Fourier Transform Ion Cylcotron Resonance Mass Spectrometry Analysis of Natural Product Biosynthesis

by

Christopher Michael Rath

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemical Biology)
in The University of Michigan
2011

Doctoral Committee:

Professor David H. Sherman, Co-chair
Assistant Professor Kristina I. Håkansson, Co-Chair
Professor Janet L. Smith
Assistant Professor Garry Dean Dotson
Assistant Professor Sylvie Garneau-Tsodikova

ACKNOWLEDGEMENTS

I would like to thank my advisors David H. Sherman and Kristina Hakansson for support and guidance. I would like to thank all of my past/present coworkers in my two labs—without your advice and support none of this would be possible.

I would like to thank my coauthors (in alphabetical order): Azad Ahmed, David L. Akey, Erica C. Anderson, Nicholas H. Bergman, Shilah Bonnet, Kyle L. Bolduc, Sarah J. Brooks, Tonia J. Buchholz, Joesph Chemler, Michael A. Christiansen, Meg Dahlgren, Yousong Ding, Jonathan Dordick, Josh Earl, Garth D. Ehrlich, Noah P. Gardner, William Gerwick, Philip C. Hanna, Margo G. Haygood, Luisa Hiller, Joanne Hothersall, Fen Z. Hu, Makato Inai, Brian K. Janes, Benjamin Janto, Andrzej Joachimiak, Joanna R. Joels, Jeffrey D. Kittendorf, Eung-Soo Kim, Youngchang Kim, Rachael Kreft, Hye Kyong Kweon, Jung Yeop Lee, Nicole B. Lopanik, Natalia Maltseva, Tyler D. Nusca, Brian F. Pfleger, Keven Renoylds, Jamie B. Scaglione, Rafay Shareef, Jennifer A. Shields, Janet L. Smith, Rachel Sullivan, Christopher M. Thomas, Robert M. Williams, Jeremy J. Wolff, and Fengan Yu.

I would also like to thank some of those people who were instrumental in getting me to graduate school: Taro Amagata, Marcy Copeland, Phil Crews, Paul Motchnik, Robbi Sera, Matt Sweeny, Brian Schmidt. I would like to thank Jeff and Tonia for additional advice and support early on in my career and Shamilya Williams for administrative support. I would also like to thank the Chemical Biology PhD Program and my committee.

I have been supported by the NIH through two training grant programs: the Chemical Biology Interface Training Program, and the Microfluidics in Biomedical Sciences Training Program. I have also received funding from Rackham in the form of Travel Grants.

Finally, none of this would be possible, or indeed worth it, without the continued foundation of support and love from my friends and family.

PREFACE

This thesis contains six chapters detailing much of my graduate research at the University of Michigan in the fields of natural product biosynthesis and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Chapter 1 is an introduction to the fields through several case studies and has been published in part as two separate review articles (NRPS/PKS Hybrid Enzymes and Their Natural Products. Christopher M. Rath, Jamie B. Scaglione, Jeffrey D. Kittendorf and David H. Sherman. In *Comprehensive Natural Products II: Chemistry and Biology*; Lew Mander, Hung-Wend Liu Editors; Elsevier: Oxford 2010; volume 1:453-492. Biosynthetic Principles in Marine Natural Product Systems. David H. Sherman, Christopher M. Rath, Jon Mortinson, Jamie B. Scaglione, and Jeffrey D. Kittendorf. In *Natural Products,*William Gerwick Editor; Text in preparation.)

Chapter 2 is a detailed investigation into extender unit processing in the pikromycin polyketide synthase and is under review as a communications in Chemistry & Biology (Acyl-CoA subunit selectivity in the terminal pikromycin polyketide synthase module: steady-state kinetics and active-site occupancy analysis by FTICR-MS. Shilah A. Bonnett*, Christopher M. Rath*, Rafay Shareef, Joanna R. Joels, Joesph Chemler, Kristina Hakansson, Kevin Reynolds, David H. Shermanb. Under Review Chemistry & Biology).

Chapter 3 is an *in vitro* biochemical investigation into a marine symbiont derived natural product pathway and has been published in Chemistry & Biology (Polyketide β-Branching in Bryostatin Biosynthesis: Identification of Surrogate Acetyl-ACP Donors for BryR, an HMG-ACP Synthase. Tonia J. Buchholz, Christopher M. Rath, Nicole B. Lopanik, Noah P. Gardner, Kristina Håkansson, and David H. Sherman. Chemistry & Biology 17:1092-1100 (2010).

Chapter 4 investigates the chemoenzymatic synthesis of cryptophycins through ester bond formation and additional elongation and processing steps—the manuscript is in preparation for submission to Journal of the American Chemical Society (Chemoenzymatic Synthesis of Cryptophycin Anticancer Agents: Non-Amino Acid Incorporation Mediated by a NRPS Module. Yousong Ding*, Christopher M. Rath*, Kyle L. Bolduc, Kristina Håkansson, David H. Sherman. Drafting for the Journal of the American Chemical Society).

Chapter 5 presents the identification of the ET-743 biosynthetic pathway from a symbiotic bacteria using a new technological platform, this manuscript has been submitted to PLOSone (Christopher M. Rath*, Benjamin Janto*, Josh Earl, Azad Ahmed, Fen Z. Hu, Luisa Hiller, Meg Dahlgren, Rachael Kreft, Fengan Yu, Jeremy J. Wolff, Hye Kyong Kweon, Michael A. Christiansen, Kristina Håkansson, Robert M. Williams, Garth D. Ehrlich, David H. Sherman. Meta-omic analysis of a marine invertebrate microbial consortium provides a direct route to identify and characterize natural product biosynthetic systems. Manuscript submitted PLOSone).

Chapter 6 presents, conclusions and future directions for my research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Natural products have provided some of our most clinically relevant drugs and continue to be a source of new leads. Indeed, our understanding of the fundamental mechanisms involved in their biosynthetic production is just beginning to develop. A FTICR-MS-centric analytical approach was applied to understand fundamental mechanistic details of natural product production, monitor chemoenzymatic generation of natural products, and characterize/identify novel natural product producing systems. Investigations in four different systems have highlighted the broad applicability of this analytical approach.

Each of the chapters of this thesis seeks to examine these three themes in different contexts. In Chapter 2 the key step of coenzyme A extender unit selection in polyketide biosynthesis is explored by *in vitro* biochemistry of the PikAIV model system. Key findings developed allow a model for catalysis to be proposed. In Chapter 3, our investigations shift to the bryostatin biosynthetic system. The biochemical characterization of BryR, the 3-hydroxy-3-methylglutaryl-CoA synthase, implicated in β- of the core ring system from the bryostatin metabolic pathway, is reported. In Chapter 4, an unusual enzyme activity of the nonribosomal peptide synthase module CrpDm2 was explored *in vitro* by FTICR-MS—resulting in chemoenzymatic synthesis of three cryptophycin analogs. In Chapter 5, the biosynthetic pathway for the approved chemotherapeutic ET-743 is characterized from a tunicate/symbiont system using a novel

workflow. Metabolite analysis, metagenomic sequencing, contig assembly, and metaproteomic analysis were used to probe the experimental system and a key enzyme activity was verified *in vitro*. Chapter 6 presents future directions based upon findings developed.

We have focused on gaining a deeper understanding of key mechanisms in natural product biosynthesis by studying defined systems *in vitro* through the application of novel, FTICR-MS centric analytical technologies. These studies have illustrated the application of chemoenzymatic methodology to generate novel analogs of active natural products, with FTICR-MS as the key tool to investigate the final product and key enzyme bound intermediates. Together, the developed tools were applied to characterize a novel biosynthetic pathway that had previously been inaccessible with current analytical technologies. Fundamentally, these developments will inform our ability to access natural products for human health concerns.

Chapter 1

Introduction

1.1 Natural products and medicine

Natural products have proved to be an exceptionally rich source of small molecule ligands for discovery and analysis of diverse molecular targets relating to human disease. From 1940 to 2006 47% of approved anticancer therapeutics were natural products or natural product derived chemical entities.[1] Between 2005-2007 thirteen natural product, natural product derived, or semi-synthetic drugs have been launched.[2] Natural products may well enjoy a resurgence in drug discovery and development efforts as recent advances such as the ability to engineer and heterologously express biosynthetic pathways may provide effective solutions to current challenges including adequate natural product supplies, reduced reagent cost, and effective analog development.

A significant number of marine natural products contain pharmacological activities that are beneficial to human health. Although there are many examples of terrestrial-derived natural product compounds that are in clinical use, including antibacterial penicillins, cephalosporins, immunosuppressive cyclosporin A, and the cholesterol-lowering HMG-CoA reductase inhibitors best known as the "statins",[3] secondary metabolites from the marine environment are just starting to reach the clinic. These molecules, which have achieved their (largely unknown) endogenous functions over the course of millions of years of evolution, offer chemical scaffolds for development of new analogs with improved or altered biological activity (**Figure 1-1**). New bioactive analogs that contain novel structural elements may be generated by both semi-synthesis and total synthesis efforts.[4,5]

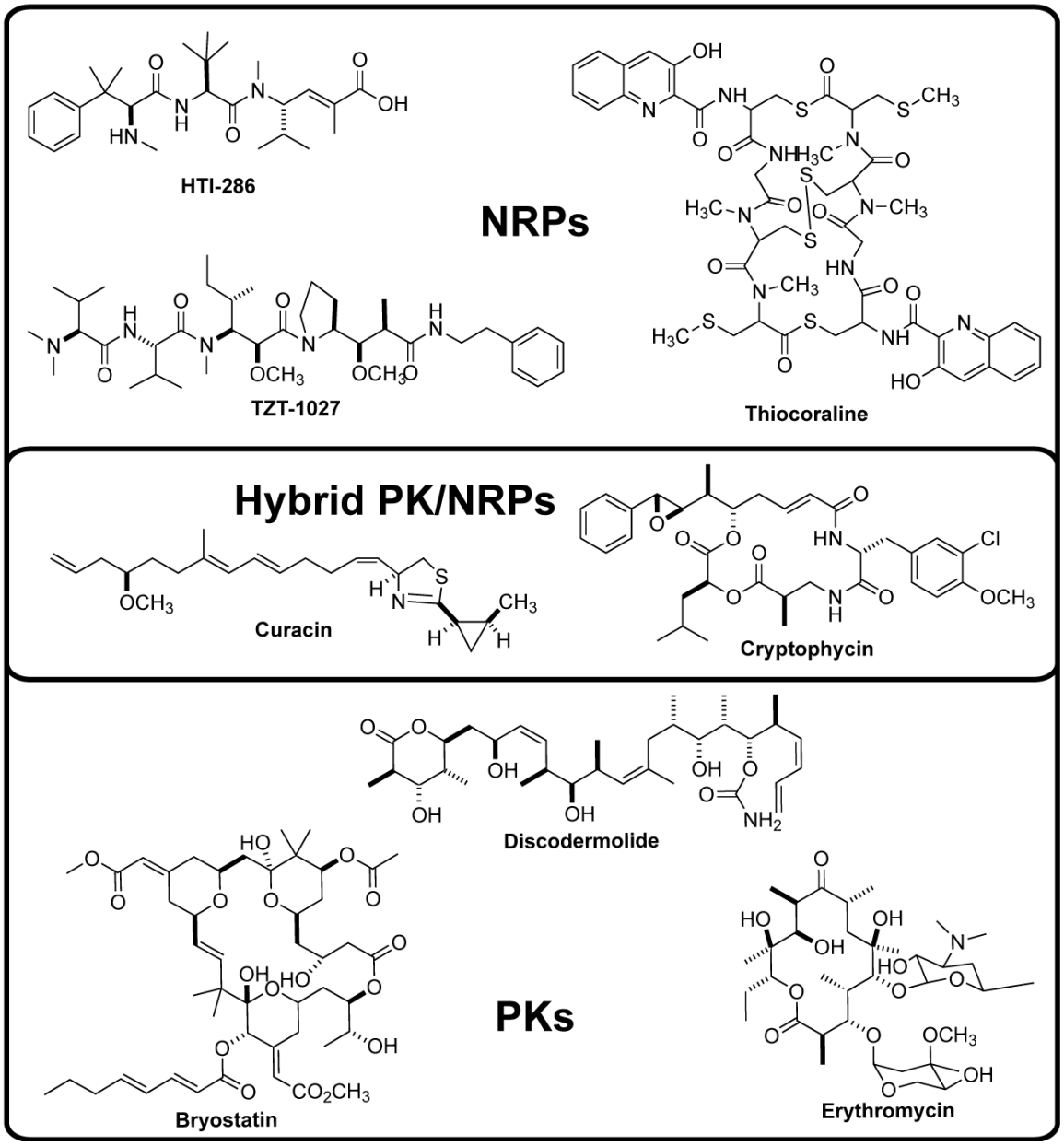


Figure 1-1. Examples of nonribosomal peptide (NRP), polyketide (PK), and hybrid (PK/NRP) natural products.

1.2 Function of polyketide synthases and nonribosomal peptide synthetases

The synthases that produce nonribosomal peptide (NRP) and polyketide (PK) natural products are equally interesting in terms of their application in natural product fermentation and as biocatalysts. The number of natural products that function in biological systems is large but represents only a small fraction of the total possible

number of small carbon-based compounds, indicating the importance of stereochemistry and functional groups in natural product functions.[6]

PKs, NRPs, and PK/NRP hybrids represent three large subclasses of highly diverse natural products with various bioactivities.[7] These natural products are generated by large mega-enzymes, polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs). Type I PKSs consist of multiple modules, with each module minimally containing three core domains: acyltransferase (AT) domain, ketosynthase (KS) domain, and thiolation (T) domain (also called acyl carrier protein (ACP) domain). Typically, one type I PKS module catalyzes a single elongation cycle for PK production (**Figure 1-2**). During elongation, the AT domain serves as the gatekeeper for specificity, responsible for selecting the appropriate acyl-CoA extender unit (e.g. malonyl-CoA, methylmalonyl-CoA) and transferring the extender unit to the sulfhydryl terminus of the phosphopantetheinyl arm on the T domain.[8] The KS domain catalyzes the decarboxylation of acyl-*S*-T to generate a carbanion that reacts with the PK intermediate linked to the T domain generated in the previous elongation cycle. The resulting ketoacyl-*S*-T becomes the substrate for the next cycle of elongation catalyzed by the subsequent elongation module. In addition to type I PKSs, there are two other PKS classes, type II PKSs and type III PKSs.[9] Unlike the type I class, type II PKSs consist of discrete enzymes that are organized as a multi-component system.[10] The type III PKSs are distinguished from the others by lack of an AT and T domain. Type III PKS systems typically use CoA substrates (i.e., malonyl-CoA), but there is precedent for their ability to accept acyl-*S*-T substrates.[9,11] ***PKSs are an important class of biosynthetic enzymes that produces a variety of polyketide natural products in both eukaryotic and prokaryotic organisms. These multi-megadalton enzymatic complexes are composed of a series of proteins—resulting in an encoded logic arranged in a linear pathway.***[12]

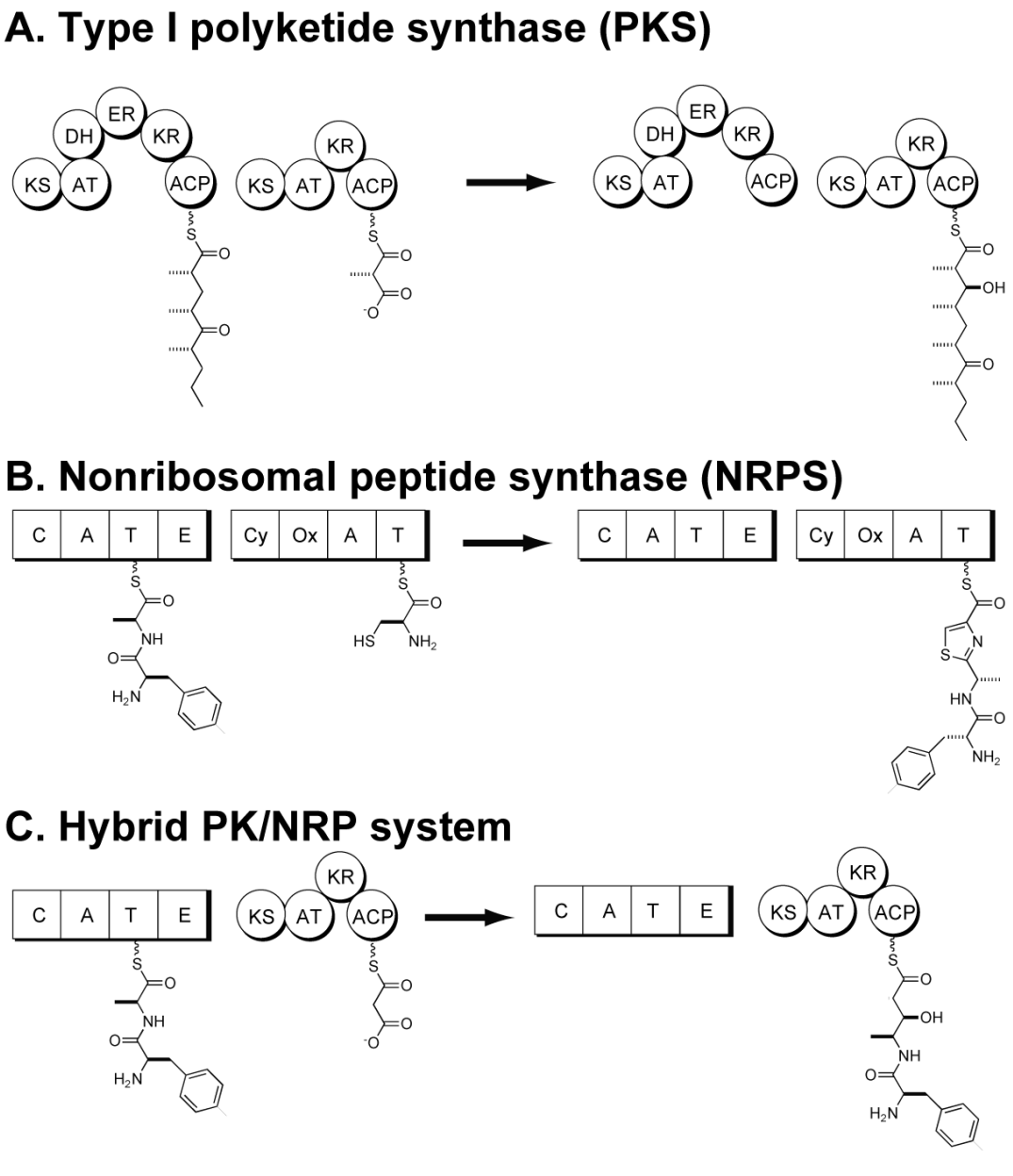


Figure 1-2. Hypothetical examples of the modular organization in polyketide synthases (PKSs), nonribosomal peptide synthases (NRPs), and hybrid PK/NRPs. (**A**) Two consecutive PKS elongation modules from a hypothetical polyketide biosynthetic pathway. The two modules catalyze the elongation of the growing polyketide intermediate by two carbons (from methylmalonyl-CoA), and subsequent β-keto reduction. (**B**) Two consecutive NRPS elongation modules from a hypothetical nonribosomal peptide biosynthetic pathway. The two modules catalyze peptide bond formation between the growing peptide intermediate and an activated cysteine residue with subsequent cyclization/oxidation of the incorporated cysteine residue to a thiazole. (**C**) Consecutive NRPS and PKS elongation modules from a hypothetical hybrid nonribosomal peptide/polyketide biosynthetic pathway. The two modules catalyze the two carbon extension (from malonyl-CoA) of the peptide intermediate and subsequent β-keto reduction.

Similar to the type I PKSs, NRPSs are comprised of multifunctional enzymes that are arranged into modules. Each NRPS module contains three core domains: adenylation (A), condensation (C), and thiolation (T) [also called peptidyl carrier protein (PCP) domain] (**Figure 1-2**).[13] The A domain is responsible for selecting and activating the natural or modified amino acid monomer. The activated amino acid monomer is covalently attached via a thioester bond to the cysteamine group of a phosphopantetheinyl arm in the holo-T domain. The condensation (C) domain catalyzes formation of a peptide bond between the amino acid monomer and the peptidyl intermediate tethered to a T domain in an adjacent module. Similar to type I PKS modules, each NRPS module performs a single elongation step of the growing peptidyl chain. In both NRPSs and PKSs, there are several additional auxiliary domains that contribute to natural product structural diversity. Ketoreductase (KR), dehydratase (DH), enoyl reductase (ER), and methyltransferase (MT) domains are commonly found in PKS modules while cyclization (Cy), *N*-MT and epimerase (E) domains are sometimes embedded within NRPS modules. These additional domains contribute significantly to the diversity and bioactivity of PKs and NRPs. Thioesterase (TE) domains, typically found at the C-terminus of the final elongation module in both PKSs and NRPSs are responsible for terminating biosynthesis. In most cases, TE domains catalyze intra-molecular macrocyclization or hydrolysis of the thioester bond between the final T domain and the PK or NRP intermediate.[14] The structures of the nascent PK and NRP products are often further modified through oxidation, glycosylation, acylation, alkylation, and halogenation reactions catalyzed by tailoring enzymes in natural product biosynthetic pathways.[15,16]

Genetic, biochemical, and structural characterization of numerous biosynthetic pathways have made much of this logic clear. Despite this progress, aspects of these systems such as the transfer of covalently bound intermediates between active-sites remain as "black-boxes". Substrate specificity during these transfers can result in dramatic differences in rates of natural versus unnatural substrate incorporation (>3 orders of magnitude).[17] If this discrimination becomes better understood, engineering of PKS pathways to provide "unnatural" products could be facilitated. Engineering of PKS systems may allow access to novel antibiotics and other bioactive molecules as a new

realm of chemical space can be accessed through low-cost, fermentation-based biosynthesis.

In the sections below, several natural product biosynthetic systems will be described as cased studies that provide an in-depth overview of genetic and biochemical mechanisms involved in assembly and tailoring of many of these beautifully complex and biologically active molecules.

1.3 Pikromycin biosynthetic pathway

PKS biosynthetic pathways, such as the PikA pathway (**Figure 1-3**) from *Streptomyces venezuelae*, are composed of a series of enzymatic domains organized into modules across multiple polypeptides. CoA-substrates are bound and used to extend a growing intermediate by at least two carbons per cycle.[12,18] A minimal module, such as PikAIV, consists of a KS, AT, and ACP, all of which covalently bind substrate with a cysteine, serine, or phosphopantetheinyl (Ppant) prosthetic group, respectively. Extension by a single minimal module (+TE), the 140 kDa PikAIV, proceeds as follows: (1) The PikAIV AT loads methylmalonate from CoA and then transfers it to the PikAIV ACP.[19] (2) The PikAIV KS accepts the hexaketide from the PikAIII ACP. (3) The PikAIV KS bound hexaketide is condensed with the methylmalonyl loaded PikAIV ACP through a Claisen-type decarboxylation to form the final heptaketide. (4) The PikAIV TE cyclizes the heptaketide to provide the aglycone 10-narbonolide (**Figure 1-3C**).

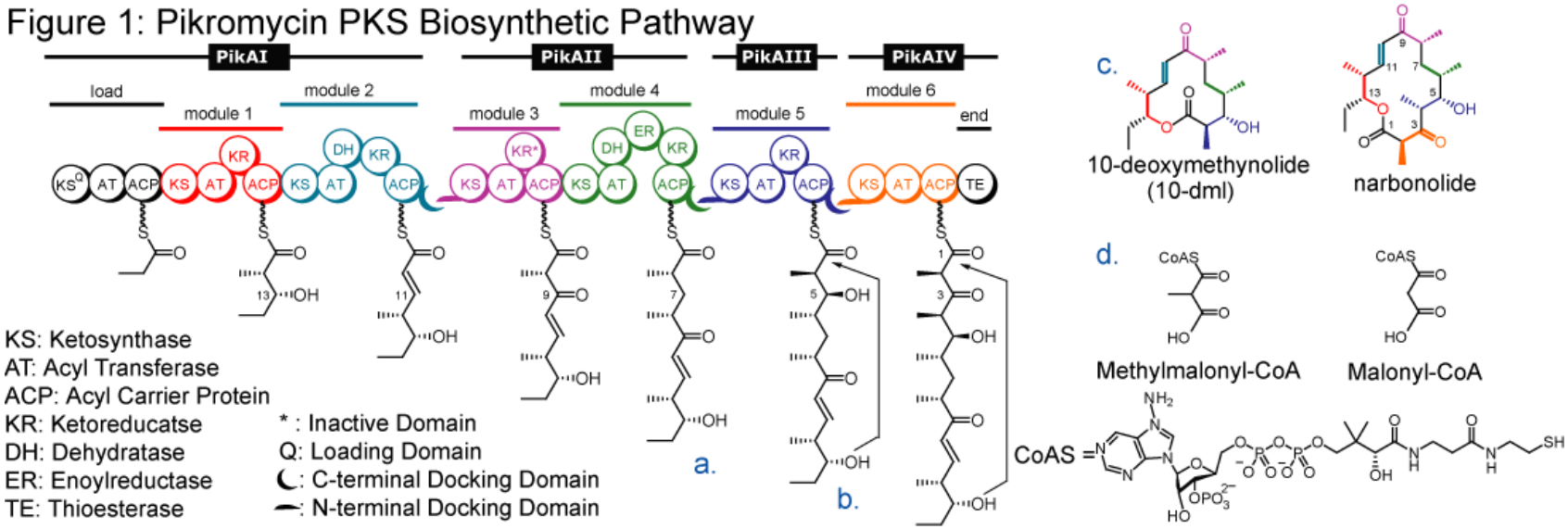


Figure 1-3. The pikromycin biosynthetic pathway. (**A**) PikA pentaketide, (**B**) PikA hexaketide (**C**) aglycone products (D) substrates excluding NADPH.

