
C8	0.081 ± 0.009	0.123 ± 0.008	0.08 ± 0.01	0.040 ± 0.007	0.004 ± 0.002
C10	0.53 ± 0.04	0.097 ± 0.010	0.70 ± 0.04	0.40 ± 0.10	0.004 ± 0.004
C12	0.49 ± 0.07	0.019 ± 0.003	0.65 ± 0.09	0.64 ± 0.08	0.00 ± 0.003
C14	0.32 ± 0.04	0.008 ± 0.008	0.39 ± 0.04	0.49 ± 0.06	0.00 ± 0.005
C16	0.085 ± 0.009	0.02 ± 0.02	0.19 ± 0.05	0.19 ± 0.03	0.00 ± 0.005
C18	0.22 ± 0.03	0.012 ± 0.003	0.32 ± 0.02	0.31 ± 0.06	0.00 ± 0.006
C20	0.054 ± 0.007	0.01 ± 0.01	0.05 ± 0.02	0.06 ± 0.01	0.00 ± 0.006
NFC	0.03 ± 0.01	0.009 ± 0.008	0.02 ± 0.01	0.037 ± 0.008	0.004 ± 0.004
NEC	0.00 ± 0.005	0.00 ± 0.0007	0.000 ± 0.004	0.000 ± 0.007	0.00 ± 0.005
NAC	0.004 ± 0.005	0.008 ± 0.006	0.03 ± 0.02	0.029 ± 0.005	0.00 ± 0.008

	M215Y	**G219D**	**A229I**	**R336A**	**F197W+M215Y**
C4	0.03 ± 0.01	0.011 ± 0.003	0.041 ± 0.003	0.024 ± 0.003	0.004 ± 0.002
C6	0.03 ± 0.01	0.009 ± 0.004	0.038 ± 0.006	0.021 ± 0.003	0.005 ± 0.002
C8	0.12 ± 0.02	0.040 ± 0.008	0.20 ± 0.01	0.10 ± 0.03	0.005 ± 0.001
C10	0.59 ± 0.05	0.17 ± 0.02	0.71 ± 0.05	0.70 ± 0.05	0.005 ± 0.003
C12	0.38 ± 0.04	0.028 ± 0.004	0.61 ± 0.02	0.69 ± 0.08	0.004 ± 0.002
C14	0.26 ± 0.04	0.006 ± 0.007	0.243 ± 0.007	0.34 ± 0.06	0.001 ± 0.003
C16	0.14 ± 0.03	0.009 ± 0.006	0.24 ± 0.06	0.09 ± 0.01	0.004 ± 0.002
C18	0.03 ± 0.02	0.010 ± 0.004	0.19 ± 0.04	0.28 ± 0.08	0.002 ± 0.002
C20	0.023 ± 0.006	0.012 ± 0.003	0.046 ± 0.006	0.05 ± 0.01	0.0037 ± 0.0007
NFC	0.020 ± 0.008	0.010 ± 0.004	0.034 ± 0.007	0.014 ± 0.005	0.004 ± 0.003
NEC	0.00 ± 0.002	0.00 ± 0.0005	0.00 ± 0.006	0.00 ± 0.003	0.00 ± 0.003
NAC	0.005 ± 0.006	0.01 ± 0.001	0.025 ± 0.008	0.019 ± 0.006	0.00 ± 0.005

CHAPTER III

Biochemical and Structural Basis for the Functional Divergence of the AMP-ligase CahJ in the Cahuitamycin Biosynthetic Pathway

3.1 Notes

This chapter is being prepared for publication in a manuscript titled "Biochemical and Structural Basis for the Functional Divergence of AMP-ligase CahJ in the Cahuitamycin Biosynthetic Pathway" with the authors: Tripathi A*, Park SR*, Sikkema AP*, Wu J, Cho H, Xi C, Smith JL, Sherman DH.
*These authors contributed equally to this work.

Author Contributions:
A.P.S. solved crystal structures of CahJ and performed ArCP loading experiments. S.R.P. cloned CahJ and performed malachite green assays. A.T. performed mutasynthesis experiments and determined the structure of cahuitamycin F. J.W. performed biofilm assays. A.P.S. and A.T. wrote the manuscript. A.P.S., A.T., J.L.S., and D.H.S. edited the manuscript.

3.2 Abstract

Cahuitamycins are potent biofilm inhibitors biosynthesized by a unique bifurcated nonribosomal peptide synthetase (NRPS) pathway in *Streptomyces gandocaensis*. Previous *in silico* analysis and chemical complementation studies suggested that an adenylation (A) domain, CahJ, acts as a gatekeeper enzyme for cahuitamycin structural diversification. On the basis of this observation, the innate substrate flexibility of CahJ was analyzed using an adenylation assay and structural characterization of CahJ in complexes containing both native and non-native substrates. The analysis also led to isolation of a new cahuitamycin

congener through targeted precursor incorporation resulting in production of a potent *Acinetobacter baumannii* biofilm inhibitor. The findings represent an effective synthetic biology approach to expand the chemical diversity of novel anti-biofilm molecules.

3.3 Introduction

Nature has traditionally served as a prolific source of drug leads for infectious diseases and other illnesses (1). Some of the most therapeutically important and best-understood natural products are nonribosomal peptides (NRPs). These molecules are produced by large multienzyme complexes called nonribosomal peptide synthetases (NRPS) (14,84). NRPSs are typically composed of repeating enzyme domains that activate and incorporate specific building blocks into a growing peptide chain. In general, each module of a traditional NRPS contains an adenylation (A), a peptidyl carrier protein (PCP, also known as a thiolation or "T" domain) and a condensation (C) domain. NRPSs assemble their peptide products through the multienzyme thiotemplate mechanism (14,84,85), in which an A domain selects and activates an amino acid substrate via ATP-dependent hydrolysis as an aminoacyl-adenylate. The reactive intermediate is then transferred to the free thiol group of the 4'-phosphopantetheinyl group on an adjacent PCP domain, releasing AMP. In peptide elongation, the upstream peptidyl-*S*-PCP and the free amino group of the downstream aminoacyl-*S*-PCP establish a peptide bond, catalyzed by the C-domain. Variations in the architecture of NRPSs can be introduced via any step involving initiation,

elongation and termination. The structural diversity of NRPs can also be expanded by additional, optional domains that tailor individual building blocks, for example an epimerization (E) domain. Several studies over the past two decades have shown that NRPS monomer choices are not limited to the 20 proteinogenic amino acids. Many different monomers, including nonproteinogenic amino acids, fatty acids, aryl acids, and hydroxy acids, have been involved in NRP biosynthesis (86). These nonproteinogenic building blocks offer potential for structural diversification of NRPs and are likely responsible for their biological activity (87).

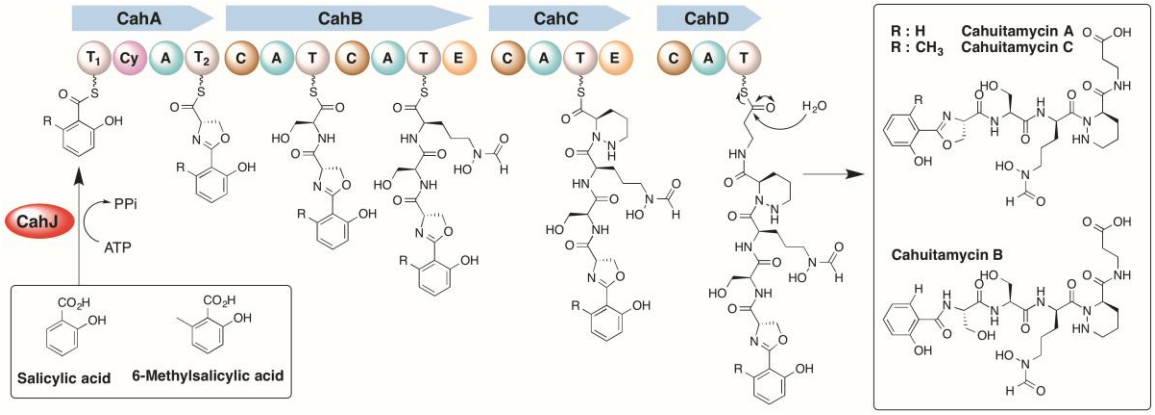

Figure 3.1 Proposed biosynthesis of cahuitamycins A-C in *S. gandocaensis*.

The CahJ-dependent adenylation of salicylic acid and 6-methylsalicylic acid in cahuitamycin assembly is an ATP-dependent process that leads to the release of pyrophosphate (PP_i). C, condensation domain; Cy, cyclization domain; E, epimerization domain T, thiolation domain.

Cahuitamycins, produced by *Streptomyces gandocaensis*, are a novel structural class that has potent anti-biofilm activity against the Gram-negative pathogenic bacterium, *Acinetobacter baumannii* (2). This multidrug-resistant microorganism is responsible for a large number of nosocomial infections, including pneumonia,

urinary tract infections, wound infections, and bacteremia with significantly high mortality rates (~60%). Biofilm formation contributes to the pathogen's high rates of antibiotic resistance (88). When in a biofilm, bacteria develop antibiotic resistance up to 1,000 fold greater than planktonic forms. Despite the significant role of biofilms in infectious diseases, there are currently no drugs that specifically target biofilms in clinical use or in clinical trials (1). The generation of structurally diverse cahuitamycin analogs could therefore facilitate efforts to develop promising anti-biofilm drug candidates for future pharmaceutical applications. Understanding cahuitamycin biosynthesis represents a critical first step toward the generation of more potent and selective drug analogs. Based on previous studies (2), the discrete A domain CahJ could represent an ideal avenue to exploit biosynthesis and create new cahuitamycin congeners. Cahuitamycin structural variants originate at the loading module, CahJ, which is a discrete aryl-AMP ligase, in a process that is analogous to yersiniabactin (89) and pyochelin (90) biosynthesis. CahJ is responsible for the selection of an appropriate aryl acid substrate and acts as the initial gatekeeper in cahuitamycin NRPS assembly (Figure 3.1). A deletion mutant of *cahI* (Δ*cahI*), which encodes a salicylate synthase, exclusively produced cahuitamycin C while the production of cahuitamycins A and B was completely abolished (2). This result confirmed the substrate selectivity of CahJ, which is able to utilize two different substrates, salicylic acid (SA) and 6-methlysalicylic acid (6-MSA), leading to formation of cahuitamycins A, B and C, respectively (Figure 3.1). These initial observations are expanded to provide a detailed biochemical analysis of CahJ and its

substrate scope, which revealed new opportunities to generate unnatural cahuitamycin analogs. The structure of CahJ in complex with both native and non-native substrates provided a rational explanation of the substrate scope. Lastly, these findings are translated towards generation of novel congeners of cahuitamycins.

3.4 Results and Discussion

3.4.1 Bioinformatics Analysis of CahJ

Based on the central role of starter unit selection in cahuitamycin diversification and biological activity, this study focused on CahJ, a member of the adenylate-forming enzyme superfamily that clades with the NRPS A domain subgroup (16). CahJ has the highest overall amino acid sequence similarity to a salicylate-AMP ligase MycE from *Sorangium cellulosum* So ce56 (91). *In silico* analysis using SMART (92) showed CahJ has an AMP-binding domain spanning residues 31-437. The closest homolog of known structure is DhbE (58% identity, PDB ID: 1MD9) (93), which activates 2,3-dihydroxybenozic acid (DHB) (Figure A.1).

3.4.2 Colorimetric Determination of CahJ Enzyme Activity

While most A domains are selective for a specific substrate, many also have the ability to adenylate a range of compounds (94,95). In an effort to investigate CahJ specificity toward acyl-substrates, we employed a nonradioactive high-throughput malachite green colorimetric assay (61). In this system, inorganic pyrophosphatase is coupled to a malachite green assay to measure the

concentration of orthophosphate (P_i) converted from pyrophosphate (PP_i) produced during aminoacyl-AMP formation by the A domain. The 59-kDa CahJ protein from *S. gandocaensis* was heterologously expressed in *E. coli* in an N-terminal His_6-tagged form to conduct the colorimetric assay. Soluble CahJ was obtained after Ni(II)-NTA affinity chromatography, followed by desalting prior to the enzyme assay (Figure A.2). SA and 6-MSA were tested to confirm their role as subunits in biosynthesis of cahuitamycins A-C. Compared with the control sample, the adenylation of SA by CahJ led to a high level of color conversion of malachite green reagent. 6-MSA also led to a comparable color conversion of the malachite reagent (Figure A.2).

3.4.3 Determination of Kinetic Parameters (k_{cat} and K_m)

The previous assay indicated that CahJ possesses an innate ability to catalyze activation of both SA and 6-MSA for loading onto the N-terminal CahA PCP, more specifically known as an aryl carrier protein (ArCP). Using this assay, steady-state kinetic parameters were determined by monitoring the release of PP_i by the adenylation reaction. The data for SA and 6-MSA were fit using the Michaelis-Menten equation (Figure 3.2). SA had a pseudo-K_m of 3.5 ± 0.3 μM and pseudo-k_{cat} of 0.107 ± 0.002 min^{-1}, while 6-MSA had a similar but slightly lower pseudo-K_m and pseudo-k_{cat} (pseudo-K_m of 1.6 ± 0.2 μM and pseudo-k_{cat} of 0.079 ± 0.002 min^{-1}). This observation established the importance of the 2-hydroxy group over 2-hydroxy-6-methyl group for catalytic efficiency.

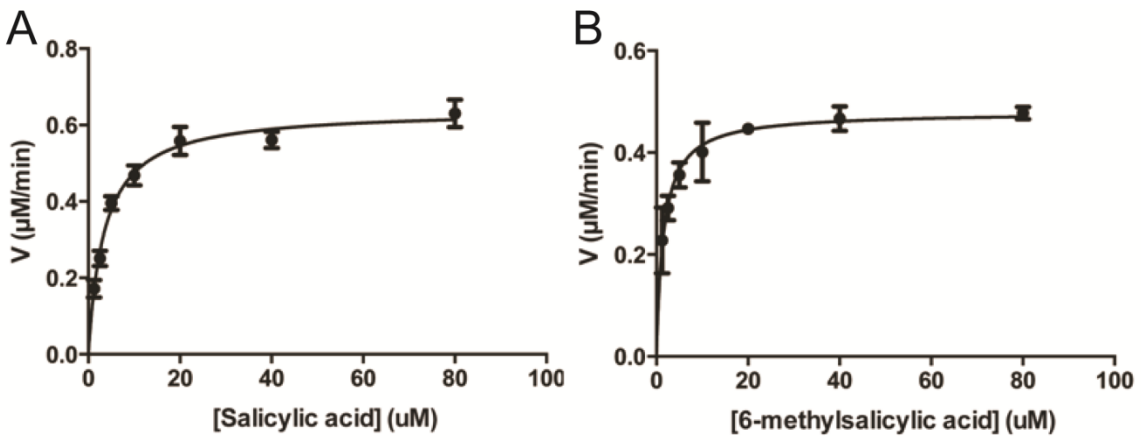

Figure 3.2 Determination of kinetic parameters of CahJ.

Steady-state kinetic data fit using the Michaelis-Menten equation to determine kinetic constants of CahJ for (A) SA and (B) 6-MSA.

Interestingly, the efficiency of CahJ in the malachite-green-based assay was ~1000 fold less than that of its closest structural and functional homologs in radiolabel-based assays of the reverse reaction (96-98). This may be due to co-purification of CahJ with tightly bound salicyl adenylate, as discovered during crystallographic studies where the adenylate was retained through multiple purification steps and was a major technical hurdle in obtaining crystal structures of other substrate complexes. Ultimately, ligand-free CahJ was obtained by partial denaturation with urea. *E. coli* does not encode a salicylate synthase based on current genome annotation, and thus the source of the salicylate was likely the rich growth media. Growth of *E. coli* on defined media resulted in co-purification with benzoyl adenylate. Turnover in the malachite green assay coupled to pyrophosphatase requires dissociation of the adenylate, which is clearly very slow.

3.4.4 Substrate Flexibility of CahJ toward Unnatural Substrates

The previous observation that CahJ activates both 6-MSA and SA (2) and the steady state kinetic analysis motivated tests of the ability of CahJ to catalyze adenylation of 32 structurally related benzoic acid substrates. As expected, CahJ exhibited 73% activity with 6-MSA relative to SA. Similarly, methylated SA derivatives, such as 4-methyl-SA (4-MSA) and 5-methyl-SA (5-MSA), showed similar activity by CahJ with the exception of 3-methyl-SA (Figure 3.3). However, replacing the hydroxyl group at the C2 position of salicylic acid to test the ability of CahJ to catalyze adenylation of benzoic acid substrates bearing nitro or acetyl functional groups abolished CahJ activity, as evidenced by 2-nitrobenzoic acid, acetyl salicylic acid, 4-methyl-2-nitrobenzoic acid, 5-methyl-2-nitrobenzoic acid, and 2-methyl-6-nitrobenzoic acid (Figure 3.3). Only 3-methyl-2-nitrobenzoic acid served as a CahJ substrate, albeit with relatively low (26%) conversion (Figure 3.3). Interestingly, halogenated benzoic acid substrates with chlorine/fluorine as a substitute for the 2-hydroxyl group and DHB displayed appreciable reaction turnover with CahJ. By contrast, CahJ showed no significant activity against substrates with extended ring systems (2-(4-chlorophenyl)-1,3-thiazolidine-4-carboxylic acid, 3-(2-hydroxyphenyl) propionic acid, etc.) or five membered ring structures (1,2,4-triazole-3-carboxylic acid, etc.). These observations can be explained by the size of the substrate-binding site (see below) (Figure 3.5).

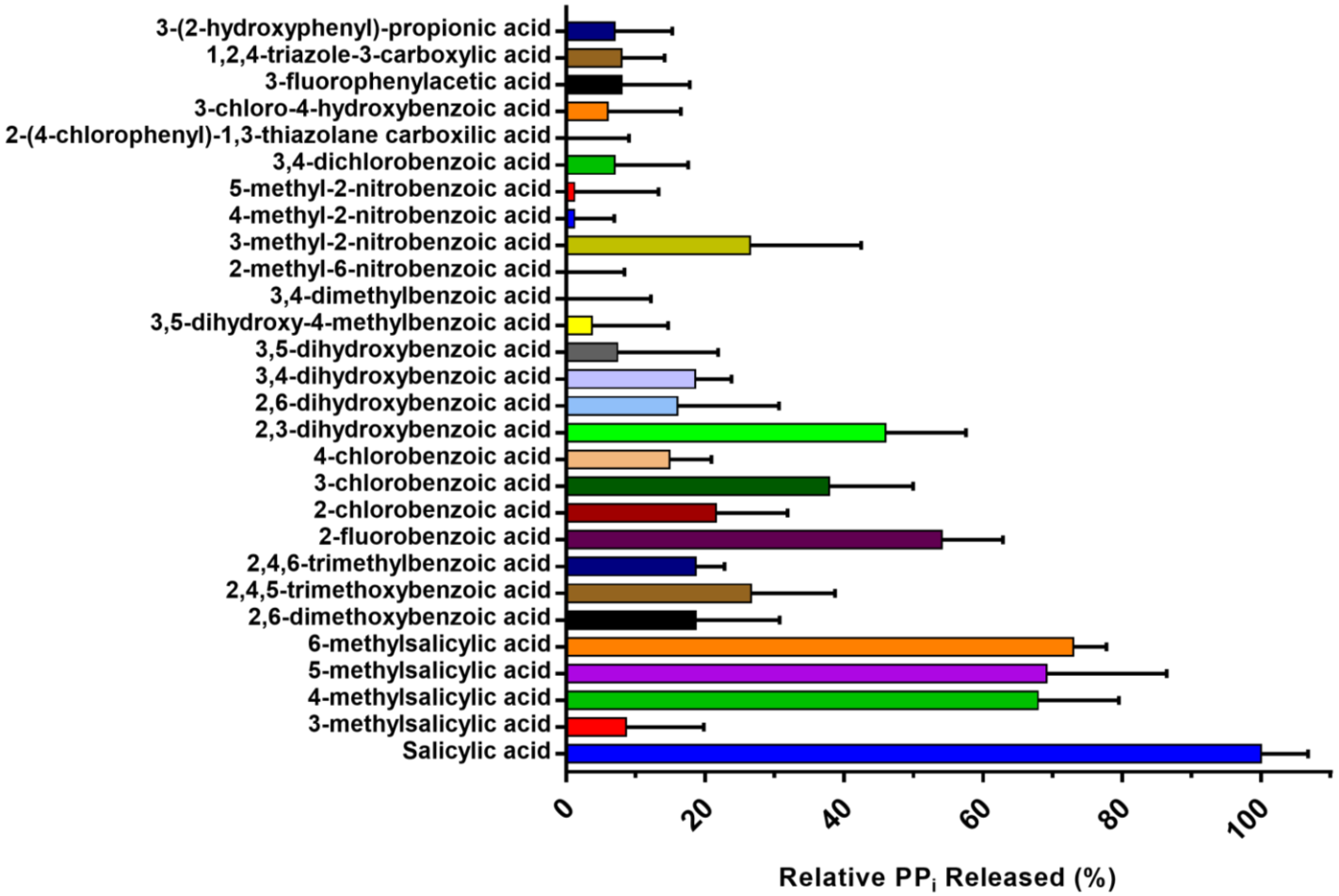

Figure 3.3 Determination of CahJ substrate specificity toward diverse benzoic acid substrates.

Relative substrate specificity of CahJ for 30 benzoic acid analogs is represented by the bars. The activity obtained from the reaction with SA is defined as 100%.

3.4.5 CahJ Loading of CahA ArCP

The ability of CahJ to transfer SA and 6-MSA onto its natural substrate, the CahA ArCP, was evaluated with an intact-protein mass spectrometry assay. CahJ effectively catalyzed loading of the CahA ArCP with both SA and 6-MSA (Figure 3.4). The substrate scope study indicated that effective aryl transfer would be limited to 5-MSA, 4-MSA, 2-fluorobenzoic acid and DHB in addition to the natural SA and 6-MSA substrates. Surprisingly, CahJ transferred all substrates tested to the ArCP including 3-methylsalicylate (3-MSA), which was a poor substrate in the malachite green assay (Figure 3.3; Figure A.3). Thus, while the fast and convenient malachite green assay sensitively discriminates substrate preferences, the results of the ArCP-dependent assay suggest that CahJ may act *in vivo* on an even broader range of substrates.

The apparently greater activity with the natural ArCP acceptor is consistent with the high affinity of CahJ for aryl adenylate intermediates. Reaction with ArCP breaks the adenylate phosphoester bond by phosphoester-thioester exchange, releasing AMP and aryl-ArCP from the enzyme whereas turnover in the malachite green assay requires dissociation of the adenylate. Taken together the assay results and purification behavior indicate that both the nucleotide and aryl moieties of the adenylate contribute to high affinity.

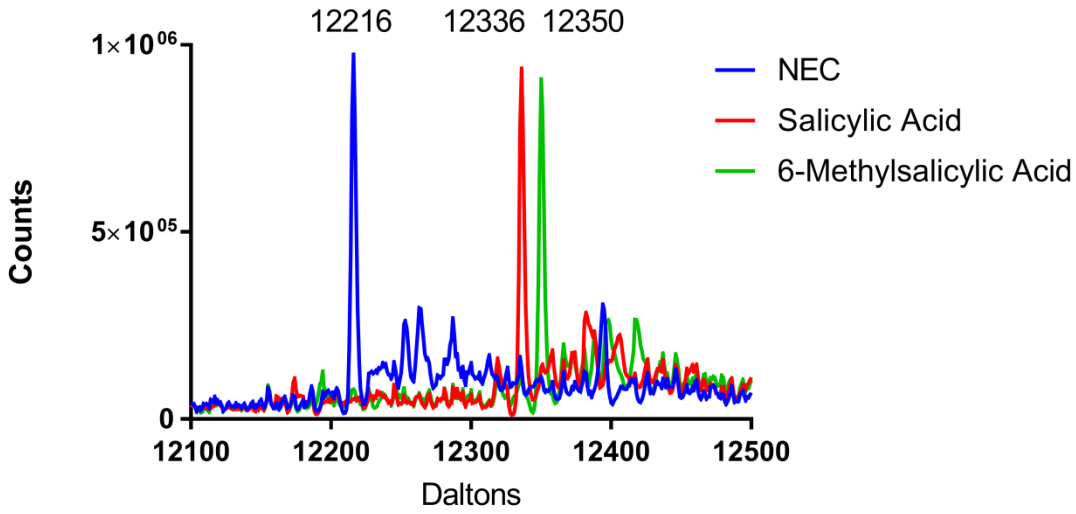

Figure 3.4 Intact protein mass spectra of CahJ loading of SA and 6-MSA onto CahA ArCP.