In other modules, further reductive domains are often present such as the KR, DH and ER, which reduce substrate to a beta-hydroxyl (as in PikA modules 1 or 5), or introduces alpha-beta unsaturation (as in module 2) or full saturation (as in module 4). Helical coiled-coil N-and C-terminal docking domains mediate the protein-protein interactions (PikAI to II, II to III, and III to IV).[20] The PikA gene cluster is unique in that it cyclizes and releases either a 12 or 14-membered ring macrolactone product in high yields (1c).[21] Additional enzymes are often present in PKS systems to perform tailoring reactions, such as oxidation and glycosylation.[22] The PikAIII/IV *in vitro* biochemistry model system in the Sherman laboratory consists of two type I PKS monomodules with four and four active-sites in series, respectively, with covalent intermediates at all but one site (PikAIII KR).[23] *In vitro* biochemical reactions of the PikAIII/IV proteins have been carried out with natural chain elongation intermediates,[17,24] methylmalonyl CoA, and NADPH. Structural biology has also provided intriguing insight into these systems, for example by illustrating a hydrophilic barrier mechanism for TE catalyzed cyclization.[25,26] The engineering of these systems, either rational or combinatorial, has much promise for generation of novel therapeutics.[27-30] By utilizing a "Legoization" strategy of mixing PKS module building blocks in a combinatorial manner, Santi and coworkers have shown that up to 72 out of 154 bimodular constructs tested were productive. Yet clearly hurdles remain to be overcome, as yields were found to be 0.023-23 mg/L in *E. coli*, much less than the g/L scale typically employed for commercial fermentation. Such methodologies have not been proven to scale for larger molecules, as in the case of heptaketides such as pikromycin, where six modules must productively interact.[17,31] ***Work in the pikromycin biosynthetic system has established the ability of in vitro biochemical investigations to inform and develop fundamental understanding as well as bio/chemical tools within the realm of natural product chemistry.***

1.4 Cryptophycin biosynthetic pathway

1.4.1 Cryptophycin isolation and biological activity

Cryptophycins, a large class of peptolides, were originally isolated from the cyanobacterium *Nostoc* sp. ATCC 53789 by researchers at Merck as a potent fungicide.

A gross structure was proposed, but Merck abandoned the project because the compounds were too toxic to be developed as antifungals.[32] Several years later, interest in the cryptophycins was renewed when a screen of the lipophilic extract of *Nostoc* sp. GSV 224 exhibited potent cytotoxic activity.[33] This activity was attributed to the cryptophycin natural products, which have since been found to have antimitotic activity and cytotoxicity toward tumor cells in culture, as well as anticancer activity against murine solid tumor models and human tumor xenografts.[34,35,36] While there are more than 25 naturally occurring analogs (in addition to the nearly structurally identical marine natural product arenastatin),[37] the major compound from both *Nostoc* sp. ATCC 53789 and *Nostoc* sp. GSV 224, cryptophycin 1 **(Figure 1-4)**, consists of four subunits: α-hydroxyoctenoic acid (Unit A), 3-chloro-*O*-methyl-D-tyrosine (Unit B), methyl-β-alanine (Unit C), and L-leucic acid (Unit D), linked in a cyclic clockwise sequence[33,38,39]. Other naturally occurring cryptophycins are analogs that differ from cryptophycin 1 by one or two of these subunits.

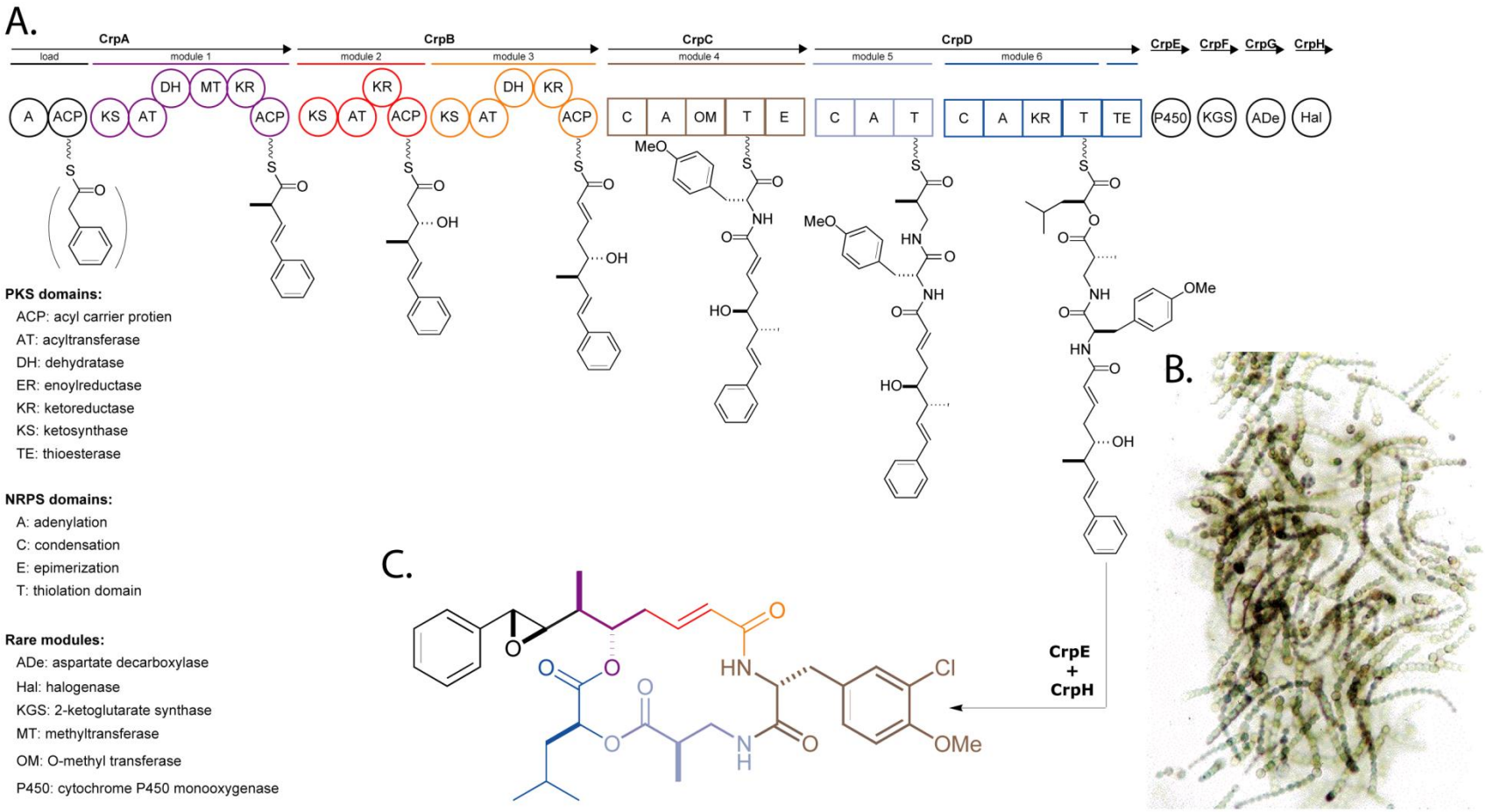


Figure 1-4. The cryptophycin biosynthetic pathway. (**A**) Cryptophycin biosynthesis, (**B**) producing organism, and (**C**) structure.

Cryptophycin 1 is one of the most potent tubulin destabilizing agents ever discovered, resulting in cellular arrest at the G2/M phase via hyperphosphorylation of Bcl-2, thereby triggering the apoptotic cascade.[40] Cryptophycins are also attractive as chemotherapeutics because they are active against multidrug-resistant tumor cell lines and are not substrates for *p*-glycoprotein pumps.[36] A synthetic analog, cryptophycin 52 (LY355703), was developed by Eli Lilly & Co. and ultimately reached phase II clinical trials; however, high production costs coupled with dose-limiting toxicity halted its development.[41] In spite of this setback, a subsequent phase II clinical trial involving patients with platinum-resistant advanced ovarian cancer concluded that the rate of disease stabilization in the absence of adverse events might justify further investigation of cryptophycin 52.[42] A second generation of cryptophycin 1 analogs with improved solubility properties has been synthesized and preclinical studies indicate a marked increase in efficacy against a variety of tumors.[43]

1.4.2 Cryptophycin gene cloning and sequence analysis

The Sherman and Moore laboratories worked collaboratively to isolate and characterize the cryptophycin gene cluster **(Figure 1-4)** using a strategy that relied on comparative metabolomic analysis.[44] In this approach, the A and KS domain sequences of *Nostoc* sp. ATCC 53789 were compared to those of *Nostoc punctiforme* ATCC 29133, a strain that does not produce cryptophycin. This comparative method resulted in the identification of six A domain sequences that were present in *Nostoc* sp. ATCC 53789 but not in *Nostoc punctiforme*. Of these six, a single A domain appeared to be a candidate for the cryptophycin pathway. From this initial lead, the 40 kb cryptophycin biosynthetic gene cluster was identified by cosmid library screening.[44]

The cryptophycin gene cluster consists of two modular PKS genes (*CrpA and CrpB*) and two modular NRPS genes (*CrpC and CrpD*). In total, these open reading frames encode seven elongation modules that contain the requisite catalytic domains for assembly of the cryptophycin macrocyclic core structure. In addition, a series of open reading frames, designated *CrpE-CrpH*, is located downstream of *CrpA-D*, and is predicted to encode enzymes that modify the nascent macrocycle to yield cryptophycin 1. These predicted enzymes include a cytochrome P450 epoxidase (*CrpE),* a putative 2-

ketoglutarate–dependent enzyme *(CrpF)*, an aspartate decarboxylase *(CrpG)*, and a flavin-dependent halogenase *(CrpH)*. In addition to characterization of the gene cluster, the Sherman and Moore groups have produced novel cryptophycin analogs by precursor-directed biosynthesis.[44] *In vitro* biochemical work has also been performed to characterize the TE domain of CrpD,[45,46] the aspartate decarboxylase, CrpG,[47] and the P450 epoxidase, CrpE.[48] These studies have unveiled the genetic blueprint of cryptophycin biosynthesis in *Nostoc* sp. ATCC 53789, thereby providing access to a set of catalytic tools for chemoenzymatic construction and modification of new cryptophycin analogs. ***This work has highlights the complexity of hybrid PKS/NRP biosynthetic systems and the potential application to supply drugs through chemoenzymatic synthesis once a deeper understanding of the biosynthetic principles inherent in the system is achieved.***

1.5 Curacin biosynthetic pathway

1.5.1 Curacin isolation and biological activity

Curacin A **(Figure 1-5)** is a mixed PK/NRP natural product with potent antiproliferative and antimitotic activity against colon, renal, and breast-cancer-derived cell lines.[49] The compound was originally isolated from strains of the tropical marine cyanobacterium *Lyngbya majuscula* discovered in Curaçao by Gerwick et al[50] and found to possess unusual structural features, including a cyclopropane group, thiazoline moiety, *cis*-alkenyl group, and terminal double bond. Curacin A has been shown to block cell cycle progression by interacting with the colchicine binding site on tubulin and inhibiting microtubule polymerization.[51] The clinical development of curacin has been hindered by its high lipophilicity; however, structural analogs having improved water solubility and potency have been recently synthesized to enable continued preclinical studies.[52,53]

1.5.2 Curacin gene cloning and sequence analysis

The Gerwick and Sherman laboratories conducted a series of isotope-labeled precursor feeding and NMR studies that established the metabolic origin of all curacin A atoms and their order of assembly.[54] The studies indicated the compound is composed of one cysteine residue, ten acetate units, and two *S*-adenosyl methionine-derived methyl groups, thus suggesting that curacin A was of mixed PK-NRP origin. Through the

creation and screening of a cosmid library from *L. majuscula* using a general PKS probe, a 64-kb gene cluster containing 14 ORFs was identified. As predicted by the precursor incorporation experiments, the curacin metabolic system **(Figure 1-5)** was found to contain nine PKS modules and one NRPS module. This biosynthetic system is unique in that all PKS multifunctional proteins, with the exception of the CurF hybrid PK/NRP, are monomodular.[54]

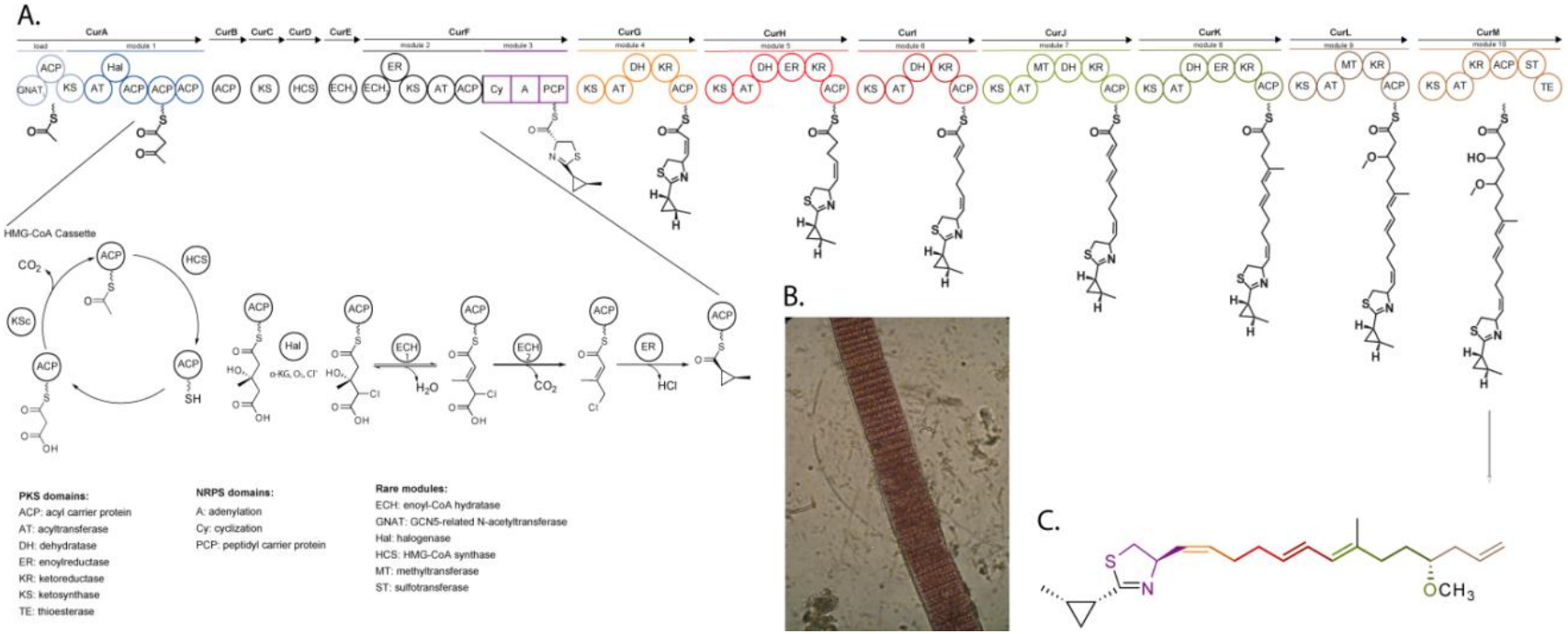


Figure 1-5. The curacin biosynthetic pathway. (**A**) curacin biosynthesis, (**B**) producing organism, and (**C**) structure.

Curacin A biosynthesis is initiated by the unique CurA PKS. Bioinformatic analysis of the AT domain located at the amino-terminus of CurA indicated homology with the *N*-acetyltransferase (GNAT) domain, PedI, from the putative pederin gene cluster.[54,55] Interestingly, a recent study established the role of an astonishing biochemical chain initiation strategy for the loading module of curacin A that involves an unusual tri-domain found at the amino-terminus of CurA. This tri-domain is comprised of an adapter domain, a GNAT domain, and an ACP domain. *In vitro* biochemical studies of the isolated tri-domain have shown that the GNAT has unprecedented bi-functional activity, as it is capable of first decarboxylating malonyl-CoA to acetyl-CoA and then directing the transfer from acetyl-CoA onto the ACP domain phosphopantetheine arm to produce the acetyl-ACP intermediate.[56]

A series of three tandem ACP domains (ACP_3) reside at the C-terminus of the CurA polypeptide that together with four ORFs encoding CurB-CurE, as well as the first two domains of CurF, was predicted to direct formation of the unique cyclopropyl ring in curacin A. Indeed, recent biochemical and structural studies confirmed that the CurE/CurF ECH_1-ECH_2 enzyme pair catalyzes successive dehydration and decarboxylation of (*S*)-HMG-ACP to generate a 3-methylcrotonyl-ACP intermediate for subsequent formation of the cyclopropane ring.[57,58] Moreover, the CurA ACP_3 domains have been shown to work synergistically resulting in enhanced catalytic output of the early chain elongation intermediate bearing the cyclopropyl ring.[59]

The remainder of the molecule is assembled by seven PKS monomodules, CurG-CurM that catalyze seven successive rounds of condensation with malonyl-CoA extender units. Furthermore, embedded methyltransferase domains in CurJ and CurL are predicted to catalyze transfer of the C-17 and O-13 methyl groups, respectively. Of final interest is the atypical biosynthetic termination mechanism that is predicted to function in both product release and decarboxylative dehydration to form the unusual terminal alkene. Like the majority of other known PK/NRP biosynthetic pathways, the final elongation module of the curacin pathway, CurM, contains a terminal thioesterase domain that was predicted to play a direct role in formation of the terminal olefin. Bioinformatic analysis of the CurM PKS monomodule also predicted the presence of a sulfotransferase (ST) domain immediately preceding the TE. ST domains are typically responsible for

transferring a sulfonate group from a donor molecule (such as 3'-phosphoadenosine-5'-phosphosulfate, PAPS) to a variety of acceptor carbohydrates, proteins, and other low-molecular weight metabolites.[60] Although STs had been previously characterized from both eubacterial and eukaryotic organisms, the presence of an ST domain within a PKS system was unprecedented.

Recent work has revealed the precise functions of the ST and TE domains in terminal olefin formation during termination of curacin biosynthesis (**Figure 1-5**).[61-62] The first step involves ST domain-catalyzed transfer of a sulfonate group (donated by PAPS) to the 3(*R*)-hydroxyl group of the ACP-bound thioester chain, followed by hydrolytic termination of curacin A biosynthesis by the TE to produce the linear free acid bearing a 3(*R*)-sulfate leaving group. High resolution X-ray crystal structure analysis of the CurM TE domain provides strong evidence that it catalyzes decarboxylation of the free acid, after which formation of the double bond would be energetically driven by elimination of the sulfate leaving group. Although it is conceivable that upon TE catalyzed hydrolysis, the decarboxylation reaction occurs spontaneously as a result of the presence of the sulfate leaving group at carbon 3, isolation of a substrate mimic bearing these two functional groups indicates that enzyme catalysis by the TE is required. Further efforts to develop this unique polyketide termination mechanism have important implications for facile conversion of fatty acid intermediates into valuable liquid fuels. ***This body of work serves as a model for the breadth and depth of knowledge that can be generated by collaboration among diverse researchers in the fields of synthetic chemistry, natural product isolation, in vitro biochemistry, structural biology and mass spectrometry within a single natural product biosynthetic platform.***

1.6 *Trans* AT domain pathways—a rich source of unusual biochemistry

1.6.1 Introduction to *trans* AT hybrid PK/NRP systems

One important subclass of hybrid PK/NRP pathways are the "*trans* AT" hybrid biosynthetic systems. Rather than containing embedded AT domains within their PKS modules, the *trans* AT systems feature a separately encoded, discrete AT domain that is responsible for loading ACP domains with the appropriate CoA substrate **(Figure 1-6)**. Interestingly, remnants of embedded ATs are found within *trans* AT hybrid pathways,

and have been proposed to act as "AT docking domains", or recognition elements, thus providing evidence of an evolutionary link between the two types of pathways.[63] ***This subclass of natural product synthases serves as an example of the great diversity in pathway architecture represented by these systems.***

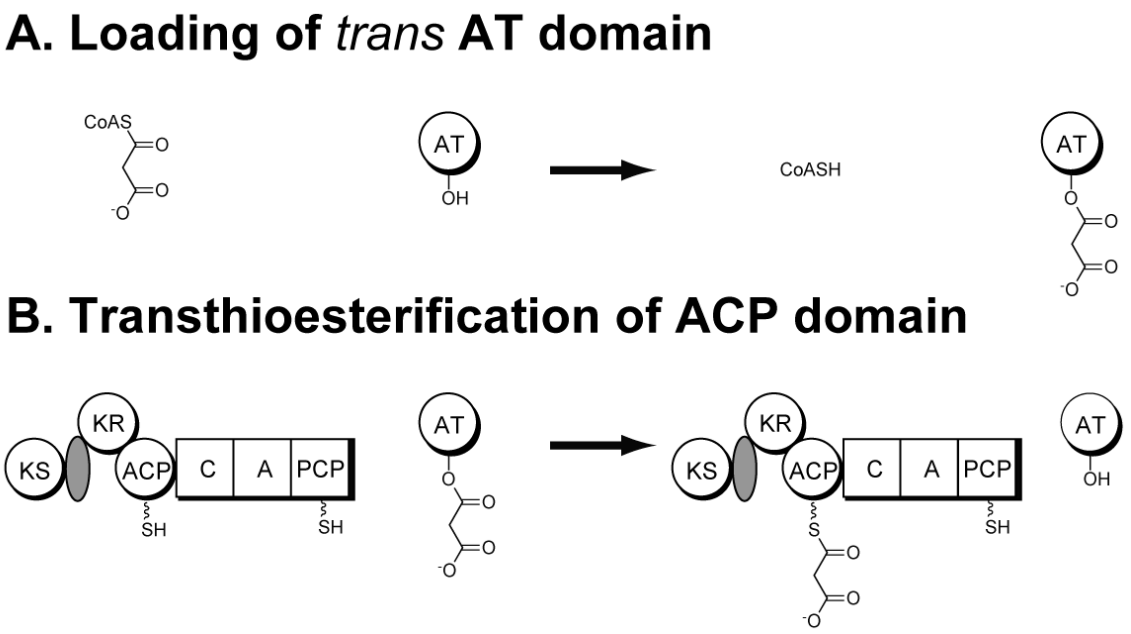


Figure 1-6. A schematic of a *trans* AT reaction scheme utilizing a hybrid PK-NRP biosynthetic module. (**A**) Loading of malonyl-CoA onto the AT-active site serine (**B**) Transfer from *trans* AT to the phosphopantetheine arm of the PKS module ACP.

Aspects of known PK/NRP *trans* AT hybrid biosynthetic pathways including products, pathway organization, *in vitro* and *in vivo* biochemistry, biological roles, and potential for future engineering efforts will be explored as examples of complex investigations into biosynthesis. Overlapping topics of interest include: hybrid PK/NRP pathways in general,[8,64] developing molecular tools to engineer these systems,[27] compounds derived from marine invertebrates and bacteria,[65] symbiotic bacteria produced secondary metabolites,[66,67] and absence of collinearity including skipping and iteration.[68]

1.6.2 Known *trans* AT hybrid PK/NRP pathways

Over the past several years, *trans* AT hybrid biosynthetic pathways have been discovered in diverse bacterial species (summarized in **Table 1-1, Figure 1-7**). Moreover, in several cases (e.g. onnamide, pederin, bryostatin and rhizoxin), the bacterial species is engaged in putative symbiotic relationships with multi-cellular hosts. These symbiotic relationships often complicate the identification of the species of origin of the hybrid PK/NRP natural product. With few exceptions, such as those involved in the biosynthesis of mycosubtilin and albicidin, the majority of *trans* AT hybrid pathways are predominantly composed of PKS elongation modules. In fact, it should be noted that *trans* AT pathways that are entirely comprised of PKS modules have been characterized: CpPKS1,[69] mupirocin,[70] macrolactam,[71] difficidin,[72] and bryostatin.[73,74]

Name	Activity	Bacterial class	Producing bacteria	host	PKS	NRPS
Albicidin	cytoxic	γ-proteobacteria	*Xanthomonas albilineans*	n/a	3	7
Chivosazol	cytoxic	δ-proteobacteria	*Sorangium cellulosum*	n/a	16	1
Disorazol	cytoxic	δ-proteobacteria	*Sorangium cellulosum*	n/a	10	1
Leinamycin	cytoxic	Actinobacteria	*Streptomyces atroolivaceus*	n/a	7	2
Onnamide	cytoxic	Unknown		sponge: *Theonella swinhoei*	?	?
Pederin	cytoxic	γ-proteobacteria	*Pseudomonas aeruginosa*	*beetle: Paederus sp.*	9	2
Rhizoxin	cytoxic	Betaproteobacteria	*Burkholderia rhizoxina*	fungi: *Rhizopus microsporus*	11	1
Antibiotic TA	antimicrobial	δ-proteobacteria	*Myxococcus xanthus*	n/a	11	1
Bacillaene	antimicrobial	Bacilli	*Bacillus subtilis*	n/a	13	2
Lankacidin	antimicrobial	Actinobacteria	*Streptomyces rochei*	n/a	5	1
Mycosubtilin	antimicrobial	Bacilli	*Bacillus subtilis*	n/a	1	7
Virginiamycin M	antimicrobial	Actinobacteria	*Streptomyces virginiae*	n/a	8	2
Thailandamide A		Bacilli	*Baccillis thailandensis*	n/a	16	1

Table 1-1. Known hybrid PK/NRP *trans* AT pathways. The known hybrid *trans* AT PK/NRP products are listed above by bioactivity, with bacterial class and latin bionomical name. For symbiont products, the host organism is also designated.