CahJ was incubated 1 hr with either SA or 6-MSA and ArCP, in both cases leading to complete conversion of holo-ArCP to aryl-ArCP. NEC is a no enzyme control. Calculated and observed masses for ionized species are listed in Table A.1.

3.4.6 CahJ Crystal Structure

For a rational understanding of CahJ substrate selectivity and subsequent ability to incorporate unnatural substrates in the cahuitamycin pathway, crystal structures of CahJ in several substrate complexes were solved (Table 3.1). The CahJ structure is similar to those of other members of the NRPS A domain subfamily (17), particularly those that act on benzoic acid derivatives (93,99,100). However, the CahJ structure is the first in the class for which salicylic acid is a natural substrate. CahJ consists of an N-terminal domain (amino acids 1-429), which contains the substrate binding site, and a C-terminal domain (430-544). ATP binds at the interface of the N- and C-terminal domains (Figure 3.5A). The C-terminal domain of related enzymes is known to move during the two-step

reaction, adopting one conformation for the adenylation reaction and another for the aryl transfer reaction (101). The C-terminal domain of CahJ occupies the aryl-transfer conformation in the structures.

3.4.7 CahJ Substrate Binding Site

CahJ binds aryl substrates in a flat, hydrophobic site between the Phe237 side chain and the Gly332-Met333 peptide. The site is ideally shaped to accommodate aromatic rings with Phe237 forming offset π-stacking interactions with the substrate aromatic ring (Figure 3.5B). The narrow shape of the binding site dictates that only aromatic, and therefore flat, substrates can be accommodated. The periphery of the binding site can be defined as a number of pockets corresponding to the positions of the aromatic ring (Figure 3.5D, Figure A.4).

The 2 position pocket is formed by Asn236, which forms a hydrogen bond with the salicylate 2-hydroxyl. Asn236 is invariant in SA and DHB-utilizing enzymes and forms the only hydrogen bond between aryl substrates and CahJ. Lack of a 2-hydroxyl group decreased activity, for example with 2-fluorobenzoic acid (Figure 3.3). The 3 position pocket is defined by the side chains of Cys241 and Leu338. The side chains form a steric block that restricts the binding of substrates with a substituent at the 3 position as evidenced by the low activity of 3-MSA in the malachite green assay.

Table 3.1 Protein production and crystal growth conditions plus data collection and refinement statistics for CahJ structures.

CahJ complex	AMP + SA	Salicyl adenylate	5-MSA adenylate	6-MSA adenylate	Benzoyl adenylate	AMP
CahJ production	Rich media, urea treatment	Rich media	Rich media, urea treatment	Rich media, urea treatment	Minimal media	Rich media, urea treatment
Added ligands	AMP, SA	none	ATP, 5-MSA	ATP, 6-MSA	none	AMP
Data Collection						
Space group	$P3_221$	$P3_221$	$P3_221$	$P3_221$	$P3_221$	$P3_221$
Cell dimensions a and b,c (Å)	122.1, 87.7	122.4, 88.1	122.0, 87.7	122.2, 88.2	121.7, 87.9	122.0, 88.0
X-ray source	APS 23ID-D	APS 23ID-D	APS 23ID-D	APS 23ID-D	APS 23ID-D	APS 23ID-D
Wavelength (Å)	1.033	1.033	1.033	1.033	1.033	1.033
d_{min} (Å)	1.55 (1.64-1.55)[1]	1.68 (1.78-1.68)	1.80 (1.91-1.80)	1.78 (1.89-1.78)	2.0 (2.12-2.0)	1.78 (1.88-1.78)
R_{merge}	0.079 (0.955)	0.120 (0.584)	0.097 (1.20)	0.090 (0.822)	0.140 (0.75)	0.098 (0.777)
Wilson B factor	24.8	29.6	33.6	31.4	28.6	27.2
Avg I/σ(I)	20.47 (2.43)	5.86 (1.54)	8.96 (1.19)	10.41 (1.50)	8.35 (1.71)	11.37 (1.76)
Completeness (%)	99.9 (99.6)	98.9 (98.4)	98.9 (99.2)	99.1 (99.3)	99.4 (99.4)	96.1 (95.1)
Multiplicity	10.02 (9.71)	3.39 (3.32)	3.40 (3.40)	3.93 (3.84)	3.39 (3.35)	4.06 (4.02)
Total observations	1,093,062 (168,861)	291,083 (45,428)	233,697 (37,664)	283,329 (44,413)	171,821 (27,161)	284,544 (44,625)
$CC_{1/2}$	0.999 (0.637)	0.985 (0.663)	0.996 (0.414)	0.997 (0.517)	0.986 (0.422)	0.996 (0.458)
CC*	0.999 (0.882)	0.996 (0.893)	0.999 (0.765)	0.999 (0.823)	0.996 (0.770)	0.999 (0.793)
Refinement						
Data range (Å)	35.60-1.55	35.76-1.68	18.51-1.80	35.76-1.78	35.64-2.0	35.68-1.78

Reflections used in refinement	109,137	85,969	69,218	72,133	50,752	70,015
R_{work}/R_{free} (%)	13.0/16.0	17.4/20.4	15.6/19.3	15.5/19.4	13.8/17.5	14.3/17.5
Twin fraction	0.29	0.04	0.12	0.17	0.49	0.47
Non-hydrogen atoms	4,701	5,001	4651	5,025	4,506	4481
protein	4,125	4,079	4056	4,083	4,070	4047
ligands	43	58	47	68	36	23
water	533	864	548	874	400	411
Amino acid residues	536	537	536	537	536	536
Deviation from ideality						
bond lengths (Å)	0.006	0.007	0.008	0.008	0.007	0.007
bond angles (°)	0.84	0.923	0.97	0.84	0.87	0.86
Average B-factor ($Å^2$)	20.3	27.5	32.1	23.8	23.1	20.3
macromolecules	18.4	24.7	30.5	21.2	22.4	19.6
ligands	27.0	23.5	29.1	21.3	14.7	17.2
solvent	35.3	40.9	44.2	36.1	30.5	27.3
Ramachandran plot						
favored (%)	96.9	97.8	98	97.6	97.0	97.2
allowed (%)	2.9	2.0	1.8	2.0	2.8	2.6
outliers (%)	0.2	0.2	0.2	0.4	0.2	0.2

[1]Values in parentheses pertain to outermost shell of data.

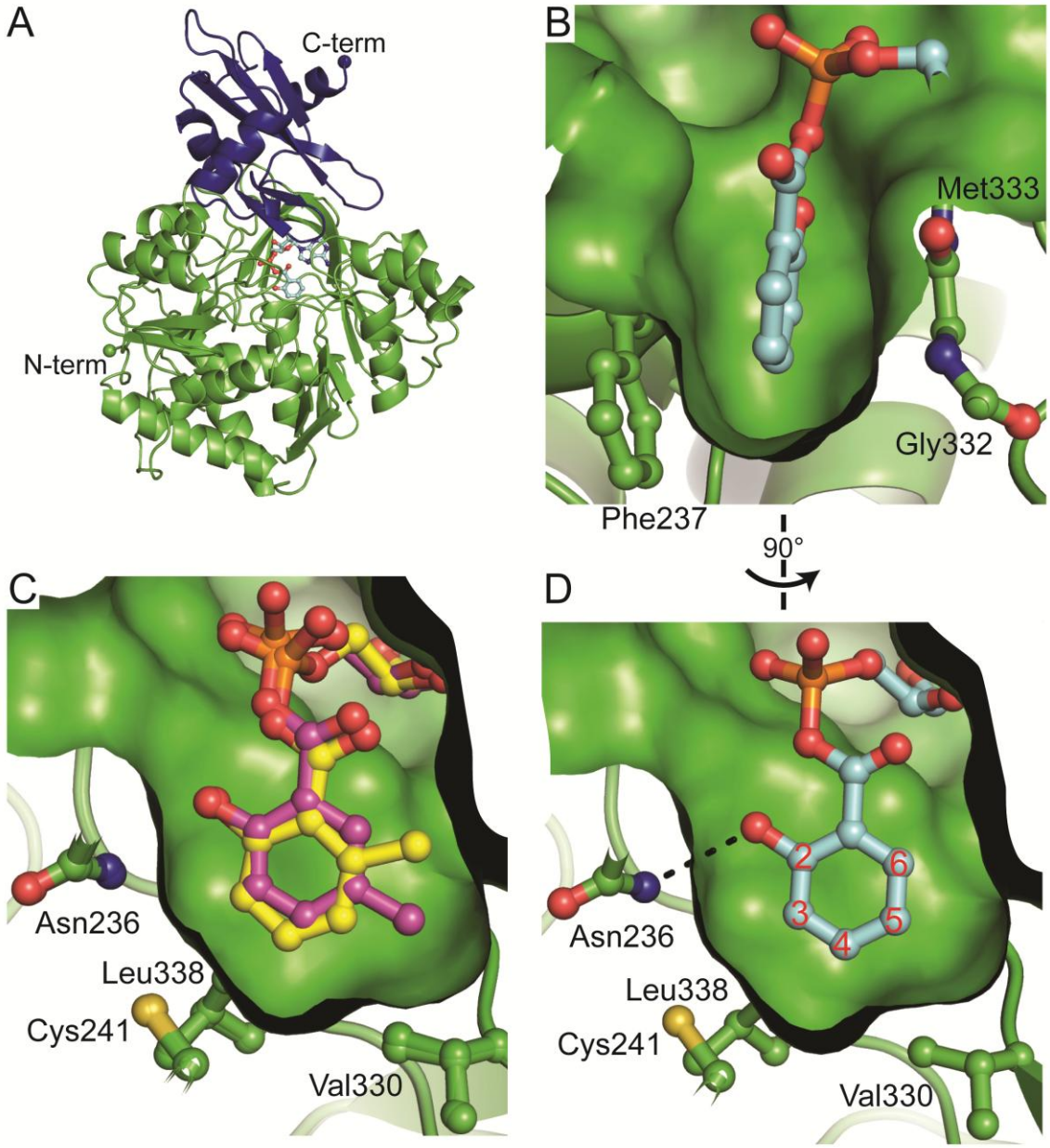
A
C-term
N-term
B
Met333
Gly332
Phe237
90°
C
Asn236
Leu338
Cys241
Val330
D
2
6
3
5
4
Asn236
Leu338
Cys241
Val330

Figure 3.5 CahJ structure and substrate binding site.

A) The overall structure of CahJ with the N-terminal domain in green and the C-terminal domain in blue. The N- and C-termini are shown as spheres. The bound SA adenylate is shown in cyan ball and stick. B) Side view of the flat, hydrophobic binding site for the substrate aromatic ring between Phe237 and the Gly332-Met333 peptide. The orientations of Phe237 and bound salicyl adenylate allow for π-stacking interactions. The protein surface is shown in green with the SA adenylate substrate in cyan. C) Overlay of 5-MSA and 6-MSA adenylates in the CahJ active site. When overlaid, the rotation of 6-MSA relative to 5-MSA (and other substrates) is clear. The observed position of 6-MSA is incompatible with a methyl substituent at the 5 position. D) Face-on view of the substrate binding site seen in part B. The 2-hydroxy group forms a hydrogen bond with Asn236 (3.0-3.2 Å among the three structures). Pockets for the binding of additional substituents are clear at the 4, 5, and 6 positions with Val330 creating a separation of the 4 and 5 position pockets. The 3 position pocket is restricted by Cys241 and Leu338.

The 4 position pocket lies between Cys241 and Leu338 and Val330, while the 5 position pocket is between Val330 and Gly307. The 4 and 5 position pockets are both hydrophobic and large enough to accommodate methyl or similarly sized substituents. In the 5-MSA adenylate structure, the methyl fits snugly into the 5 position pocket (Figure 3.5C). In addition, substrates with substituents at either the 4 or 5 position had generally high activity in the malachite green assay. Thus, it is anticipated that Cl substituent of 3-cholobenzoic acid does not bind in the restricted 3 position, but rather occupies the 5 position.

The 6 position pocket is a bounded by invariant Gly307 and Gly308, and is not large enough to accommodate a methyl group without a slight rotation (~5°) of the aromatic ring, as seen in the 6-MSA adenylate structure (Figure 3.5C). This rotation places the 6-methyl substituent close to the 5 position pocket, and would

create a steric clash with Val330 if a 5-methyl substituent were present simultaneously This indicates a likely mutually exclusive relationship between the two sites. A similar relationship appears to exist between the 4 position and 5 positions, in which CahJ would accommodate a 4 substituent by a slight rotation of the aryl ring. This would result in the binding site poorly accommodating a 5 substituent.

The CahJ active site is virtually identical to those of family members that act on DHB, including DhbE (93), BasE (99) and EntE (100,102). The active sites of all four enzymes contain pockets at the 4, 5, and 6 positions of the aromatic ring. Other enzymes in the SA/DHB family are expected to possess similar substrate flexibility, but only CahJ has been interrogated with a wide panel of substrates. EntE possessed substrate promiscuity in tests with two non-natural substrates, but neither contained methyl substituents (103).

3.4.8 Expanding the CahJ Active Site

None of the three substrates induced structural changes to the binding site. The stable active site structure may provide an opportunity to engineer the individual binding pockets without significantly altering the overall active site structure, which may allow further expansion of the substrate range of CahJ and the creation of a broader range of cahuitamycin congeners, and ultimately clinically efficacious cahuitamycin analogs. The best engineering potential is found at the 3 position pocket where the Cys241 and Leu338 that form the steric block are

naturally variable and are replaced by Ser and Val, respectively, in enzymes that act on 2,3-dihydroxybenzoic acid (93). This indicates that amino acid substitutions will be well tolerated at these positions, allowing for the removal of the steric block and, therefore, activity with substrates containing a 3 substituent (Figure A.5B). Val330 restricts the size of substituents at the 4 and 5 positions at the bottom of the binding site. Val330 is invariant among enzymes of the subfamily, but its replacement with Gly could significantly expand the binding site and allow CahJ to accommodate bicyclic ring substrates (Figure A.5C).

3.4.9 Generation of a New Cahuitamycin Analog through a Rational Mutasynthesis Approach

Insights from the biochemical and structural analyses were applied toward generation of novel cahuitamycin congeners. The *in vitro* substrate 4-MSA was exogenously introduced to the Δ*cahI* strain and was incorporated to produce a new cahuitamycin analog, cahuitamycin F (**1**; Figure 3.6). This new metabolite was isolated by reverse-phase HPLC (RPHPLC) and the HRESIMS $[M+H]^+$ ion peak at *m/z* 650.2706 provided a molecular formula of $C_{28}H_{39}N_7O_{11}$ (Figure A.6), requiring 13° of unsaturation. Extensive 1D and 2D NMR data were acquired for **1** by Dr. Ashootosh Tripathi and indicated the expected structural similarity with **2** (Figure 3.6; Table A.2), including 8 methyl/methane carbons, 10 methylene carbons and 9 carbonyls/quaternary carbons, similar to the carbon backbone of reported cahuitamycins with a clear difference at the phenyl ring system.

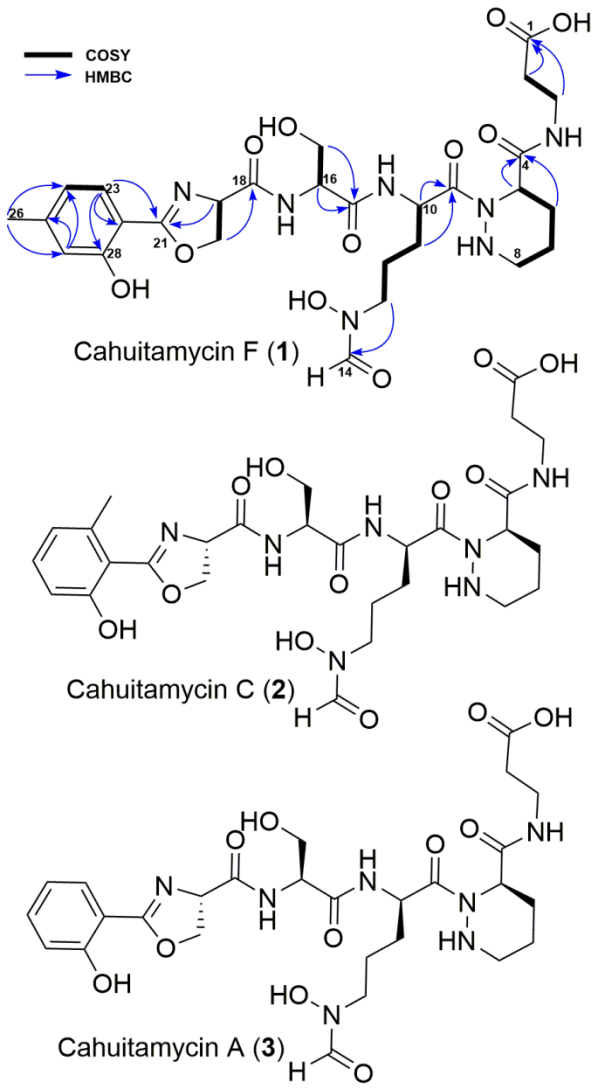

Figure 3.6 Structures of cahuitamycins F, C and A.

Cahuitamycins F, C and A (**1–3**) with absolute stereochemistry. Cahuitamycins F shows key HMBC and COSY correlations.

3.4.10 Biological Activity of cahuitamycin F (1)

Cahuitamycin F (**1**) was next tested for its ability to inhibit biofilm formation of *A. baumannii* using crystal violet based static biofilm assay followed by optical density measurements. The assay was conducted using cahuitamycin A (**3**) as a positive control and the result from this assay showed that **1** is able to inhibit biofilm formation with only a minimal effect on bacterial growth (Figure 3.7). The calculated half-maximal inhibitory concentration (IC_{50}) value for **1** was 18.3 µM, which is similar to the IC_{50} of **3** (15.6 µM) against *A. baumannii* biofilm formation. It was previously observed that methylation at the 5 or 6 position of the terminal hydroxyphenyl group caused an increase in biofilm inhibition compared to cahuitamycin A (**3**) (2). Methylation at the 4 position causes a slight decrease in activity, indicating that the benefits of methylation are position specific. Cahuitamycin F (**1)** showed negligible impact on the growth of *A. baumannii*, consistent with the activity of other cahuitamycins. The data suggest that the terminal 2-hydroxybenzoyl-oxazoline group represents a key pharmacophore of the cahuitamycins. This insight is important for future medicinal chemistry efforts geared toward increasing the efficacy of this new class of natural product biofilm inhibitors.

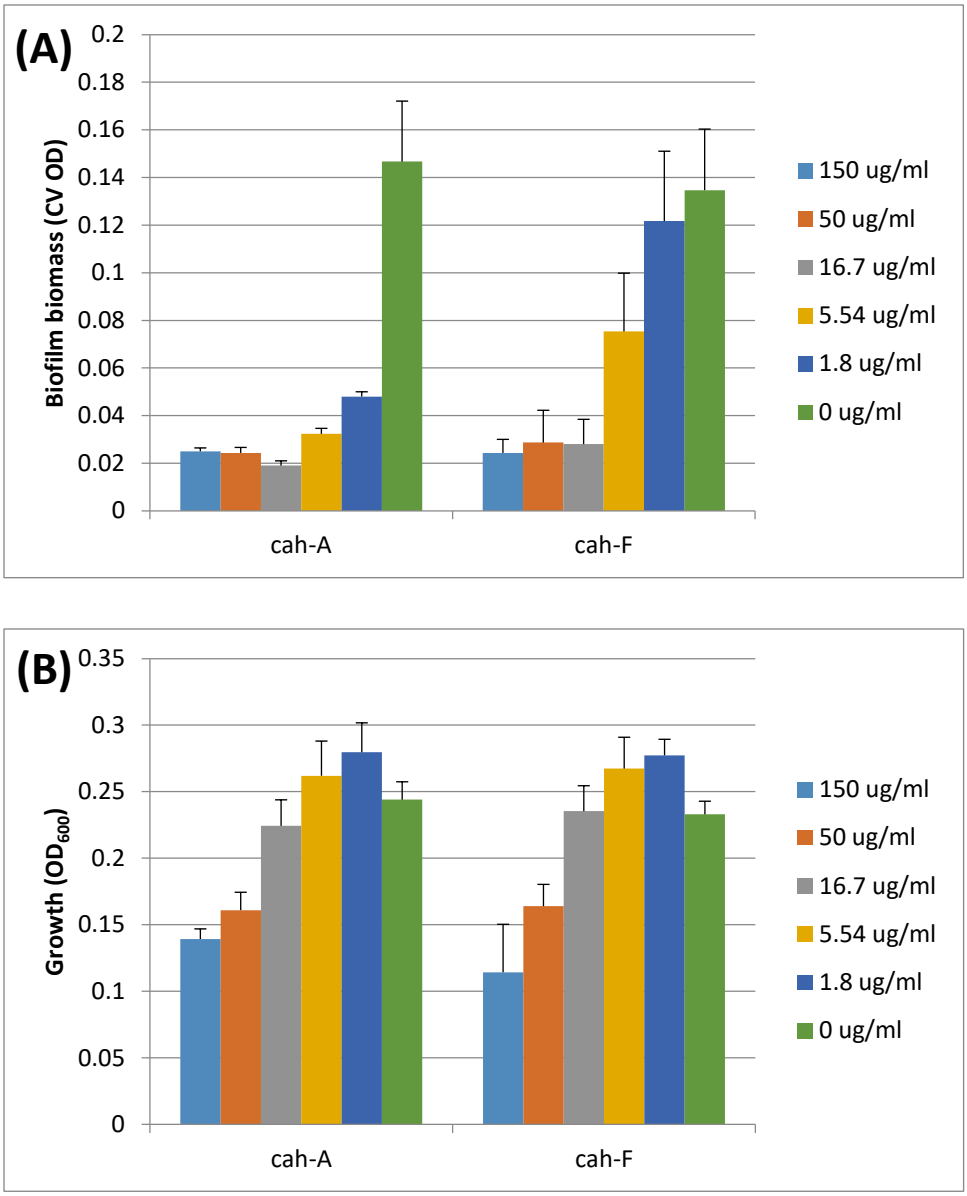

Figure 3.7 Biological activity of cahuitamycins A (**3**) and F (**1**).

A) Inhibition of biofilm formation and (B) growth of *A. baumannii* in the presence of cahuitamycins A and F. Results are the average of three replicates ± standard deviation.

3.5 SIGNIFICANCE

This study describes the structural and biochemical characterization of CahJ, a promiscuous NRPS adenylation domain from *Streptomyces gandocaensis.* CahJ

natively selects salicylic acid and 6-methylsalicylic acid as starter units for the cahuitamycin biosynthetic pathway. First CahJ steady-state kinetic parameters were determined for SA (pseudo-K_m of 3.5 ± 0.3 μM, pseudo-k_{cat} of 0.107 ± 0.002 min^{-1}) and 6-MSA (pseudo-K_m of 1.6 ± 0.2 μM, pseudo-k_{cat} of 0.079 ± 0.002 min^{-1}) (Figure 3.2). It was then demonstrated that CahJ is capable of activating a range of benzoic acid derivatives and transferring them to the N-terminal ArCP of CahA. The first crystal structures of CahJ complexes with both natural and unnatural substrates provided insight to the structural basis of substrate flexibility and development of a structure-based explanation of the biochemical data. The combined biochemical and structural studies where then used to guide efforts to generate new cahuitamycin congeners, leading to the generation of a novel cahuitamycin analog, cahuitamycin F, via mutasynthesis. This study also provides an effective roadmap for future protein engineering to alter CahJ substrate selectivity to generate new clinically efficacious cahuitamycin analogs.