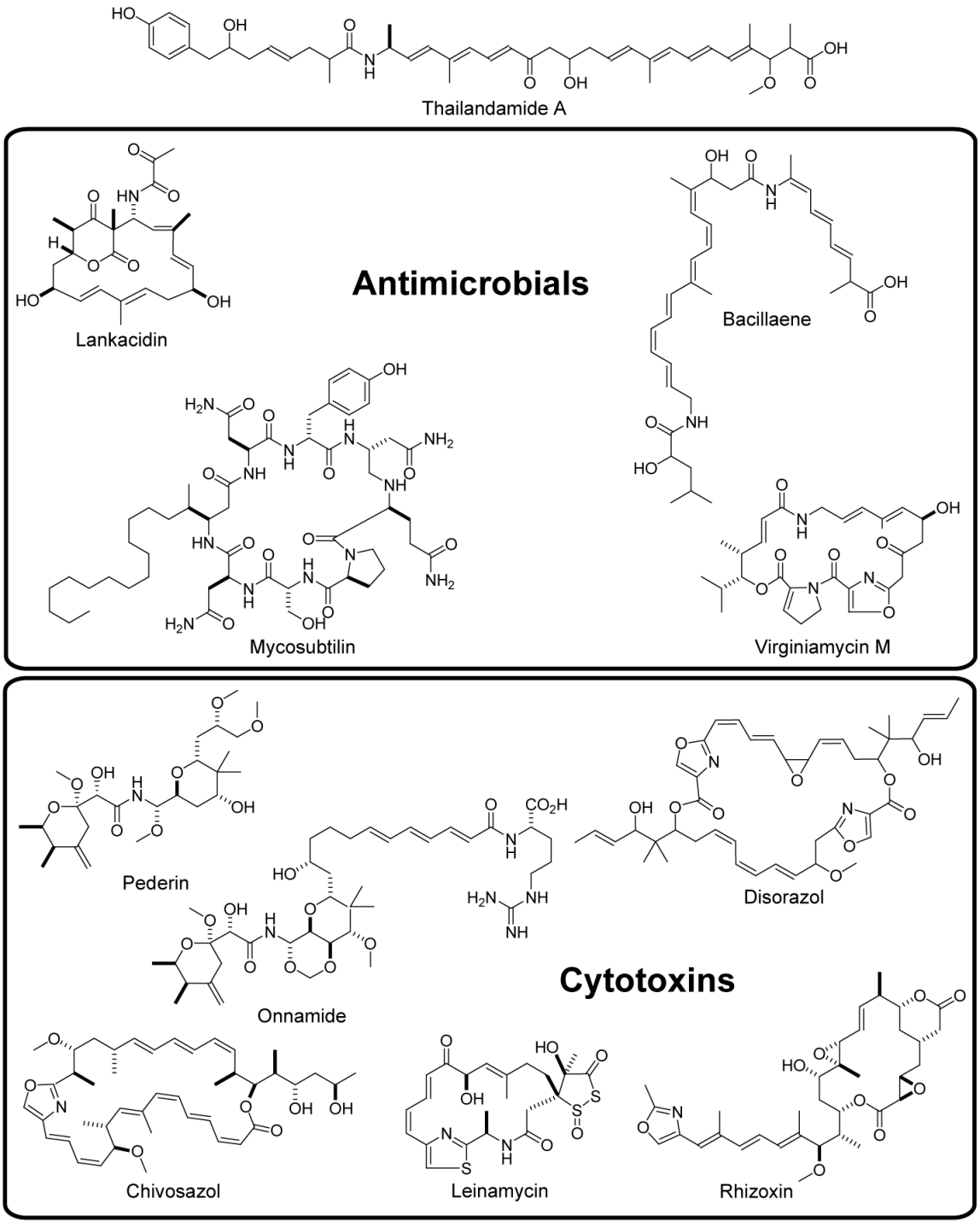


Figure 1-7. *Trans* AT hybrid PK/NRP biosynthetic pathways grouped by bioactivity.

1.6.3 Biological activity and structure of *trans* AT hybrid PK/NRPs

PK/NRP derived natural products that are assembled from *trans* AT biosynthetic pathways can be broadly classified as either antimicrobials or cytotoxic

chemotherapeutics based on their biological activity (**Table 1-1, Figure 1-7)**. Certain antimicrobial compounds, such as albicidin and mycosubtilin, have well defined activities while others, such as bacillaene, chivosazol, and lankacidin, have yet to be rigorously characterized in terms of biochemical targets. Given their ability to cause damage to rapidly growing cells, leinamycin, onnamide, and disorazol, each has potential as an anticancer therapeutic.

1.6.4 *In vivo* analysis of *trans* AT hybrid PK/NRP systems

To date, at least thirteen *trans* AT hybrid PK/NRP biosynthetic pathways have been described in the literature (**Table 1-1/Figure 1-7**). Intriguingly, aside from the presence of the *trans* AT domain, each of these pathways displays multiple deviations from the typical PK/NRP modular organization and composition. In fact, only five of these pathways are collinear with genetic organization, as is typically observed in bacterial PKS or NRPS systems. Module splits, whereby the domains of a single module are divided among multiple polypeptides, are frequently observed within *trans* AT hybrid pathways. 3-hydroxy-3-methylglutaryl synthase (HMGS) cassettes are also commonly found in these biosynthetic pathways. These cassettes are responsible for the insertion of β-branch points into the middle of the growing polyketide chain.[75] Repeated "tandem" domains, and unusual or uncharacterized enzymatic domains are also often present in biosynthetic pathways that employ *trans* AT domains.[68] Additional non-standard features, such as iteratively acting modules or inactive modules, can often be inferred from the chemical structure of the natural product.

Putative *trans* AT biosynthetic pathways can be linked to a specific natural product through comprehensive bioinformatic analysis which is used to determine domain compositions and predicted acyl-[76] or peptidyl-substrate[77] incorporation of AT or A domains, respectively.[78] However, in non-linear pathways, or in pathways that skip or iterate elongation modules, these predictions are challenging and can be misleading. Pathway assignment is typically obtained through genetic disruption and complementation. For example, if the inactivation of a key biosynthetic gene results in a non-producing phenotype, the pathway → product link is verified. However, this genetic approach is not a viable option for bacteria that are not culturable in the laboratory (often

the case for symbionts) or for microorganisms that are not amenable to genetic manipulation. In such cases, the final proof may require complete pathway reconstitution in a heterologous host–a task not yet accomplished for any symbiont pathway. Alternatively, detailed biochemical studies that provide direct evidence for conversion of a specific natural product biosynthetic intermediate for its cognate enzyme offers key information to correlate pathways from unculturable marine microbial symbionts. Studies on the *trans* AT and β-branching pathways in the bryostatin biosynthetic pathway have offered unique insights into this important marine natural product with anticancer and neuroprotective activity (**Chapter 3**).[74,79]

1.6.5 *In vitro* characterization of *trans* AT hybrid PK/NRP pathways

Given the many unusual features of the *trans* AT hybrid PK/NRP biosynthetic pathways, the precise sequence of compound assembly and the exact role of specific domains cannot always be easily ascertained from either sequence analysis or *in vivo* biochemistry. A more direct approach toward understanding these issues is to perform detailed *in vitro* biochemical investigations employing recombinant enzymes. Using defined assay conditions, detailed enzymology studies can provide important details of these hybrid mega-synthetases. Additionally, heterologous expression and purification of recombinant proteins enables the possibility of gaining key structural data, and might eventually inform new avenues toward pathway re-engineering *in vitro*. It is becoming increasingly apparent that rigorous *in vitro* examination of enzymes from a few select pathways has dramatically improved our understanding of the role of unusual domains and architecture in these pathways.[75,80-86]

1.6.6 Evolution, biology, and symbiosis of *trans* AT hybrid PK/NRP systems

Elucidation of the biological roles of PKS, NRPS and hybrid PK/NRP natural products that are produced by *trans* AT PK/NRP synthetases is a rapidly emerging field of research, particularly in relation to developing models of microbial symbiosis in natural product biosynthesis. Why do organisms expend so much energy and genome composition to generate these elaborate natural products? While the chemical ecology of some of these compounds has been explored, considerable work remains to understand

the role these compounds serve for the producing organism.[87] Indeed, even the identity of the organism (host versus symbiont) that is responsible for natural product biosynthesis is an area of intense interest. Macroscopic eukaryotes including insects, plants, marine sponges, and tunicates (**Chapter 5**) have long been recognized as sources of diverse natural products. Yet time and time again, experimental evidence strongly suggests that associated microorganisms are responsible for natural product biosynthesis, especially when similar compounds are isolated from taxonomically diverse producers. Cell separation experiments, as performed in the marine sponge *Theonella swinhoei,* have shown that the isolated fraction of bacteria co-localized with secondary metabolite production.[88] The complexity of this problem becomes evident when it is recognized that up to 40% percent of the mass of a sponge may be composed of bacteria, fungi, and other microorganisms.[89]

1.6.7 Onnamide and pederin biosynthetic pathways

1.6.7.1 Onnamide and pederin biological activity and structure of *trans* AT hybrid PK/NRP

Both onnamide and theopederin, a close analog of pederin, **(Figure 1-8)** are nanomolar inhibitors of protein synthesis, leading to induction of ribotoxic stress response, p38 kinase activity and apoptosis. These activities were discovered during a screen for activators of transforming growth factor β (TGF-β) expression.[90] It has been hypothesized that these compounds may directly bind to the eukaryotic ribosome, thus resulting in downstream activation of apoptotic pathways.[90]

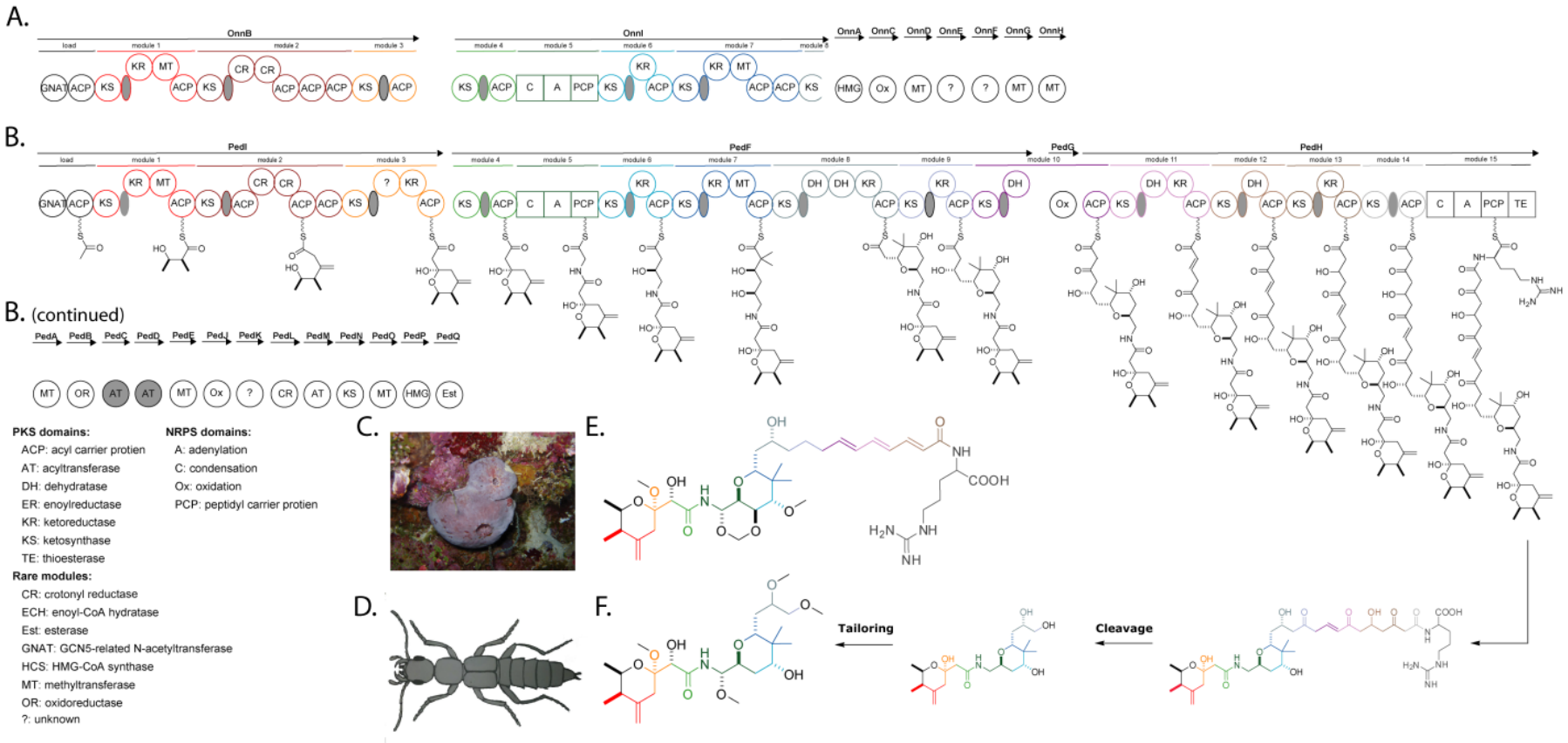


Figure 1-8. Onnamide and pederin biosynthesis. The onnamide (**A**), and pederin (**B**) biosynthetic pathways are displayed with predicted intermediates. The onnamide (**C**) and pederin (**D**) symbiont macroscopic host organisms are also displayed, as are the final products onnamide (**E**) and pederin (**F**).

Onnamide and pederin each contain a similar carbon backbone with two tetrahydropyran rings linked through an amide bond. Furthermore, each features an exocyclic double bond. Differences between the two natural product compounds include the presence of an additional hydroxyl group in onnamide that results in formation of a third six-membered ring. Onnamide also contains a longer, conjugated side chain that terminates in an arginine residue. The structural similarity shared between onnamide and pederin was proposed as evidence for the existence of a related symbiont producer in natural product biosynthesis[91,92] well before Piel obtained the gene clusters from the producing organisms.[93,94]

1.6.7.2 Onnamide and pederin *in vivo* biochemistry of *trans* AT hybrid PK/NRP

The biosynthesis of pederin and onnamide is discussed together due to the close chemical and biosynthetic similarities that are shared between the compounds and pathways.[93-95] The role of these pathways **(Figure 1-8)** in settling the long-standing debate over the source of natural products in marine invertebrates (e.g. sponges, tunicates) is discussed below. Both pederin and onnamide are initiated from a PKS elongation module (OnnB/PedI) that begins with a GNAT loading domain. A similar initiation mechanism is found in the curacin biosynthetic pathway.[56] The domain composition and arrangement is identical for each pathway over the first twelve domains, encompassing two PKS elongation modules. Interestingly, two unusual domains exist within these first twelve domains, and have been annotated as crotonyl-CoA reductases. Divergence in biosynthesis of the two molecules occurs beyond this point. The onnamide biosynthetic machinery proceeds with a tandem triple ACP, whereas pederin biosynthesis continues with a tandem di-ACP. In addition, the final PKS module of OnnB contains only KS and ACP domains, whereas the final module in PedB features a KS domain, a domain of unknown function, a KR domain, and finally an ACP domain. The subsequent polypeptides, OnnI or PedF, share a high degree of similarity, with the arrangement of the first twelve domains being identical. Briefly, each polypeptide begins with a PKS elongation module, followed by an NRPS module, and a second PKS module. At this point in the biosynthetic pathways, divergence is observed. Here, OnnI terminates with tandem ACP domains, followed by a KS domain. It should be noted that the DNA

sequence of the onnamide biosynthetic pathway is presumably incomplete, and therefore OnnI does not likely represent the terminus of the metabolic system. This incomplete sequence is because out of a 500,000 member clone library, only one cosmid containing the incomplete onnamide gene cluster was isolated.[94] In comparison, PedF continues with a single ACP domain that is followed by two additional PKS modules, the first of which contains a tandem DH didomain. At this point, two biosynthetic possibilities have been proposed. First, PedG catalyzes hydrolysis of the chain elongation intermediate from PedI and yields the pederin nascent intermediate (pre-tailored). Alternatively, chain extension continues through PedH, resulting in a product having a very similar structure to onnamide. This "onnamide" type intermediate would then presumably be cleaved to yield the beetle-derived product, pederin. The domain arrangement of PedH involves a presumed oxygenase, PedG, of the previous split module, PedF. PedH continues with four PKS elongation modules prior to the final arginine incorporating NRPS module that terminates with TE. Additional discrete proteins are also present in the pederin pathway. These include PedA/E/Q (methyltransferases), PedB (oxidoreductase), PedC/D/M (*trans* AT tri-domain), Ped J (oxidase), PedK (unknown function), Ped L (crotonyl-CoA reductase), PedN (KS), PedP (HMG-ACP synthase), and PedO (esterase). Several proteins with high sequence similarity are found in the onnamide pathway. These include OnnA (HMG-ACP synthase), OnnC (oxidase), OnnD/G/H (methyltransferases), and OnnE/F (unknown function).[93-95] As a matter of caution, it is important to note that these pathway assignments are considered putative, as they have not been confirmed as the metabolite source through heterologous expression or through detailed biochemical analysis of corresponding purified proteins. ***Research on the pederin and onnamide biosynthetic pathways illustrates the challenges and rewards of working with symbiont host systems—one of the emerging areas in studying natural product biosynthesis.***

1.6.8 Evolution, biology, and symbiosis of *trans* AT hybrid PK/NRP systems

One of the mysteries that have intrigued natural product chemists for years is how structurally similar natural products can be isolated from evolutionarily distinct hosts (e.g. marine sponges, myxobacteria). These discoveries have led to the hypothesis that microorganisms are the likely producers of marine invertebrate-derived natural products.

Support for this hypothesis was offered by Piel, who isolated, characterized and comparatively analyzed the DNA encoding the biosynthetic pathways of pederin and onnamide from the rove beetle and a marine sponge, respectively. The isolation and subsequent screening of DNA from the gut bacteria of the pederin source (rove beetle) eventually lead to the identification of the pederin pathway that originated from an unculturable symbiont.[93] Subsequently, Piel hypothesized that a similar symbiotic relationship accounted for the existence of a homologous pathway for onnamide production in the marine sponge *Theonella swinhoei.* Screening of a *T. swinhoei* metagenomic DNA library lead to the identification of a biosynthetic pathway that share high similarity with that of pederin.[96] The relationship of pederin and onnamide represents a fascinating example in which similar natural product biosynthetic gene clusters are derived from unique, phylogenetically distinct strains from widely disparate macroorganisms. Extensive gene sequencing has identified that *Pseudomonas aeruginosa* is the beetle endosymbiont responsible for pederin production.[97] To date, the microbial symbiont of the sponge has yet to be determined, but the DNA appears to be of bacterial origin.[95] Further questions regarding evolution and divergence of the two pathways, and host diversity remain to be fully explored.

Recent work to investigate the evolution of *trans* AT PKS systems, the main class of synthase isolated to date from marine microbial symbiotic organisms, may help frame the pertinent biological questions. Piel recently grouped known *trans* AT synthases and subjected them to multiple amino acid sequence alignments.[78] Interestingly, only the KS and MT domains showed conservation in all examined sequences. *Trans* AT KS specific clades formed in an intriguing manner. Domains did not cluster based on whether or not they were from the same gene cluster, as seen for *cis* ATs, but rather based on what final extension unit is generated. From this analysis, several insights are gained. First, it is possible to predict *trans* AT PKS product structure to a reasonable degree of accuracy, even in cases where pathways have reductive or β-branching domains acting *in trans*. Such a prediction was illustrated for thailandamide.[78] Secondly, and perhaps most importantly, Piel *et al* revealed that *trans* AT systems have likely evolved through a very different mechanism compared to their *cis* AT PKS counterparts. In *trans* AT pathways, horizontal gene transfer and recombination appear to be the driving force, as opposed to

recombination alone. This suggests that *trans* AT PKS pathways may be classified separately from their *cis* AT counterparts, much in the sense that PKS and NRPS pathways are to each other.[78]

The rapidly growing number of complete genome sequences from free-living and symbiotic bacteria, as well as from environmental samples is expected to lead to an increase in the number of characterized *trans* AT hybrid PK/NRP pathways.[78] Previous comparative studies made with significantly smaller amounts of sequence data began to show evidence of distinct clades forming between methylmalonyl-CoA *cis* ATs, malonyl-CoA *cis* ATs and *trans* AT modular PKS domains.[80] Interestingly, it appears that *trans* AT PKS or PK/NRP systems may be under-represented in current databases, as relatively larger numbers of *trans* AT PKS systems have been observed (relative to *cis* AT PKS systems) in random sequencing of bacterial strains.[98]

1.7 Technologies for probing biosynthetic pathways

1.7.1 DNA sequencing strategies in PK/NRP systems

When searching for bioactive compounds in nature, or the genes that direct their biosynthesis, it is best to survey a relatively large pool of structural or genetic diversity. However, given that 16S rRNA gene sequence analysis suggests that less then 0.1% of bacterial species collected in a marine sample are amenable to traditional laboratory culturing techniques, alternative approaches become necessary.[99] One strategy involves the collection of whole environmental DNA (eDNA) and subsequent screening for biosynthetic gene clusters based on homology to known genes. The disadvantage of this approach is that it typically requires densely populated DNA libraries (e.g. pederin and onnamide). Even having access to large DNA libraries does not ensure successful identification of desired gene clusters, as was demonstrated in the recent search for the discodermolide biosynthetic pathway from the sponge *Discodermia disoluta.* Screening of more than 150,000 cosmids produced over 4 GB of DNA sequence data, but failed to identify the gene cluster.[100] Interestingly 90% of the DNA sequenced appeared to be bacterial in nature thereby suggesting that the sponge does contain a diverse microbial community with high biosynthetic potential.[100] Because of this promise, investigators

continue to pursue the development of techniques to enable the efficient manipulation and screening of huge pools of DNA, such as clone pooling in semi-liquid medium.[101]

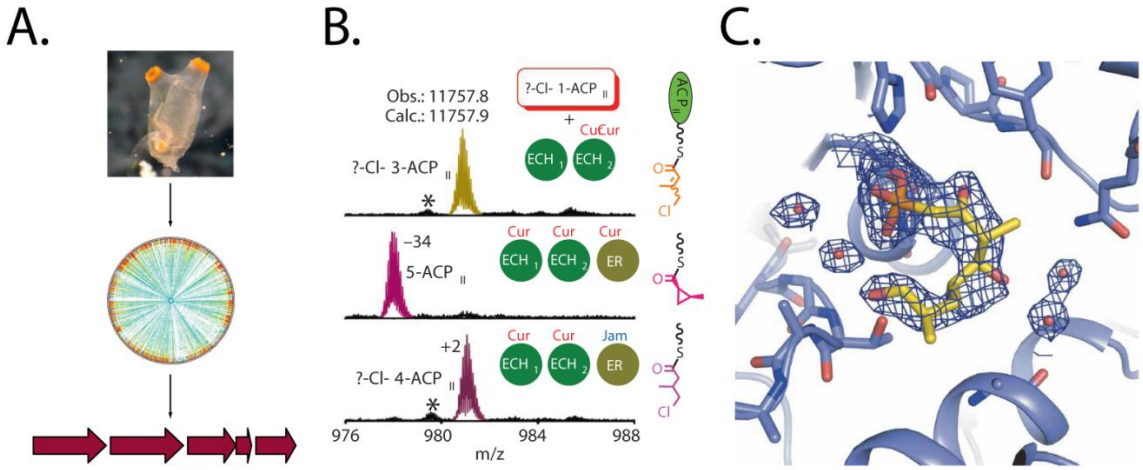


Figure 1-9. Technologies for improved analysis of natural product biosynthetic systems. (**A**) Metagenomic sequencing may be applied to symbiont-host systems to rapidly identify genes (**Chapter 5**), which can then be probed *in-silico* for biosynthetic genes. **(B)** Mass spectrometry allows for the direct interrogation of enzyme bound intermediates in biosynthetic pathway as illustrated above for curacin. (**C**) Covalent probes and X-ray crystallography allow for active site specificity determining structures to be mapped in biosynthetic pathways, as illustrated for the pikromycin thioesterase.