3.6 EXPERIMENTAL PROCEDURES

3.6.1 General experimental procedures

Optical rotation measurements were obtained on a Perkin-Elmer 241 Polarimeter calibrated using a Rudolph Quartz Control Plate Calibration Standard at sodium D line (at +11.502°). Ultraviolet spectra were obtained on a UV-visible Molecular Devices SpectraMax M5 spectrophotometer using 1-ml cuvettes with 1.0 cm path lengths at room temperature in solvent methanol (MeOH). Spectrophotometric assays were performed on Molecular Devices SpectraMax M5 384 variable

wavelength spectrometer. All NMR spectra were acquired on a Varian INOVA 600MHz spectrometer at the NMR Facility, Life Sciences Institute, University of Michigan. HRESIMS spectra were measured at the University of Michigan core facility in the Department of Chemistry using an Agilent 6520 Q-TOF mass spectrometer equipped with an Agilent 1290 HPLC system. RP-HPLC was performed using Econosil C18 10 mm 22×250-mm column and Agilent ZORBAX RX-C8 5 mm 9.4×250-mm column and a solvent system of MeCN and H_2O. The LC-MS analysis of HPLC fractions was performed on a Shimadzu 2010 EV APCI spectrometer.

3.6.2 Protein Expression, Purification

The *cahJ* gene was amplified by PCR from genomic DNA of *S. gandocaensis* using the primer SR235 (5'-GCCGCGCGGCAGCCATATGCTCGATGGGTGGGTG-3') and SR236 (5'-TCGAGTGCGGCCGCAAGCTTCAGTGGACACCGCCGTC -3'). Cloning of *cahJ* gene into pET28b (Novagen) for expression of N-terminal His_6-tagged CahJ was performed using Gibson assembly (104) master mix (New England Biolabs). The amplified PCR product of *cahJ* and pET28b, which had been cut by *Eco*RI and *Hind*III, were incubated at 50°C for 2 h. Following incubation, the samples were transformed into *E. coli* DH5α and isolated. The accuracy of *cahJ* in pET28b was confirmed by restriction digestion and sequencing. Synthetic codon-optimized DNA encoding the CahA ArCP was purchased from IDT, inserted into the pMCSG7 vector using LIC cloning (105), and verified by sequencing.

3.6.3 Protein Expression and Purification

Recombinant plasmid pET28b-*cahJ* was utilized to transformed into *Escherichia coli* BL21(DE3)/pRARE cells. The resulting transformant was grown at 37 °C in 1 liter of LB broth containing kanamycin (50 μg/ml) until OD_{600} reached 0.6~1.0. Then, isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.2 mM to induce gene expression and cells were cultured at 18°C overnight. Cells were harvested by centrifugation at 5,000 × g for 10 min at 4°C to harvest cells. The cell pellet was resuspended in 70 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, 10% glycerol, 0.1mM EDTA (pH8.0), 0.2mM TCEP, pH7.6) and lysed by sonication. Cell debris was removed by centrifugation for 20 min at 45,000 × g. The supernantant was mixed with 1 ml of Ni-NTA agarose (Qiagen) for 1 h at 4 °C and loaded onto an empty column. The column was washed with 50-100 ml of wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, 10% glycerol, 0.2mM TCEP, pH7.6). The bound His_6-tagged proteins were eluted with elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, 10% glycerol, 0.2mM TCEP, pH 8.0). The CahJ (59 kDa) proteins were further purified and concentrated with 30K size exclusion filters (Amicon). Buffer exchange with storage buffer (50 mM NaH_2PO_4, 10% glycerol, pH7.3) using a PD-10 column (GE Healthcare) was used to lower the sample's salt concentration. Protein concentration was determined using NanoDrop spectrophotometer. The molar extinction coefficient for CahJ at A_{280} is 52,370 $M^{-1}cm^{-1}$.

3.6.4 Expression and Purification of CahA ArCP and CahJ for crystallization

Cultures were grown in either Terrific Broth (TB) medium containing 4% glycerol or in M9 minimal medium supplemented with nutrient mix (AthenaES) and L-methionine (Sigma) as indicated. Ampicillin (100 µg/mL, pMCSG7), spectinomycin (50 µg/mL, pRARE), and kanamycin (50 µg/mL, pET28b) were used with the corresponding plasmids in *E. coli* cultures. *E. coli* strain BAP1 (65) was transformed with pMCSG7-*cahA_ArCP*, and strain BL21-AI (Thermofisher) was transformed with pET28b-*cahJ*. Both strains contained the pRARE plasmid (106). Cultures were grown at 37°C in 500 mL TB with 4% glycerol to an OD_{600} of 1.5, cooled to 20°C for at least 1 hr, induced with IPTG (final concentration 0.2 mM), and grown for an additional 14-18 hr before harvesting by centrifugation. Cell pellets were stored at -20°C until used for protein purification.

Cell pellets were resuspended in approximately 5 mL Buffer A (50 mM HEPES pH 7.6, 250 mM NaCl, 5% glycerol) containing 15 mM imidazole per 1 g cell paste. All purification steps were performed at 4°C. Resuspended cells were incubated with DNase (4 mg), lysozyme (10 mg), and $MgCl_2$ (4 mM) for 30 min, then lysed by three passes through an Avestin EmulsiFlex-C3 homogenizer. The lysate was clarified by centrifugation at 30,000xg in a Sorvall F21-8x50y rotor. The soluble fraction was filtered through 0.45 µm and 0.22 µm filters and then loaded onto a 5-mL HisTrap Ni NTA column (GE Healthcare) and eluted with an imidazole gradient up to 300 mM in Buffer A. The eluate was concentrated to

approximately 5 mL using a 50 kDa Amicon Ultra-15 membrane, centrifuged at 20,000xg for 30 min, then loaded onto a HiLoad 16/600 Superdex 200 column (GE Healthcare) equilibrated with Buffer B (20 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol) and isocratically eluted. Fractions were pooled, concentrated as above to 20-30 mg/mL, centrifuged at 20,000xg for 30 min, flash frozen in 20-50 µL aliquots with liquid N_2, and stored at -80°C. CahJ co-purified with bound SA adenylate from cells grown in rich media and with bound benzoyl adenylate from cells grown in minimal media. These co-purifying ligands were removed from CahJ by inserting the following steps while the protein was bound to the HisTrap Ni NTA column, prior to elution with imidazole: 10 CV linear gradient of 0 to 500 mM urea in Buffer A; 5 CV wash; 10 CV linear gradient of 500 mM to 0 urea in Buffer A containing 1 mM AMP; and 5 CV wash with Buffer A containing 30 mM imidazole.

CahA ArCP was purified identically to CahJ except that the buffer was 100 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 5 mM TCEP. The eluate was concentrated using a 10 kDa Amicon Ultra-15 to approximately 10 mg/mL. CahA was then incubated for 16 hr at 30°C with the phosphopantetheinyl transferase Sfp (66) at a 1:40 ratio in a reaction mix containing 16 mM ATP, 4 mM pantethine, 4 mM DDT, and 0.2 mg/mL of CoaA, CoaD, and CoaE (67). After incubation, protein was centrifuged at 20,000 xg for 30 min and then loaded onto a HiLoad 16/600 Superdex 75 column (GE Healthcare) equilibrated with 100 mM Tris pH 7.5, 250 mM NaCl, 5% glycerol, 5 mM TCEP and isocratically eluted.

Fractions were pooled, concentrated to 10 mg/mL, centrifuged at 20,000 xg for 30 min, flash frozen in 20-30 µL aliquots with liquid N_2, and stored at -80°C.

3.6.5 Crystallization

Frozen CahJ stock solution was thawed on ice and glycerol was removed by buffer exchange into 20 mM Tris pH 7.5, 150 mM NaCl using a 50K Amicon Ultra 0.5 membrane. After addition of the appropriate substrates (final concentration for all 1.5 mM, Table 3.1), the mixture was diluted to a final concentration of 20 mg/mL CahJ, incubated on ice for 30 min, and clarified by centrifugation at 20,000xg for 30 min at 4°C. Crystals grew in approximately one week at 20°C by vapor diffusion in 1.5-µL sitting drops containing equal volumes of protein stock and reservoir solution (0.08-0.14 M sodium cacodylate pH 6.5, 1.5 - 2.0 M sodium acetate). Crystals were harvested in nylon loops, cryoprotected in reservoir solution containing 20% glycerol, and flash cooled in liquid N_2.

3.6.6 Diffraction Data Collection and Structure Determination

Diffraction data were collected at 100 K on GM/CA beamline 23ID-D at the Advanced Photon Source (APS) at Argonne National Laboratory (Argonne, IL). For each crystal, 360^o of data were collected at an X-ray energy of 12 keV with 0.2^o rotation and 0.2 sec exposure per image. Data were processed using XDS (68). The structure of CahJ with bound AMP and salicylic acid was solved by molecular replacement (MR) with Phaser (69) in the PHENIX (70) software suite from the DhbE structure (93) (58% identity, PDB code: 1MD9) with the C-terminal

domain removed. The initial model from MR was modified with AutoBuild (71) to generate a 75% complete model of CahJ. The remaining portion of model was completed manually using Coot (72). Refinement was performed using phenix.refine (73). Structures of CahJ with other ligands were solved by difference Fourier methods as the crystals were isomorphous. Restraints for all substrates, except AMP, were generated using eLBOW (74). Images and figures were prepared using PyMOL (82).

3.6.7 Determination of Substrate Specificity and Kinetic Parameters

The substrate selectivity of CahJ was determined using a non-radioactive colorimetric assay (61). All assays were performed in 384-well reaction plates. To perform the assays, benzoic acid substrates were prepared in DMSO and 10 µl was transferred in triplicate to a 384-well reaction plate. Reactions were initiated by adding the mixture (40 µl) containing recombinant CahJ (6 µM), inorganic pyrophosphatase (0.2 U/ml), ATP (0.5 mM), NaCl (100 mM), $MgCl_2$ (10 mM), substrates (20 µM), and 50 mM Tris–HCl (pH 7.5) and incubated at RT. The reactions were terminated after 10 min by the addition of molybdate/malachite green reagent (10 µl). After 10 min of color development, the released P_i concentration was monitored at 600 nm on a microplate reader (SpectraMax®). For the determination of K_m and k_{cat} for formation of SA adenylate (SA-AMP) and 6-MSA adenylate (6-MSA-AMP), reactions were done with varying concentrations of SA and 6-MSA (1.25 µM, 2.5 µM, 5.0 µM, 10.0 µM, and 20.0 µM) in the same condition as described above. The reactions were incubated at

RT and stopped after 0, 5 and 10 min by the addition of molybdate/malachite green reagent (10 µl). A reaction mixture without the recombinant CahJ domain was used as a negative control. The experiments were carried out in triplicate for each substrate concentration.

3.6.8 CahJ Mass Spectrometry Assays with ArCP

Reactions in 25 µL mixtures (50 mM Tris pH 7.8, 100 mM NaCl, 10 mM $MgCl_2$, 500 µM ATP, 2 µM CahJ, and 20 µM ArCP) were initiated by addition of 500 µM substrate in DMSO (final concentration 0.5% DMSO), incubated at 25°C for 1 hr, quenched by the addition of 10% formic acid, and stored at 4°C prior to analysis. 1 µL of reaction mixture was used for LC-MS analysis (Agilent 6545 Q-TOF equipped with Agilent 1290 HPLC system). Samples were separated for intact protein analysis by RP-HPLC (Aeris widepore C4 column 3.6 µM, 50 X 2.10 mm) with a 10 min gradient of 10-95% acetonitrile with 0.1% formic acid at a flow rate of 0.5 mL/min. Data were analyzed using the Agilent Mass Hunter Qualitative Analysis software with the maximum entropy deconvolution algorithm. A reaction mixture without CahJ was used as a negative control.

3.6.9 Generation of Cahuitamycin Analog through Mutasynthesis

Mutant strain of Δ*cahI* (DHS334) was first pre-cultured in 3ml R2YE liquid medium for 15 days at 28 °C and then 3ml of the seed culture was used to inoculate 100 ml of the same medium, followed by cultivation for 15 days at 28 °C. Salicyclic acid (as positive control) and 4-methylsalicylic acid were added

every alternate day to separate 100 mL cultures of Δ*cahI* at a final concentration of 500 µM for 10 days. The products were first extracted using Amberlite XAD-16. The resin was separated and subjected to organic extraction using MeOH:EtOAc (1:1) for LC-MS analysis as described above.

The substrate-fed crude extract was then further purified by RP-HPLC on a gradient of 10–75% ACN and was followed by ultraviolet–visible photodiode array detection at 215 nm to yield semi-pure compounds **1** (4.3 mg). Compounds were again subjected to re-purification over RP-HPLC on isocratic condition of 35% MeOH (0.1% FA) using C-8 column to get compounds **1** (2.1 mg).

Cahuitamycin F (**1**). Amorphous powder; ultraviolet (ACN: H_2O) λ_{max} 204, 240, 252 and 304 nm; ^{1}H and ^{13}C NMR, see Table A.2; HRESIMS m/z 650.2706 $[M+H]^+$ (calculated for $C_{28}H_{40}N_7O_{11}$, 650.2786).

Advanced Marfey's analysis (107,108) was pursued to ascertain the absolute stereochemistry of **1**. Acid hydrolysis of cahuitamycin F was conducted followed by derivatization with 1-fluoro-2,4-dinitrobenzene-5-alanine amide (FDAA). The derivatized product was then analyzed using LC-ESI-MS chromatogram to compare *m/z* 357.27, 382.32 and 400.34 for Ser, Pip and *N*-OH-Orn (hydrolyzed product of *N*-OH-*N*-fOrn), respectively. The analysis clearly revealed the absolute configuration of the moieties in the hydrolysate of **1** to be L-Ser, L-Ser, D-Pip and D-*N*-OH-Orn, respectively.

3.6.10 Minimum Inhibitory Concentration Test

Cultures of *A. baumannii* ATCC 17978 grown overnight were diluted into fresh 10% Mueller Hinton II broth to get 10^5 colony-forming units per ml by determining optical density at OD_{600}. A volume of 100 µl of diluted cultures were incubated in individual wells of 96-well plate containing 1% of DMSO and various concentrations of cahuitamycins A (**3**) and F (**1**). The plate was incubated at 30°C with shaking at 150 r.p.m. The optical density of each well was monitored every 1 hr up to 24 hrs using a microtiter plate reader (Synergy HT, Bio-Tek, Winooski, VT)

Appendix A

Supporting Information for Chapter III

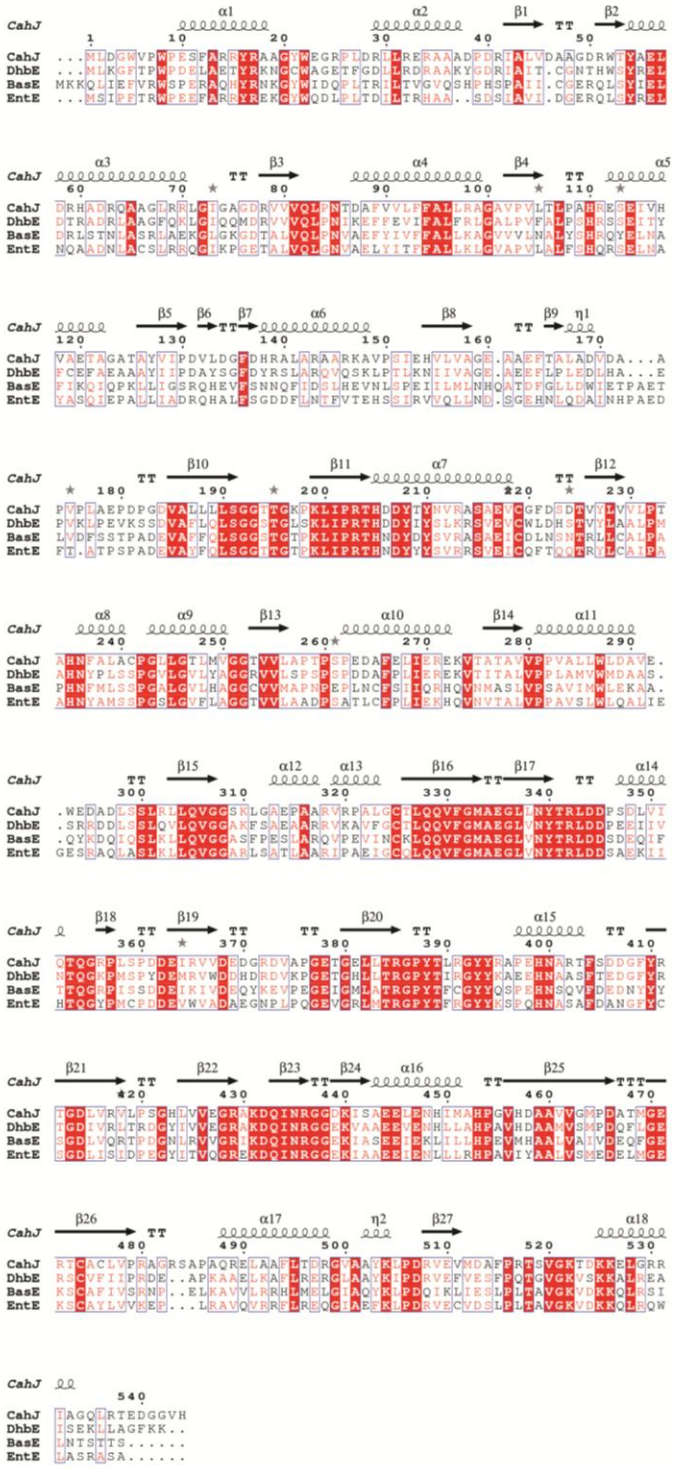

Figure A.1 CahJ sequence alignment.

Secondary structure of CahJ shown above alignment. GenBank accession codes are as follows: CahJ (AMK48234.1), DhbE (AAN15214.1), BasE (AKA30931.1), EntE (ADB98052.1).

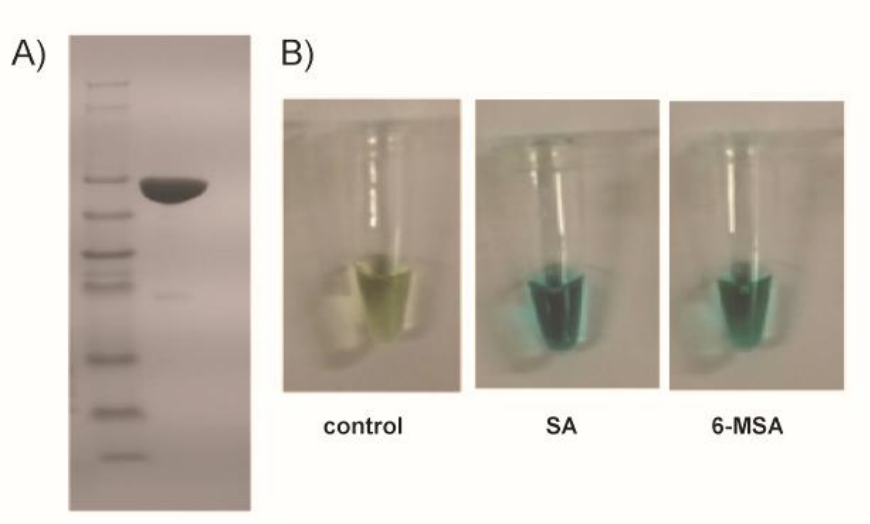

Figure A.2 CahJ purification and activity.

A) Coomassie blue-stained SDS-PAGE gel of CahJ. B) Representative data of CahJ activity for SA and 6-MSA by malachite green assay. SA and 6-MSA samples show an obvious color change from yellow to green compared to control.

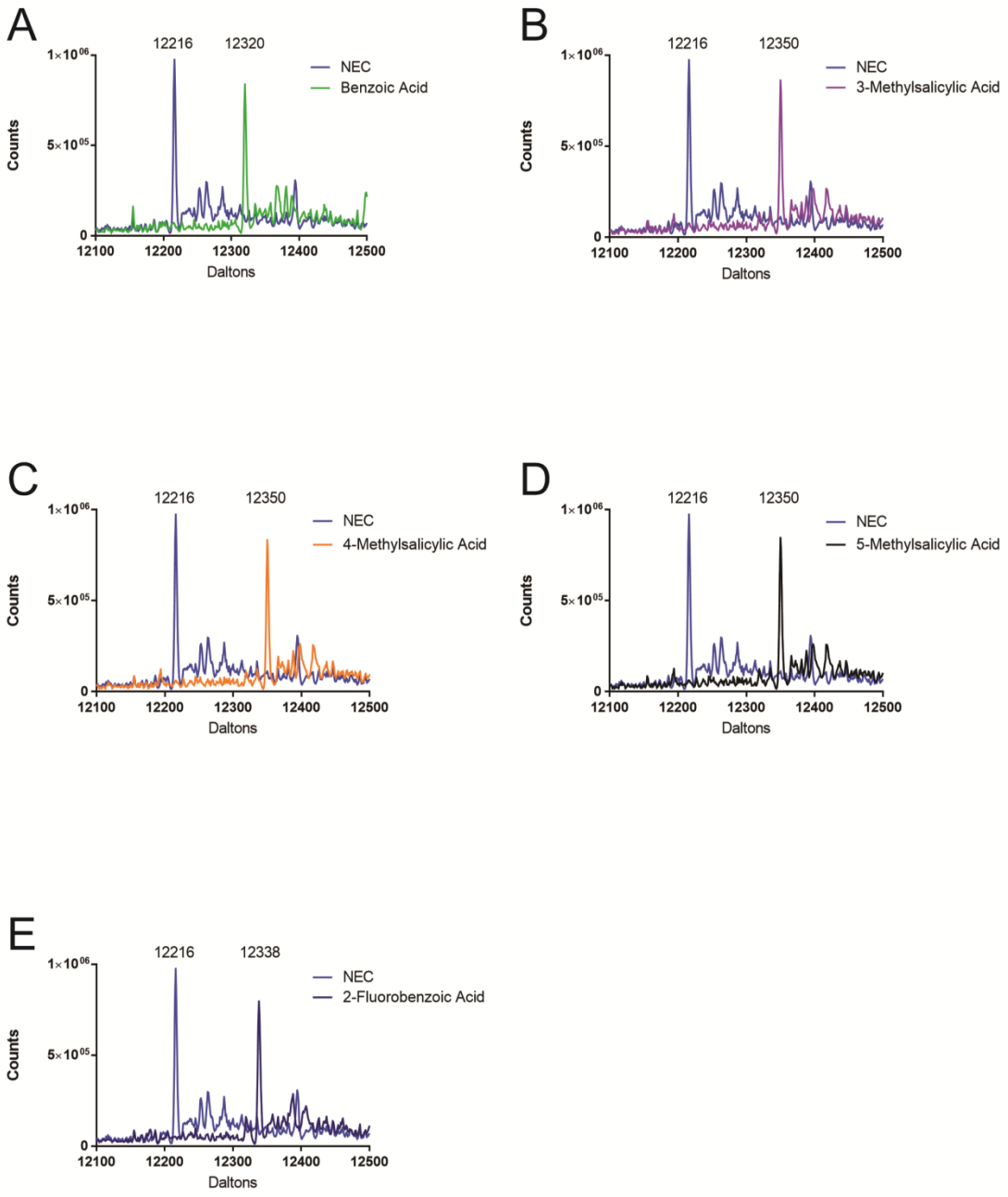

Figure A.3 Intact protein MS of CahJ loading of unnatural substrates onto CahA ArCP.

CahJ was incubated 1 hr with benzoic acid (A), 3-MSA (B), 4-MSA (C), 5-MSA (D), or 2-fluorobenzoic acid and ArCP, in all cases there was complete conversion of holo-ArCP to substrate loaded ArCP. NEC is a no enzyme control. Calculated and observed masses for ion species can be found in Table A.1

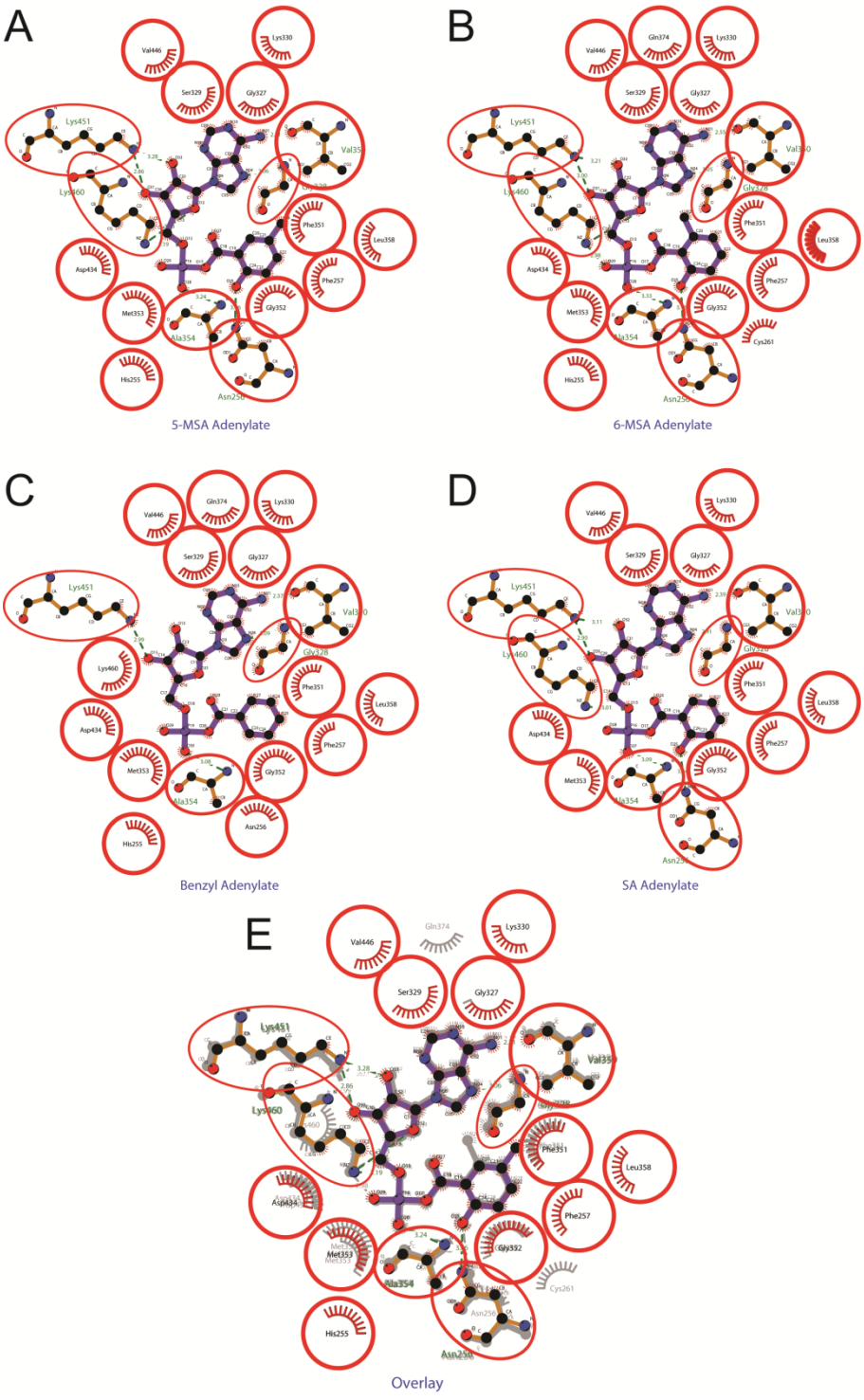

Figure A.4 CahJ substrate interactions.

CahJ interaction with substrate adenylates, generated by LigPlot+ (109). A) 5-MSA adenylate. B) 6-MSA adenylate. C) Benzyl adenylate. D) SA adenylate. E) Overlay of the four previous LigPlots. Red circles indicate contacts observed in multiple substrate complexes.

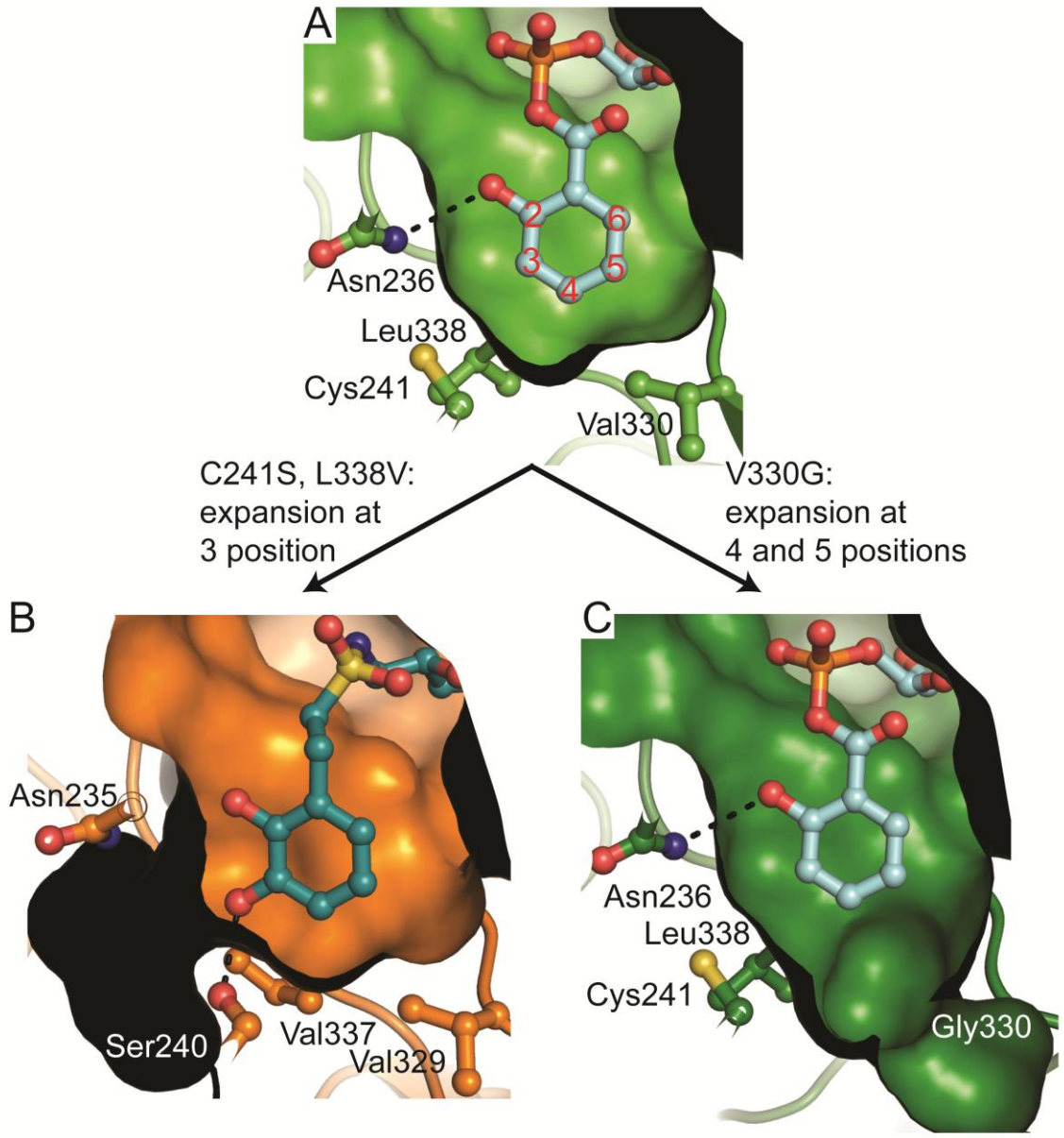

Figure A.5 Engineering potential of the CahJ substrate binding site.

A) The native CahJ binding site with bound SA adenylate (protein in green, SA adenylate in cyan). B) EntE binding site with bound DHB analog (100). Substitution of CahJ Cys241 with Ser and Leu338 with Val is expected to open the 3 position pocket in a similar fashion (protein in orange, substrate in deep teal). C) Model of the CahJ substrate binding site with the V330G substitution. The replacement opens a significant volume at the 4 and 5 position pockets and could accommodate bicyclic rings (protein in forest green, substrate in cyan).

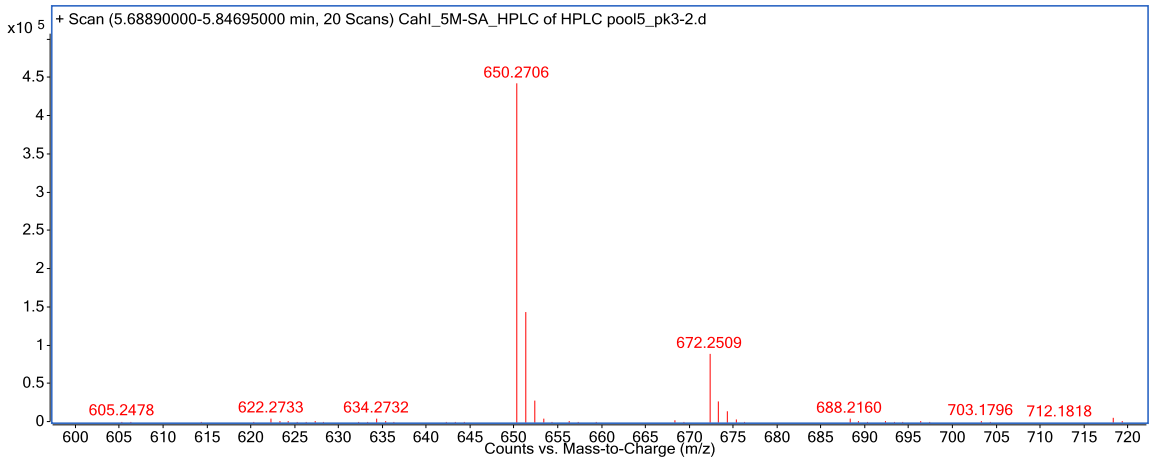

Figure A.6 HRRESIMS $[M+H]^+$ chromatogram of cahuitamycin F (1).

Table A.1 Calculated and observed ion masses from intact protein MS assay.

Species	**Calculated mass (Da)**	**Observed Mass (Da)**
holo-ArCP	12216	12216
Salicylyl ArCP	12336	12336
6-MSA ArCP	12350	12350
Benzyl ArCP	12320	12320
3-MSA ArCP	12350	12350
4-MSA ArCP	12350	12350
5-MSA ArCP	12350	12350
2-Fluorobenzyl ArCP	12338	12338

Table A.2 NMR spectroscopic data for cahuitamycins F.

	Cahuitamycin F (1)			
	δ_C	**δ_H, multi (*J* in Hz)**	**COSY**	**HMBC**
1	174.1			
2	36.5	2.37, m	3	1
		2.41, m		
3	36.9	3.37, m	2	1
		3.51, m	2	4
4	167.3			
5	50.8	4.41, dd (4.5, 9.5)	6	4, 9
6	28.1	1,78, m	5, 7	
		1.98, m	5, 7	
7	20.9	1.92, m	6, 8	5
		1.99, m	6, 8	
8	52.5	3.57, m	7	9
		3.62, m	7	6
9	173.2			
10	54.4	4.30, dt (6.0, 14.0)	11	9, 11, 15
11	29.1	1.62, m	12, 10	
		1.84, m	10	
12	23.5	1.63, m	11, 13	13, 10
		1.68, m	11, 13	10
13	51.1	3.43, t (6.4, 13.0)	12	11, 14
14	163.1	7.78, s		
15	172.5			
16	56.9	4.47, m	17	15, 18
17	62.6	3.89, d (5.0)	16	15
18	173.8			
19	68.9	5.10, dd (7.5, 10.0)	20	18
20	70.4	4.59, dd (7.5, 9.0)	19	18, 21
		4.68, dd (8.0, 9.0)	19	21
21	168.7			
22	146.5			
23	129.3	7.58, d (7.0)	24	21, 22, 28
24	121.6	6.77, d (7.0	23	26, 27
25	108.6			
26	21.3	2.31, s		24, 25, 27
27	117.7	6.82, s	28	24, 25, 26
28	160.1			

Chapter IV

Structural and Biochemical Studies of the Pivalyl- Generating Loading Module from the Apratoxin A Biosynthetic Pathway

4.1 Notes

AprA di-domain acetoacetyl-ACP methylation activity data provided by Meredith Skiba.

4.2 Abstract

Natural products are a source of interesting chemical moieties and novel enzymatic reactions. Apratoxin A, from the marine cyanobacterium *Moorea bouillonii,* contains a rare natural *t*-butyl group, and its creation likely involves novel enzymatic reactions. The *t*-butyl is introduced as part of a pivalyl group by the biosynthetic loading module, the AprA protein. By sequence analysis, AprA is a GNAT-based loading module. GNAT-based loading modules decarboxylate malonyl-CoA and transfer the resulting acetyl group to an associated acyl carrier protein (ACP). AprA contains two additional methyltransferase (MT) domains, MT1 and MT2, which are proposed to convert the acetyl-ACP to pivalyl-ACP. Here, data are presented showing that the AprA loading module diverged from other GNAT-based loading modules. A GNAT-MT2 crystal structure reveals that the GNAT domain is truncated relative to other GNAT domains and is likely vestigial. Biochemical experiments showed that AprA does not possess acyl-CoA

decarboxylation activity, but retains acyltransfer activity and preferentially utilizes acetyl- and propionyl-CoA as substrates. The ArpA MT2 domain is similar to an extension-module *C*-methyltransferase (CMT) from the curacin A pathway, and catalyzes di-methylation of acetoacetyl-ACP. These findings provide a tantalizing hint at the *in vivo* function of AprA and a foundation for future investigations of AprA.

4.3 Introduction

Marine cyanobacteria are a rich source of secondary metabolites with great potential for pharmaceutical development (1). Apratoxin A is a cytotoxic natural product of *Moorea bouillonii* with potent antitumor activity (110,111). It interferes with biogenesis of proteins destined for the secretory pathway by blocking cotranslational translocation across the endoplasmic reticulum membrane through interaction with Sec61, the central subunit of the protein translocation channel (5,52,53). Although the total synthesis of apratoxin A has been reported (112), its ten chiral centers present a significant challenge to exploration of chemical space in the development of derivatives with altered bioactivity (Figure 4.1). The biosynthetic pathway is thus an attractive target for further diversification of this interesting molecule.

Apratoxin A is produced by a hybrid modular polyketide synthase (PKS) / nonribosomal peptide synthetase (NRPS) (Figure 4.1) (4). One of the most striking chemical features of apratoxin A is a *t*-butyl group generated by the

loading module that initiates biosynthesis, based on biosynthetic gene cluster annotation (4) (Figure 4.2). The initiation step is an important source of natural product diversity that can proceed though a number of biosynthetic schemes (15). The apratoxin loading module consists of the single polypeptide, AprA, which belongs to a family of loading modules defined by the presence of a GCN5-related N-acetyltransferase (GNAT)-like (hereafter referred to as GNAT) domain. Previous work characterized a remarkable dual catalytic activity in the GNAT of the curacin A biosynthetic pathway in which malonyl-CoA is decarboxylated and the resulting acetyl group is transferred to the loading module acyl carrier protein (ACP) (35). GNAT-based loading modules occur in secondary metabolite biosynthetic pathways in diverse biological sources including uncultured symbionts (44,47,113,114), betaproteobacteria (48,115), myxobacteria (49), and cyanobacteria (4,50,51); however, only the curacin A loading module has been characterized (35). The GNAT and ACP domains constitute the minimal GNAT-based loading module, but many of the approximately dozen annotated GNAT-based loading modules contain additional domains. The AprA loading module contains three additional domains and has the following domain organization: adaptor region (AR) - methyltranferase 1 (MT1) - GNAT - methyltransferase 2 (MT2) - ACP.

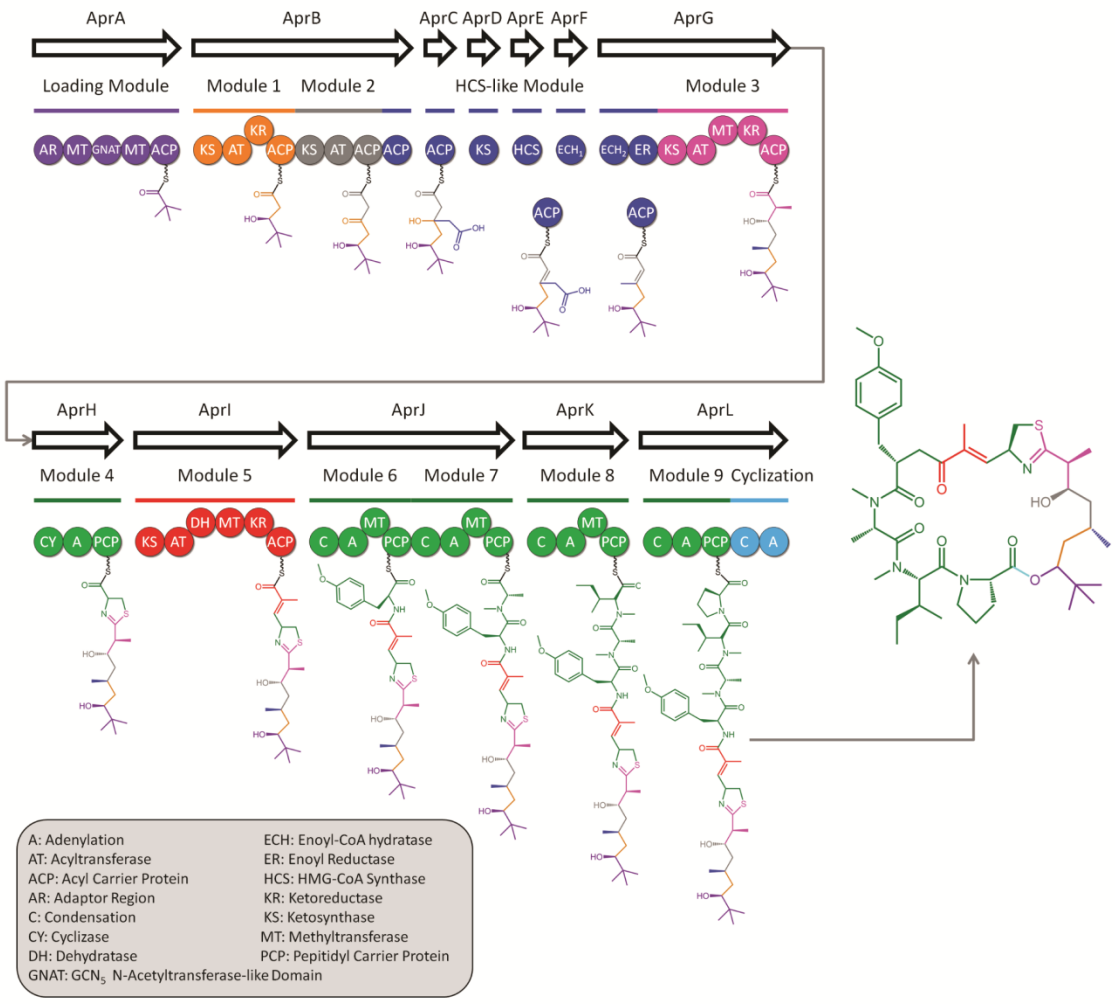

Figure 4.1 The apratoxin A biosynthetic pathway.

The domains and products of the pathway modules are color coded: purple loading module, orange/gray/pink/red PKS extension modules, blue β-branching cassette, green NRPS extension modules, and cyan offloading module.

The loading-module GNAT domains are presumed to decarboxylate malonyl-CoA and to transfer acetyl to the adjacent ACP, based on the established activity of the CurA GNAT (35) and their high level of sequence conservation (typically 30-40% sequence identity) (Figure B.3). However, the AprA GNAT is an exception, as it has poor sequence conservation (~20% identity to other GNATs,

with several gaps in the alignment) (Figure 4.3; Figure B.3) and appears to lack the conserved His and Thr amino acids that are essential to the decarboxylation activity of the CurA GNAT (35). This suggests that the AprA GNAT may function differently from the other loading module GNATs.

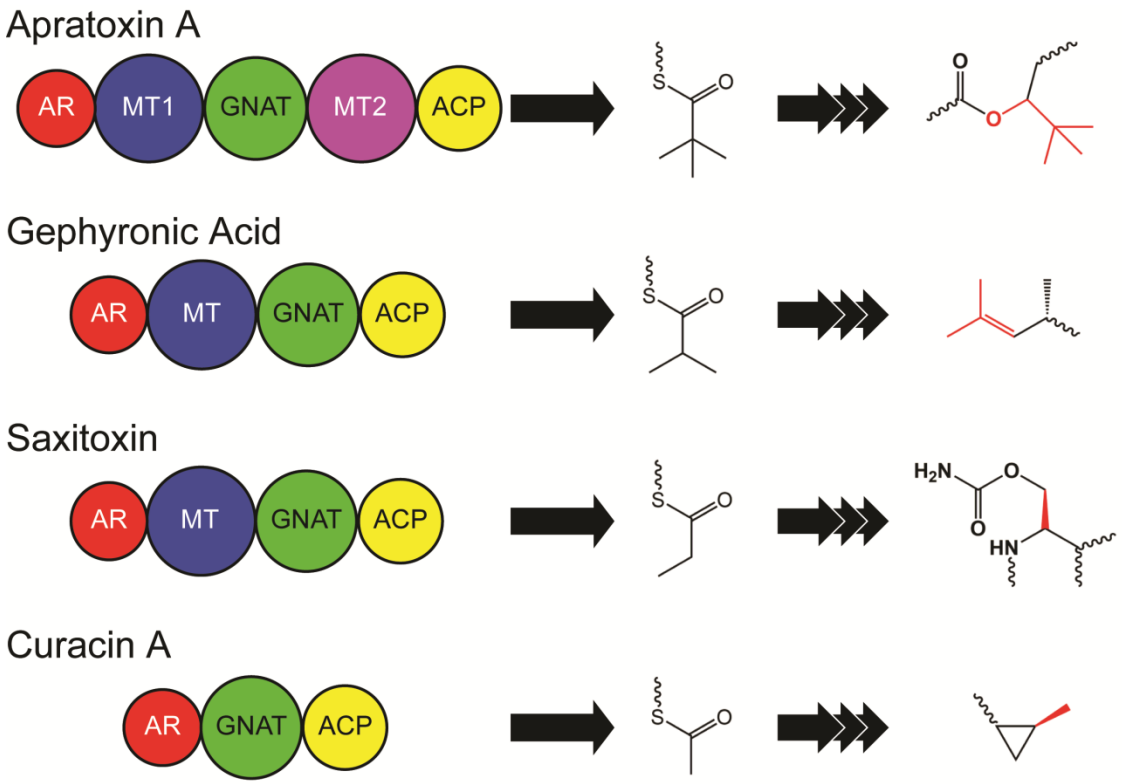

Figure 4.2 Representative GNAT-based loading modules from four biosynthetic pathways.

From left to right: the domain composition of the loading modules, the chemical product of the loading module, and location of the loading module product, shown in red, in the final natural product structure.

Many GNAT-based loading modules contain an AR domain. Although it is highly conserved, the AR domain has no identified function and has no obvious homology with any protein of known structure or function. The AR domain is proposed to bind the phosphopantetheine (Ppant) cofactor of the ACP and facilitate its interaction with the GNAT domain (35). In several loading modules, the AR domain is followed by a putative MT domain, (4,49,50,116) based on conservation of sequence motifs for type I *S*-adenosylmethionine (SAM) dependant methyltransferases (Figure B.2) (117). Interestingly, while the MT domains from the saxitoxin (SxtA) and gephyronic acid (GphF) loading modules are ~50% identical (Figure 4.3C), the SxtA MT is proposed to perform a mono-methylation, whereas feeding studies indicate that the GphF MT performs a di-methylation (49). The determinants of mono- or di-methylation are unknown. If the AprA GNAT loads acetyl on the AprA ACP, then three methylation reactions are needed to generate the *t*-butyl group. A second methyltransferase domain (MT2) following the GNAT is unique to the AprA loading module. MT2 has low sequence identity to other loading module MT domains, including AprA MT1 (Figure 4.3C), and is most closely related to MT domains in PKS extension modules. Recent studies showed that these MT domains methylate β-keto substrates linked to ACP or *N*-acetylcysteamine (NAC) (118,119) (54). The core domains of AprA MT2 and the MT domain from a curacin A extension module (CurJ MT), for which the crystal structure was recently reported (54), are approximately 40% identical (Figure B.4).

A

GphF	30%	51%	21%	-	23%
SxtA	48%	58%	21%	-	32%
CurA	44%	-	24%	-	29%
PDB	-	-	2REE CurA GNAT (21%)	5THZ CurJ CMT (33%)	3LMO (29%)

B

GNAT	AprA	SxtA	GphF
SxtA	22.0		
GphF	19.5	42.8	
CurA	22.0	45.8	36.5

C

MT	AprA MT1	GphF MT	SxtA MT
GphF MT	50.0		
SxtA MT	58.8	49.8	
AprA MT2	26.9	16.7	21.9

Figure 4.3 Bioinformatics analysis of the AprA loading module.