The rise of rapid and inexpensive whole genome DNA sequencing is also expected to have a profound impact on to the ability to access data from environmental samples of microbial consortia (**Figure 1-9**). Early experiments have demonstrated that there are seven diverse biosynthetic pathways in a single strain of *Salinispora tropica*.[91] Genome mining has also been applied in the search for rhizoxin pathway homologs in other source strains.[92] In the near future the ready access to inexpensive, high-throughput DNA sequencing will undoubtedly enable direct targeting of PKS and NRPS pathways from diverse metagenomic samples (**Chapter 5**).[102]

1.7.2 Mass spectrometry in PK/NRP systems

High performance mass spectrometry, particularly FTICR-MS experiments[103] have been conducted to characterize a wide variety of enzyme bound intermediates in PKS, NRPS, and hybrid PK/NRP pathways. Application of this technology has greatly enabled analysis of enzyme kinetics using radiolabel-free approaches by identifying and

characterizing intermediates with a high degree of sensitivity and selectivity.[103] FTICR-MS has also been applied toward the screening of new pathways. This work has relied on a phage display to express protein segments encoding thiolation domains, which are then identified in a high-throughput manner using loss of the phosphopantetheine prosthetic group as a specific signal.[104] New hybrid MS approaches, developed further by the Dorrestein laboratory are also highly innovative. By using MALDI imaging to localize marine natural products to a specific location, micro-manipulation can be employed to simplify the environmental sample prior to whole genome sequencing.[105] Finally, the combination of multiple analytical techniques such as enzyme kinetics, FTICR-MS, and X-ray crystallography, enables the enzymology of diverse PK/NRP systems to be explored in remarkable detail as was recently demonstrated in the curacin biosynthetic pathway β-branching cassette (**Figure 1-9**).[56]

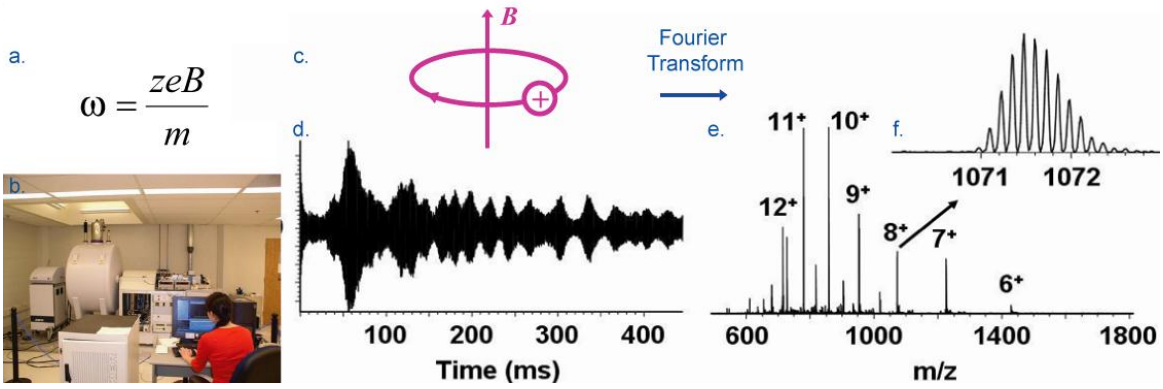


Figure 1-10. FTICR-MS methodology. (**A**) Descriptive equation for FTICR-MS (**B**) An FTICR-MS instrument (**C**) Cyclotron motion (**D**) Image current (**E**) Fourier transformed mass spectrum (**F**) Isotopic resolution of a single protein peak.[106]

FTICR-MS is a powerful analytical technique which allows for the mass to charge ratio, m/z, of an analyte to be measured with high resolution and mass accuracy. Ions are introduced into an ultra-high vacuum within a strong magnetic field, where they begin to orbit in a cyclotron motion (**Figure 1-10C**). This motion allows for determination of the m/z ratio as it is dependent only on the number, z, of elementary charges, e, and the magnetic field, B (**Figure 1-10C**). An RF pulse then excites molecules into a coherent packet that generates an image current over time **(Figure 1-10D**). This time domain signal is then Fourier transformed into the frequency domain, and calibrated (**Figure 1-**

10E). High resolution spectra enable isotopic splitting to be determined. This high resolution allows for unambiguous determination of analyte charge and conversion of the m/z ratio to mass at a low-ppm level of accuracy **(Figure 1-10F**).[107-109] The experimental configuration utilized allows for various fragmentation strategies to be employed, including collisionally induced dissociation (CID), infrared multiphoton dissociation (IRMPD), electron capture dissociation (ECD), electron capture dissociation (ETC), and electron detachment dissociation (EDD)—which can together offer complementary structural information.[110-113]

Peptides, often derived from enzymatic digestion of proteins, are a frequent target of MS/MS based fragmentation approaches described above. Indeed, this ability to identify peptides in the gas phase is a key step in many proteomics workflows. The nomenclature is typically described with the Biemann nomenclature of a, b, c and x, y, z ions for backbone fragmentation (**Figure 1-11**). In the gas-phase upon slow multi-step activation (e.g. IRMPD and CID) b- and y- ions are typically observed as well as b- ions that have undergone subsequent CO loss (a-ions). All of the other possible fragments can be observed with more exotic fragmentation techniques (e.g. ECD, EDD, ETD).

CID is the most commonly implemented fragmentation strategy across all instrument types. In CID, b-ions are often drawn as acylium ions, substantial experimental evidence suggests that they actual form cyclic products.[114] The location of basic residue can strongly effect fragmentation behavior with either "fixed-charge" or "mobile-proton directed" pathways dominating. Indeed, such chemical properties are manifested in residue specific fragmentation behavior.[115] This behavior can be both helpful in that presence or absence of key fragments can be used as supporting evidence of a CID MS/MS identification. Alternatively, such behavior in CID can be frustrating in that incomplete product ion coverage is often observed—thus prompting a need for alternative fragmentation strategies (e.g. ECD, EDD, ETD).

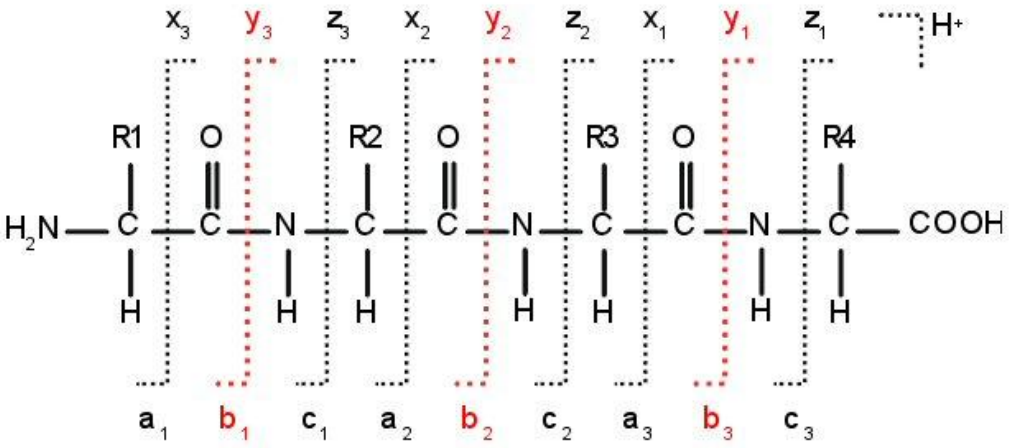


Figure 1-11. Peptide fragmentation nomenclature. Peptide backbone fragmentation resulting in a, b, c and x, y, z, ions is illustrated in the positive ion mode.

Electrospray ionization (ESI) "gently" introduces analytes, including large biomolecules, into the mass spectrometer in a distribution of several different charge states by means of electrostatic nebulization.[116] The degree to which an analyte is ionized during the transition to gas phase strongly affects the amount of analyte entering the mass spectrometer and therefore final signal. This ionization process is greatly affected by acidity/basicity, which, in turn, affects the degree of the total charge an analyte can compete for. Hydrophobicity, which determines analyte location relative to the surface of the evaporating droplets is also an important contributor to the ionization process. The sum of these and other factors is termed ionization efficiency and is one of the primary challenges in ESI-MS quantification.[117,118]

Enzyme kinetics by mass spectrometry is a rapidly developing field. Mass spectrometry is often thought of as a qualitative method, however, it also provides detailed relative quantification as demonstrated by recent advances in proteomics,[117] or even absolute quantification as in certain targeted phosphoproteomics applications.[118] The Leary group has implemented detailed steady-state kinetic methods to monitor both product and enzyme bound intermediates.[121-124] Kelleher's group examines thioester templated NRPS by top-down methodology to monitor enzyme-substrate intermediates.[125] They have promoted several assays, including substrate screening, active-site characterization, and facile gas-phase cleavage of phosphopantetheine-bound substrates.[86,103,126] The yersiniabactin system is a prime example of the power of this

top-down methodology, as both kinetic and chemical data could be gathered together by monitoring enzyme substrate intermediates. A limitation of that study was that kinetic parameters were relative in nature (% v_{max}) with no replicates performed due to the amount of time required for sample preparation (>1 hr).[127] Similarly, the nature of active-site transfer between homodimers in mixed PK/NRP systems was examined and transfer between dimers was found to occur.[15] MS investigations in other NRPS, mixed, or PKS systems have mostly focused on identification of enzyme-bound intermediates.[128-131]

1.7.3 Structural biology in PK/NRP systems

X-ray crystallography and NMR spectroscopy based structural analysis of biosynthetic enzymes can enhance our fundamental understanding of how these protein machines manufacture diverse natural products. The rate-limiting step for X-ray crystallography often resides in protein crystallization, however, high-throughput techniques to rapidly clone and express diverse gene constructs offer one viable strategy for overcoming this problem.[132] Fundamental questions, such as the nature of PK/NRP systems and whether they have a dimeric (PKS) or monomeric (NRPS) quaternary structure can be addressed by structural analysis, as was demonstrated with recent crystallographic work in the type I iterative fatty acid synthases.[133-138] The most structurally well characterized natural product biosynthetic system is the erythromycin (DEBS) modular PKS, which has been examined by X-ray crystallography through a series of excised catalytic domains and didomains.[139] One limiting factor in the complete structural determination of these megasynthases appears to be the overall flexibility of the ACP domains. Currently, structural information of ACPs has been derived by solution phase NMR studies.[140-142] Emerging approaches for studying large protein complexes such as cryo-EM may also be integrated with high resolution X-ray structural information as a step towards full understanding of these fascinating multi-component biochemical machines.

1.8 Summary

By studying the coupled PKS and NRPS programming of PK/NRP pathways a vast realm of biosynthetic space can be explored. Likewise, products of these pathways present a large pool of novel chemical entities that have been selected during evolution by providing an advantage to the producer/host. By leveraging these two pools of diversity, it is possible to access new tools to treat human health conditions within the sphere of cancer, immunomodulatory, infectious diseases and other areas, as illustrated above. These clinically relevant marine derived PK/NRP products represent a potent and expanding source of clinical leads. The coupling of this biosynthetic and chemical diversity is enabling us to take steps towards bypassing the traditional drawbacks of natural products research by providing facile access to metabolites through fermentation, and modification of existing products through pathway engineering.

Through further investigations of the unusual biosynthetic capabilities of the emerging class of non-canonical *trans* AT PKSs, we can hope to both expand our repertoire of capabilities while expanding our fundamental understanding of the flexibility of PK/NRP biosynthesis. Other developing topics such as the role of symbiosis in marine natural product biosynthesis seem to be uniquely located within unusual biosynthetic systems. It is encouraging to look forward with the hope of applying modern techniques in molecular biology, biochemistry, and analytical chemistry to further dissect and manipulate the mixed PK/NRP natural products through rational design of their PK/NRP biosynthetic machinery.

Portions of this chapter have been previously published in:

NRPS/PKS HYBRID ENZYMES AND THEIR NATURAL PRODUCTS
Christopher M. Rath, Jamie B. Scaglione, Jeffrey D. Kittendorf and David H. Sherman. In *Comprehensive Natural Products II: Chemistry and Biology*; Lew Mander, Hung-Wend Liu Editors; Elsevier: Oxford 2010; volume 1:453-492.

BIOSYNTHETIC PRINCIPLES IN MARINE NATURAL PRODUCT SYSTEMS
David H. Sherman, Christopher M. Rath, Jon Mortinson, Jamie B. Scaglione, and Jeffrey D. Kittendorf. In ***Natural Products: A Textbook***, William Gerwick Editor; Text in preparation.

NIH support for research on PK/NRP systems in the Sherman laboratory is gratefully acknowledged through grants GM076477, CA108874, and ICBG U01TW007404.

1.9 References

1. Newman, D. J.; Cragg D. A. *J Nat Prod,* **2007**, *70,* 461.

2. Butler, M. S. *Nat Prod Rep,* **2008**, *25*, 475.

3. Keller, K. B. *et al. Am Journal Crit Care,* **2005**, *14*, 338.

4. Nicolaou, K. C.; *et al*. *J Am Chem Soc,* **2000**, *122*, 9939.

5. Wohlleben, W.; Pelzer, S. *Chemistry & Biology,* **2002**, *9,* 1163.

6. Dobson, C. M. *Nature,* **2004**, *432*, 824.

7. Fischbach, M. A.; Walsh, C. T. *Chem Rev,* **2006**, *106,* 3468.

8. Walsh, C. T. *Science,* **2004**, *303*, 1805.

9. Austin, M. B.; Joel, N. P. *Nat Prod Rep*, **2003**, *20,* 79.

10. Hopwood, D. A. *Chem Rev,* **1997**, *97*, 2465.

11. Gruschow, S.; *et al*. *Chembiochem,* **2007**, *8,* 863.

12. Staunton, J.; Weissman, K. J. *Nat Prod Rep,* **2001**, *18,* 380.

13. Marahiel, M. A.; Stachelhaus, T.; Mootz, H. D. *Chem Rev*, **1997**, *97*, 2651.

14. Kopp, F.; Marahiel, M. *Current Op Biotech,* **2007**, *18,* 513.

15. Hicks, L. M.; *et al*. *Chem Biol,* **2004**, *11*, 327.

16. Lang, *et al. Nat Product Rep,* **2008**, 71, 1595.

17. Aldrich, C. C.; Beck, B. J.; Fecik, R. A.; Sherman, D. H. *J Am Chem Soc,* **2005**, *127*, 8441.

18. Hill, A. M. *Nat Prod Rep,* **2006**, *23*, 256.

19. Mercer, A.C.; Burkart, M.D. *Nat Prod Rep*, **2007**, *24*, 750.

20. Buchholz, T. J.; *et al. ACS Chem Biol*, **2009**, *4*, 41.

21. Kittendorf, J. D.; *et al. Chem Biol*, **2007**, *14*, 944.

22. Sherman, D. H. *et al. J Biol Chem*, **2006**, *281*, 26289.

23. Xue, Y.; Zhao, L.; Liu, H. W.; Sherman, D. H. *Proc Natl Acad Sci USA*, **1998**, *95*, 12111.

24. Beck, B. J.; *et al*. *J Am Chem Soc*, **2003**, *125*, 12551.

25. Akey, D. L.; *et al. Nat Chem Biol*, **2006**, *2*, 537.

26. Tang, G.-L.; Cheng, Y.-Q.; Shen, B. *J Biol Chem*, **2007**, *282*, 20273.

27. Kittendorf, J. D.; Sherman, D.H. *Cur Op Biotech*, **2006**, *17*, 597.

28. Rix, U.; Fischer, C.; Remsing, L. L.; Rohr, J. *Nat Prod Rep*, **2002**, *19*, 542.

29. Menzella, H. G.; Carney, J. R.; Santi, D. V. *Chem & Biol*, **2007**, *14*, 143.

30. Fortman, J. L.; Sherman, D. H. *Chembiochem*, **2005**, *6*, 960.

31. Menzella, H. G.; *et al. Nat Biotech*, **2005**, *23*, 1171.

32. Schwartz, R. E.; *et al. J Ind Microbiol*, **1990**, *5*, 113.

33. Trimurtulu, G.; *et al. J Am Chem Soc*, **1994**, *116*, 4729.

34. Corbett, T. H.; *et al*. *J Exp Ther Oncol*, **1996**, *1*, 95.

35. Panda, D.; *et al*. *Biochem*, **1997**, *36*, 12948.

36. Smith, C. D.; *et al*. *Cancer Res*, **1994**, *54*, 3779.

37. Kobayashi, M.; *et al. Tet Let*, **1994**, *35*, 7969.

38. Subbaraju, G. V.; Golakoti, T.; Patterson, G. M.; Moore, R. E. *J Nat Prod*, **1997**, *60*, 302.

39. Chaganty, S.; *et al*. *J Nat Prod*, **2004**, *67*, 1403.

40. Lu, K.; *et al*. *Cancer Chemother Pharmacol*, **2001**, *47*, 170.

41. Edelman, M. J.; *et al*. *Lung Cancer,* **2003**, *39*, 197.

42. D'Agostino, G.; *et al*. *Int J Gynecol Cancer,* **2006**, *16*, 71.

43. Liang, J.; *et al*. *Invest New Drugs,* **2005**, *23*, 213.

44. Magarvey, N. A.; *et al*. *ACS Chem Biol,* **2006**, *1*, 766.

45. Seufert, W.; Beck, Z. Q.; Sherman, D. H. *Angew Chem Int Ed Engl,* **2007**, *46*, 9298.

46. Beck, Z. Q.; *et al*. *Biochem,* **2005**, *44*, 13457.

47. Beck, Z. Q.; Burr, D. A.; Sherman, D. H. *Chembiochem,* **2007**, *8*, 1373.

48. Ding, Y.; Seufert, W. H.; Beck, Z. Q.; Sherman, D. H. *J Am Chem Soc,* **2008**, *130*, 5492.

49. Verdier-Pinard, P.; *et al*. *Mol Pharmacol,* **1998**, *53*, 62.

50. Gerwick, W. H.; *et al*. *J Org Chem,* **1994**, *59*, 1243.

51. Blokhin, A. V.; *et al*. *Mol Pharmacol,* **1995**, *48*, 523.

52. Wipf, P.; Reeves, J. T.; Balachandran, R.; Day, B. W. *J Med Chem,* **2002**, *45*, 1901.

53. Wipf, P.; Reeves, J. T.; Day, B. W. *Curr Pharm Des,* **2004**, *10*, 1417.

54. Chang, Z.; *et al*. *J Nat Prod,* **2004**, *67*, 1356.

55. Piel, J.; Wen, G.; Platzer, M.; Hui, D. *Chembiochem,* **2004**, *5*, 93.

56. Gu, L. *et al*. *Science,* **2007**, *318*, 970.

57. Gu, L.; *et al*. *J Am Chem Soc,* **2006**, *128*, 9014.

58. Geders, T. W.; *et al*. *J Biol Chem,* **2007**, *282*, 35954.

59. Gu, L.; *et al*. *Agnew Chem Int Ed*, **2011**, *12*, 2795.

60. Negishi, M.; *et al*. *Arch Biochem Biophy,* **2001**, *390*, 149.

61. Gu, L.; *et al*. *J Am Chem Soc*, **2009**, *131*, 6033.

62. Gehret, J.J; *et al*. *J Biol Chem*, **2011**, *27*, epub ahead of print.

63. Tang, G.-L.; Cheng, Y.-Q.; Shen, B. *Chem & Biol*, **2004**, *11*, 33.

64. Du, L.; Shen, B. *Cur Op Drug Disc Devel*, **2001**, *4*, 215.

65. Fortman, J. L.; Sherman, D. H. *Chembiochem*, **2005**, *6*, 960.

66. Piel, J. *Nat Prod Rep*, **2004**, *21*, 519.

67. Schmidt, E. W. *Nat Chem Biol*, **2008**, *4*, 466.

68. Moss, S. J.; Martin C.J.; Wilkinson B. *Nat Prod Rep*, **2004**, *21*, 575.

69. Zhu, G.; *et al. Gene*, **2002**, *298*, 79.

70. El-Sayed, A. K.; *et al. Chem & Biol*, **2003**, *10*, 419.

71. Ogasawara, Y.; *et al. Chem & Biol*, **2004**, *11*, 79.

72. Chen, X.-H.; *et al. Bacteriol*, **2006**, *188*, 4024.

73. Sudek, S.; *et al. J Nat Prod*, **2007**, *70*, 67.

74. Lopanik, N. B.; *et al. Chem & Biol*, **2008**, *15*, 1175.

75. Calderone, C. T. *Proc Nat Acad Sci USA*, **2006**, *103*, 8977.

76. Yadav, G.; Gokhale, R. S.; Mohanty, D. *J Mol Biol*, **2003**, *328*, 335.

77. von Dohren. H.; Dieckmann, R.; Pavela-Vranic, M. *Chem & Biol*, **1999**, *6*, R273.

78. Nguyen, T.; *et al. Nat Biotech*, **2008**, *26*, 225.

79. Buchholz, T. J.; *et al. Chem & Biol*, *17*, 1092.

80. Cheng, Y. Q. *Proc Nat Acad Sci USA*, **2003**, *100*, 3149.

81. Tang, G. L.; Cheng, Y. Q.; Shen, B. *J Nat Prod*, **2006**, *69*, 387.

82. Aron, Z. D.; *et al*. *J Am Chem Soc*, **2005**, *127*, 14986.

83. Reddick, J. J.; Antolak, S. A.; Raner, G. M. *Biochem Biophysical Res Com,* **2007**, *358*, 363.

84. Hansen, D. B.; *et al. J Am Chem Soc,* **2007**, *129*, 6366.

85. Aron, Z. D. *Chembiochem,* **2007**, *8*, 613.

86. Dorrestein, P. C.; *et al. Biochem,* **2006**, *45*, 12756.

87. Fischbach, M. A.; Walsh, C. T.; Clardy, J. *Proc Nat Acad Sci USA,* **2008**, *105*, 4601.

88. Carole A. Bewley, Faulkner, D.J.. *Ang Chem Int Ed*, **1998**, *37*, 2162.

89. Haygood, M.G.; Schmidt, E.W.; Davidson, S. K.; Faulkner, D.J. *J Microbial Biotech,* **1999**, *1*, 33.

90. Lee, K.-H.; *et al. Cancer Sci*, **2005**, *96*, 357.

91. Perry, N. B. *J Am Chem Soc,* **1988**, *110*, 4850.

92. Sakemi, S.; *et al*. *J Am Chem Soc,* **1988**, *110*, 4851.

93. Piel, J. *Proc Nat Acad Sci USA,* **2002**, *99*, 14002.

94. Piel, J. *et al. Proc Nat Acad Sci USA,* **2004**, *101*, 16222.

95. Piel, J.; *et al. J Nat Prod*, **2005**, *68*, 472.

96. Piel, J.; *et al. Proc Nat Acad Sci USA,* **2004**, *101*, 16222.

97. Piel, J.; Hofer, I.; Hui, D. *J Bacteriol,* **2004**, *186*, 1280.

98. Li, Z.-F.; *et al. Sys Ap Microbiol,* **2007**, *30*, 189.

99. Webster, N. S.; Wilson, K. J.; Blackall, L. L.; Hill, R. T. *Appl Environ Microbiol,* **2001**, *67*, 434.

100. Schirmer, A.; *et al. Ap Environ Microbiol*, **2005**, *71*, 4840.

101. Hrvatin, S.; Piel, J. *J Microbiol Meth,* **2007**, *68*, 434.

102. Piel, J.; Hui, D.; Fusetani, N. Matsunaga, S.. *Environ Microbiol*, **2004**, *6*, 921.

103. Dorrestein, P. C.; Kelleher, N. L. *Nat Prod Rep*, **2006**, *23*, 893.

104. Yin, J.; *et al. Chem & Biol*, **2007**, *14*, 303.

105. Esquenazi, E.; *et al. Mol Biosys*, **2008**, *4*, 562.

106. Hakansson, K.; Cooper, H. J.; Hudgins, R. R.; Nilsson, C. L. *Cur Org Chem*, **2003**, *7*, 1503.

107. Marshall, A.G.; Grosshans, P.B. *Anal Chem*, **1991**, *63*, A215.

108. Amster, I. J. *J Mass Spec*, **1996**, *31*, 1325.

109. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom Rev*, **1998**, *17*, 1.

110. Laskin, J.; Futreil, J.H. *Mass Spectrom Rev*, **2003**, *22*, 158.

111. Little, D. P. *Anal Chem*, **1994**, *66*, 2809.

112. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F.W. *J Am Chem Soc*, **1998**, *120*, 3265.