A) Schematic of AprA with estimated domain boundaries shown at top. Solid lines indicate high-confidence portions, dashed lines indicate low-confidence portions. Domain coloring corresponds to Figure 4.2. Sequence identities to loading module domains of other pathways are shown below. The bottom row indicates the structure in the PDB with greatest sequence identity to each AprA domain. B) Pairwise identity matrix of the GNAT domains shown in A. Green indicates a good alignment. Loading module GNAT domains typically have ~40% sequence identity, AprA is an outlier. C) A pairwise identity matrix of the MT domains shown in A. Blue indicates a good alignment. See Figures B.1 to B.4 for sequence alignments.

Among GNAT-type loading modules, the dual MT domains and the truncated GNAT are unique to AprA, and may represent a novel biosynthetic loading scheme and a novel route to a natural product *t*-butyl group. Here we present investigations of the AprA loading module decarboxylation, acyltransfer, and methyltransfer activities as well as the X-ray crystal structure of a GNAT-MT2 di-domain.

4.4 Results and Discussion

4.4.1 Bioinformatics Analysis

Extensive bioinformatics analysis was performed on AprA to assess the function of each domain and the boundaries between the 5 domains. A well conserved AR domain exists in about half of GNAT-based loading modules, but it is not similar to any protein of known structure or function (Figure 4.3A and Figure B.1). The MT1 domain is also well conserved among loading modules that contain them but has very weak similarity of any MT of known structure (Figure 4.3A and Figure B.2). The high sequence conservation of the AR and MT1 domains also exists in the linker region, making it difficult to identify the boundaries of the domains. The putative AprA GNAT domain has weak similarity to other loading module GNAT domains (Figure 4.3A and B), and the C-terminal region is highly dissimilar. (Figure B.3). The dissimilarity in the C-terminal half of the GNAT domain continues into a poorly conserved region at the N-terminus of MT2. MT2 has strong similarity to PKS extension module CMT domains, but not other loading module MT domains such as MT1 (Figure 4.3C). The alignment of AprA

MT2 and PKS CMT domains (Figure B.4) informed the structural study of the CurJ CMT as published recently in *ACS Chemical Biology* (54). The uncertainly in the GNAT C-terminal region and the MT2 N-terminal region led to uncertainty in the boundary between the GNAT and MT2 domains.

Based on the bioinformatics analysis, a wide screen of AprA variants consisting of both individual domains and multidomains was carried out. Ultimately four variants were purified; full length AprA, AprA ΔACP, AprA di-domain, and AprA ACP (Table B.1). The ΔACP variant consisted of the AR, MT1, GNAT, and MT2 domains and the di-domain variant consisted of the GNAT and MT2 domains. The ACP domain was the only individual domain that could be purified. No excision points were identified for production of other single domains.

4.4.2 Decarboxylation Activity of AprA

To determine if AprA has decarboxylation activity through the use of alternative catalytic amino acids, we assessed the ability of the AprA to decarboxylate malonyl- and methylmalonyl-CoA. AprA di-domain was incubated with malonyl- or methylmalonyl-CoA for 4 hours and quantified by RP-HPLC. Under the conditions used, no acetyl- or propionyl-CoA products were detected (Figures 4.4A and B). The GNAT domain from the CurA loading module was used as a positive control and showed complete decarboxylation (Figures 4.4C and D). Previously the CurA GNAT was shown to catalyze the decarboxylation of malonyl-CoA with a k_{cat} of 1.8 per sec using the same assay (35). Thus, the AprA

GNAT is unlikely to be a decarboxylase, consistent with the lack of decarboxylase-essential His and Thr amino acids.

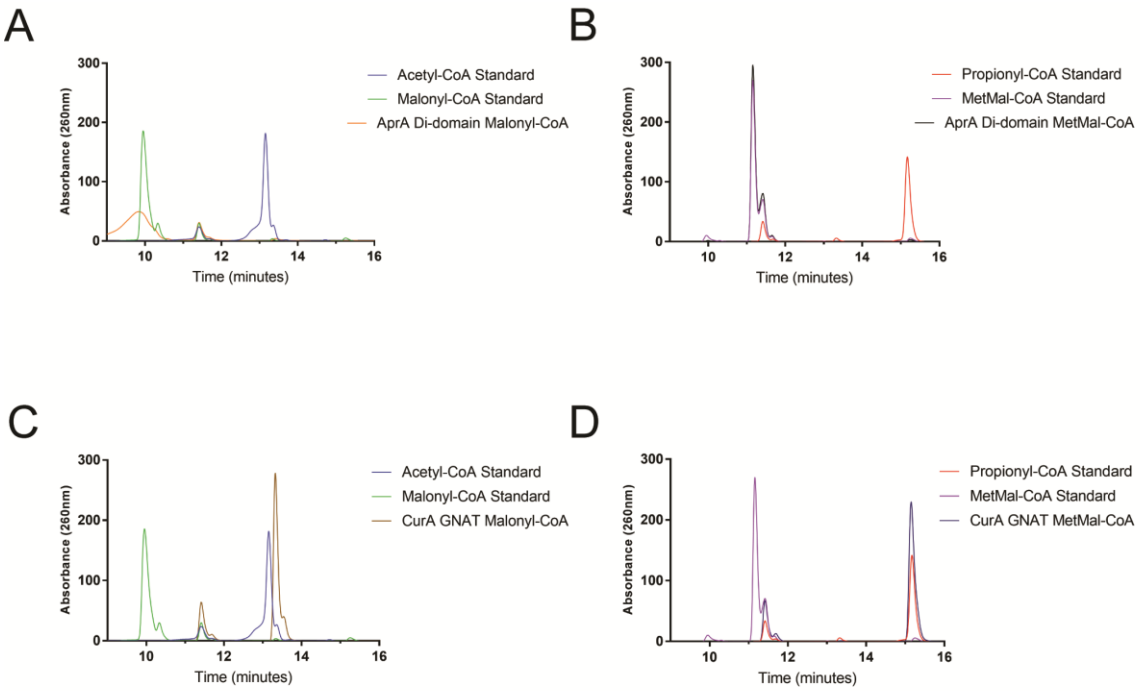

Figure 4.4 HPLC assay of AprA di-domain decarboxylation activity.

The AprA di-domain was incubated with either malonyl-CoA (A) or methylmalonyl-CoA (B) for four hours at 30°C to assess decarboxylation activity. The CurA GNAT was used as a positive control for both malonyl-CoA (C) and methylmalonyl-CoA (D). The peak at ~11.5 min is CoA.

4.4.3 Acyltransfer Activity of AprA

Next, an MS-based assay was developed to test whether AprA possesses acyltransfer activity and to determine which acyl-CoA is the preferred substrate. AprA ΔACP or di-domain was incubated with AprA ACP and acyl-CoA for 3 hours with time points taken approximately every 30 min. Reactions were monitored by Ppant ejection MS (83). The percent loaded ACP was determined by dividing the amount of loaded ACP by the total amount of ACP (holo plus loaded ACP). Both

AprA ΔACP and AprA di-domain displayed acyltransfer activity with a clear substrate preference: acetyl-CoA > propionyl-CoA >> malonyl-CoA > methylmalonyl-CoA (Figures 4.5A and B). A competition experiment performed with an equimolar mixture of the four acyl-CoAs showed the same trend for both AprA variants (Figure 4.5D). Thus, acetyl-CoA is the preferred *in vitro* substrate of those tested, with only very slow transfer of the negatively charged acyl groups. The two AprA variants had similar rates of acyltransfer for each acyl-CoA (Figure 4.5C). These data show that the charged, bulky carboxylic acid containing malonyl and methylmalonyl-CoA are much poorer acyltransfer substrates then are either the uncharged acetyl- or propionyl-CoA. Combined with the apparent lack of decarboxylation activity this indicates that neither malonyl- nor methylmalonyl-CoA are likely to be the *in vivo* substrate of AprA.

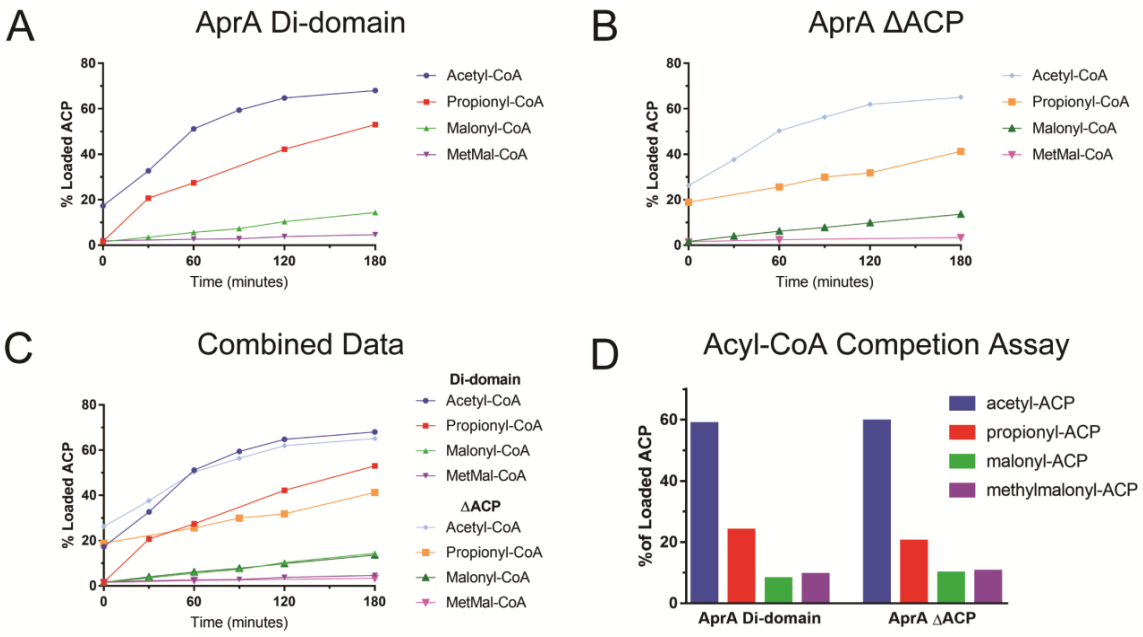

Figure 4.5 AprA acyltransfer activity.

AprA di-domain (A) or AprA ΔACP (B) was incubated with either acetyl-, propionyl-, malonyl, or methylmalonyl-CoA at 30°C with time points taken at 0, 30, 60, 90, 120, and 180 min. The % Loaded ACP indicates the conversion of holo-ACP to the appropriate acyl-ACP, as evaluated by LC/MS with Ppant ejection. C) AprA di-domain and AprA ΔACP had nearly identical activity with all substrates tested. D) Competition assays with both AprA di-domain and AprA ΔACP with an equimolar mixture of acyl-CoAs. The substrate selectivity was identical to the individual assays with a strong preference for acetyl-CoA.

4.4.4 Crystallization and overall structure of the AprA Di-domain

The AprA di-domain used for activity assays (amino acids 503-1050) was highly stable, but did not crystallize. The recent crystal structure of CurJ CMT (54), which is homologous to AprA MT2, suggested the possibility that the MT2 C-terminus was earlier in the sequence than previously thought. A new AprA di-domain (503-1022) was also highly soluble and crystallized readily. An initial crystal form grew as large hexagonal rods in a few hours at both 4°C and 20°C, but diffracted to only ~6Å and was recalcitrant to optimization. More extensive

screening produced a second crystal form of thin plates that grew in a few days (Figure 4.6B). The addition of SAM to the crystal drop generally improved both the size and morphology of the crystals. The second crystal form diffracted to ~2.5Å and was used to determine the di-domain structure (Table 4.1). The di-domain consists of two distinct, noninteracting domains connected by a short tether (Figure 4.6A). Overlay with the CurA GNAT reveals that the AprA GNAT is highly truncated (Figure 4.6C). In contrast, the AprA MT2 and the CurJ CMT are highly similar throughout the domain (Figure 4.6D).

4.4.5 Comparison of the AprA GNAT and CurA GNAT domains

Approximately half of the CurA GNAT domain is not present in the AprA GNAT (Figure 4.6C). Moreover, the positions of active site tunnels for the ACP and CoA Ppant groups in the CurA GNAT are blocked by loops in the AprA GNAT (Figure 4.7A). A topology diagram of the AprA and CurA GNAT domains shows that almost half of the central β-sheet, which is conserved in all GNAT superfamily members, is not present in the AprA GNAT (Figure 4.7B). A structure-based sequence alignment also clearly indicates that both decarboxylation-essential amino acids (His and Thr) in the CurA GNAT are absent in the AprA GNAT (Figure B.3). This leads to the conclusion that the AprA GNAT domain does not contain the catalytic machinery to perform either the decarboxylation or acyltransfer reactions. Therefore, my working hypothesis is that the GNAT domain is vestigial. Because the acyltransfer assay showed that the di-domain

catalyzes the acyltransfer reaction, I further hypothesize that the MT2 domain performs the acyltransfer reaction.

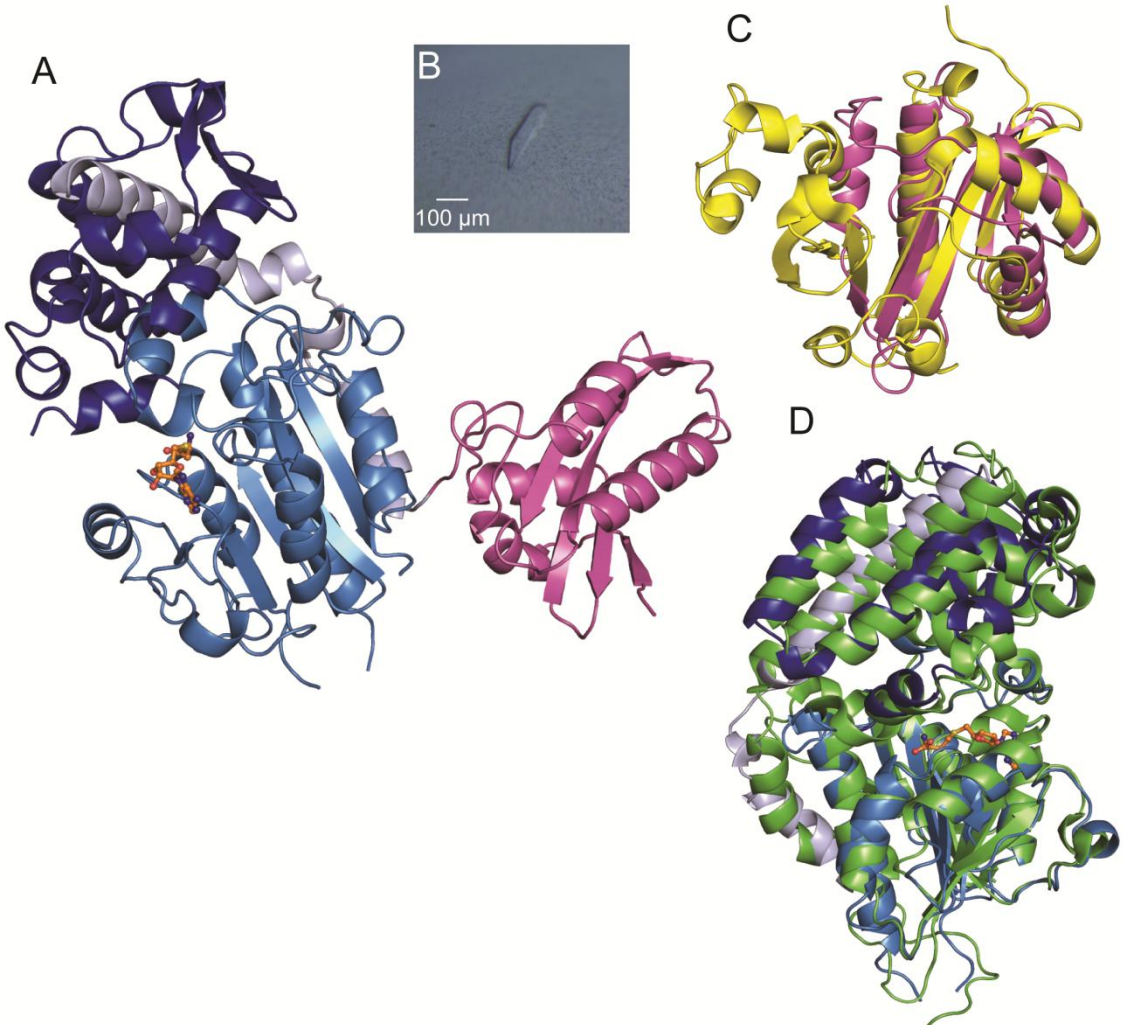

Figure 4.6 The structure of the AprA GNAT-MT2 di-domain.

A) Overall structure of the AprA di-domain, with magenta GNAT, light blue MT2 seatbelt helix, dark blue lid domain, sky blue core domain, and bound SAH in orange ball-and-stick. B) AprA di-domain crystal. C) Structural alignment of AprA GNAT domain (magenta) and the CurA GNAT domain (yellow). The overall RMSD is 2.7 Å for 66 Cα atoms. D) Structural alignment of the core domains of the AprA MT2 domain (blues) and the CurJ CMT domain (green) with SAH in orange ball-and-stick. The overall RMSD is 1.1 Å for 225 Cα atoms.

4.4.6 The AprA MT2 domain structure

Although the AprA GNAT domain is highly divergent from its closest homologues, the AprA MT2 domain is highly similar to the CurJ CMT (54) (Figure 7D). The similarity includes an N-terminal “seatbelt” helix, which in the CurJ CMT brings the N- and C-termini close together and is proposed to be important for proper domain arrangement in PKS extension modules (54). However, this N-terminal region of the CurJ MT has low sequence identity with the AprA MT2 (Figure B.4). This was a major source of uncertainty in the original analysis of AprA domain boundaries. All designs of single-domain AprA fragments assumed that GNAT had a similar length to the CurA GNAT when in fact the AprA GNAT is 75 amino acids shorter.

Strong electron density in the MT2 active site was easily fit as the *S*-adenosylhomocysteine (SAH) demethylation product of SAM, which was included in the crystallization solution (Figure 4.8A). The homocysteine portion of SAH was poorly ordered in the crystal structure. The homocysteine portion of SAH is known to interact with the GxGxG motif and was positioned accordingly. Superposition of the AprA MT2 and CurJ CMT structures shows good agreement of the SAH positions (Figure 4.8B).

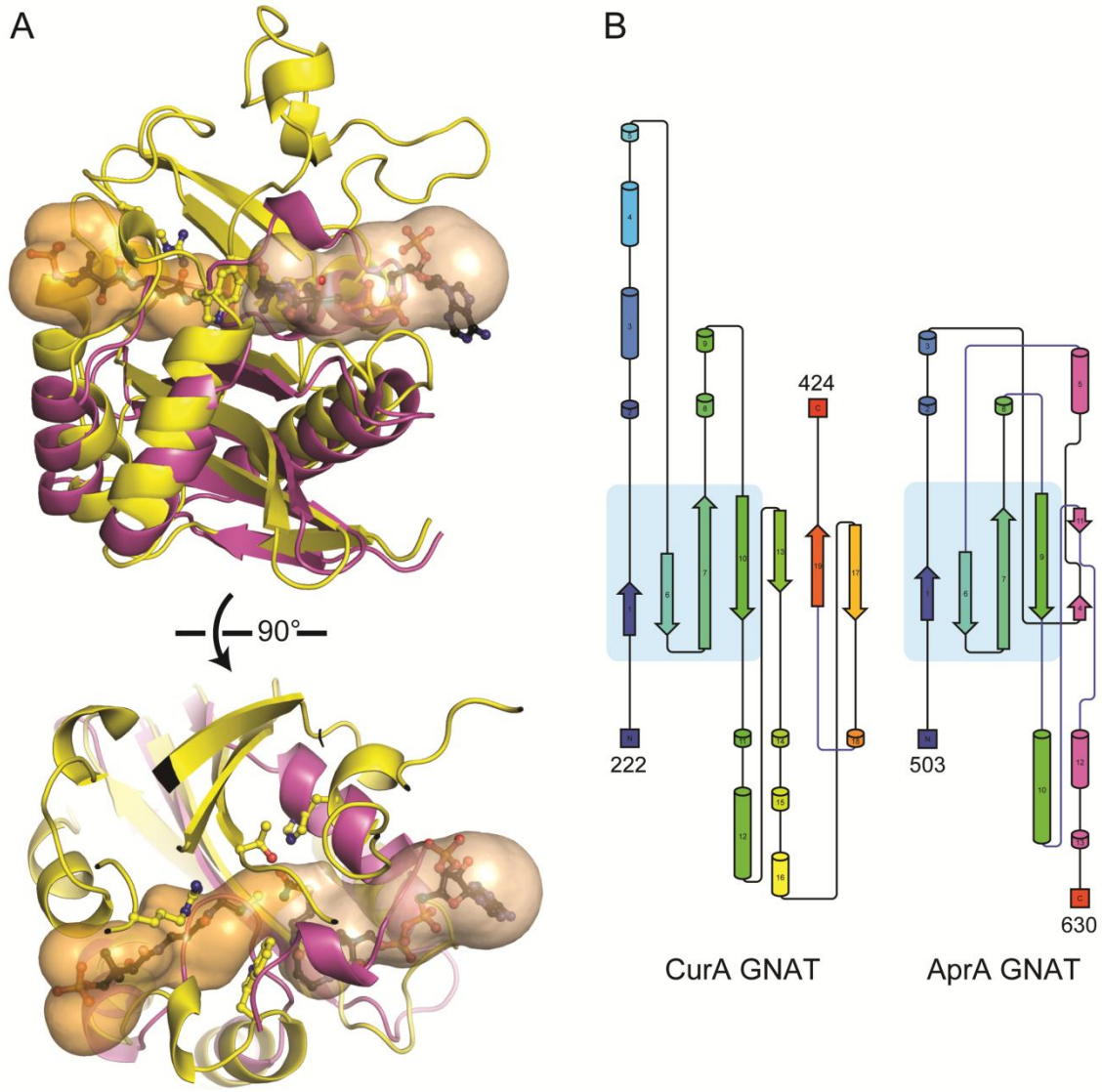

Figure 4.7 Comparison of the AprA and CurA GNAT domains.

A) Structural alignment of the AprA GNAT domain (magenta) and the CurA GNAT domain (yellow) with important CurA residues shown in ball and stick, the CurA bound acetyl-CoA and modeled ACP Ppant in black ball-and-stick, the CurA acetyl-CoA binding tunnel in wheat and the CurA Ppant binding tunnel in orange. B) Topology diagrams of the CurA GNAT and the AprA GNAT domains. Portions of the AprA GNAT that align to the CurA GNAT are shown in identical colors. The portion of the central β-sheet that is conserved between the structures is shown in a light blue box. Unique portions of the AprA GNAT structure are shown in magenta.

AprA MT2 has a long tunnel running along a cleft between the core and lid domains covered by a connecting “junction” helix (Figure 4.9A). The junction helix of the CurJ CMT is flexible and is proposed to interact with bound substrates (54). The junction helix of AprA MT2 also appears to be flexible and adopts a slightly different conformation than the CurJ CMT junction helix (Figure 4.9B). Two Phe side chains of the MT2 junction helix that face the active site, Phe793 and Phe797, are both poorly ordered (Figure 4.9C), supporting the earlier proposal. Further evidence of substrate induced conformation changes can be seen with the core and lid domains. With the core domains of CurJ CMT and AprA MT2 aligned (RMSD 0.6Å) it can also be seen that the lid domains are in slightly different positions relative to the core domains (Figure 4.6D). The lid domains themselves are also highly similar (RMSD 1.7Å). This reinforces the previous observation that the core and lid domains move relative to each other and that this movement likely changes the substrate tunnel in response to substrate binding (54). The His and Glu catalytic dyad amino acids shown to be critical for methylation by the CurJ CMT are present in identical positions in AprA MT2 (Figure 4.9A) (54).