113. Budnik, B. A.; Haselmann, K.F.; Zubarev, R.A. *Chem Phys Let*, **2001**, *342*, 299.

114. Paizs, B.; Suhai, S. *Mass Spec Rev*, **2005**, *24*, 508-.

115. Huang, Y.; *et al. Anal Chem*, **2005**, *77*, 5800.

116. Fenn, J. B.; *et al. Science*, **1989**, *246*, 64.

117. Cech, N. B.; Enke, C. G. *Anal Chem*, **2000**, *72*, 2717.

118. Cech, N. B.; Enke, C. G. *Mass Spectrom Rev*, **2001**, *20*, 362.

119. Lill, J. *Mass Spectrom Rev*, **2003**, *22*, 182.

120. Cutillas, P. R.; *et al. Proc Natl Acad Sci USA*, **2006**, *103*, 8959.

121. Ge, X.; *et al. Anal Chem*, **2001**, *73*, 5078.

122. Gao, H.H.; Leary, J.A. *J Am Soc Mass Spec*, **2003**, *14*, 173.

123. Pi, N.; Meyers, C. L.; Pacholec, M.; Walsh, C. T.; Leary, J. A. *Proc Natl Acad Sci USA*, **2004**, *101*, 10036.

124. Pi, N.; Yu, Y.; Mougous, J. D.; Leary, J. A. *Protein Sci*, **2004**, *13*, 903.

125. Dorrestein, P. C.; *et al. Biochem*, **2006**, *45*, 1537.

126. McLoughlin, S. M.; *et al. Biochem*, **2005**, *44*, 14159.

127. McLoughlin, S. M.; Kelleher, N. L. *J Am Chem Soc*, **2004**, *126*, 13265.

128. Schnarr, N. A.; Chen, A. Y.; Cane, D. E.; Khosla, C. *Biochem*, **2005**, *44*, 11836.

129. Hong, H. H.; *et al. FEBS J*, **2005**, *272*, 2373.

130. Zhai, H.; *et al*. *J Am Soc Mass Spectrom*, **2005**, *16*, 1052.

131. Hicks, L. M.; *et al*. *ACS Chem Biol*, **2006**, *1*, 93.

132. Dunlap, W. C.; *et al. Cur Med Chem*, **2006**, *13*, 697.

133. Smith, S. *Chem & Biol*, **2002**, *9*, 955.

134. Maier, T.; Jenni, S.; Ban, N. *Science*, **2006**, *311*, 1258.

135. Jenni, S.; *et al. Science*, **2007**, *316*, 254.

136. Leibundgut, M.; Jenni, S.; Frick, C.; Ban, N. *Science*, **2007**, *316*, 288.

137. Smith, S.; Tsai, S-C. *Nat Prod Rep*, **2007**, *24*, 1041.

138. Maier, T.; Leibundgut, M.; Ban, N. *Science*, **2008**, *321*, 1315.

139. Khosla, C.; *et al. Annu Rev Biochem*, **2007**, *76*, 195.

140. Alekseyev, V. Y.; *et al. Protein Sci*, **2007**, *16*, 2093.

141. Mercer, A. C.; Burkart, M. D. *Nat Prod Rep*, **2007**, *24*, 750.

142. Zhou, Z. *Proc Nat Acad Sci USA*, **2007**, *104*, 11621.

Chapter 2

Acyl-CoA subunit selectivity in the terminal pikromycin polyketide synthase module: steady-state kinetics and active-site occupancy analysis by FTICR-MS

2.1 Introduction

Polyketides are a structurally diverse class of natural products that function as antifungals (amphotericin B), immunosuppressives (FK506), antibiotics (erythromycin A) and other important pharmaceuticals.[1] The medicinal value of these compounds has inspired efforts to design novel molecules by reprogramming the polyketide synthase (PKS) pathways responsible for their assembly. Toward this end, it is crucial that we develop a deeper understanding of the chemical processes encoded by these systems.[2,3] One current gap in our knowledge of modular PKSs is the mechanistic basis for substrate processing and discrimination towards acyl-coenzyme A (CoA) extender units (**Figure 1, steps I-II**).

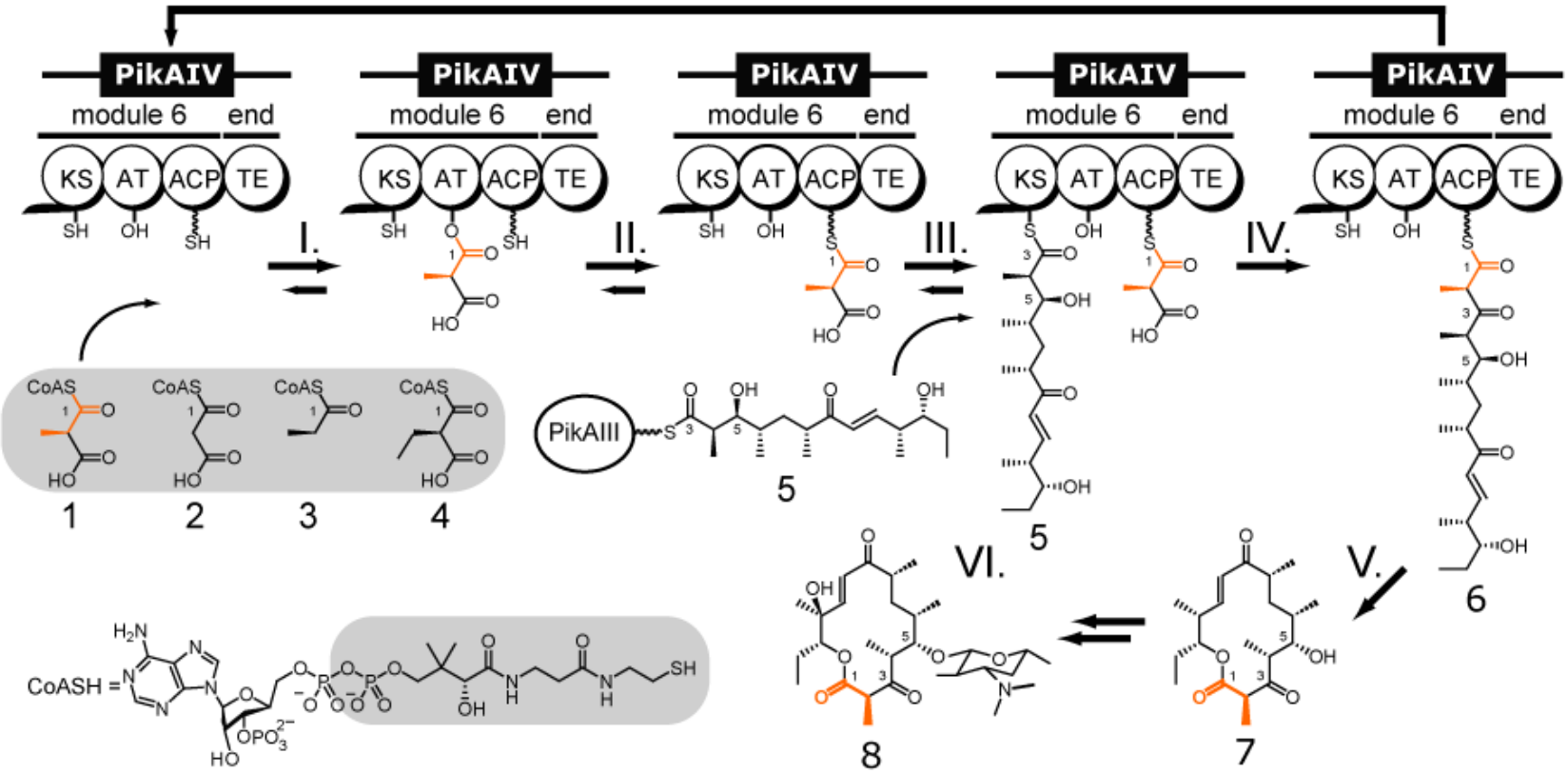


Figure 2-1. Catalytic cycle for PikAIV. Acyl-CoA extender units (**1** native, **2-4** unnatural) are loaded onto the AT active-site serine (step I) and undergo transthioesterification to the ACP phosphopantetheine (grey portion of CoAS, step II). The hexaketide chain elongation intermediate (**5**) condenses with the MM-CoA extender unit (steps III-IV) to form the heptaketide (**6**) on the ACP prior to TE cyclization[28] to form narbonolide (7, step V) processed to pikromycin (8, step VI).[29] Reversible steps are noted by a backwards arrow. Off-pathway reactions (ex. hydrolysis from the active-site residues) are not illustrated.

Previous investigations have led to the proposal that acyltransferase (AT)-bound extender units are stable in modular PKS and the related fatty acid synthase (FAS) systems, with deacylation occurring only in the presence of a specific thiol acceptor (e.g. CoA or panthetheine).[4-7] However, in the 6-deoxyerythronolide B synthase (DEBS) PKS, modules act as methylmalonyl-CoA (MM-CoA) hydrolases based on loss of radioactivity from [1-^{14}C]-MM-CoA (1) labeled proteins.[8,9] Thus, in the absence of a chain elongation intermediate, the extender unit may be hydrolytically released from the protein as methylmalonate (MM). Similar mechanisms have been proposed in FAS systems.[10,11] Fundamental aspects of this process including the specific site and rates of hydrolysis, catalytic domains involved, and the molecular basis for acyl-CoA extender unit selectivity have not been reported.

In the current study, we investigated the fate of the acyl-CoA extender unit (**Figure 1, Step I-II**) and the role of the four catalytic domains (KS-AT-ACP-TE) of PikAIV (pikromycin (Pik) PKS module 6). Two complementary assays enabled us to probe this system including 1) a fluorescent assay using ThioGlo-1 to monitor the steady-state kinetics of extender unit uptake/free-CoA release, and 2) a Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS) method for directly monitoring active-site occupancy. Covalently linked intermediates at the KS (C207), AT (S652), ACP (S980 holo), and TE (S1196) were assessed by FTICR-MS with the data leading to a new mechanistic hypothesis for acyl-CoA processing.[12-14] These assays are complementary in that the steady-state kinetic data provides a direct read-out of catalysis in the system by monitoring substrate utilization, while the FTICR-MS assay rigorously interrogates the chemical occupancy of the enzymatic machinery.

Four extender units were utilized in our analysis, including native MM-CoA, malonyl-CoA (M-CoA), propionyl-CoA (P-CoA), and ethylmalonyl-CoA (EM-CoA). The use of specific PKS variants in addition to the holo (phosphopantetheinylated) wild-type (WT) PikAIV (S980 holo), including: dKS (C207A/S980 holo), dAT (S652A/S980 holo), apo (S980 apo) dTE (S1196A/S980 holo), and apo/dTE (S980 apo/S1196A) enabled us to assess the importance of individual domains in extender unit processing and hydrolytic activity (**Figure 2-2**).[15]

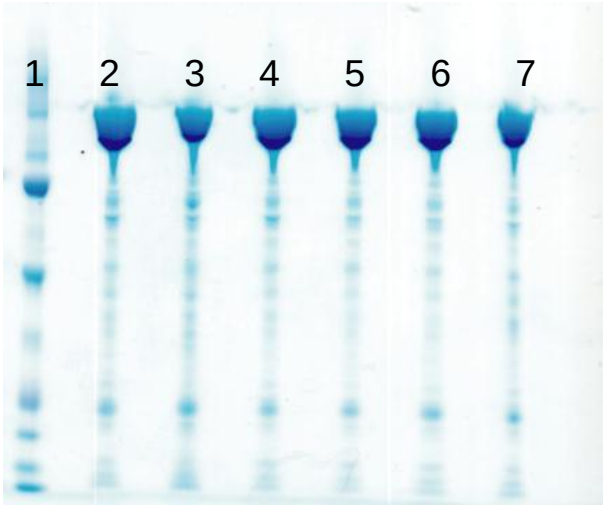


Figure 2-2. Expression of PikAIV variants. 8-12% NuPAGE SDS gel with Simply Blue Safe Stain (Invitrogen). Lanes are MWM (1), PikAIV WT (2), PikAIV Apo (3), PikAIV dKS (4), PikAIV dAT (5), PikAIV dTE (6), and PikAIV Apo/dTE (7).

2.2 Results

In our PikAIV functional assay, loading of MM-CoA (**Figure 1, Step I**) was very rapid compared to hydrolysis (**Table 2-1**) or 10-deoxymethynolide production (3.3 +/-0.4 min^{-1}).[16] Rates were directly monitored as loss of free AT (S652) active-site hydroxyl (35 +/- 23 s^{-1}), and the build-up of AT (S652) active-site bearing MM (21 +/-11 s^{-1}) (**Figures 2-3 and 2-4**) by rapid-quench and FTICR-MS.[17] Observed rates were similar

to values reported for mammalian FAS-AT reactions with radiolabeled substrates (43-150 s^{-1}).[10,11]

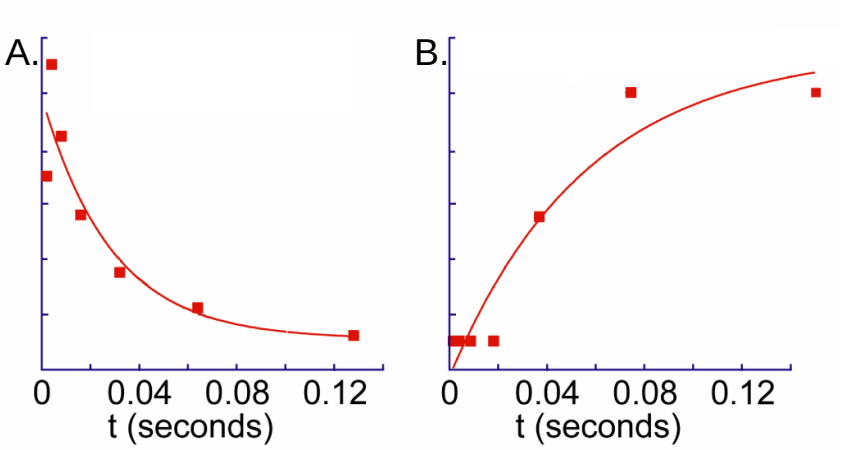


Figure 2-3. Fitting of rapid-quench time points. Loss of AT-OH (35 +/- 23 S^{-1}, **A**) and build-up of AT-MM (21 +/-11 s^{-1}, **B**) fitted to a single exponential curve for determination of transient kinetic rates. The y-axis is in arbitrary units reflecting the ratio of AT-IS peptide intensity to AT-MM or AT-OH.

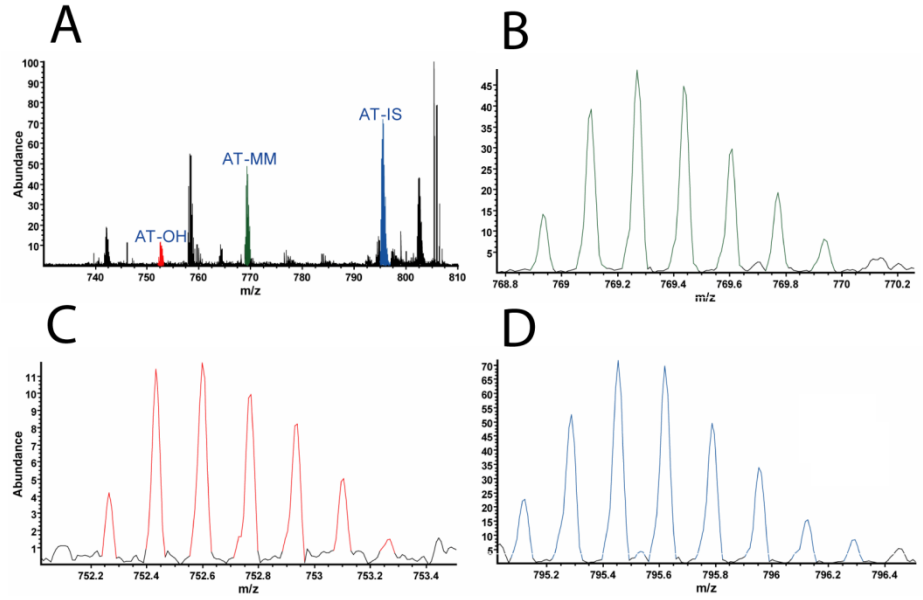


Figure 2-4. Example spectra for PikAIV KS-AT transient kinetic analysis. The mass spectrum (**A**) as well as zoomed AT-OH (**B**), AT-MM (**C**), and AT-IS (**D**) insets are shown.

The rates of MM-CoA hydrolysis (**Table 2-1**) were determined for the PikAIV variants under steady-state conditions in the absence of chain elongation intermediates with the ThioGlo-1 assay (**Figure 2-1, Steps I-II only**).[16] Values were determined under identical *in vitro* biochemical conditions as compared to previous analysis of the PikAIV system.[15] Holo PikAIV and the dKS (C207A) variant had the highest k_{cat} values (the dKS mutant had a slightly higher value, however, the reason for this is unclear). When the AT was inactivated (dAT, S652A) all hydrolytic activity was abolished. Inactivation of the ACP (S980 apo) or the dTE (S1196A) resulted in a cumulative effect for the apo-ACP/dTE (S980 apo/S1196A).

PikAIV	**k_{cat}(min^{-1})**	**% Decrease k_{cat}**
WT	1.04±0.08	0
dKS	1.18±0.07	-13
dAT	NA	100
Apo	0.81±0.08	21
dTE	0.80±0.10	23
Apo/dTE	0.70±0.01	32

Table 2-1. MM-CoA extender unit hydrolysis rates of PikAIV. Apparent k_{cat} was determined using Michaelis-Menton kinetics by ThiolGlo-1 assay with four replicates to calculate the standard deviation

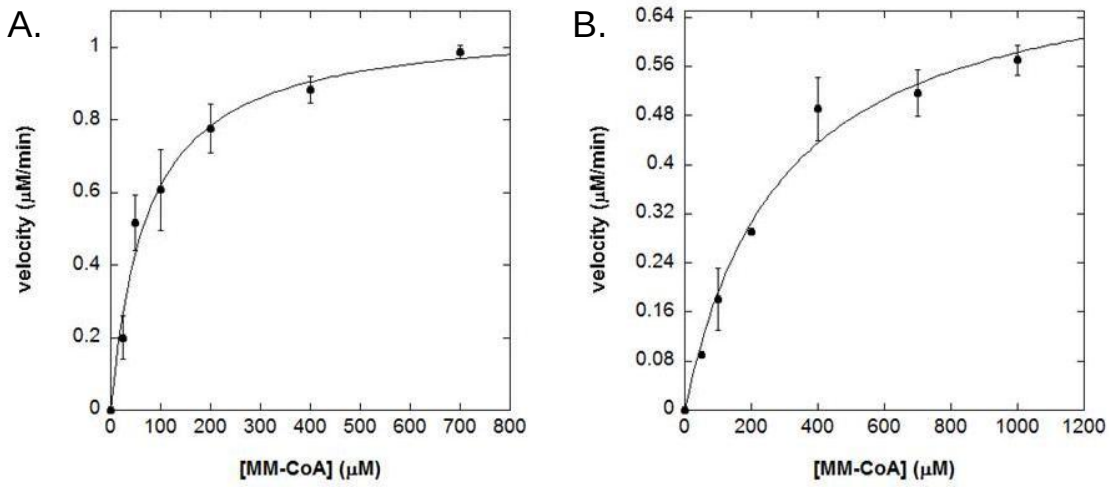


Figure 2-5. ThioGlo-1 plate reader assay for steady state kinetic analysis. Wild-type PikAIV (A) and apo PikAIV (B) in the presence of MM-CoA.

The analysis above revealed that the AT domain was the major site of hydrolysis accounting for ~70-80% of activity. The ACP and TE domains are responsible for the remaining 20-30% of the observed extender unit hydrolysis activity. MM is released as the free acid since the rate of triketide lactone formation from CoA extender units in PikAIV (0.00005-0.0001 min^{-1}) is substantially slower than the rate of hydrolysis determined in this study (**Table 2-1**).[18]

PikAIV	Substrate	KS%	AT%	ACP%	TE%
WT	MM-CoA	0±0	100±0	64±19	4±0
dKS	MM-CoA	X	100±0	51±9	1±0
dAT	MM-CoA	0±0	X	0±0	0±0
Apo	MM-CoA	0±0	100±0	X	0±0
dTE	MM-CoA	0±0	100±0	87±13	X
Apo/dTE	MM-CoA	0±0	100±1	X	X
WT	M-CoA	0±0	0±0	3±5	1±0
WT	P-CoA	0±0	0±0	19±13	0±0
WT	EM-CoA	0±0	90±10	52±46	6±2

Table 2-2. Extender unit active-site occupancy analysis by FTICR-MS with enzyme variants and alternative substrates. % is the apparent active-site loading by comparing free active-site and covalent +MM active-site peptides, Four replicates were run to calculate %RSD.

PikAIV active-site occupancy was determined by FTICR-MS under identical conditions to the ThioGlo-1 steady-state kinetic analysis (**Table 2-2**).[16] The KS active-site (C207A) was not loaded for any PikAIV variants. The AT active-site S652 was saturated (AT:100%) when active, and no downstream loading of the ACP or TE occurred in the dAT (S652A) variant form of PikAIV. Moderate loading on to the ACP (S980 holo) was seen with WT (ACP:64%) and dKS (ACP:51%, C207A), while more MM accumulated in the dTE (S1196A) variant (ACP:87%). In the apo ACP variant (S980 apo) the AT was saturated (AT:100%) with no downstream TE (S1196) loading observed.

The ThioGlo-1 steady-state kinetic analysis and MS active-site occupancy data for PikAIV support an *in vitro* biochemical model in the bacterial type I modular PKS where native extender unit MM-CoA is loaded onto the AT active-site (**Figure 2-1, step I**), with substantial hydrolysis occurring directly at this site. AT-bound MM is also transferred to the ACP (**Figure 2-1, step II**) and TE active-site, where further hydrolysis occurs.

The ability of PikAIV to select different acyl-CoA extender units (**Figure 2-1, compounds 1-4**) was monitored by active-site occupancy using FTICR-MS (**Table 2-2**). For all species, no significant loading was observed on the KS active-site (C207). For the disfavored (based on predicted AT-domain specificity/observed product formation)[19] M-CoA substrate, no loading on the AT domain (S652) was detected. Low amounts of loading were also observed on the ACP (ACP:3%, S980 holo) and TE (TE:1% S1196) active-sites. Similarly, for the "dead-end" non-extendable P-CoA substrate, loading onto the AT active-site (S652) was not detected, but low levels were observed on the ACP (ACP:19%, S980 holo). For EM-CoA, which is a rare extender unit in PKS biosynthesis, a high level of loading onto the PikAIV AT (AT:90%, S652), ACP (ACP:52%, 980 holo), and TE (TE: 6%, S1196) active-sites was evident by FTICR-MS.

This occupancy data with M-CoA and P-CoA demonstrated that alternate extender units can be loaded onto an AT and transferred to the adjacent ACP. The high level AT loading with MM-CoA and EM-CoA led us to reason that a level of selectivity may be realized through hydrolytic activity of the disfavored substrate from the AT and ACP active sites.

To test this hypothesis, hydrolysis was monitored by the ThioGlo-1 steady-state assay (**Table 2-3**). When incubated in the presence of MM-CoA a baseline rate of hydrolysis was observed for holo PikAIV (WT), with a 27% lower rate for the apo variant, consistent with the role of the ACP and TE in facilitating hydrolysis (**Table 2-3**). For holo PikAIV (WT), a 5-fold increase in the rate of hydrolysis occurs upon substitution with M-CoA, due to acylation and subsequent deacylation with this non-preferred extender unit. For apo PikAIV a 10-fold increase in hydrolysis was observed with M-CoA. Thus, in the case of M-CoA loaded PikAIV, slow transfer to the ACP domain could contribute to the attenuated rate of M-CoA hydrolysis observed for the holo compared to apo. Overall, these data support the hypothesis that substrate discrimination against malonyl-CoA is mediated by subunit loading followed by hydrolysis at the AT domain. If the substrate were not loaded, then the rate of enzyme-catalyzed hydrolysis would be negligible.

Substrate	**WT Rate (μM/min)**	**Apo Rate (μM/min)**	**Ratio of Rates WT/Apo**
MM-CoA	1.13±0.12	0.83±0.04	1.4
M-CoA	5.17±0.33	9.36±0.37	0.55
MM-CoA + M-CoA	1.45±0.31	1.79±0.10	0.81
P-CoA	2.16±0.26	1.23±0.04	1.7
EM-CoA	ND	ND	

Tables 2-3. Acyl-CoA extender unit hydrolysis rates for PikAIV. Apparent rate was determined by the ThiolGlo-1 assay with four replicates to calculate %RSD.

The ability of PikAIV to select the correct extender unit from a mixture was tested by employing equimolar amounts of M-CoA and MM-CoA (**Table 2-3**). The observed total rate for this mixture was substantially closer to that observed for MM-CoA alone with the ThioGlo-1 assay. Only MM-loaded active-site residues were detected by FTICR-MS in this competition experiment (data not shown). Thus, in the presence of MM-CoA the futile turnover of M-CoA may be reduced due to the comparatively slow rate of MM hydrolysis from the saturated PikAIV AT active site.

Studies on the non-extendable P-CoA subunit offer additional insights into the function of PikAIV. Hydrolysis rates toward this subunit were found to be at an intermediate value between MM-CoA and M-CoA (**Table 2-3**). The PikAIV holo ACP species exhibited a faster hydrolytic rate than the corresponding apo protein. Whether P-CoA is loaded directly *in vivo*, or occurs from spontaneous decarboxylation of MM-CoA, it is likely that additional editing mechanisms have been developed to off-load this, and other dead end intermediates (ex. PikAV TEII).[20]

The AT occupancy data for the EM-CoA (**Table 2-2**) is similar to MM-CoA suggesting that there is slow hydrolysis of this extender unit. Indeed neither the apo nor holo PikAIV variants resulted in detectable levels of hydrolytic activity of EM-CoA when analyzed in our ThioGlo-1 steady-state assay (**Table 2-3**). This suggests that relatively slow turnover (loading of extender unit and subsequent hydrolysis) is occurring with this extender unit. The ACP occupancy data indicated that PikAIV could accept EM-CoA as an alternate extender unit (**Figure 2-7**). A lack of selectivity between EM-CoA and MM-CoA for AT units is not unexpected. A number of polyketide products, such as monensin, are generated in various analog forms using either MM-CoA or EM-

CoA extender units at a specific point during chain elongation.[21-22] The EM and MM AT domain sequence motifs have a high degree of sequence similarity that have been predictive of AT acyl-CoA subunit selectivity in type I modular PKSs.[23-24] Moreover, it has been shown that an EM specific AT-domain substitution engineered into DEBS PKS module 5 can utilize MM-CoA (under limiting EM-CoA levels).[23-24] However, the observation that there is no significant hydrolysis of the EM-loaded PikAIV module is unexpected. A plausible hypothesis is that a relatively constrained AT active-site allows hydrolysis of smaller extender units but not larger ones, high resolution crystal structures with bound substrate models could be utilized to test this. Structural studies are likely to provide further insight into the mechanistic basis for subunit hydrolysis in PKS AT domains.

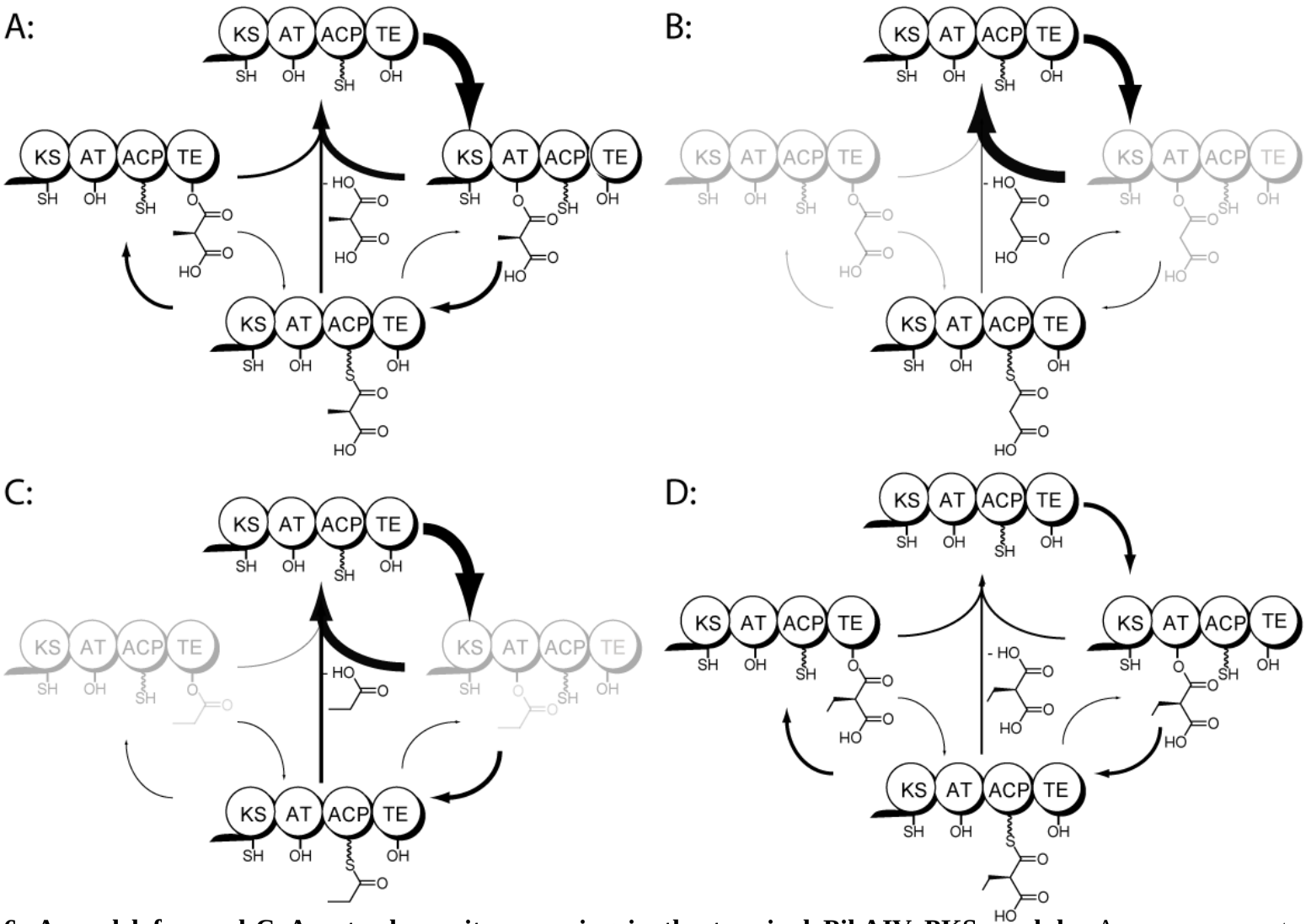


Figure 2-6. A model for acyl-CoA extender unit processing in the terminal PikAIV PKS module. Arrows represent proposed flux through the system based on Thiolglo steady-state kinetic analysis. Presence/absence of intermediates, as indicated by black/grey coloration, was determined from FTICR-MS analysis of active-site occupancy.

The FT active-site occupancy and ThioGlo-1 steady state acyl-CoA loading data enabled articulation of a new model for type I modular PKS biosynthesis based on the Pik system (**Figure 2-6**). For MM-CoA against PikAIV, loading occurs faster than hydrolysis, saturating the AT with transfer to the ACP and TE domains, which also contribute to hydrolysis (**Figure 2-6A**). For M-CoA, loading of this disfavored extender unit occurs, but it is rapidly removed by a high rate of hydrolysis. A small degree of M is transferred to the PikAIV ACP, competing with hydrolysis (**Figure 2-6B**). The dead-end P-CoA is similarly disfavored and is readily removed by hydrolysis (**Figure 2-6C**). In contrast, the unnatural substrate EM-CoA is loaded at PikAIV AT, ACP, and TE active-sites and exhibits a slow rate of hydrolysis (**Figure 2-6D**).

The unexpected finding that PikAIV could load the EM-CoA extender unit led us to investigate if the enzyme could produce the C2-ethyl narbonolide analog. This experiment serves as an example of how mechanistic insights (e.g. high levels of active-site loading with EM-CoA) can be applied to chemoenzymatic synthesis. The pikromycin SNAC-hexaketide chain elongation intermediate[16] was loaded onto PikAIV in the presence of EM-CoA. The reaction was extracted with organic solvent and product formation was monitored by LC FTICR-MS (**Figure 2-7**). The reaction of PikAIV, SNAC-hexaketide, and EM-CoA (**Figure 2-7E**) led to a new peak with the expected MH^+ of C2-ethyl narbonolide at a mass error of only 5 ppm. This peak has a similar elution profile compared to both the narbonolide authentic standard (**Figure 2-7A**), and chemoenzymatically generated narbonolide (**Figure 2-7C**). This product peak is absent in the no enzyme control reactions (**Figure 2-7B/D**). The relatively "noisy" extracted ion

chromatogram is due to the poor ionization efficiency of narbonolide. In addition, long elution profiles are noted due to the use of a protein compatible 300 Å C8 column.

These data strongly suggest that in addition to successfully loading EM-CoA, PikAIV can also extend and cyclize it into a natural product. Thus, when engineering PKS pathways, the use of rare extender units lacking evolved selectivity may be a viable strategy to generate novel polyketide analogs. In a similar experiment using M-CoA as the extender unit, the corresponding C2-desmethyl narbonolide analog failed to be generated by PikAIV (data not shown), suggesting that acyl-CoA extender unit selectivity correlates with the ability to make the corresponding product.

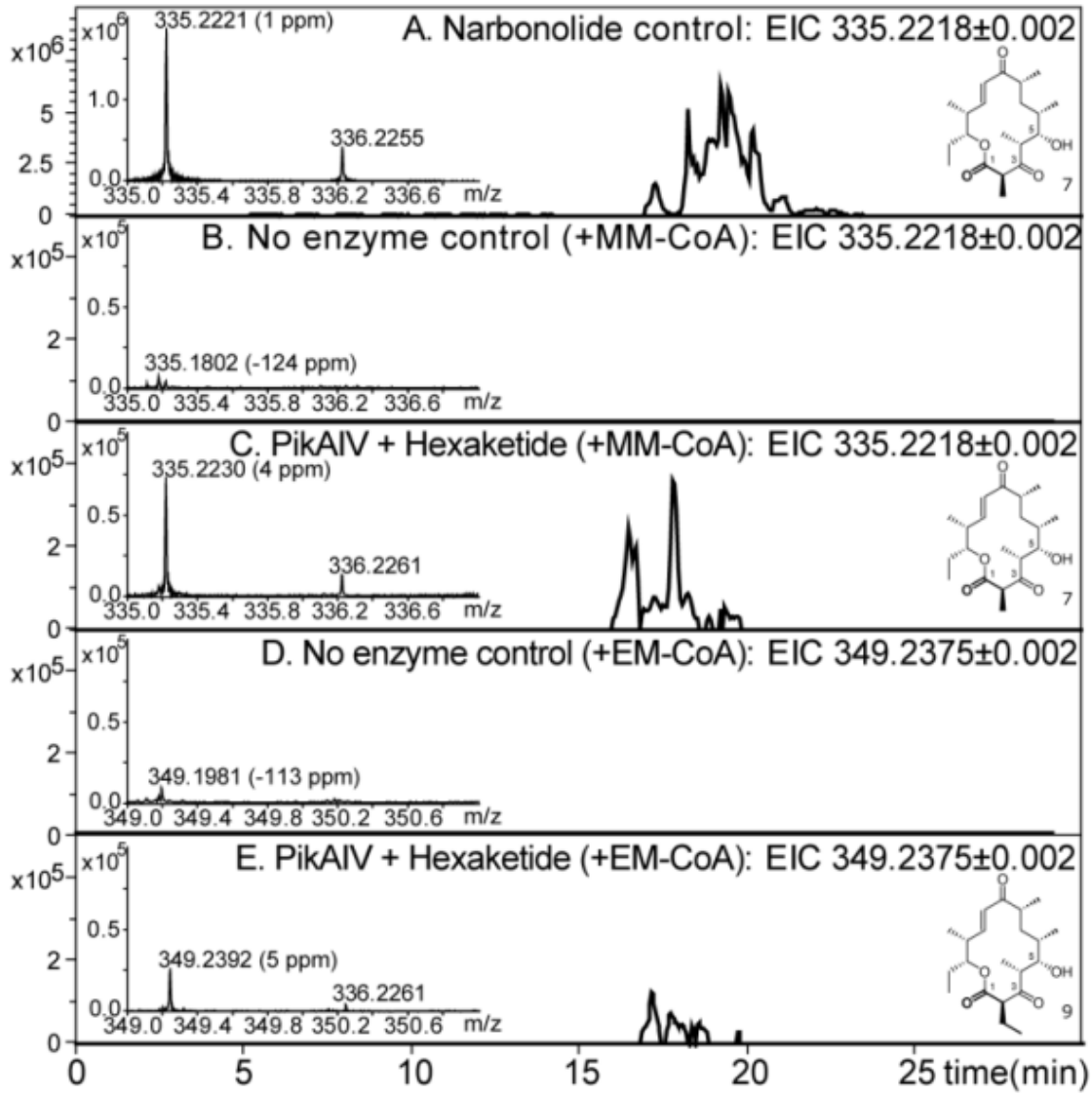


Figure 2-7. PikAIV catalyzed production of narbonolide and 2-ethyl narbonolide from MM-CoA and EM-CoA extender units with SNAC-hexaketide. A narbonolide positive control (**A**), +MM-CoA no enzyme control (**B**), +MM-CoA WT PikAIV reaction (**C**), +EM-CoA no enzyme control reaction (**D**), and a +EM-CoA WT PikAIV reaction (**E**) are presented. Extracted ion chromatograms were generated on an FTICR-MS to ±0.002 Da (6 ppm) and are shown as time versus intensity. The average mass spectrum from 16-20 min is inset with intensity versus mass/charge and error in ppm as compared to the expected product. The observed isotopic distribution closely matches the theoretical spectra (data not shown).

2.3 Discussion

The results presented in this study have demonstrated that the PikAIV PKS (and presumably other PKS systems) contains an intrinsic acyl-CoA hydrolytic editing process. A unique insight from this investigation is that the initial acyl-CoA loading step occurs for extender units regardless of their ability to be incorporated into natural product. Thus, final subunit occupancy is determined by reduced rates of hydrolysis *in vitro* (and presumably precursor pool levels *in vivo*), ultimately assuring proper acylation of the PKS module. Our ongoing analysis of other PKS monomodules, PikAIII and DEBS, has demonstrated similar hydrolytic activities and is indicative of a general metabolic process for AT subunit selectivity. Therefore, we may be able to exploit this characteristic by focusing on the use of unnatural extender units that undergo slow rates of hydrolysis. Efforts to tailor this hydrolytic activity, for example, by altering the active-site of the KS-AT domains, could present new strategies for generating novel engineered natural products. As shown with EM-CoA, extender unit active-site occupancy and a slow rate of hydrolysis correlates with generation of a new macrocyclic natural product analog. Future efforts will focus on applying the complementary dual-assay system to effectively determine kinetic rates and other biochemical details, including polyketide subunit selection and β-keto group processing (**Figure 2-1, Steps 3-4**), as well as docking domain interactions and module → module transfer of chain elongation intermediates.[17] In principle, this dual-assay system is a powerful tool for exploring catalysis in PKS systems. Further verification of this *in vitro* model in other modular systems will help generalize these findings. In the future, the FTICR-MS active-site analysis could be

applied to directly extend this *in vitro* model toward investigation of biosynthesis of complex natural products *in vivo*.[25]

2.4 Supplement

Materials

Unless otherwise noted, all chemicals, including acyl-CoAs, were purchased from Sigma. ThioGlo-1, [10-(2,5-dihydro-2,5-dioxo-1H-pyrrol-1-yl)-9-methoxy-3-oxo-methyl ester], was purchased either from Calbiochem or Covalent Associates. [1-^{14}C]-malonyl-CoA was from Moravek. Hexaketide-SNAC was synthesized as previously described.[16] Briefly, 10-deoxymethynolide was purified from a large scale fermentation of *Streptomyces veneszulae* SC1016. After organic extraction and HPLC purification, the compound was reduced at the keto position, and the macrolide ring was opened. The free acid was then activated with SNAC, and the compound was then oxidized immediately prior to use.

Cloning and protein expression

The construction of PikAIV mutants with KS, AT and TE catalytic domains individually inactivated has been described previously.[15] The ACP domain was activated by expressing the protein in BAP1 cells to give the holo-form.[26] The ACP domain was inactivated (apo form) by expression in the presence of excess iron, thus inhibiting a promiscuous phosphopantetheinyl transferase enzyme in *E. coli*. Expression and purification of each of the PikAIV mutant proteins was achieved according to procedures described elsewhere,[15] and the resulting recombinant proteins were purified to > 90%

homogeneity as determined by SDS-PAGE (**Figure 2-2**). Protein concentrations were determined by the Quant-iT Assay kit (Invitrogen) or the BCA assay (Pierce) with BSA as a standard. Cloning and expression of the PikAIV KS-AT construct has been previously reported.[25] The protein was overexpressed and purified using standard Ni-NTA chromatography.

For transient kinetic analysis, an internal standard peptide (IS) was generated through overexpression of a fusion construct. Briefly, the DNA encoding the PikAIV active site peptide: VWQHHGITPEAVIGHSQGEIAAAYVAGALTLDDAARSK was amplified from the plasmid containing PikAIV with Phusion DNA polymerase. This DNA was then cloned into the vector pMCSG7-MOCR with LIC technology. pMCSG7-MOCR is a variant of pMCSG7 with the protein MOCR[27] in frame (NCBI PT703G, courtesy Clay Brown LSI) and upstream of the LIC cloning/TEV cleavage site. A fusion partner was utilized, as the PikAIV AT-IS peptide did not overexpress to a suitable level, potentially due to proteolysis. The MOCR-AT-IS protein was then purified by standard Ni-NTA methodology, prior to TEV cleavage. The AT-IS peptide could be purified from MOCR and TEV by RP-HPLC using a 4000A PLRP-S column and a gradient of 0-100% 0.1% formic acid and acetonitrile + 0.1% formic acid (**Figure 2-8**). Fractions containing the AT-IS were collected, and concentration was determined by the BCA assay. The AT-IS was also characterized by MS/MS as described below. The final PikAIV AT-IS peptide contains three additional residues from the TEV cleavage site for a sequence of: SNAVWQHHGITPEAVIGHSQGEIAAAYVAGALTLDDAARSK.

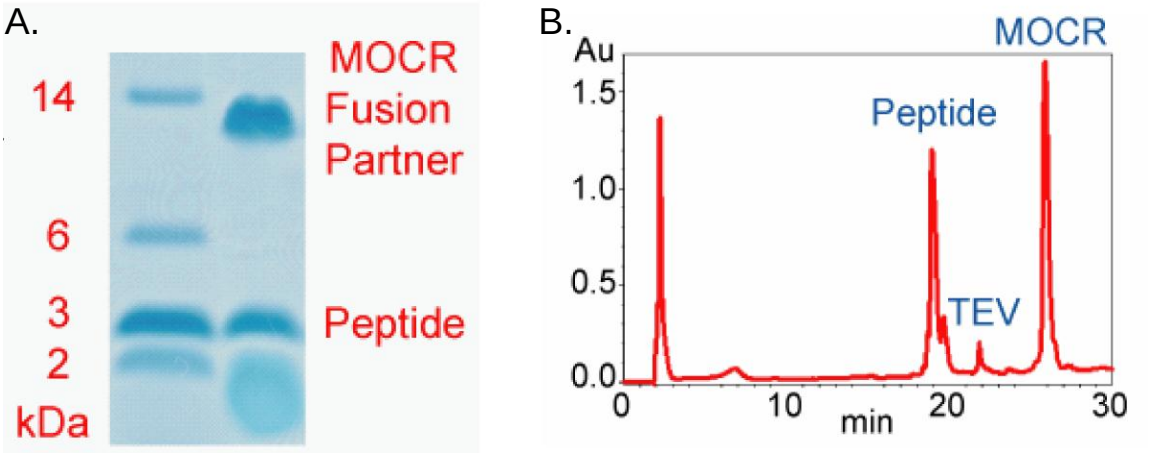


Figure 2-8. SDS-PAGE (A) and RP-HPLC (B) analysis of the PikAIV AT-IS peptide.

The *S. collinus ccr* gene was excised as a 1.3-kb *Nde*I-*Hind*III fragment from pHL18[22] and subcloned into pET28a to produce pCH1, which was used to transform *E. coli* BL21(DE3). The resulting transformant was used to inoculate 200 mL of LB medium supplemented with 50 µg/mL of kanamycin, grown at 37 °C to an optical density (OD_{600}) of 0.7, induced with 0.1 mM IPTG, and grown for a further 3 hrs. CCR was purified by Ni-affinity chromatography (according to standard protocols) and dialyzed against 50 mM Tris-HCl at pH 7.2. Ethylmalonyl-CoA was produced by incubation of 1 µM CCR, 15 mM crotonyl-CoA, 16 mM NADPH, and 200 mM $NaHCO_3$ in 50 mM Tris-HCl at pH 7.5. Reaction progress was assessed spectrophotometrically at 340 nm.

Kinetic analysis of hydrolytic activity

All reactions were carried out at 30 ^{o}C in the presence of 400 mM sodium phosphate (pH 7.2), 5 mM NaCl, 20% glycerol, 0.5 mM TCEP (pH 7.5), varying concentration of methylmalonyl-CoA and either 0.5 or 1 µM of protein. A series of controls were conducted in parallel, including the use of boiled protein, to ensure hydrolysis of acyl-

CoAs was due to enzymatic activity. Aliquots (50 µL) were withdrawn at specific time points and added to a well of a black 96-well plate containing an equal volume of dimethyl sulfoxide (DMSO) to quench the reaction. 100 µL of a 200 µM ThioGlo-1 solution (in DMSO) was added to each well, and the plate was incubated in the dark at room temperature with gentle shaking for 20 min. All samples were analyzed by fluorescence (excitation 378 nm and emissions at 480 nm) using a Gemini XPS microplate spectrofluorometer (Molecular Devices). The relative fluorescent unit (RFU) was converted to concentration using a standard CoA curve. Kinetic parameters were calculated from the average of at least three sets of triplicates with the Michaelis-Menten equation using the curve fitting software Kaleidagraph 4.03 (Synergy Software, Reading, PA) (**Figure 2-5**).

Substrate specificity and competition assays

The hydrolytic activity of wild type, apo and dAT PikAIV proteins were examined in the presence of 1 mM malonyl-CoA or 1 mM ethylmalonyl-CoA under conditions described above. Aliquots (25 µL) were withdrawn at specific time points and added to a well of a black 96-well plate containing 75 µL of DMSO to quench the reaction. 100 µL of a 200 µM ThioGlo-1 solution (in DMSO) was added to each well, and the plate was incubated in the dark at room temperature with gentle shaking for 20 min and analyzed by fluorescence as described above. For the competition assays, the wild type and apo PikAIV proteins were incubated in the presence of 1 mM methylmalonyl-CoA and 1 mM malonyl-CoA.

FTICR-MS analysis of active site occupancy

PikAIV (1 μM) was reacted with acyl-CoA extender units under saturating conditions (1 mM) in the presence of 400 mM sodium phosphate (pH 7.2), 5 mM NaCl, 20% glycerol, and 1 mM TCEP (pH 7.5). Reactions were incubated at 25 ^{o}C for 10 minutes, followed by a 2.5 fold dilution in 50 mM ammonium bicarbonate with tris-base added to pH 8. Trypsin was present at an enzyme to substrate ratio of 1:10. Proteolysis was allowed to proceed for 15 min at 37 ^{o}C followed by addition of formic acid to pH 4. 30 min and 45 min digests yielded similar results, suggesting that hydrolysis from the active site peptides is insignificant compared to other sources of experimental error. Reactions were frozen at -20 ^{o}C until analysis. 50 μL of sample (20 pmol / 3 μg of protein) was injected onto a Jupiter C4 2x250 mm 300 μm column (Phenomenex) using an Agilent 1100 LC system with a flow rate of 200 μL/min and a gradient of 2-98% acetonitrile over 40 min. 0.1% formic acid was added to the water and acetonitrile solvents. A divert valve was utilized for online desalting. The LC was coupled to an FTICR-MS (APEX-Q with Apollo II ion source and actively shielded 7T magnet; Bruker Daltonics). Data were gathered from m/z 200–2,000 in positive ion mode. Electrospray was conducted at 2,600 V with 1 scans per spectra utilizing 0.33 s external ion accumulation in a hexapole and 1 ICR cell fills prior to excitation and detection. Data were analyzed using DECON2LC (Pacific Northwest National Labs), VIPER (Pacific Northwest National Labs), and Data Analysis (Bruker Daltonics). Similar ionization efficiencies were assumed between the loaded and unloaded form, as no functional groups which either introduce or remove a charge site in positive mode ESI conditions (<pH 3) are different between the loaded and unloaded forms. Changes in overall mass and hydrophobicity may have an impact, but

on the relatively large peptides monitored this is likely less significant than other experimental variation in this semi-quantitative method.

Transient Kinetic Analysis

PikAIV KS-AT didomain (2 μM) was mixed with an equal volume of acyl-CoA extender units (2 mM) in a Kintech rapid-quench apparatus equilibrated to 30 ^{o}C for a two-fold dilution. Each reagent was in the following buffer: 400 mM sodium phosphate (pH 7.2), 5 mM NaCl, 20% glycerol, and 1 mM TCEP (pH 7.5). Data were recorded at time points 2, 4, 5, 8, 32, 64, and 128 ms in triplicate. Each reaction was immediately quenched in 1 M HCl in 6 M urea then heated for 3 min at 90 ^{o}C. Each reaction was then frozen in liquid nitrogen. After all reactions were conducted, they were simultaneously thawed, diluted to 2 M urea in 50 mM ammonium bicarbonate and the pH was adjusted to 8 with tris-base. 2 μL of Lys-C (Roche) was then added to each reaction for a final concentration of 0.2 mg/mL. The samples were then incubated for 15 minutes at 37 ^{o}C after which the pH was reduced to 4 with 10% formic acid. Samples were desalted with Handee Microspin columns (Pierce) packed with 20 μL of 300 Å polymeric C18 resin (Vydac). Samples were loaded onto the columns and washed with 30 column volumes of 0.1% formic acid prior to elution with 10 column volumes of 50% acetonitrile plus 0.1% formic acid. Intact protein samples were analyzed by FTICR-MS (APEX-Q with Apollo II ion source and actively shielded 7T magnet; Bruker Daltonics). Data were gathered from m/z 200–2,000 utilizing direct infusion electrospray ionization in positive ion mode. Electrospray was conducted at 3,600 V with 24 scans per spectra utilizing 1 s external ion accumulation in a hexapole and 4 ICR cell fills prior to excitation and detection. The

external quadrapole was set to only allow ions from 740-815 m/z to reach the FTICR mass analyzer. Data were processed in Data Analysis (Bruker Daltonics) and Midas (NHMFL). All identified species were accurate to 20 ppm with external calibration. The PikAIV AT active site, and PikAIV AT MM loaded active site were quantified by total peak height for each isotope in comparison to the PikAIV AT active site internal standard peptide, which was added at 2 μM during sample preparation. This peptide contains the additional three residues SNA- at the N-terminus from the TEV cleavage site.

Transient kinetic analysis in a rapid-quench apparatus enabled a series of time-point samples to be generated. These samples were then processed and analyzed by FTICR-MS to generate a loading curve (**Figure 2-3**). By normalizing to an internal standard peptide (ATIS), %RSD values were improved by approximately 10-fold to 5-20%, which is sufficient for transient kinetic analysis. The data were fitted to a single exponential curve, the simplest possible model for the data, and rates were determined. Fitting a simple single exponential model is also appropriate given the experimental error in this measurement. We found that the initial rate of direct AT loading for the native substrate is dramatically faster than the overall rate of hydrolysis or catalysis in the system, and thus the simple model that the data is fitted to and the modest %RSD values are sufficient for this interpretation. Sample spectra from such an experiment are shown (**Figure 2-4**). Due to the absence of detected AT active-site bound serine intermediates this analysis could not be performed for P-CoA or M-CoA. EM-CoA analysis was not performed due to limited substrate availability.