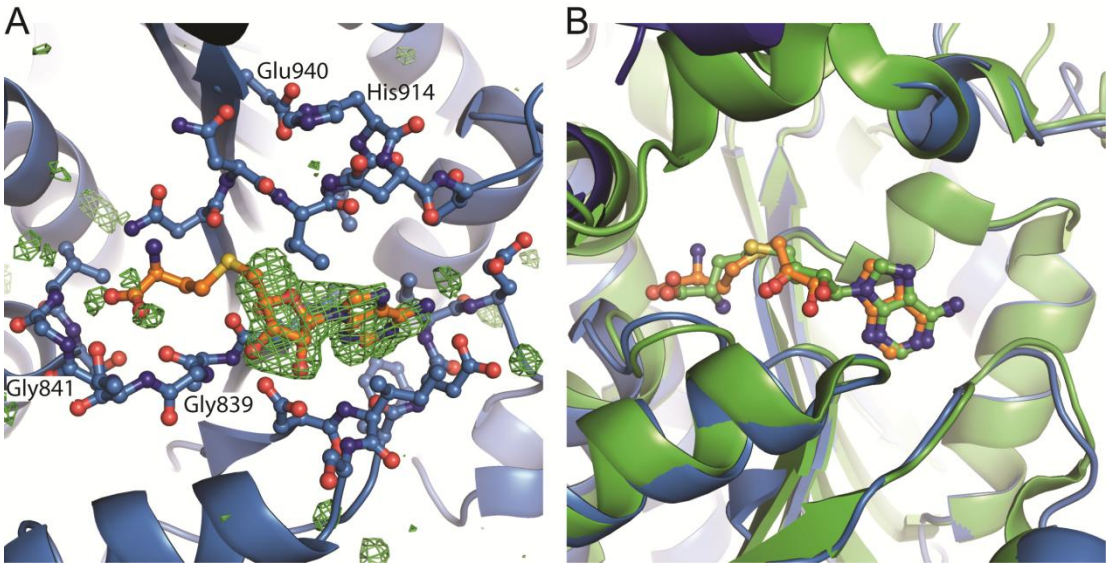

Figure 4.8 The AprA MT2 SAM binding site.

A) Fo-Fc omit map at 3σ generated with simulated annealing shown as green mesh. The homocysteine portion of SAH is poorly ordered. Protein shown in sky blue ball-and-stick or cartoon. SAH shown in orange ball-and-stick. His914 and Glu940 constitute the catalytic dyad. Gly839 and Gly841 make up part of the GxGxG SAM binding motif, Gly837 is obscured by the SAH ribose. B) Structural alignment of the core domains of the AprA MT2 domain (sky blue) and the CurJ CMT domain (green). SAH is shown as ball-and-stick in orange (AprA MT2) and green (CurJ CMT).

4.4.7 Methyltrasferase Activity of AprA

Prior to determining the crystal structure of the AprA di-domain, substantial effort was put into demonstrating the ability of the AprA MT domains to perform SAM-dependent methylation. Based on prior hypotheses of AprA function, efforts were focused on acetyl- and propionyl-ACP as substrates. No methylation was detected with either substrate. Methylation of malonyl or methylmalonyl groups was also not detected, consistent with previous reports on natural product MT domains (54,118,119).

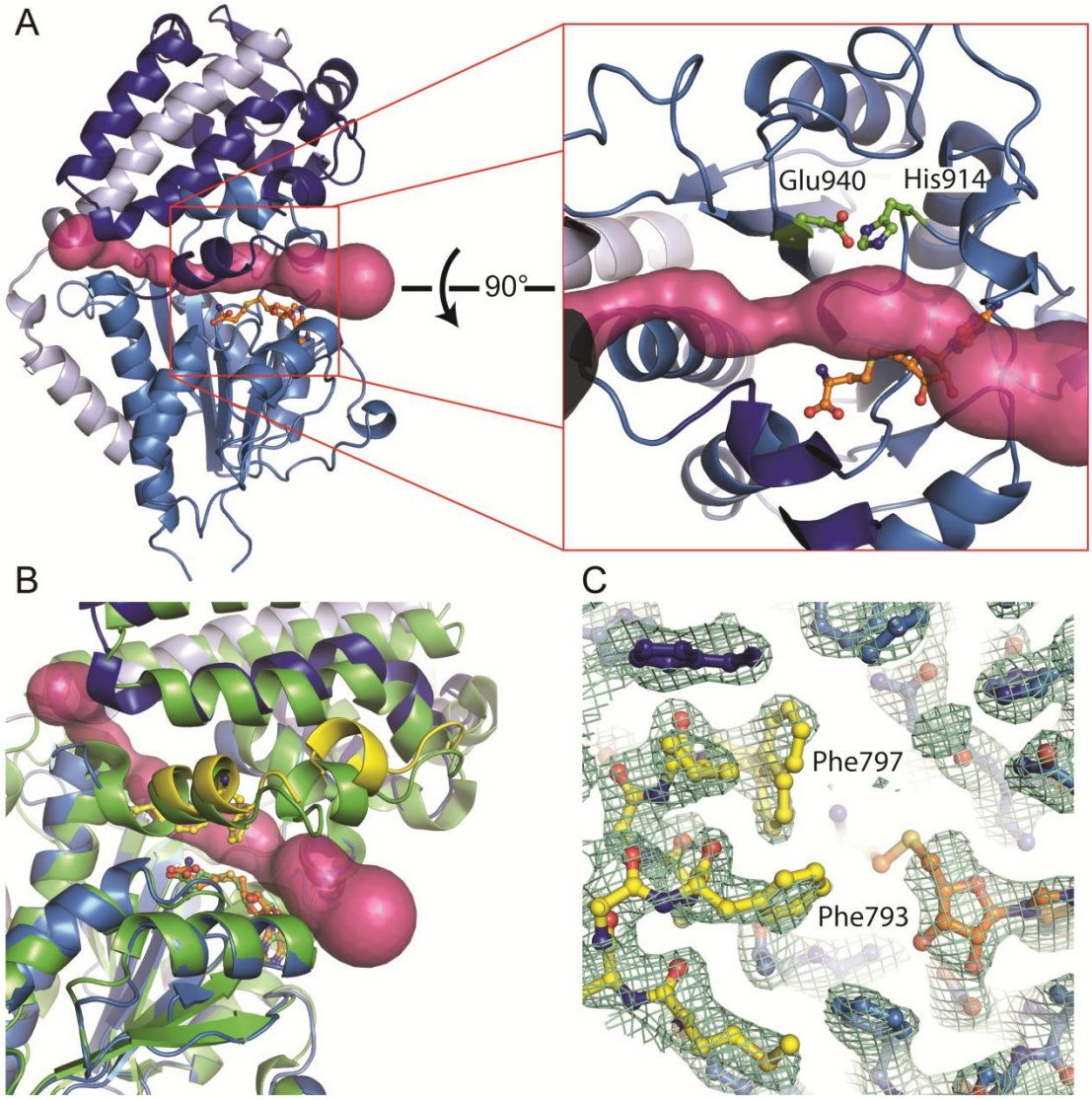

Figure 4.9 The AprA MT2 substrate binding tunnel and conserved active site.

A) The active site sits at the interface between the core and lid domains. The seatbelt helix of MT2 is shown in light blue, the lid domain in blue, the core domain in sky blue, the bound SAH in orange ball-and-stick, and the catalytic dyad in chartreuse ball-and-stick. The active site tunnel is shown as a warm pink surface created with Caver (120). B) The junction helix (yellow) adopts a new conformation in AprA MT2 compared to the CurJ CMT (green). C) The junction helix appears to be flexible, as demonstrated by two junction helix residues, Phe793 and Phe797, which face the active site and are poorly ordered. Phe797 is best modeled as two conformations with each at 50% occupancy. 2Fo-Fc map contoured at 1σ shown in green-cyan mesh.

Table 4.1 AprA di-domain X-ray Data Collection and Refinement Statistics

Data Collection	
Space group	*C*2
Cell dimensions a,b,c (Å)	152.0, 54.6, 109.7
α,β,γ (°)	90, 131.9, 90
X-ray source	APS 23ID-B
Wavelength (Å)	1.033
d_{min} (Å)	2.50 (2.59-2.50)[1]
R_{merge}	0.129 (1.22)
Wilson B factor	25.0
Avg I/σ(I)	11.18 (1.45)
Completeness (%)	98.0 (99.0)
Multiplicity	6.8 (7.0)
Total observations	159,576 (16,137)
$CC_{1/2}$	0.997 (0.512)
CC*	0.999 (0.823)
Refinement	
Data range (Å)	49.21-2.50
Reflections used in refinement	23,084
R_{work}/R_{free}	17.9/24.8
Number of non-hydrogen atoms	4,336
protein	4,107
ligands	27
water	202
Amino acid residues	508
Deviation from ideality	
bond lengths (Å)	0.008
bond angles (°)	1.05
Average B-factor	38.4
macromolecules	38.6
ligands	65.6
solvent	34.1
Ramachandran plot	
favored (%)	96.2
allowed (%)	3.6
outliers (%)	0.2

[1]Values in parentheses pertain to outermost shell of data.

The high structural similarity of the MT2 domain to the CurJ CMT domain (54), especially in their active sites, prompted the testing of acetoacetyl-ACP as a potential methylation substrate. Surprisingly, when incubated overnight with acetoacetyl-ACP and SAM, the AprA di-domain (503-1022) completely converted acetoacetyl to dimethyl-acetoacetyl (Figure 4.10). Currently, it is not known if acetoacetyl is the *in vivo* substrate of AprA, but these data indicate that the methyltransfer reaction requires a more reactive carbon then the acetyl or propionyl groups previously hypothesized.

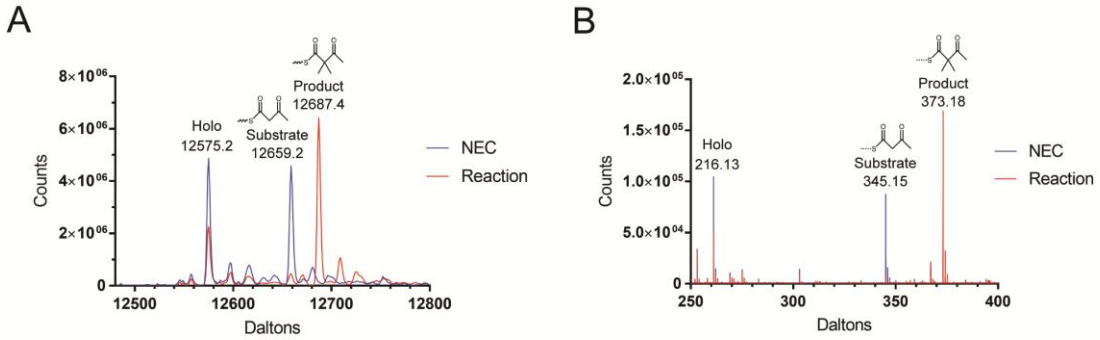

Figure 4.10 Methylation activity of AprA di-domain.

LC-MS analysis of AprA di-domain methyltransfer activity showing intact protein MS (A) and Ppant ejection MS (B). For the assay, Sfp-loaded acetoacetyl-ACP was incubated at 30°C with AprA di-domain and SAM for 17 hr. See Table B.2 for calculated and observed masses. NEC: No Enzyme Control.

4.5 Conclusion and Future Directions

The data presented here show that the AprA loading module diverged from other GNAT-based loading modules. Initial experiments showed that AprA does not possess the acyl-CoA decarboxylation activity previously observed with the CurA GNAT. In addition, AprA was shown to prefer acetyl- and propionyl-CoA

substrates for the acyltransfer reaction *in vitro*. It was therefore hypothesized that either acetyl- and propionyl-ACP was the initial substrate for the methyltransferase domains. Extensive efforts to elicit methylation on either acetyl- or propionyl-ACP were unsuccessful.

The recently solved structure of the CurJ CMT (54) provided valuable insights into the AprA MT2 domain that allowed for the determination of the AprA di-domain structure and subsequent demonstration of acetoacetyl-ACP di-methylation. These recent findings provide a tantalizing hint at the *in vivo* function of AprA, and indicate that previous thinking on the function of AprA requires significant re-evaluation. The following is a brief list of points that must be addressed to begin to understand AprA.

The biggest outstanding question is, what are *in vivo* substrates of AprA? Previously it was thought that the GNAT domain transferred an acetyl or propionyl group from CoA to the AprA ACP followed by three or two, respectively, SAM-dependent methylations to generate the *t*-butyl group. The revelation that acetoacetyl-ACP is an MT2 methylation substrate indicates that a more activated carbon is required for methylation. This leads to a number of questions regarding a synthetic scheme for generating both an activated carbon and the final *t*-butyl group. Obviously much more study of both the acyltransfer and methylation reactions is required.

As had been noted previously, the sequence alignments for the N-terminal portions of MT2 domain and most of the GNAT domain were poor. This led to an unclear boundary between the two domains. The di-domain structure has allowed for the domain boundaries between the two domains to be assigned. Additionally, the lack of interaction between the domains indicates that they could be separable. Producing individual domains would allow for testing of the hypothesis that MT2 catalyzes the acyltransfer reaction in place of a catalytically dead, vestigial GNAT domain.

The AprA MT1 domain is conserved in a number of other GNAT-based loading modules that apparently require methylation to generate their products (Figure 4.2). Thus far, no biochemical characterization of these MT domains has been reported. Our demonstration of methylation of acetoacetyl-ACP by the AprA MT2 again indicates that the previously proposed reaction schemes are likely to be incorrect. AprA MT1 likely plays a role in the generation of the *t*-butyl group, but with the current number of outstanding questions it is difficult to ascribe MT1 a function. Investigating MT1 function through the study of the MT domain of a homologous loading module may significantly reduce the complexity of experiments by eliminating MT2. These and many more questions about AprA function must be addressed to fully understand the function of this remarkable enzyme.

4.6 Methods and Materials

4.6.1 Cloning and Sequence Alignments

AprA residues 2-1138 were amplified from *Moorea bouillonii* fosmid 11H10 (4) by PCR and inserted into pMCSG7 by LIC (105) to create pMCSG7-*AprA* and verified by sequencing. Subsequent constructs were created by subcloning from pMCSG7-*AprA* using identical methods. See Table B.1 for full list of constructs and primers. Sequence alignments were created using Jalview (76,77) using the Clustal O (78).

4.6.2 Bacterial strains, media and culture conditions

Cultures were grown in Terrific Broth (TB) medium containing 4% glycerol. Ampicillin (100 µg/mL, pMCSG7), spectinomycin (50 µg/mL, pRARE), and chloramphenicol (35 µg/mL, pG-KJE8) were used with the corresponding plasmids in *E. coli* cultures. *E. coli* strain BAP1 (65) was transformed with pMCSG7-*AprA_ACP*. Strain BL21(DE3) containing the pRARE (106) and pG-KJE8 (Takara Bio) plasmids was transformed with pMCSG7-*AprA* and pMCSG7-*AprA_ΔACP*. Remaining plasmids were transformed into strain BL21(DE3) containing no additional plasmids. Cultures were grown at 37°C in 500 mL TB to an OD_{600} of 1.5, cooled to 20°C for at least 1 hr, induced with IPTG (final concentration 0.2 mM), and grown for an additional 14-18 hr before harvesting by centrifugation. Cell pellets were stored at -20°C until used for protein purification.

4.6.3 Purification of AprA Proteins

Cell pellets were resuspended in approximately 5 mL Buffer A (50 mM HEPES pH 7.0, 50 mM $AmSO_4$, 10% glycerol) containing 15 mM imidazole per 1 g cell paste. All purification steps were performed at 4°C. Resuspended cells were incubated with DNase (4 mg), lysozyme (10 mg), and $MgCl_2$ (4 mM) for 30 min, then lysed by three passes though an Avestin EmulsiFlex-C3 homogenizer and clarified by centrifugation at 30,000 xg. The soluble fraction was loaded onto a 5-mL HisTrap Ni NTA column (GE Healthcare) in Buffer A and eluted with an imidazole gradient from 30-300 mM. The eluted protein was further purified via size exclusion chromatography with a HiLoad 16/600 Superdex 200 column (GE Healthcare) in Buffer A. Final fractions were pooled, concentrated, flash cooled with liquid N_2, and stored at -80°C. AprA ACP was purified identically to AprA enzymes except the initial buffer was 100 mM Tris pH 7.5, 500 mM NaCl, 10% glycerol, 5 mM TCEP and the size exclusion chromatography buffer was 100 mM Tris pH 7.5, 250 mM NaCl, 5% glycerol, 5 mM TCEP.

4.6.4 Purification of AprA di-domain for Crystallization

AprA di-domain (503-1022) for crystallization was purified identically to other AprA proteins except Buffer C (100 mM Tris pH 7.9, 500 mM NaCl, 5% glycerol) was used in the initial HisTrap Ni NTA column, and steps of TEV protease cleavage and second HisTrap Ni NTA column were added to the protocol. The eluted protein from the first HisTrap Ni NTA was cleaved with TEV protease (1:30 ratio TEV:AprA di-domain with 2 mM DTT) and dialyzed into Buffer D (50 mM

Tris pH 7.9, 100 mM NaCl, 10% glycerol). The digested protein was run over a second 5-mL HisTrap Ni NTA column. The cleaved AprA protein in the flow-through fraction was purified via size exclusion chromatography with a HiLoad 16/600 Superdex 200 column (GE Healthcare) in Buffer E (50 mM Tris pH 7.9, 100 mM NaCl). Final fractions were pooled, concentrated, flash cooled with liquid N_2, and stored at -80°C.

4.6.5 Purification of CurA GNAT

Plasmid pMCSG7-*GNAT*$_L$ (35) was used to express CurA GNAT, which was purified identically to the AprA di-domain protein except the buffer for the HisTrap Ni NTA column was 100 mM Tris pH 7.9, 300 mM NaCl, 10% glycerol, TEV protease cleavage was done in Buffer F (20 mM Tris pH 7.9, 300 mM NaCl, 10% glycerol). The tag-free protein was further purified via size exclusion chromatography with a HiLoad 16/600 Superdex 200 column (GE Healthcare) in Buffer F. Final fractions were pooled, concentrated, flash cooled with liquid N_2, and stored at -80°C.

4.6.6 Crystallization

AprA di-domain (503-1022) was crystallized from a 5 mg/mL stock solution in Buffer E with 1 mM SAM using micro seeding from seed stocks generated using Seed Beads (Hampton). Crystals grew in 2-4 days at 20°C by vapor diffusion in 1.7 µL sitting drops containing 0.75 µL protein stock, 0.2 µL seed stock, and 0.75 µL reservoir solution (0.01-0.05M trimethylamine N-oxide, 12-17% PEG 8000,

0.12M Tris pH 7.5). Crystals were harvested in nylon loops, cryoprotected with reservoir solution containing 10% glycerol, and flash cooled in liquid N_2.

4.6.7 Diffraction Data Collection and Structure Determination

Diffraction data were collected at 100 K on GM/CA beamline 23ID-B at the Advanced Photon Source (APS) at Argonne National Laboratory (Argonne, IL). For each crystal, 360º of data were collected using 0.2º rotation and 0.2 sec exposure per image (Table 4.1). Data were processed using XDS (68). The structure of AprA di-domain with bound SAH was solved by molecular replacement (MR) with Phaser (69) in the PHENIX software suite (70) using search models created with sculptor (121) from the CurJ CMT (33% identity, PDB code: 5THY) (54) and CurA GNAT (24% identity, PDB code: 2REE) (35) as search models. The initial model from MR was modified with AutoBuild (71) to generate a 77% complete model of AprA di-domain. The remaining portion of model was completed manually using Coot (72). Refinement was performed using phenix.refine (73). Images and figures were prepared using PyMOL (82) and CAVER (120). Topology diagrams were generated using Pro-origami (122). The structure was validated with MolProbity (81).

4.6.8 CoA Decarboxylation Assay

The decarboxylation activity of AprA proteins was investigated using a modified version of the decarboxylation assay previously used to characterize the CurA GNAT (35). Briefly, 10 μM of enzyme (di-domain (503-1050)) was combined with

500 µM of either malonyl of methylmalonyl-CoA in reaction buffer (10 mM $MgCl_2$, 5 mM TCEP, 100 mM Tris pH 7.5). Reactions were run at 30°C for 4 hours and then quenched with 10% formic acid. HPLC analysis was performed on an Agilent 1290 HPLC system by RP-HPLC (Luna C18(2) 5µ C18 column 100Å, 250 X 4.6 mm) with a 30 min gradient of 3-90% methanol with 10mM ammonium acetate at a flow rate of 1.0 mL/min. CurA GNAT was used as a positive control.

4.6.9 Acyl Transfer Activity Assay

The acyl transfer activity of AprA proteins was investigated using a mass spectrometry based phosphopantetheine (Ppant) ejection assay (83). 25 µL reaction mixtures consisted of 100 mM Tris pH 7.5, 10 mM $MgCl_2$, 5 mM TCEP, 500 µM acyl-CoA, 10 µM AprA enzyme (di-domain (503-1050) or ΔACP), and 50 µM AprA holo-ACP. Reactions were incubated at 30°C until quenched with 10% formic acid, time points were taken at 0, 30, 60, 90, 120, and 180 min. The competition assay was run using a 1:1:1:1 mix of acetyl-, propionyl-, malonyl-, and methylmalonyl-CoA (500 µM each) for 180 min. A reaction mixture without an AprA enzyme was used as a negative control. Reaction mixtures were prepared for mass spectrometry (MS) by binding to a spin column packed with 40 µ C4 resin (Analtech B265A BME) equilibrated with water containing 0.1% formic acid. Samples were washed twice with 120 µL water containing 0.1% formic acid, and eluted with 30 µL acetonitrile with 0.1% formic acid. Samples were spun at 4,000 xg for 2 min after each step. 10 µL of prepared sample was injected onto an Agilent 6520 Q-TOF mass spectrometer equipped with an Agilent 1290 HPLC

system. Samples were separated for MS analysis by RP-HPLC (Aeris widepore C4 column 3.6 µM, 50 X 2.10 mm) with a 10 min gradient of 10-95% acetonitrile with 0.1% formic acid at a flow rate of 0.5 mL/min and Ppant ejection data was (83). Ppant ejection data were analyzed using the Agilent Mass Hunter Qualitative Analysis software with the maximum entropy deconvolution algorithm. The percent loaded ACP with each of the four acyl groups was calculated for each time point using the area of the ejected Ppant peaks and the following equation:

$$\%\ Loaded\ ACP_{acyl} = \frac{ACP_{acyl}}{ACP_{holo}+ACP_{acetyl}+ACP_{propionyl}+ACP_{malonyl}+ACP_{methylmalonyl}}$$

4.6.10 Methyltransferase Activity Assay

Methyltransferase activity was assessed using an MS-based Ppant ejection assay similar to that used to assess acyl transfer activity. 50 µL reaction mixtures consisted of 50 mM HEPES pH 7.4, 1 mM $MgCl_2$, 500 µM acetoacetyl-CoA, 25 µM AprA Di-domain (503-1022), 10 µM Sfp (66), and 100 µM apo AprA ACP. Reactions were incubated at 30°C for 17 hrs and then quenched with 1% formic acid. A reaction mixture without enzyme was used as a negative control. 0.5 µL of reaction mixture was injected onto an Agilent 6545 Q-TOF mass spectrometer equipped with an Agilent 1290 HPLC system and separated for MS analysis by RP-HPLC (Aeris widepore C4 column 3.6 µM, 50 X 2.10 mm) with a 10 min gradient of 10-95% acetonitrile with 0.1% formic acid at a flow rate of 0.5 mL/min. Intact protein and phosphopantetheine ejection data were analyzed using the

Agilent Mass Hunter Qualitative Analysis software with the maximum entropy deconvolution algorithm.

Appendix B

Supplementary Information for Chapter IV

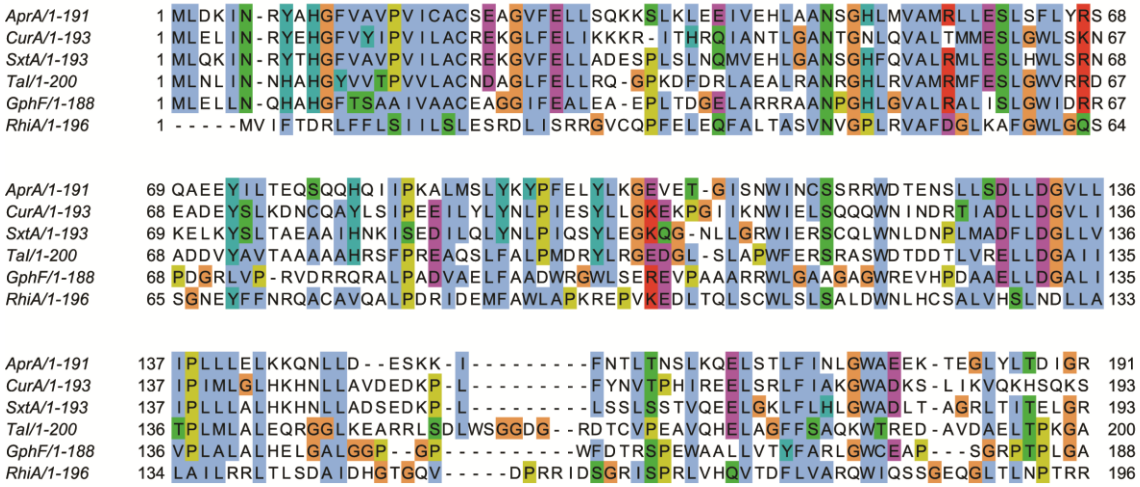

Figure B.1 Sequence alignment of loading module AR domains.

Amino acid numbers in the figure correspond to the AR domain within each multidomain protein. Protein abbreviations (GenBank accession codes) are as follows: CurA - curacin A (AEE88289.1), SxtA - saxitoxin (ABI75094.1), TaI - myxovirecin (WP_011553948.1), GphF - gephyronic acid (AHA38199.1), RhiA - rhizoxin (WP_013435483.1).

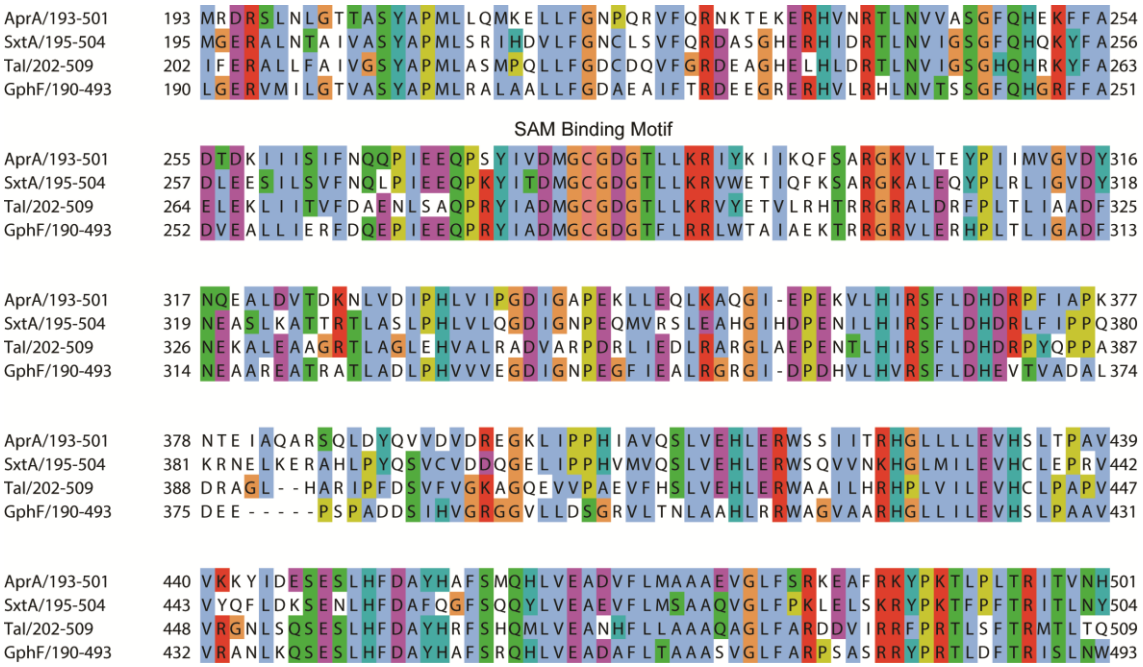

Figure B.2 Sequence alignment of AprA MT1 and loading module MT domains.

The SAM binding consensus sequence is indicated. Amino acid numbers in the figure correspond to the MT domain within each multidomain protein. Protein abbreviations (GenBank accession codes) are as follows: SxtA - saxitoxin (ABI75094.1), TaI - myxovirecin (WP_011553948.1), GphF - gephyronic acid (AHA38199.1).

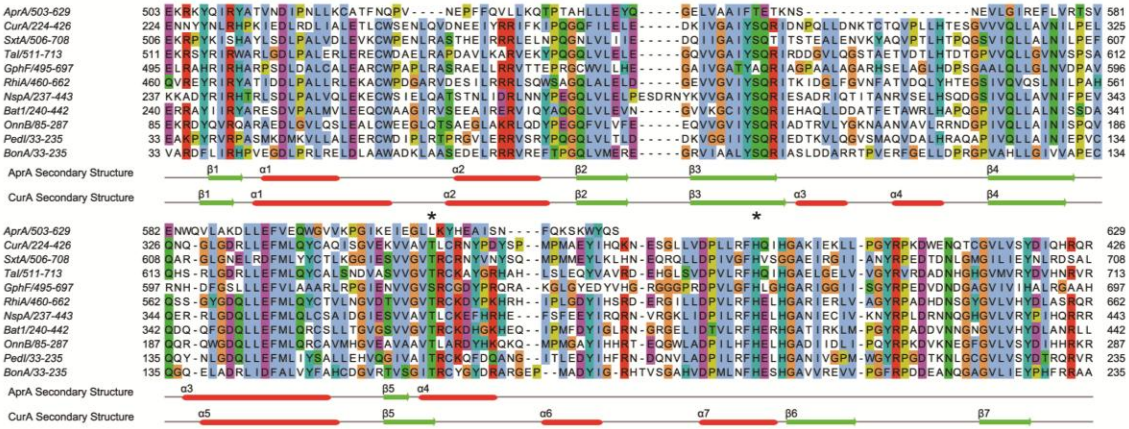

Figure B.3 Sequence alignment of loading module GNAT domains with AprA and CurA secondary structure notations.

The conserved decarboxylation residues identified in the CurA GNAT are marked with asterisks. Amino acid numbers in the figure correspond to the GNAT domain within each multidomain protein. Protein abbreviations (GenBank accession codes) are as follows: CurA - curacin A (AEE88289.1), SxtA - saxitoxin (ABI75094.1), TaI - myxovirecin (WP_011553948.1), GphF - gephyronic acid (AHA38199.1), RhiA - rhizoxin (WP_013435483.1), NspA - nosperin (ADA69237.1), Bat1 - batumin (WP_052451043.1), OnnB - onnamide (AAV97870.1), PedI - pederin (AAR19304.1), BonA - bongkrekic acid (AFN27480.1).

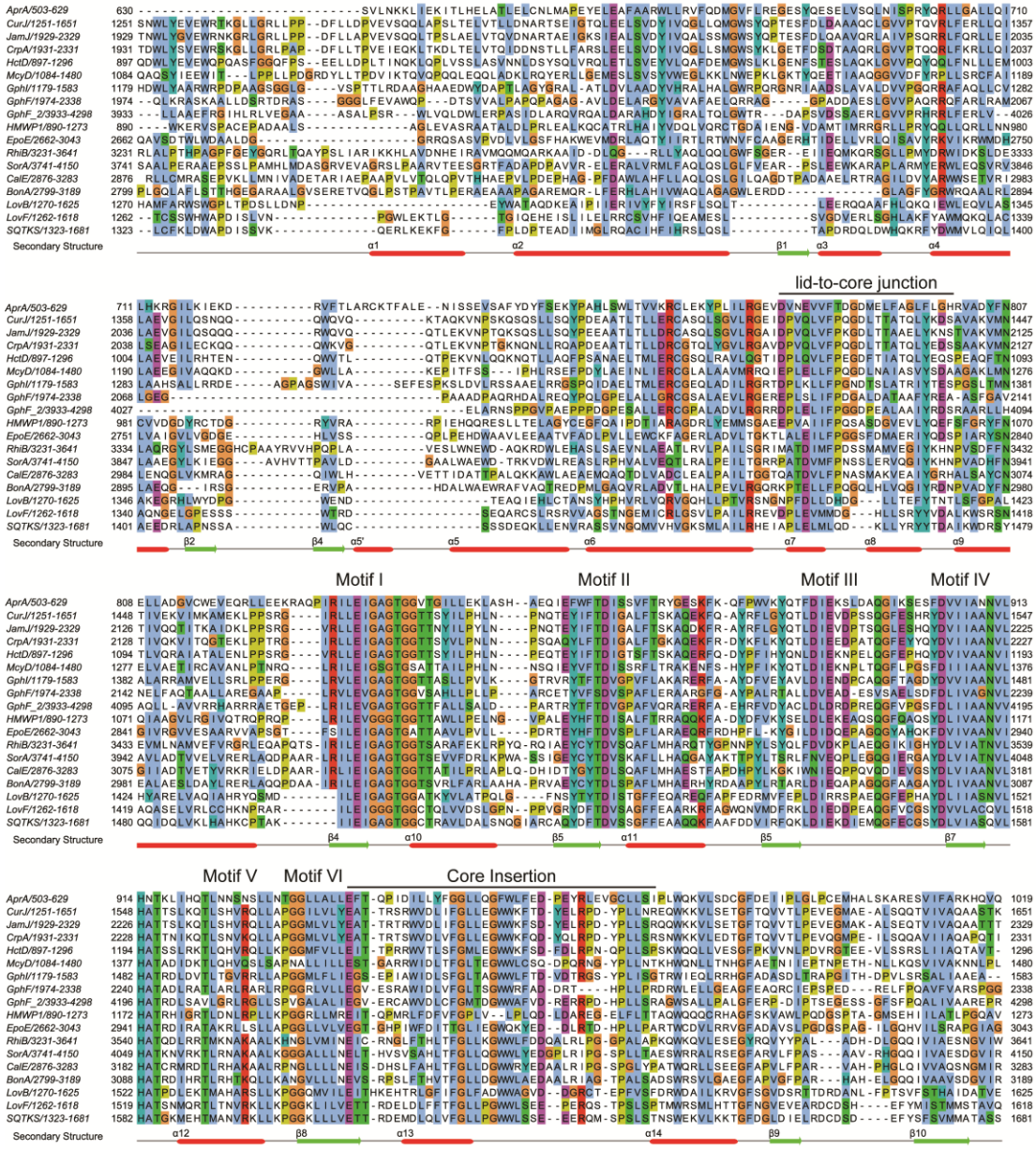
lid-to-core junction
Motif I
Motif II
Motif III
Motif IV
Motif V
Motif VI
Core Insertion
Secondary Structure

Figure B.4 Sequence alignment of the AprA MT2 domain and PKS *C*-MT domains with AprA secondary structure notation.

MT motifs are labeled. Amino acid numbers in the figure correspond to the MT domain within each multidomain protein. Pathway abbreviations (GenBank accession codes) are as follows: Cur - curacin A (HQ696500.1), Jam - jamacamide (AY522504.1), Crp - cryptophycin (EF159954.1), Hct - hectochlorin (AY974560.1), Mcy - microcystin (KC699835.1), Gph - gephyronic acid (KF479198.1), HMWP1- yersiniabactin (AE009952.1), Epo - epothilone (AF217189.1), Rhi - rhizoxin (AM411073.1), Sor - sorangicin (HM584908.1), Cal - calyculin (AB933566.1), Bon - bongkrekic acid (AFN27480.1), Lov - lovastatin (AF151722.1, AH007774.2), SQTKS - squalestatin synthase (AY217789.1).

Table B.1 AprA Fragments, Expression Plasmids, Expression Strains, and Primers

Protein	Plasmid	Start	End	Forward Primer	Reverse Primer	Expression Strain
AprA	pMCSG7-*AprA*	2	1138	TACTTCCAATCCAATGCACTAGATAAAATAAATCGTTATGCTCATGGGTTTGTAGC	TTATCCACTTCCAATGCTAACTAAAATTAATATCTTCCCGTTCTGAAAGATACTCTGC	BL21(DE3)+pRARE+pG-KJE8
AprA ACP	pMCSG7-*AprA_ACP*	1058	1138	TACTTCCAATCCAATGCTGAAATTTTTGAACAGGAATGTC	TTATCCACTTCCAATGTTAACTAAAATTAATATCTTCCCGTTCTG	Bap1
AprA ΔACP	pMCSG7-*ΔACP*	2	1034	TACTTCCAATCCAATGCACTAGATAAAATAAATCGTTATGCTCATGGGTTTGTAGC	TTATCCACTTCCAATGTTAATTTTCTGTCAAATTTTGTTTAATTTTTTC	BL21(DE3)+pRARE+pG-KJE8
AprA di-domain for assay (503-1050)	pMCSG7-*503-1050*	503	1050	TACTTCCCATCCAATGCAGAAAAGCGCAAATATCAGATACGGTATGCAACT	TTATCCACTTCCAATGTTATGAATTATTAATACTAATAAAATCGAATTCTGC	BL21(DE3)
AprA di-domain crystallized (503-1022)	pMCSG7-*503-1022*	503	1022	TACTTCCCATCCAATGCAGAAAAGCGCAAATATCAGATACGGTATGCAACT	TTATCCACTTCCAATGTTATGTTTTCTCCTGTACTTGATGC	BL21(DE3)

Table B.2: Expected and Observed Ion Masses for AprA Mass Spectrometry Experiments

Species	**Intact protein calculated mass (Da)**	**Intact Protein Observed Mass (Da)**	**Ppant ejection calculated mass (Da)**	**Ppant ejection observed mass (Da)**
apo ACP	12235.0	None	N/A	N/A
holo ACP	12575.1	12575.2	261.13	216.13
acetoacetyl ACP	12659.1	12659.2	345.15	345.15
methyl-acetoacetyl ACP	12673.1	None	359.16	None
dimethyl-acetoacetyl ACP	12687.1	12687.4	373.18	373.18

Chapter V

Conclusions and Future Directions

Conclusions

5.1 Overview

Initiation is a critical step in natural product biosynthesis. The studies presented here broadened our understanding of the initiation steps though structural and biochemical investigations of three biosynthetic pathways: the olefin synthase of the marine cyanobacterium *Synechococcus sp.* PCC 7002, the cahuitamycin pathway of *Streptomyces gandocaensis*, and the apratoxin pathway of the marine cyanobacterium *Moorea bouillonii*. Observations are presented supporting a proposed mechanism of fatty acid chain-length selectivity used by the olefin synthase FAAL enzyme from *Synechococcus sp.* PCC 7002. Additionally, a rational structure-based explanation of substrate selectivity by the naturally promiscuous AMP-ligase CahJ from the cahuitamycin pathway was developed. Lastly, the structure of the GNAT-MT2 di-domain as well as the first activity data from the poorly understood AprA loading module from the apratoxin pathway is presented. These investigations have potential to form the basis of useful applications of the biosynthetic pathways.

5.2 An electrostatic fatty acid length selection mechanism by SynFAAL

The first enzyme of olefin synthase pathways is an FAAL domain, which activates a fatty acids using ATP and transfers them to an ACP. *In vivo* studies indicated high selectively for stearic acid, since only 1-nonadecene was observed (3,30). The structure of the FAAL domain from the *Synechococcus* PCC 7002 olefin synthase pointed to an electrostatic mechanism that selects against C16 fatty acids and a substrate tunnel that is unable to accommodate fatty acids longer than 18 carbons. Combined with data showing that fatty acids shorter than 16 carbons are rare in *Synechococcus* PCC 7002 (29), the preference for C18 substrates provides an explanation for the high *in vivo* selectivity of the enzyme. This was further supported with mutagenesis experiments to decrease the electrostatic charge of the substrate tunnel, which resulted in improved hexadecanoic acid utilization. In addition, mutagenesis to introduce larger amino acid side chains to occlude the fatty acid binding tunnel demonstrated that the substrate selectivity of the enzyme could be tailored to accommodate different substrates.

5.3 Understanding the substrate selectivity of the CahJ

The naturally promiscuous AMP-ligase CahJ natively utilizes SA and 6-MSA as substrates. The biochemical and structural investigations of CahJ presented here provided both the scope of substrate utilization and a rational explanation of substrate utilization. In brief, the CahJ substrate binding site consists of a flat, hydrophobic site surrounded by binding pockets that correspond to substituent

positions of the central aromatic ring. The 2 position contains an Asn that forms a hydrogen bond with the 2-hydroxy group of salicylic acid. The 3 position is sterically restricted by Leu and Cys side chains and does not accommodate substituents well (Figure 3.5). The binding pockets at the 4, 5, and 6 positions are all large enough to accommodate methyl substituents; however, slight rotations of the aromatic ring are required for each substituent to fit into its binding pocket, making it difficult for multiple substituents to be present simultaneously. This understanding guided the creation of a new cahuitamycin congener, cahuitamycin F, which has anti-biofilm activity similar to cahuitamycin A.

5.4 Investigations into AprA, the loading module of the apratoxin A biosynthetic pathway

AprA, the loading module of the apratoxin biosynthetic pathway is a GNAT-based loading module, homologs of which are found at the beginning of approximately a dozen pathways. The structural data presented here shows that the GNAT domain is truncated and likely not catalytically competent. An AprA di-domain consisting of the GNAT and MT2 domains retains acyltransfer activity indicating that the MT2 domain may perform an acyltransfer reaction. If so, this would constitute the discovery of new chemistry for an enzyme with a type I methyltransferase fold. Also presented here is data showing the AprA GNAT-MT2 di-domain performs the SAM dependent methylation of acetoacetyl-ACP to

form dimethyl-acetoacetyl-ACP, the first demonstration of methylation activity for a GNAT-based loading module.

Future Directions

5.5 Olefin production by the intact olefin synthase

The biological production of length-specific olefins has potential applications in a number of contexts, including chemical feedstocks, lubricants, surfactants, and biofuels. The successful biological production of olefins for commercial use will require an understanding of the olefin synthase pathways that produces them. Olefin synthases are composed of three sections; the initiation FAAL domain, a canonical PKS extension module, and the terminal ST and TE domains. Previous studies of olefin synthases have elucidated the function of the ST and TE domains (32-34). The canonical PKS extension module has not been studied directly, but can be assumed to function in manner consistent with other PKS extension modules. With the additional findings presented here, all of the individual pieces of olefin syntheses are now reasonably well understood. The next step toward biological olefin production is investigating the full-length olefin synthase. Studies with the full-length olefin synthase are critical to determine if results showing the ability to control to length of the fatty acid selected by the FAAL domain can be propagated through the remainder of the olefin synthase, leading to the production of shorter olefin chains. This work would likely encounter technical hurdles; first, full-length olefin synthase is a very large protein (~300 kDa) that has not been previously purified and could be difficult to

produce in significant quantities. Second, the detection of free olefins at low concentrations to monitor the reaction could be challenging. This issue may be circumvented by monitoring the turnover of a co-factor such as NADPH by the KR domain, but the ideal method would detect olefin production directly. An assay of olefin production would allow the olefin synthase to be engineered to optimize olefin production for both chain-length specificity and yield. For the production of olefins to be practical for large-scale use they must produced *in vivo*. The best candidate would be likely a cyanobacterium because olefin synthases are natively found in cyanobacteria and should therefore be well expressed in cyanobacteria, and cyanobacteria are capable of carbon fixation and would require minimal nutrient input.

5.6 Investigating short-chain FAAL enzymes

Another route of investigation that furthers the goal of length-specific olefin production would be to characterize a short-chain FAAL enzyme. The data presented here showed that SynFAAL had low activity with fatty acid substrates shorter than 10 carbons. Instead of engineering SynFAAL to utilize short-chain fatty acids, it may be more effective to swap the SynFAAL for an FAAL that natively uses shorter fatty acids. A potential candidate would be JamA, which activates hexanoic acid (58,59). A methodical investigation of JamA substrate selectivity could be combined with investigations of the full-length olefin synthase by the replacement of SynFAAL with JamA to access olefin chain lengths not possible with the SynFAAL. If successful, the replacement of SynFAAL with other

FAAL enzymes, such as the medium-chain length enzyme ColA, may increase olefin production.

5.7 Engineering the CahJ Substrate Binding Site

A key factor in the success of the cahuitamycins as biofilm inhibitors may be the ability to easily produce new congeners. The work presented showed the native substrate range of CahJ and indicated that many more cahuitamycin congeners could be produced. However, as the biological mechanism of cahuitamycin activity is revealed, starter molecules incompatible with wild type CahJ may be needed for the creation of desired cahuitamycin congeners. Therefore, protein engineering to expand and/or alter CahJ substrate selectivity could have important future applications. Understanding the substrate binding requirements of CahJ has provided a roadmap for engineering to expand its substrate range. The 3 position pocket should be amenable to protein engineering given the natural variability at that position (Figure A.5B). Val330 at the bottom of the active site separates the 4 and 5 pockets and a glycine replacement would open a large binding volume (Figure A.5C). It may also be possible to restrict currently open binding pockets to create variants of CahJ with high selectivity for a desired substrate, thereby increasing production of a desired cahuitamycin congener while reducing the production of unwanted cahuitamycin congeners.

5.8 Continued biochemical investigations of AprA

The insights provided by CurJ CMT rapidly lead to the structure of the AprA GNAT-MT2 di-domain by indicating that the di-domain C-terminus should be shortened by nearly 30 residues. The structure then provided valuable insights into AprA function. However, the story is far from complete. Foremost among the open questions is the identity of the native substrate for the acyltransfer reaction. Without the initial substrate as a starting point it will be difficult to propose an accurate reaction scheme that generates the final *t*-butyl-containing pivalyl-ACP. For example, the current biochemical data indicate that acetoacetyl-ACP is good substrate for dimethylation by MT2. Unfortunately, there is no clear route for the conversion of dimethyl-acetoacetyl-ACP to pivalyl-ACP. This leaves two possibilities, that acetoacetyl-ACP is not the native substrate, or that there is a cryptic conversion of dimethyl-acetoacetyl- to pivalyl-ACP. Both possibilities present interesting opportunities for the discovery of novel biochemical reactions. Knowing which acyl group is initially loaded should provide a base upon which the pivalyl-ACP can be built, and lead to testable hypotheses that guide the course of future investigations. Feeding studies of the apratoxin A producer, *Moorea bouillonii*, would also provide additional data by placing constraints on the number of possible mechanisms that must be considered. The MT1 domain of AprA has not yet been assayed successfully, but this domain likely plays some role in the generation of pivalyl-ACP and also warrants continued biochemical study.