LC-FTICR-MS Analysis of Product Formation

Chain elongation unit and extender unit product formation was examined by LC-FTICR-MS and confirmed by LC-MS/MS. PikAIV (1 μM) was reacted with CoA extender units under saturating conditions (1 mM) and SNAC-hexaketide (1 mM) in the presence of 400 mM sodium phosphate (pH 7.2), 5 mM NaCl, 20% glycerol, 1 mM TCEP. The 100 μL reactions were incubated overnight at room temperature. Samples were extracted with chloroform (3:1 ratio) and concentrated under N_2. The sample was reconstituted in 200 μL of MeOH and 50 μL of this sample was analyzed on a Zorbax C8 300 Å 2x50 mm 5 μm column (Phenomenex). A gradient was generated on an Agilent 1100 HPLC. The following conditions were used: 0 min (90,10), 5 min (90,10), 20 min (2,98), 24 min (2,98) and 25 min (98,2). Values are provided as Time (%A, %B) (min), with the total run time of 30 min. Flow was at 0.2 mL / min. A column heater was operated at 50^{o}C. Flow was diverted for the first 5 min of the run. Buffer A consisted of 0.1% formic acid in DDI water. Buffer B consisted of 0.1% FA in acetonitrile.

FTICR-MS was performed on an APEX-Q (Apollo II ion source 7T magnet, Bruker Daltonics). Data were gathered by ESI in positive ion mode (2,400 V, m/z 150–1,000, transient 128 K, 1 scan/spectrum) with external ion accumulation, dynamic trapping (0.33 s), and 1 ICR cell fill per spectrum. External calibration utilized HP-mix (Agilent). Product peaks were detected over multiple samples and runs.

PikAIV Active Site Occupancy by LC-FTICR MS

PikAIV active site occupancy was determined by LC-FTICR-MS after terminating the reaction under steady-state kinetic conditions by the addition of trypsin. The ratios of the loaded and unloaded forms were determined and reported from four replicates. A single replicate of PikAIV WT + EM-CoA is shown as an example of the raw data (**Figure 2-9**). This figure is intended to highlight the complex nature of the experimental dataset. **Figures 2-9A1 and 2-9A4** indicate the high level of sample complexity even from the digest of a single protein. **Figures 2-9B1, 2-9C1, 2-9D1, 2-9E1, 2-9F1, 2-9G1,** and **2-9H1** indicate the high quality of the LC separation based upon temporal resolution of the active site peptides. **Figures 2-9B2, 2-9C2, 2-9D2, 2-9E2, 2-9F2, 2-9G2,** and **2-9H2** illustrate the separate charge states monitored and relative intensity over the respective elution windows. **Figures 2-9B3, 2-9C3, 2-9D3, 2-9E3, 2-9F3, 2-9G3,** and **2-9H3** illustrate the high resolution characteristics of the data and the ability to monitor the specific isotopic peaks for each species. **Figures 2-9B4, 2-9C4, 2-9D4, 2-9E4, 2-9F4, 2-9G4,** and **2-9H4** illustrate the results of the automated deconvolution software in terms of charge state, elution over multiple scans, and mass error.

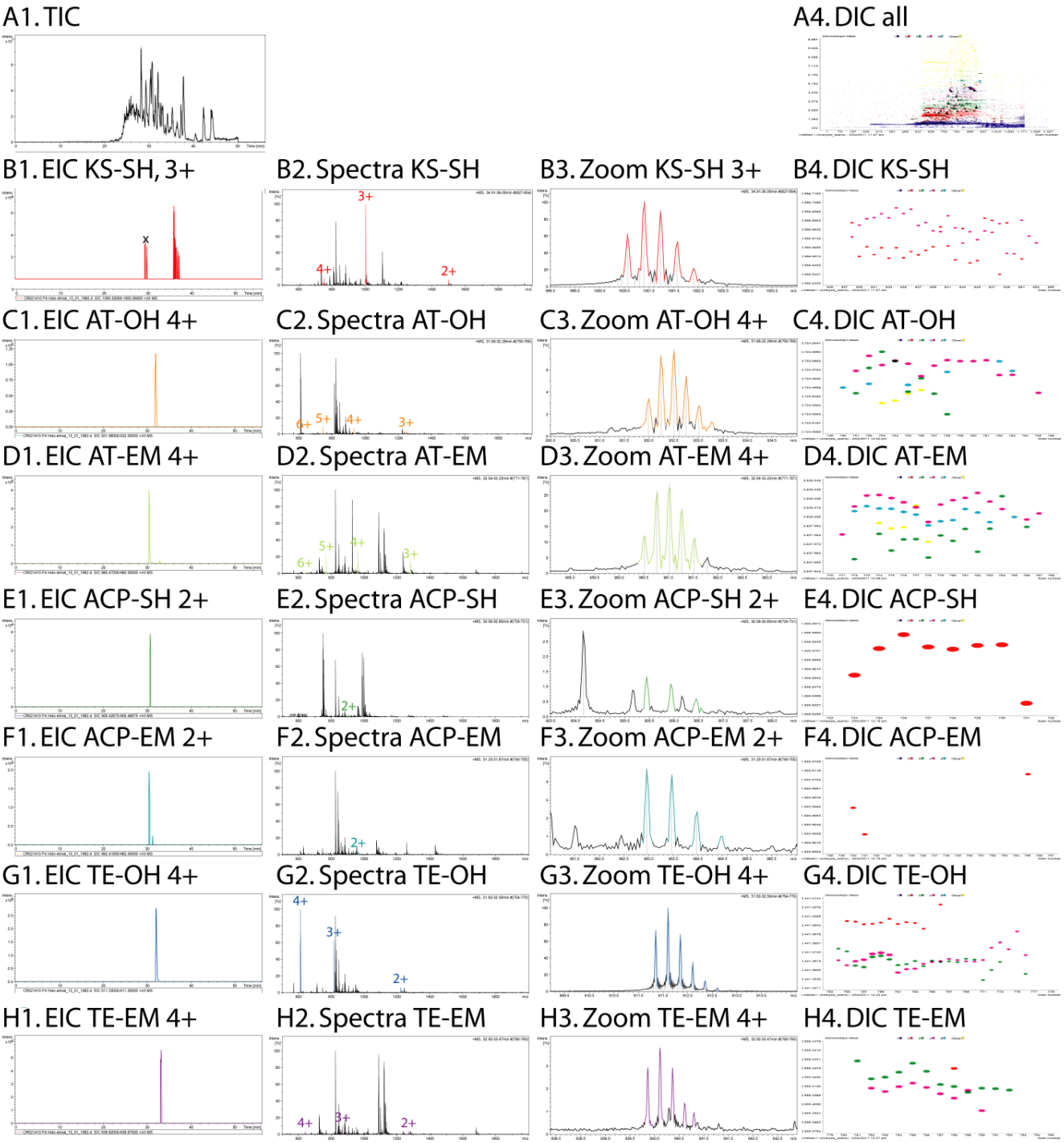


Figure 2-9. Sample data: PikAIV WT + EM-CoA active site occupancy by LC-FTICR-MS. Data are displayed for the free KS-SH active site (**B**), the free AT-OH active site (**C**), the loaded AT-EM active site (**D**), the free ACP-SH active site (E), the loaded ACP-EM active site (**F**), the free TE-OH active site (**G**), and the loaded TE-EM active site (**H**). The total ion chromatogram (TIC, **A1**) is shown from 5-55 min (the first five min of the run are not recorded) in absolute intensity. Extracted ion chromatograms (**EIC**) are shown ±20 ppm for all active site loaded and unloaded species from **5-55** min in absolute intensity normalized to the largest peak (**B1, C1, D1, E1, F1, S5G1, H1**).

Average mass spectrum over the eluting active site species are displayed from 500-2,000 m/z in normalized intensity with the charge state and location of the active site ions noted in color (**B2, C2, D2, E2, F2, G2, H2**). A 5 m/z unit zoom is shown for the most abundant charge state for each of the active sites identified in m/z versus normalized intensity (**B3, C3, D3, E3, F3, G3, H3**). A deconvoluted ion chromatogram (DIC) is shown for all ions present as generated from DECON2LC and VIPER. Data are displayed as scan versus deconvoluted monoisotopic mass with charge state indicated in color (1+ 2+ 3+ 4+ 5+ other, **A4**). Denconvoluted ion chromatograms are shown corresponding to each active site species as scan number versus mono-isotopic molecular weight with charge state indicated in color **(B4, C4, D4, E4, F4, G4, H4**).

MS/MS confirmation of active site peptides

All active site peptides (PikAIV KS, AT, ACP, TE) reported in this paper have been confirmed by CID MS/MS fragmentation. MM, M, P, and EM loaded species have also been investigated. Active-site peptides monitored with b- and y- ion sequence coverage observed are provided below in **Table 2-4**. For the ACP active-site including the holo and acyl- loaded species, the phosphopantetheine ejection ions were also observed.[30] Peptide parent and product ion assignments by FTICR-MS are within 20 ppm and within 0.3 Da by iontrap-MS. All active site peptide assignments were also validated based upon the presence or absence of the species in specific reactions. For example, the loaded AT-EM active site was not observed when MM-CoA was added to the sample and the free KS-SH active site was not observed in the dKS active site cysteine to alanine variant.

				Sequence	**Coverage**
				IAYSLGLEGPAVTVD-	
KS	C-207	-	FTICR	TACSSSLVALHLALK	b_5,b_7,b_8,b_8,b_9,b_9, y_{18}^{2+},y_{21}^{2+},y_{22}^{2+},y_{23}^{2+},y_{24}^{2+},y_{25}^{2+}
				VWQHHGITPEAVIG-	
				HSQGEIAAAYVAGA-	
AT	S-652	-	FTICR	LTLDDAAR	b_7,b_{11}^{2+},b_{12}^{2+},b_{13}^{2+},b_{19}^{2+},b_{20}^{2+},b_{21}^{2+},b_{22}^{2+},b_{23}^{2+},y_5,y_6,y_7,y_8,y_{10},y_{11}
		MM	FTICR		y_3,y_4,y_5,y_6,y_7,y_8
		EM	FTICR		y_4,y_5,y_6,y_7
ACP	S-980	Apo	FTICR	EIGFDSLTAVDFR	y_2, y_4,y_5,y_6,y_7,y_8,y_9,y_{10},y_{11}
		Holo	Iontrap		$Ppant_1$ (261.1) , $Ppant_2$ (359.1), Apo-18, Apo+80
		Holo-MM	Iontrap		$Ppant_1$ (361.1) , $Ppant_2$ (459.1), Apo-18, Apo+80
		Holo-P	Iontrap		$Ppant_1$ (317.1) , Apo+80
		Holo-EM	FTICR		$Ppant_1$ (375.156) , Apo+80
				AAGDAPVVLLGHSG-	
TE	S-1196	-	FTICR	GALLAHELAFR	b_4,b_5,b_7,b_8,b_{12},y_2,y_3,y_4,y_6,y_7,y_8,y_9,y_{10}
		MM	FTICR		b_5,y_4,y_5,y_6
		EM	FTICR		b_5,y_4,y_5,y_6

Table 2-4. Active site peptides monitored and MS/MS confirmation. Only b-ions, y-ions, and phosphopantetheine ejection ions are shown.[30] b- and y- primary sequence ions are noted. $Ppant_1$ refers to the phosphopantetheine elimination ion 1 at m/z 261 ($C_{11}H_{21}N_2O_3S^+$) for the holo free –SH cofactor. $Ppant_2$ refers to the phosphopantetheine elimination ion 2 at m/z 359 ($C_{11}H_{23}N_2O_7PS^+$) for the holo free –SH cofactor. Since the Ppant ions contain any ACP loaded substrate, the mass changes in the case of loaded MM, P, or EM.

Portions of this chapter have been previously published in:

Acyl-CoA subunit selectivity in the terminal pikromycin polyketide synthase module: steady-state kinetics and active-site occupancy analysis by FTICR-MS. Shilah A. Bonnett,[#] Christopher M. Rath,[#] Rafay Shareef, Joanna R. Joels, Joesph Chemler, Kristina Hakansson, Kevin Reynolds, David H. Sherman. Under Review in *Chemistry & Biology*. ([#]Authors contributed equally to this work)

NIH support for research on PKS/NRPS systems in the Sherman laboratory is gratefully acknowledged through grants GM076477, CA108874, ICBG U01TW007404, and the Hans W. Vahlteich Professorship (to DHS).

2.5 References

1. Walsh, C.T. *Science*, **2004**, *303*, 1805.
2. Wu, N.; Cane, D.E.; Khosla, C. *Biochemistry*, **2002**, *41*, 5056.
3. Li, S.J.; Podust, L.M.; Sherman, D.H. *J Am Chem Soc*, **2007**, *129*, 12940.
4. Smith, S.; Tsai, S.C. *Nat Prod Rep*, **2007**, *24*, 1041.
5. Tang, Y.; *et al. Chem Biol*, **2007**,*14*, 931.
6. Tang, Y.; *et al. Proc Nat Acad Sci USA*, **2006**,*69*, 11124.
7. Serre, L.; *et al. J Biol Chem*, **2005**, *270*, 12961.
8. Roberts, G.A.; Staunton, J.; Leadlay, P.F. *Eur J Biochem*, **1993**, *214*, 305.
9. Marsden, A.; *et al. Science*, **1994**, *263*, 378.
10. Yuan, Z.Y.; Hammes, G.G. *J Biol Chem*, **1985**, *260*, 13532.
11. Cognet, J.A.H.; Hammes, G.G. *Biochemistry*, **1983**, *22*, 3002.
12. Dorrestein, P.C.; Kelleher, N.L. *Nat Prod Rep*, **2006**, *23*, 893.
13. Gu, L.; *et al.* D.H. *Science*, **2006**, *318*, 970.
14. Schnarr, N.A.; Chen, A.Y.; Cane, D.E.; Khosla, C. *Biochemistry*, **2005**,*44*, 11836.
15. Kittendorf, J.D.; *et al. Chem Biol*, **2007**, *14*, 944.
16. Aldrich, C.C.; Beck, B.J.; Fecik, R.A.; Sherman, D.H. *J Am Chem Soc*, **2005**, *127*, 8441.

17 McLoughlin, S.M.; Kelleher, N.L. *J Am Chem Soc*, **2004**, *126*, 13265.

18. Beck, B.J.; *et al. J Am Chem Soc*, **2003**, *125*, 4682.

19. Haydock, S.; *et al. FEBS Let*, **1995**, *374*, 246.

20. Kim, B.S.; *et al. J Biol Chem*, **2002**, *277*, 48028.

21. Oliynyk, M.; *et al. Mol Microbiol*, **2003**, *49*, 1179.

22. Liu, H.; Reynolds, K.A. *J Bacteriol*, **1999**,*181*, 6806.

23. Stassi, D.L.; *et al. Proc Nat Acad Sci USA*, **1998**,*95*, 7305.

24. Suo, Z.; Chen, H.; Walsh, C.T. *Proc Nat Acad Sci USA*, **2000**, 97, 14188.

25. Buchholz, T.J.; *et al. ACS Chem Biol*, **2009**, *4*, 41.

26. Pfeifer, B.A.; *et al. Science*, **2001**, *291*, 1790.

27. Walkinshaw, M.D.; *et al. Mol Cell*, **2002**, *9*, 187.

28. Akey, D.L.; *et al. Nat Chem Biol*, **2006,** *2*, 537.

29. Li, S.J.; *et al. Proc Nat Acad Sci USA*, **2009**, *106*, 18463.

30. Dorrestein, P.C; *et al. Biochemistry* **2006**, *45*, 12756.

Chapter 3

Polyketide β-branching in bryostatin biosynthesis: identification of surrogate acetyl-ACP donors for BryR, an HMG-ACP synthase

3.1 Introduction

The bryostatins are antifeedant polyketide natural products produced by a bacterial symbiont of the marine bryozoan *Bugula neritina*.[1] They are highly potent protein kinase C (PKC) modulators,[2] and, as such, bryostatin 1 (**Figure 3-1**) has been investigated in numerous clinical trials as a potential anticancer agent.[3] Separately, the neuroprotective activity of PKC activators has recently been demonstrated in preclinical studies where bryostatin 1 was able to rescue memory loss after postischemic stroke.[4] Additional studies suggest that bryostatin 1 (and a synthetic analog) may be able to reduce the levels of A-β, a toxic peptide implicated in Alzheimer's disease.[5,6] However, like many marine-derived natural products, fulfilling the promise of these initial studies may be hindered by the low abundance of bryostatins available from either natural sources or chemical synthesis.[1,7] The intriguing biological activities and lingering supply questions motivate our continued study of the bryostatin biosynthetic pathway (**Figure 3-1**). Increasing our knowledge of the molecular mechanisms employed may help open the door to new methods of bryostatin production as well as the generation of related bryostatin analogs. Herein, we report the biochemical characterization of BryR, the 3-hydroxy-3-methylglutaryl (HMG)-CoA synthase (HMGS) homolog implicated in β-branching at C-13 and C-21 of the core bryostatin ring system (**Figure 3-1**).

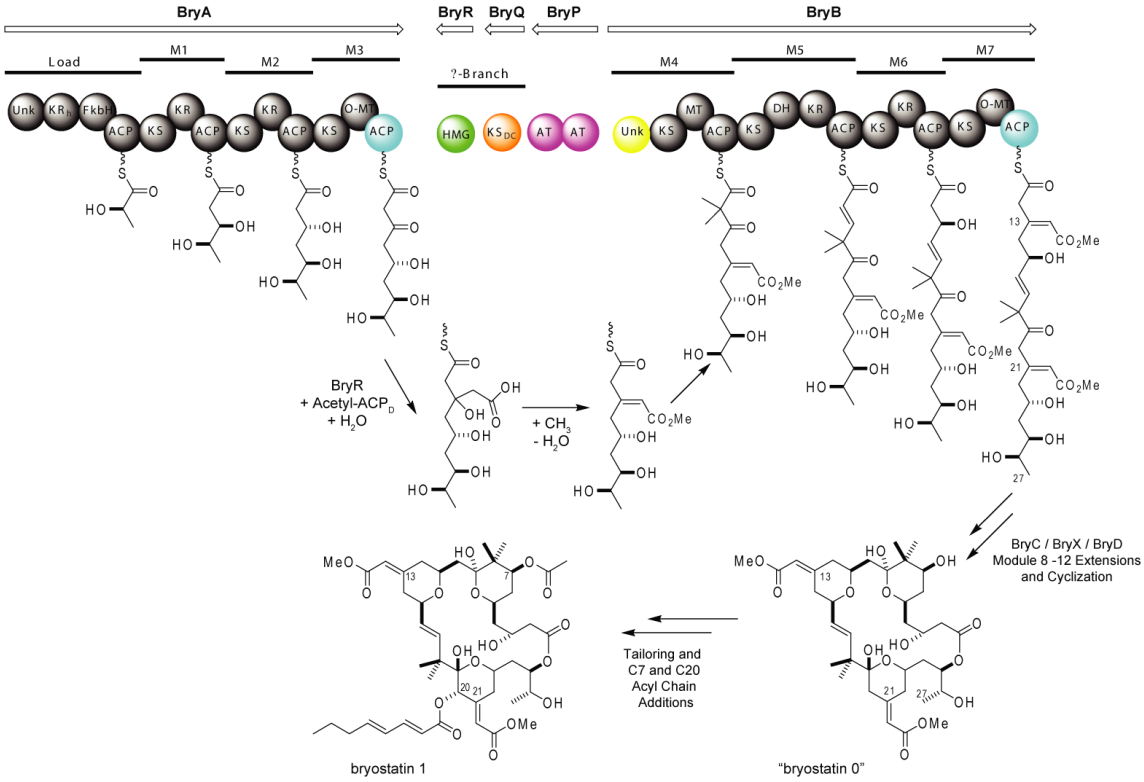


Figure 3-1. Portions of the pathway utilized in β-branching are highlighted with color in this depiction of the bryostatin biosynthetic pathway. BryC, BryX and BryD are not shown.[1] Though the complete structure of the PKC modulator, bryostatin 1 is shown here, the full suite of bryostatin molecules contains acyl chain variability at both C7 and C20. ACP, acyl carrier protein; AT, acyltransferase; DH, dehydratase; FkbH, homolog to FkbH;[38] HMGS, HMG-CoA synthase homolog; KR, ketoreductase; KS, ketosynthase; KS_{DC}, decarboxylative ketosynthase; MT, methyltransferase; Unk, domain with unknown function.

Polyketide metabolites are produced by diverse bacterial taxa, including soil-dwelling bacteria, cyanobacteria, and bacterial symbionts living within insects or marine invertebrates, and are all generated by decarboxylative condensation reactions of simple coenzyme A (CoA) building blocks.[8-10] Polyketides with variable levels of reduction at the β-ketone position are built by type I polyketide synthases (PKSs). Type I PKSs are composed of a linear arrangement of covalently fused catalytic domains within large, multifunctional proteins. A unidirectional assembly line process is used to generate a linear intermediate that is often off-loaded as a cyclized lactone product. Sets of domains grouped together to accomplish a single round of extension are termed modules. The

number, arrangement, and architecture of modules within type I systems serve as a blueprint to determine the core structure of the natural product.[9,11,12]

While methylation at the α position relative to the carbonyl group is well-characterized,[8,11] alkylation at the β position (e.g. β-branching) is less commonly observed, but can introduce further functional group complexity into polyketides. Significant genetic and biochemical evidence has been obtained to demonstrate that β-position alkyl side-chains are typically introduced through an "HMGS cassette" of enzymes/domains performing reactions similar to those observed in mevalonate biosynthesis (**Figure 3-2**).[13-17] This set of enzymes typically contains three discrete proteins; the HMGS homolog, a decarboxylative ketosynthase (Cys to Ser active site variant, KS_{DC}), and a donor acyl carrier protein (ACP_D) upon which acetyl-ACP_D (Ac-ACP_D) is typically generated (**Figure 3-3**). Additionally, one or two enoyl-CoA hydratase (ECH) homologs responsible for dehydration and decarboxylation transformations (ECH_1 and ECH_2, respectively) may be present as discrete proteins or embedded domains in larger, multifunctional proteins (**Figure 3-2**). Finally, many of the pathways contain tandem acceptor ACPs (ACP_A) at the site of modification. Full HMGS cassettes have been shown to install methyl branch points in bacillaene, curacin, jamaicamide, and mupirocin,[14,15,18,19] and hypothesized for methylation in pederin,[20] and virginiamycin M.[21] Methoxymethyl and ethyl branches are added to the growing myxovirescin molecule in a similar fashion.[17,22] However, the identity of the AT/KS pair responsible for generating the propionyl-ACP_D remains unconfirmed.[23] In some cases, the methyl branch points are elaborated further by neighboring domains (the action of the nearby halogenase and enoyl reductase domains convert the β-position in the mature curacin to a cyclopropyl ring while jamaicamide contains a vinyl chloride)[24] (**Figure 3-4**). One notable exception to the HMGS-mediated chain-branching strategy was recently found in the rhizoxin biosynthetic pathway where a PKS-mediated Michael addition is employed in the generation of a γ-lactone.[25]

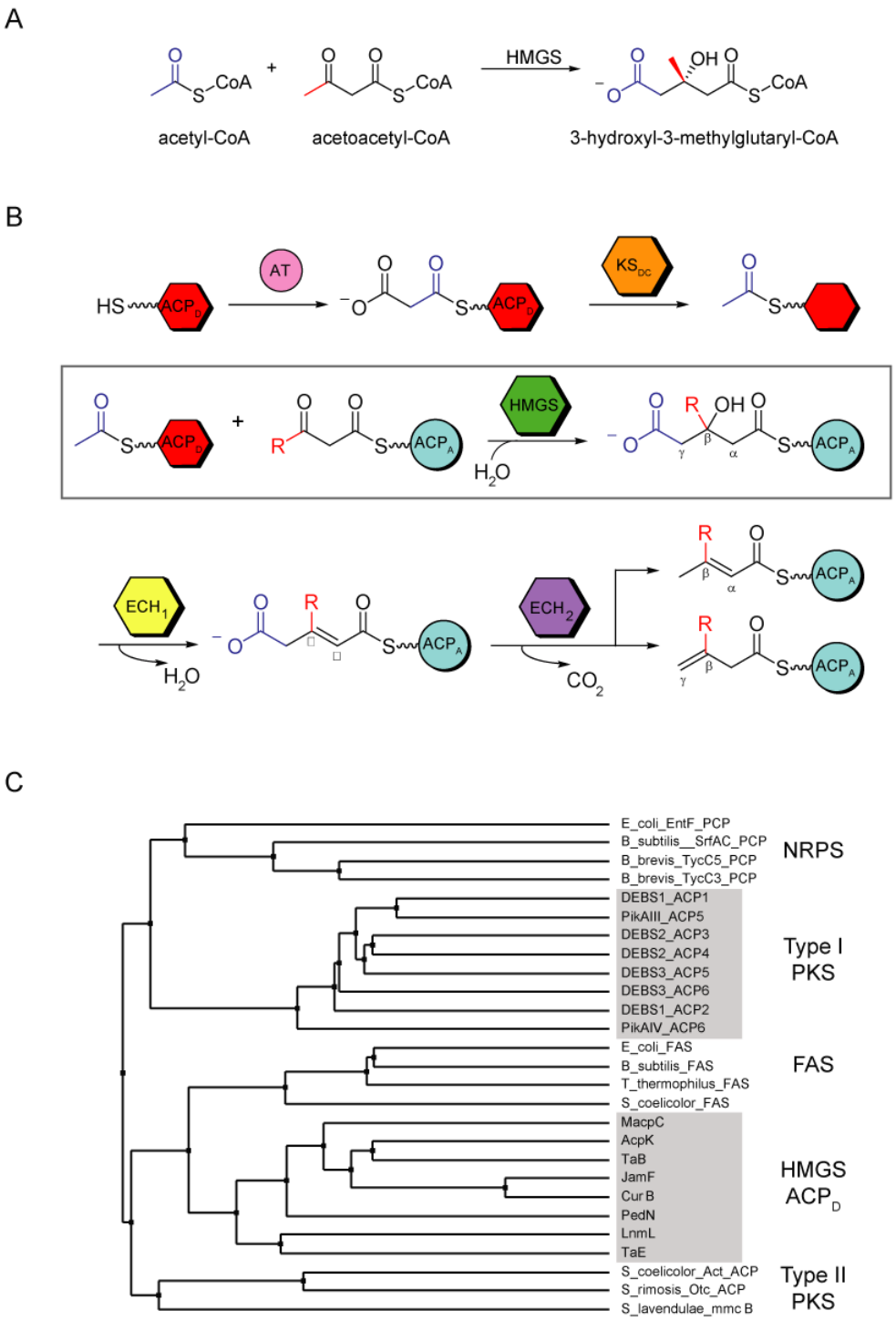


Figure 3-2. Proteins and/or domains involved in HMG generation. (**A**) HMG generation in the mevalonate pathway and (**B**) during polyketide β-branching in PKS and mixed biosynthetic pathways. The covalent, enzyme-bound intermediate from the reaction being analyzed in this paper is boxed. (**C**) The HMGS cassette ACP_D subclass of acyl carrier proteins can be observed in the phylogenetic tree generated using Jalview software (average distance BIOSUM2).[39]