References

1. Newman, D. J., and Cragg, G. M. (2016) Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* **79**, 629-661
2. Park, S. R., Tripathi, A., Wu, J., Schultz, P. J., Yim, I., McQuade, T. J., Yu, F., Arevang, C. J., Mensah, A. Y., Tamayo-Castillo, G., Xi, C., and Sherman, D. H. (2016) Discovery of cahuitamycins as biofilm inhibitors derived from a convergent biosynthetic pathway. *Nat Commun* **7**, 10710
3. Mendez-Perez, D., Begemann, M. B., and Pfleger, B. F. (2011) Modular synthase-encoding gene involved in alpha-olefin biosynthesis in Synechococcus sp. strain PCC 7002. *Appl Environ Microbiol* **77**, 4264-4267
4. Grindberg, R. V., Ishoey, T., Brinza, D., Esquenazi, E., Coates, R. C., Liu, W. T., Gerwick, L., Dorrestein, P. C., Pevzner, P., Lasken, R., and Gerwick, W. H. (2011) Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* **6**, e18565
5. Liu, Y., Law, B. K., and Luesch, H. (2009) Apratoxin a reversibly inhibits the secretory pathway by preventing cotranslational translocation. *Mol Pharmacol* **76**, 91-104
6. Dutta, S., Whicher, J. R., Hansen, D. A., Hale, W. A., Chemler, J. A., Congdon, G. R., Narayan, A. R., Hakansson, K., Sherman, D. H., Smith, J. L., and Skiniotis, G. (2014) Structure of a modular polyketide synthase. *Nature* **510**, 512-517
7. Whicher, J. R., Dutta, S., Hansen, D. A., Hale, W. A., Chemler, J. A., Dosey, A. M., Narayan, A. R., Hakansson, K., Sherman, D. H., Smith, J. L., and Skiniotis, G. (2014) Structural rearrangements of a polyketide synthase module during its catalytic cycle. *Nature* **510**, 560-564
8. Smith, S., and Tsai, S. C. (2007) The type I fatty acid and polyketide synthases: a tale of two megasynthases. *Nat Prod Rep* **24**, 1041-1072
9. Lowry, B., Li, X., Robbins, T., Cane, D. E., and Khosla, C. (2016) A turnstile mechanism for the controlled growth of biosynthetic intermediates on assembly line polyketide synthases. *ACS Cent Sci* **2**, 14-20
10. Khayatt, B. I., Overmars, L., Siezen, R. J., and Francke, C. (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One* **8**, e62136
11. Hill, A. M., and Staunton, J. (2010) Type I Modular PKS. *Comprehensive Natural Products Ii: Chemistry and Biology, Vol 1: Natural Products*

Structural Diversity-I: Secondary Metabolites: Organization and Biosynthesis, 385-452
12. Smith, J. L., Skiniotis, G., and Sherman, D. H. (2015) Architecture of the polyketide synthase module: surprises from electron cryo-microscopy. *Curr Opin Struct Biol* **31**, 9-19
13. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., and Huson, D. H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res* **33**, 5799-5808
14. Walsh, C. T. (2016) Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat Prod Rep* **33**, 127-135
15. Moore, B. S., and Hertweck, C. (2002) Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat Prod Rep* **19**, 70-99
16. Schmelz, S., and Naismith, J. H. (2009) Adenylate-forming enzymes. *Curr Opin Struct Biol* **19**, 666-671
17. Conti, E., Stachelhaus, T., Marahiel, M. A., and Brick, P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J* **16**, 4174-4183
18. Butcher, R. A., Schroeder, F. C., Fischbach, M. A., Straight, P. D., Kolter, R., Walsh, C. T., and Clardy, J. (2007) The identification of bacillaene, the product of the PksX megacomplex in Bacillus subtilis. *Proc Natl Acad Sci U S A* **104**, 1506-1509
19. Schwecke, T., Aparicio, J. F., Molnar, I., Konig, A., Khaw, L. E., Haydock, S. F., Oliynyk, M., Caffrey, P., Cortes, J., Lester, J. B., and et al. (1995) The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc Natl Acad Sci U S A* **92**, 7839-7843
20. Zhang, Z., Zhou, R., Sauder, J. M., Tonge, P. J., Burley, S. K., and Swaminathan, S. (2011) Structural and functional studies of fatty acyl adenylate ligases from E. coli and L. pneumophila. *J Mol Biol* **406**, 313-324
21. Goyal, A., Verma, P., Anandhakrishnan, M., Gokhale, R. S., and Sankaranarayanan, R. (2012) Molecular basis of the functional divergence of fatty acyl-AMP ligase biosynthetic enzymes of Mycobacterium tuberculosis. *J Mol Biol* **416**, 221-238
22. Arora, P., Goyal, A., Natarajan, V. T., Rajakumara, E., Verma, P., Gupta, R., Yousuf, M., Trivedi, O. A., Mohanty, D., Tyagi, A., Sankaranarayanan, R., and Gokhale, R. S. (2009) Mechanistic and functional insights into fatty acid activation in Mycobacterium tuberculosis. *Nat Chem Biol* **5**, 166-173
23. Guillet, V., Galandrin, S., Maveyraud, L., Ladeveze, S., Mariaule, V., Bon, C., Eynard, N., Daffe, M., Marrakchi, H., and Mourey, L. (2016) Insight into structure-function relationships and inhibition of the Fatty Acyl-AMP Ligase (FadD32) orthologs from Mycobacteria. *J Biol Chem* **291**, 7973-7989
24. Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., Duggan, B. M., Di Marzo, V., Sherman, D. H., Dorrestein, P. C., Gerwick, L., and Gerwick, W. H. (2015) Combining mass spectrometric metabolic profiling with genomic analysis: a powerful

approach for discovering natural products from cyanobacteria. *J Nat Prod* **78**, 1671-1682
25. Edwards, D. J., Marquez, B. L., Nogle, L. M., McPhail, K., Goeger, D. E., Roberts, M. A., and Gerwick, W. H. (2004) Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chem Biol* **11**, 817-833
26. Drake, E. J., and Gulick, A. M. (2011) Structural characterization and high-throughput screening of inhibitors of PvdQ, an NTN hydrolase involved in pyoverdine synthesis. *ACS Chem Biol* **6**, 1277-1286
27. Trivedi, O. A., Arora, P., Sridharan, V., Tickoo, R., Mohanty, D., and Gokhale, R. S. (2004) Enzymic activation and transfer of fatty acids as acyl-adenylates in mycobacteria. *Nature* **428**, 441-445
28. Marrakchi, H., Laneelle, M. A., and Daffe, M. (2014) Mycolic acids: structures, biosynthesis, and beyond. *Chem Biol* **21**, 67-85
29. Coates, R. C., Podell, S., Korobeynikov, A., Lapidus, A., Pevzner, P., Sherman, D. H., Allen, E. E., Gerwick, L., and Gerwick, W. H. (2014) Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. *PLoS One* **9**, e85140
30. Mendez-Perez, D., Herman, N. A., and Pfleger, B. F. (2014) A desaturase gene involved in the formation of 1,14-nonadecadiene in Synechococcus sp. strain PCC 7002. *Appl Environ Microbiol* **80**, 6073-6079
31. Lennen, R. M., and Pfleger, B. F. (2013) Microbial production of fatty acid-derived fuels and chemicals. *Curr Opin Biotechnol* **24**, 1044-1053
32. McCarthy, J. G., Eisman, E. B., Kulkarni, S., Gerwick, L., Gerwick, W. H., Wipf, P., Sherman, D. H., and Smith, J. L. (2012) Structural basis of functional group activation by sulfotransferases in complex metabolic pathways. *ACS Chem Biol* **7**, 1994-2003
33. Gehret, J. J., Gu, L., Gerwick, W. H., Wipf, P., Sherman, D. H., and Smith, J. L. (2011) Terminal alkene formation by the thioesterase of curacin A biosynthesis: structure of a decarboxylating thioesterase. *J Biol Chem* **286**, 14445-14454
34. Gu, L., Wang, B., Kulkarni, A., Gehret, J. J., Lloyd, K. R., Gerwick, L., Gerwick, W. H., Wipf, P., Hakansson, K., Smith, J. L., and Sherman, D. H. (2009) Polyketide decarboxylative chain termination preceded by o-sulfonation in curacin a biosynthesis. *J Am Chem Soc* **131**, 16033-16035
35. Gu, L., Geders, T. W., Wang, B., Gerwick, W. H., Hakansson, K., Smith, J. L., and Sherman, D. H. (2007) GNAT-like strategy for polyketide chain initiation. *Science* **318**, 970-974
36. Akey, D. L., Gehret, J. J., Khare, D., and Smith, J. L. (2012) Insights from the sea: structural biology of marine polyketide synthases. *Nat Prod Rep* **29**, 1038-1049
37. Lau, J., Cane, D. E., and Khosla, C. (2000) Substrate specificity of the loading didomain of the erythromycin polyketide synthase. *Biochemistry* **39**, 10514-10520
38. Bisang, C., Long, P. F., Cortes, J., Westcott, J., Crosby, J., Matharu, A. L., Cox, R. J., Simpson, T. J., Staunton, J., and Leadlay, P. F. (1999) A chain

initiation factor common to both modular and aromatic polyketide synthases. *Nature* **401**, 502-505
39. Xue, Y., Zhao, L., Liu, H. W., and Sherman, D. H. (1998) A gene cluster for macrolide antibiotic biosynthesis in Streptomyces venezuelae: architecture of metabolic diversity. *Proc Natl Acad Sci U S A* **95**, 12111-12116
40. Silakowski, B., Nordsiek, G., Kunze, B., Blocker, H., and Muller, R. (2001) Novel features in a combined polyketide synthase/non-ribosomal peptide synthetase: the myxalamid biosynthetic gene cluster of the myxobacterium Stigmatella aurantiaca Sga15. *Chem Biol* **8**, 59-69
41. Ligon, J., Hill, S., Beck, J., Zirkle, R., Molnar, I., Zawodny, J., Money, S., and Schupp, T. (2002) Characterization of the biosynthetic gene cluster for the antifungal polyketide soraphen A from Sorangium cellulosum So ce26. *Gene* **285**, 257-267
42. Ikeda, H., Nonomiya, T., Usami, M., Ohta, T., and Omura, S. (1999) Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in Streptomyces avermitilis. *Proc Natl Acad Sci U S A* **96**, 9509-9514
43. Hagen, A., Poust, S., de Rond, T., Yuzawa, S., Katz, L., Adams, P. D., Petzold, C. J., and Keasling, J. D. (2014) In vitro analysis of carboxyacyl substrate tolerance in the loading and first extension modules of borrelidin polyketide synthase. *Biochemistry* **53**, 5975-5977
44. Piel, J., Wen, G., Platzer, M., and Hui, D. (2004) Unprecedented diversity of catalytic domains in the first four modules of the putative pederin polyketide synthase. *Chembiochem* **5**, 93-98
45. Salah Ud-Din, A. I., Tikhomirova, A., and Roujeinikova, A. (2016) Structure and functional diversity of GCN5-related N-acetyltransferases (GNAT). *Int J Mol Sci* **17**
46. Neuwald, A. F., and Landsman, D. (1997) GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. *Trends Biochem Sci* **22**, 154-155
47. Fisch, K. M., Gurgui, C., Heycke, N., van der Sar, S. A., Anderson, S. A., Webb, V. L., Taudien, S., Platzer, M., Rubio, B. K., Robinson, S. J., Crews, P., and Piel, J. (2009) Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat Chem Biol* **5**, 494-501
48. Partida-Martinez, L. P., and Hertweck, C. (2007) A gene cluster encoding rhizoxin biosynthesis in "Burkholderia rhizoxina", the bacterial endosymbiont of the fungus Rhizopus microsporus. *Chembiochem* **8**, 41-45
49. Young, J., Stevens, D. C., Carmichael, R., Tan, J., Rachid, S., Boddy, C. N., Muller, R., and Taylor, R. E. (2013) Elucidation of gephyronic acid biosynthetic pathway revealed unexpected SAM-dependent methylations. *J Nat Prod* **76**, 2269-2276
50. Kellmann, R., Mihali, T. K., Jeon, Y. J., Pickford, R., Pomati, F., and Neilan, B. A. (2008) Biosynthetic intermediate analysis and functional

homology reveal a saxitoxin gene cluster in cyanobacteria. *Appl Environ Microbiol* **74**, 4044-4053
51. Chang, Z., Sitachitta, N., Rossi, J. V., Roberts, M. A., Flatt, P. M., Jia, J., Sherman, D. H., and Gerwick, W. H. (2004) Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. *J Nat Prod* **67**, 1356-1367
52. Huang, K. C., Chen, Z., Jiang, Y., Akare, S., Kolber-Simonds, D., Condon, K., Agoulnik, S., Tendyke, K., Shen, Y., Wu, K. M., Mathieu, S., Choi, H. W., Zhu, X., Shimizu, H., Kotake, Y., Gerwick, W. H., Uenaka, T., Woodall-Jappe, M., and Nomoto, K. (2016) Apratoxin A shows novel pancreas-targeting activity through the binding of Sec 61. *Mol Cancer Ther* **15**, 1208-1216
53. Paatero, A. O., Kellosalo, J., Dunyak, B. M., Almaliti, J., Gestwicki, J. E., Gerwick, W. H., Taunton, J., and Paavilainen, V. O. (2016) Apratoxin kills cells by direct blockade of the Sec61 protein translocation channel. *Cell Chem Biol* **23**, 561-566
54. Skiba, M. A., Sikkema, A. P., Fiers, W. D., Gerwick, W. H., Sherman, D. H., Aldrich, C. C., and Smith, J. L. (2016) Domain organization and active site architecture of a polyketide synthase C-methyltransferase. *ACS Chem Biol*
55. Kang, M. K., and Nielsen, J. (2016) Biobased production of alkanes and alkenes through metabolic engineering of microorganisms. *J Ind Microbiol Biotechnol*
56. Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., and Keasling, J. D. (2008) Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr Opin Biotechnol* **19**, 556-563
57. Ramaswamy, A. V., Sorrels, C. M., and Gerwick, W. H. (2007) Cloning and biochemical characterization of the hectochlorin biosynthetic gene cluster from the marine cyanobacterium Lyngbya majuscula. *J Nat Prod* **70**, 1977-1986
58. Zhu, X., Liu, J., and Zhang, W. (2015) De novo biosynthesis of terminal alkyne-labeled natural products. *Nat Chem Biol* **11**, 115-120
59. Dorrestein, P. C., Blackhall, J., Straight, P. D., Fischbach, M. A., Garneau-Tsodikova, S., Edwards, D. J., McLaughlin, S., Lin, M., Gerwick, W. H., Kolter, R., Walsh, C. T., and Kelleher, N. L. (2006) Activity screening of carrier domains within nonribosomal peptide synthetases using complex substrate mixtures and large molecule mass spectrometry. *Biochemistry* **45**, 1537-1546
60. Kuhn, M. L., Alexander, E., Minasov, G., Page, H. J., Warwrzak, Z., Shuvalova, L., Flores, K. J., Wilson, D. J., Shi, C., Aldrich, C. C., and Anderson, W. F. (2016) Structure of the essential Mtb FadD32 enzyme: a promising drug target for treating tuberculosis. *ACS Infect Dis* **2**, 579-591
61. McQuade, T. J., Shallop, A. D., Sheoran, A., Delproposto, J. E., Tsodikov, O. V., and Garneau-Tsodikova, S. (2009) A nonradioactive high-throughput assay for screening and characterization of adenylation

domains for nonribosomal peptide combinatorial biosynthesis. *Anal Biochem* **386**, 244-250
62. Li, W., Gu, S., Fleming, J., and Bi, L. (2015) Crystal structure of FadD32, an enzyme essential for mycolic acid biosynthesis in mycobacteria. *Sci Rep* **5**, 15493
63. Leger, M., Gavalda, S., Guillet, V., van der Rest, B., Slama, N., Montrozier, H., Mourey, L., Quemard, A., Daffe, M., and Marrakchi, H. (2009) The dual function of the Mycobacterium tuberculosis FadD32 required for mycolic acid biosynthesis. *Chem Biol* **16**, 510-519
64. Winters, K., Parker, P. L., and Van Baalen, C. (1969) Hydrocarbons of blue-green algae: geochemical signfficance. *Science* **163**, 467-468
65. Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E., and Khosla, C. (2001) Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science* **291**, 1790-1792
66. Quadri, L. E., Weinreb, P. H., Lei, M., Nakano, M. M., Zuber, P., and Walsh, C. T. (1998) Characterization of Sfp, a Bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **37**, 1585-1595
67. Worthington, A. S., and Burkart, M. D. (2006) One-pot chemo-enzymatic synthesis of reporter-modified proteins. *Org Biomol Chem* **4**, 44-46
68. Kabsch, W. (2010) Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132
69. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674
70. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221
71. Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J., and Adams, P. D. (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* **64**, 61-69
72. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501
73. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68**, 352-367
74. Moriarty, N. W., Grosse-Kunstleve, R. W., and Adams, P. D. (2009) electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr D Biol Crystallogr* **65**, 1074-1080

75. Krissinel, E., and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774-797
76. Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004) The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427
77. Thompson, J. D., Muller, A., Waterhouse, A., Procter, J., Barton, G. J., Plewniak, F., and Poch, O. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics* **7**, 318
78. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539
79. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* **32**, W665-667
80. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., and Baker, N. A. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* **35**, W522-525
81. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21
82. Schrodinger, LLC. (2015) The PyMOL Molecular Graphics System, Version 1.8.
83. Meluzzi, D., Zheng, W. H., Hensler, M., Nizet, V., and Dorrestein, P. C. (2008) Top-down mass spectrometry on low-resolution instruments: characterization of phosphopantetheinylated carrier domains in polyketide and non-ribosomal biosynthetic pathways. *Bioorg Med Chem Lett* **18**, 3107-3111
84. Fischbach, M. A., and Walsh, C. T. (2006) Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. *Chem. Rev.* **106**, 3468-3496
85. Sieber, S. A., and Marahiel, M. A. (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chemical reviews* **105**, 715-738
86. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. in *Nucleic Acids Research*
87. Strieker, M., Tanović, A., and Marahiel, M. A. (2010) Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol.* **20**, 234-240
88. Boucher, H. W., Talbot, G. H., Bradley, J. S., Edwards, J. E., Gilbert, D., Rice, L. B., Scheld, M., Spellberg, B., and Bartlett, J. (2009) Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1-12

89. Miller, D. A., Luo, L., Hillson, N., Keating, T. A., and Walsh, C. T. (2002) Yersiniabactin synthetase: a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of Yersinia pestis. *Chemistry & biology* **9**, 333-344
90. Patel, H. M., and Walsh, C. T. (2001) In vitro reconstitution of the Pseudomonas aeruginosa nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and N-methyltransferase activities. *Biochemistry* **40**, 9023-9031
91. Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M. O., Bartels, D., Bekel, T., Beyer, S., Bode, E., Bode, H. B., Bolten, C. J., Choudhuri, J. V., Doss, S., Elnakady, Y. A., Frank, B., Gaigalat, L., Goesmann, A., Groeger, C., Gross, F., Jelsbak, L., Jelsbak, L., Kalinowski, J., Kegler, C., Knauber, T., Konietzny, S., Kopp, M., Krause, L., Krug, D., Linke, B., Mahmud, T., Martinez-Arias, R., McHardy, A. C., Merai, M., Meyer, F., Mormann, S., Muñoz-Dorado, J., Perez, J., Pradella, S., Rachid, S., Raddatz, G., Rosenau, F., Rückert, C., Sasse, F., Scharfe, M., Schuster, S. C., Suen, G., Treuner-Lange, A., Velicer, G. J., Vorhölter, F.-J., Weissman, K. J., Welch, R. D., Wenzel, S. C., Whitworth, D. E., Wilhelm, S., Wittmann, C., Blöcker, H., Pühler, A., and Müller, R. (2007) Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nat Biotechnol* **25**, 1281-1289
92. Letunic, I., Doerks, T., and Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research* **43**, D257-260
93. May, J. J., Kessler, N., Marahiel, M. A., and Stubbs, M. T. (2002) Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A* **99**, 12120-12125
94. Garneau, S., Dorrestein, P. C., Kelleher, N. L., and Walsh, C. T. (2005) Characterization of the formation of the pyrrole moiety during clorobiocin and coumermycin A1 biosynthesis. *Biochemistry* **44**, 2770-2780
95. Tang, G.-L., Cheng, Y.-Q., and Shen, B. (2007) Chain initiation in the leinamycin-producing hybrid nonribosomal peptide/polyketide synthetase from Streptomyces atroolivaceus S-140. Discrete, monofunctional adenylation enzyme and peptidyl carrier protein that directly load D-alanine. *J. Biol. Chem.* **282**, 20273-20282
96. Gehring, A. M., Mori, I. I., Perry, R. D., and Walsh, C. T. (1998) The nonribosomal peptide synthetase HMWP2 forms a thiazoline ring during biogenesis of yersiniabactin, an iron-chelating virulence factor of yersinia pestis. *Biochemistry* **37**, 17104
97. Quadri, L. E., Keating, T. A., Patel, H. M., and Walsh, C. T. (1999) Assembly of the Pseudomonas aeruginosa nonribosomal peptide siderophore pyochelin: In vitro reconstitution of aryl-4, 2-bisthiazoline synthetase activity from PchD, PchE, and PchF. *Biochemistry* **38**, 14941-14954
98. Quadri, L. E., Sello, J., Keating, T. A., Weinreb, P. H., and Walsh, C. T. (1998) Identification of a Mycobacterium tuberculosis gene cluster

encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem Biol* **5**, 631-645
99. Drake, E. J., Duckworth, B. P., Neres, J., Aldrich, C. C., and Gulick, A. M. (2010) Biochemical and structural characterization of bisubstrate inhibitors of BasE, the self-standing nonribosomal peptide synthetase adenylate-forming enzyme of acinetobactin synthesis. *Biochemistry* **49**, 9292-9305
100. Sundlov, J. A., and Gulick, A. M. (2013) Structure determination of the functional domain interaction of a chimeric nonribosomal peptide synthetase from a challenging crystal with noncrystallographic translational symmetry. *Acta Crystallogr D Biol Crystallogr* **69**, 1482-1492
101. Gulick, A. M. (2009) Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem Biol* **4**, 811-827
102. Sundlov, J. A., Shi, C., Wilson, D. J., Aldrich, C. C., and Gulick, A. M. (2012) Structural and functional investigation of the intermolecular interaction between NRPS adenylation and carrier protein domains. *Chem Biol* **19**, 188-198
103. Rusnak, F., Faraci, W. S., and Walsh, C. T. (1989) Subcloning, expression, and purification of the enterobactin biosynthetic enzyme 2,3-dihydroxybenzoate-AMP ligase: demonstration of enzyme-bound (2,3-dihydroxybenzoyl)adenylate product. *Biochemistry* **28**, 6827-6835
104. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Meth* **6**, 343-345
105. Stols, L., Gu, M., Dieckman, L., Raffen, R., Collart, F. R., and Donnelly, M. I. (2002) A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Expr Purif* **25**, 8-15
106. Whicher, J. R., Smaga, S. S., Hansen, D. A., Brown, W. C., Gerwick, W. H., Sherman, D. H., and Smith, J. L. (2013) Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem Biol* **20**, 1340-1351
107. Fuji, K., Ikai, Y., Mayumi, T., Oka, H., Sujuki, M., and Harada, K. I. (1997) A nonempirical method using LC/MS for determination of the absolute configuration of constituent amino acids in a peptide: elucidation of limitations of Marfey's method and of its separation mechanism. *Anal. Chem.* **69**, 3346-3352
108. Marfey, P. (1984) Determination of d-amino acids. II. Use of a bifunctional reagent, 1,5-difluoro-2,4dinitrobenzene. *Carlsberg Res. Commun.* **49**, 591-596
109. Laskowski, R. A., and Swindells, M. B. (2011) LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* **51**, 2778-2786
110. Luesch, H., Yoshida, W. Y., Moore, R. E., Paul, V. J., and Corbett, T. H. (2001) Total structure determination of apratoxin A, a potent novel

cytotoxin from the marine cyanobacterium Lyngbya majuscula. *J Am Chem Soc* **123**, 5418-5423
111. Chen, Q. Y., Liu, Y., and Luesch, H. (2011) Systematic Chemical Mutagenesis Identifies a Potent Novel Apratoxin A/E Hybrid with Improved in Vivo Antitumor Activity. *ACS Med Chem Lett* **2**, 861-865
112. Chen, J., and Forsyth, C. J. (2003) Total synthesis of apratoxin A. *J Am Chem Soc* **125**, 8734-8735
113. Kampa, A., Gagunashvili, A. N., Gulder, T. A., Morinaka, B. I., Daolio, C., Godejohann, M., Miao, V. P., Piel, J., and Andresson, O. (2013) Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proc Natl Acad Sci U S A* **110**, E3129-3137
114. Piel, J., Hui, D., Wen, G., Butzke, D., Platzer, M., Fusetani, N., and Matsunaga, S. (2004) Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. *Proc Natl Acad Sci U S A* **101**, 16222-16227
115. Moebius, N., Ross, C., Scherlach, K., Rohm, B., Roth, M., and Hertweck, C. (2012) Biosynthesis of the respiratory toxin bongkrekic acid in the pathogenic bacterium Burkholderia gladioli. *Chem Biol* **19**, 1164-1174
116. Simunovic, V., Zapp, J., Rachid, S., Krug, D., Meiser, P., and Muller, R. (2006) Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting acyltransferases. *Chembiochem* **7**, 1206-1220
117. Liscombe, D. K., Louie, G. V., and Noel, J. P. (2012) Architectures, mechanisms and molecular evolution of natural product methyltransferases. *Nat Prod Rep* **29**, 1238-1250
118. Stevens, D. C., Wagner, D. T., Manion, H. R., Alexander, B. K., and Keatinge-Clay, A. T. (2016) Methyltransferases excised from trans-AT polyketide synthases operate on N-acetylcysteamine-bound substrates. *J Antibiot (Tokyo)* **69**, 567-570
119. Wagner, D. T., Stevens, D. C., Mehaffey, M. R., Manion, H. R., Taylor, R. E., Brodbelt, J. S., and Keatinge-Clay, A. T. (2016) alpha-Methylation follows condensation in the gephyronic acid modular polyketide synthase. *Chem Commun (Camb)* **52**, 8822-8825
120. Chovancova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., Gora, A., Sustr, V., Klvana, M., Medek, P., Biedermannova, L., Sochor, J., and Damborsky, J. (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* **8**, e1002708
121. Bunkoczi, G., and Read, R. J. (2011) Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr D Biol Crystallogr* **67**, 303-312
122. Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., and Stuckey, P. J. (2011) Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics* **27**, 3315-3316