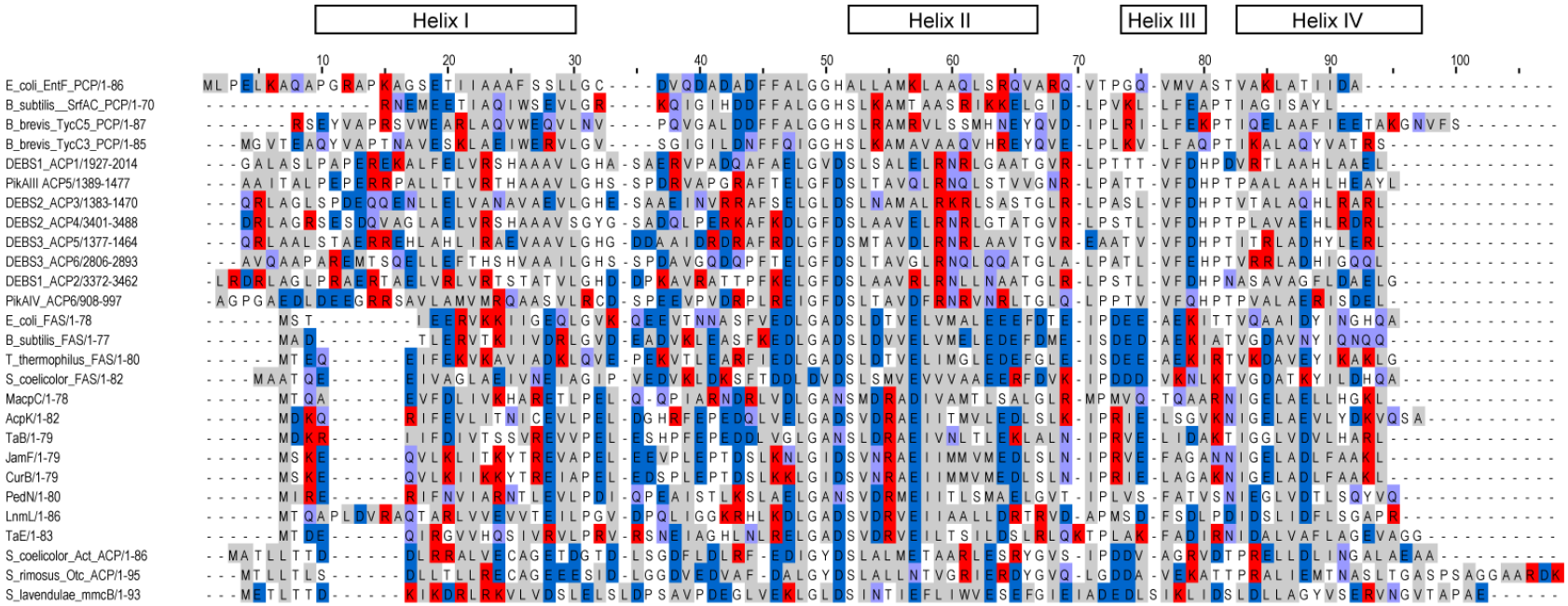


Figure 3-3. Various acyl carrier protein subclasses. The multiple sequence alignment (Clustal) was generated using Jalview software.[40] Numbering is based on MacpC. Helix designations are predicted from alignment with DEBS1_ACP2 structure.[40] Basic residues are colored in red, acidic in blue, and hydrophobic in grey.

Partial HMGS cassettes have been identified in the onnamide, difficidin, psymberin, leinamycin,[20,23,26,27] and bryostatin (missing $ACP_{D,}$ ECH_1 & ECH_2)[1] biosynthetic pathways (**Figure 3-4**). Lack of complete gene cluster sequencing or annotation is one possible explanation for the presence of a partial HMGS cassette. This is likely the case for the onnamide, virginiamycin, difficidin and bryostatin systems where either firm pathway boundaries have yet to be determined for contiguous pathways or the pathway is possibly dispersed across the genome. In other instances (leinamycin, bryostatin), product formation is unlikely to involve enzymatic transformations by the ECH homologs (dehydration and decarboxylation). Alternately, functions performed by the canonical set of HMGS cassette members might be catalyzed by alternative domains/enzymes within the pathway. For example, the leinamycin pathway does not include a KS_{DC}, as LnmK fulfills this role as an acyltransferase/decarboxylase to generate the acyl donor propionyl-LnmL.[23] Similarly, the β-methoxylacylidene moieties found in the bryostatins are hypothesized to be the result of a β-γ dehydration (whereas the dehydration performed by ECH_1 enzymes typically occurs across the α-β positions).[13] The N-terminal domains of unknown function found on BryB and BryC are candidates to catalyze these transformations, found immediately downstream of both HMGS modification sites in the bryostatin pathway.[1]

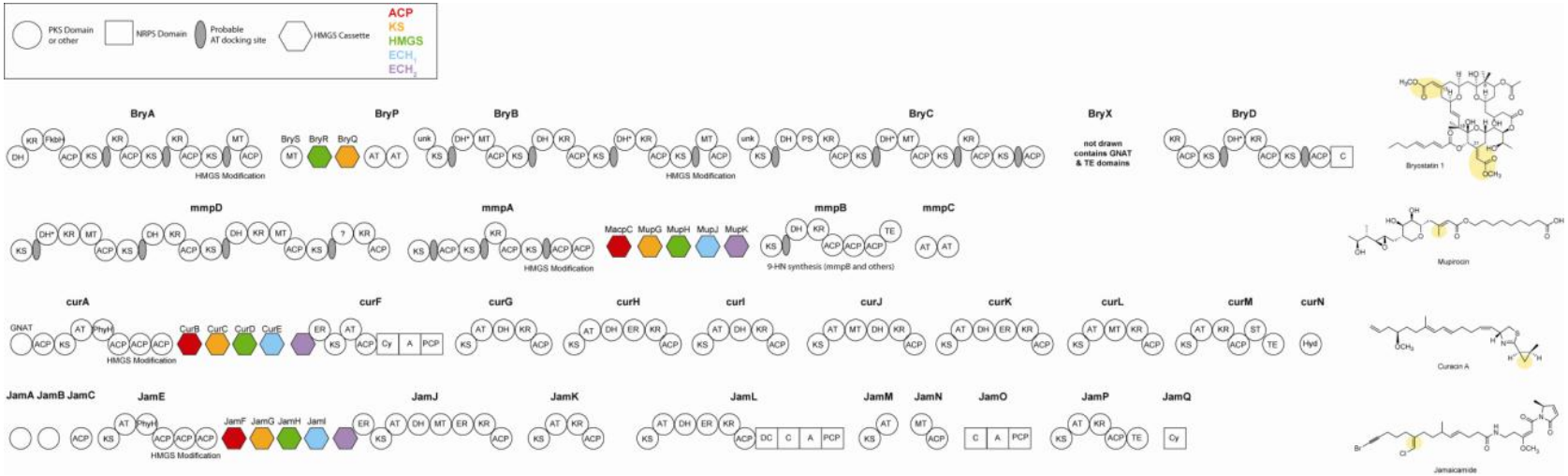


Figure 3-4. HMGS cassette-containing biosynthetic pathways featured in this report. Members of the HMGS cassettes are colored, and sites of HMGS modification are marked. β-branch points are shaded yellow in the compound structures. *Abbreviations*: A – Adenylation, ACP – acyl carrier protein, AT – acyltransferase, C – condensation, Cy – cyclization, DH – dehydratase, ER – enoyl reductase, FkbH – FkbH homolog, GNAT – GCN5-related *N*-acetyltransferase, KS – ketosynthase, KR – ketoreductase, MT – methyltransferase, PCP – peptidyl carrier protein, PhyH – phytanoyl-CoA dioxygenase, PS – pyrone synthase, TE – thioesterase, unk – unknown function, * - inactive domain.

The type I PKS biosynthetic gene cluster (*bry*) presumed responsible for the synthesis of the bryostatins has been identified and sequenced from two sibling unculturable bacterial symbiont species of "*Ca.* Endobugula sertula/*B. neritina*".[1] The shallow-water North Carolina (NC) sibling species appears to be located within a contiguous DNA fragment approximately 77 kb in length, whereas the deep-water California (CA) species is split between two or more locations on the chromosome. Apart from the transposition of the HMGS cassette and AT enzymes, the two sequences exhibit >99.5% identity at the DNA level. In the current study we have focused on BryR from the NC species of "*Ca.* Endobugula sertula/*B. neritina*" to elucidate its role in the β-branching process (e.g. formation of HMG-ACP) during bryostatin assembly.

The canonical HMGS cassette activities have been elucidated through *in vitro* biochemistry or *in vivo* gene disruption studies in the bacillaene, curacin/jamaicamide and myxovirescin pathways.[13]. A key step for selectivity in the HMGS cassette appears to be the HMGS reaction itself.[18] Biochemical studies of PksG, the HMGS homolog of the bacillaene pathway, revealed that the enzyme only accepts the acetyl group when presented on AcpK, its cognate ACP_D.[18] In addition, gene deletion studies of the myxovirescin HMGS cassette enzymes indicate that the two HMGS homologs present (TaC/TaF) utilize separate ACP_Ds (TaB/TaE).[16] The ability of PksG to accept a model substrate, acetoacetyl (Acac)-ACP_A, was also examined in vitro—illustrating that the HMGS-ACP_D proteins interact.[18] The role of protein-protein interactions in mediating biosynthetic processes for polyketide β-branching, and its role in chemical diversification motivated our current studies, described below.

3.2 Results

Based on the reported activities of the previously characterized secondary metabolite HMGS homologs PksG and TaC,[18,22] BryR is likely to be involved in the β-branching at C-13 and C-21 of the bryostatins (**Figure 3-1**). To date, no discrete ACP for the HMGS cassette of the bryostatin pathway has been located in either the NC or CA *bry* cluster sequences. The possibility exists that BryR, like its primary metabolism counterparts, may be able to use acetyl-CoA as the acyl donor in its reaction.[13] However, the presence of a KS-type (BryQ) decarboxylase, whose presumed role is to

generate acetyl-ACP from malonyl-ACP, makes this an unlikely scenario. A discrete ACP upstream of the *bry* cluster was identified adjacent to genes that encode proteins likely involved in fatty acid biosynthesis (Bry FAS ACP). Though other fatty acid synthase (FAS) ACPs have not been reported as part of HMGS cassettes, no other endogenous ACP_D candidates were evident in or near the *bry* cluster. Therefore, in the absence of a Bry ACP_D, we sought to identify surrogate acetyl donors for substrate loading of BryR.

Several types of ACPs were surveyed (discrete ACP_Ds from HMGS-cassettes, type II PKSs and bacterial FASs, and excised ACPs from type I PKSs) in search of suitable ACP_D partners for BryR (**Figures 3-2 and 3-3**). The unmodified (apo-) and phosphopantetheine (Ppant)-containing (holo-) forms of the ACPs were overexpressed in *E. coli* and purified. The Ac- or Acac-modified ACPs were generated by loading the apo-ACPs *in vitro* using Sfp or Svp (flexible phosphopantetheinyl transferases (PPANTases), respectively).[28,29] Modified ACPs were separated from unreacted CoAs before testing. To assess the ability of BryR to catalyze HMG formation using the surrogate acyl carriers (Ac-ACP_D + Acac-ACP_A → HMG-ACP_A), we monitored the enzymatic activity of BryR when paired with different Ac-ACP_D and Acac-ACP_A substrates (**Figure 3-5**).

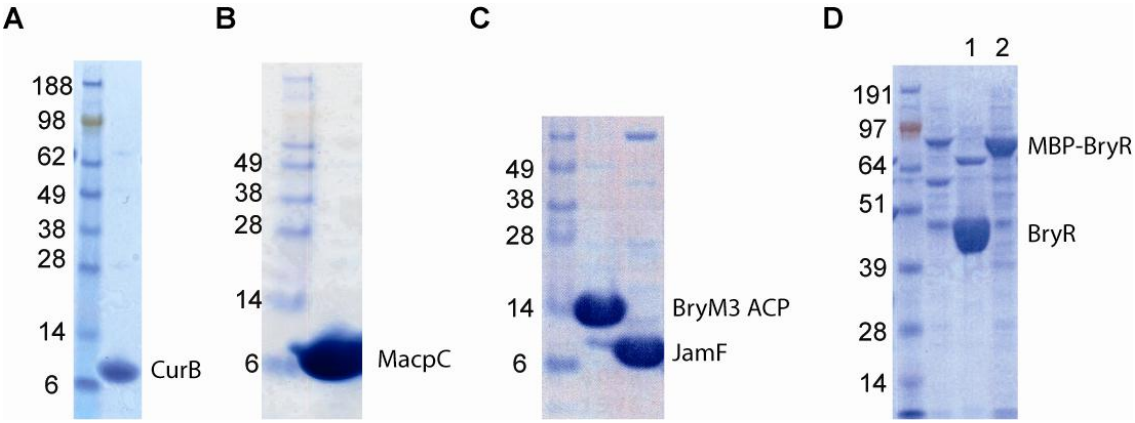


Figure 3-5. SDS-PAGE analysis of purified proteins. Apparent molecular weight of the SeeBlue Plus2 molecular weight marker (Invitrogen) is shown for reference. (**A-C**) Proteins were run on a NuPAGE 12% SDS-PAGE gel using MES buffer. (**D**) Proteins were run on a NuPAGE 4-12% SDS-PAGE gel using MOPS buffer. Lane 1 – MBP-BryR; Lane 2 – BryR after TEV protease cleavage.

In primary metabolism, HMG-CoA synthase (HMGS) catalyzes the condensation of C2 of acetyl-CoA onto the β-ketone of acetoacetyl-CoA to form 3-hydroxyl-3-

methylglutaryl-CoA and free CoASH.[30-32] A number of secondary metabolite pathways have been identified over the past five years that perform a similar reaction, although they appear to use ACP-tethered acyl groups as opposed to acyl-CoA substrates. By analogy to primary metabolism HMGSs, the first step in the BryR enzyme mechanism should be acetylation of the active site cysteine in the enzyme.[33] Subsequently, the C2 of acetate reacts with the β-keto group of the Acac-ACP_A substrate to form HMG (or a related molecule during biosynthesis) (**Figure 3-2**). These steps were assessed by both Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) (**Figures 3-6A, 3-7, and 3-10A-B**) and radio-SDS PAGE (**Figure 3-6B**) with the surrogate ACP_D MacpC. Substrate transfer from Ac-ACP (FTICR-MS) or [1-^{14}C]-Ac-ACP (radio-SDS PAGE) to BryR was confirmed only when a member of the discrete HMGS-cassette ACP_D group (**Figure 3-2**) was paired with BryR (**Figure 3-6**). By FTICR-MS, we were also able to observe loss of the Ac-ACP_D species and its conversion to holo-ACP_D in the presence of BryR (**Figures 3-6A, 3-7,** and **3-10AB**). Similar conversions were seen for two additional surrogate ACP_Ds: CurB (**Figures 3-8** and **3-10C-D**) and CurB (**Figures 3-9,** and **3-10 E-F**). Control reactions helped establish confidence in our FTICR-MS results (**Figures 3-11** and **3-12**).

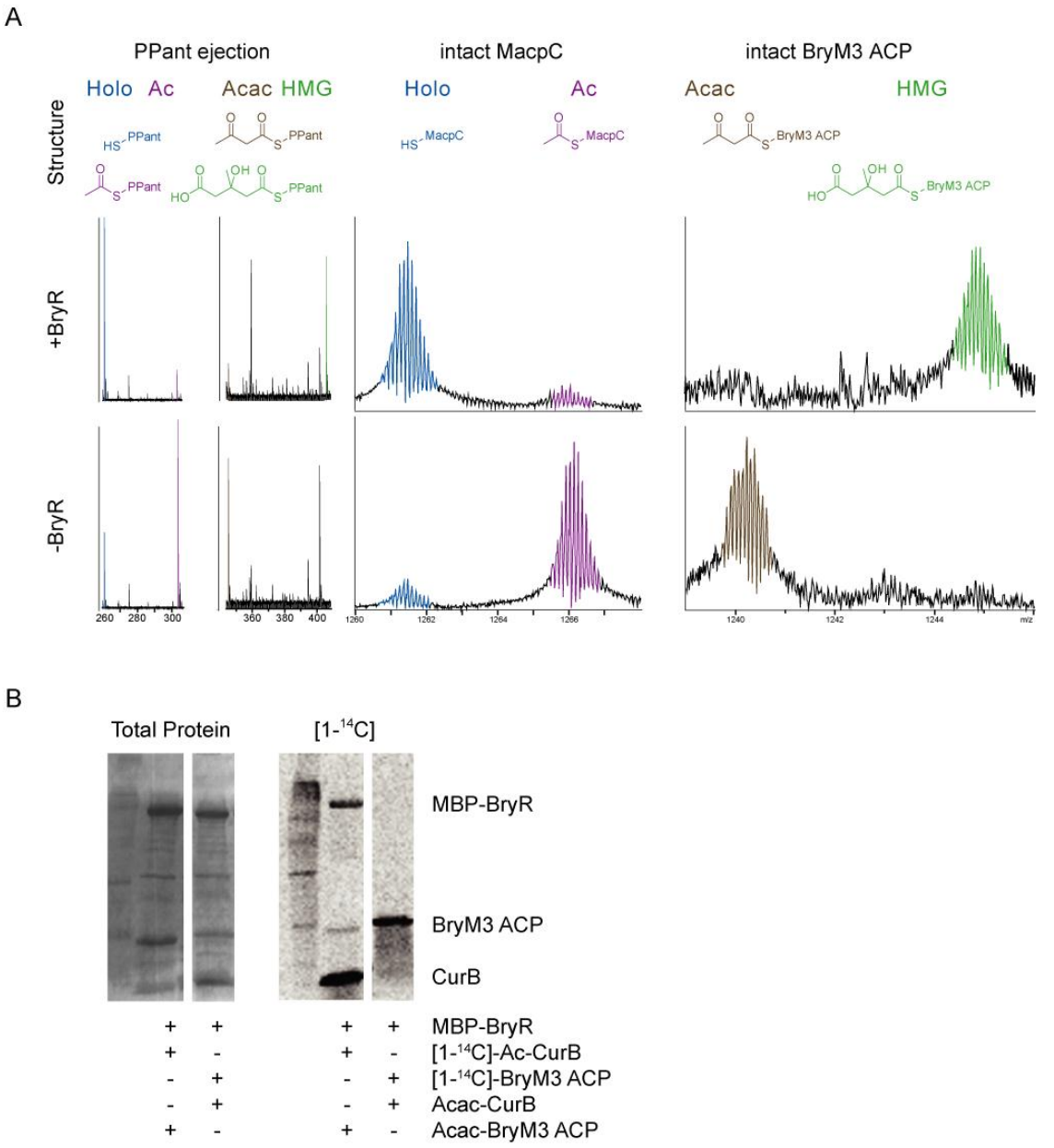


Figure 3-6. BryR catalyzed generation of HMG-BryM3 ACP from Ac-MacpC. (**A**) As monitored by FTICR-MS, data are presented as m/z versus abundance. Ppant ejection assay data from the entire charge state distribution are presented. Ppant ejection peaks are in the +1 charge state. Intact donor and acceptor ACP data are also illustrated. Holo- and Ac-MacpC peaks are shown in the +13 charge state. Acac- and HMG-BryM3 are shown in the +14 charge state. Particular charge states illustrated are representative of the entire charge state envelope. (**B**) Rado-TLC monitored acetyl transfer from [1-^{14}C]-Ac-CurB to Acac-BryM3 ACP to form HMG-BryM3 ACP monitored by radio-SDS PAGE.

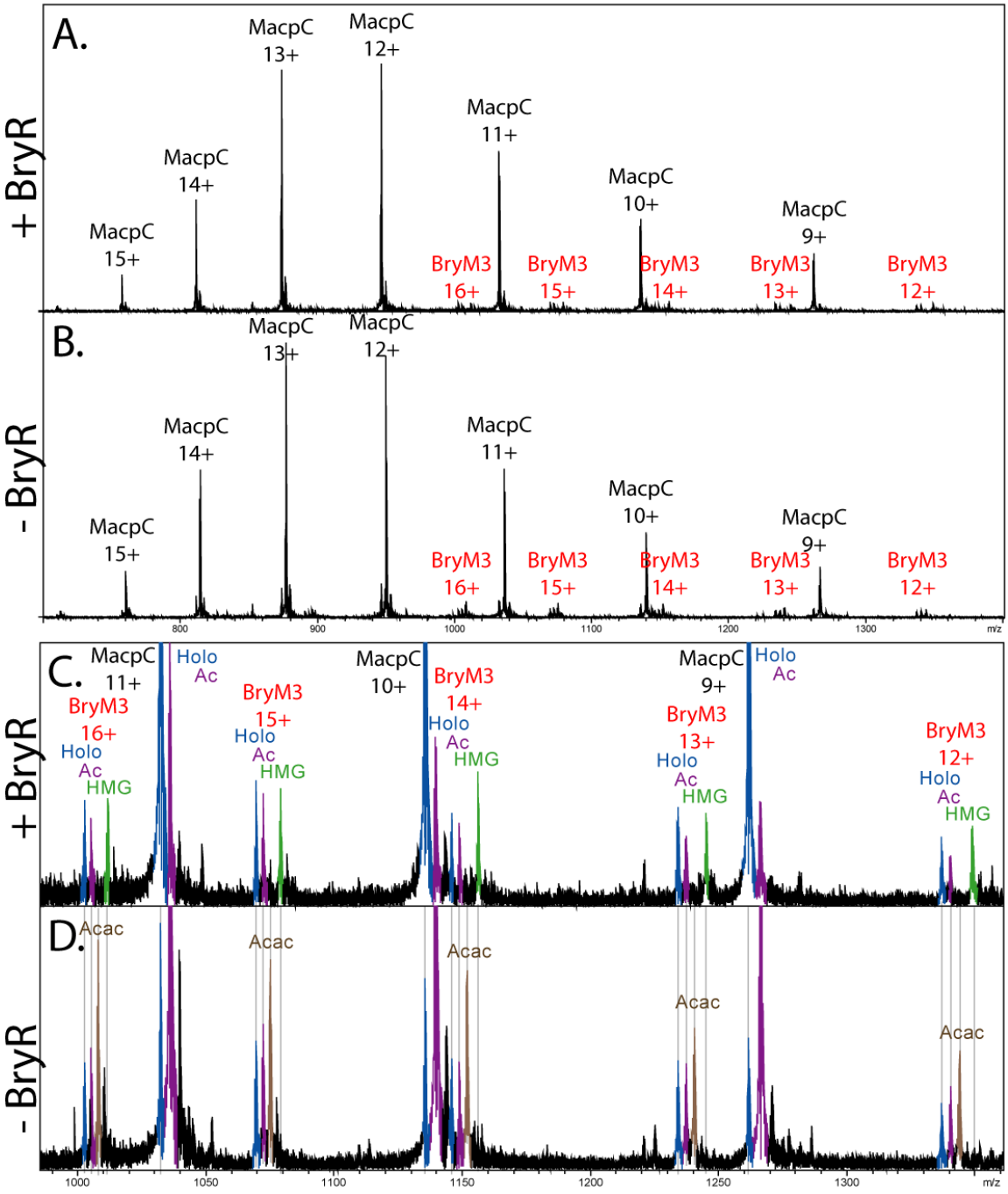


Figure 3-7. BryR catalyzed generation of HMG-BryM3 ACP from Ac-MacpC and Acac-BryM3 as monitored by FTICR-MS. Intact donor and acceptor ACP data are illustrated. Data are presented as m/z versus abundance. (**A**) BryR + Ac-MacpC + Acac-BryM3, Full spectrum. (**B**) Ac-MacpC + Acac-BryM3, Full spectrum. (**C**) BryR + Ac-MacpC + Acac-BryM3, zoom. (**D**) Ac-MacpC + Acac-BryM3, zoom. See **Figure 3-11** for Ppant ejection data and **3-12** and **3-13** for authentic standards.

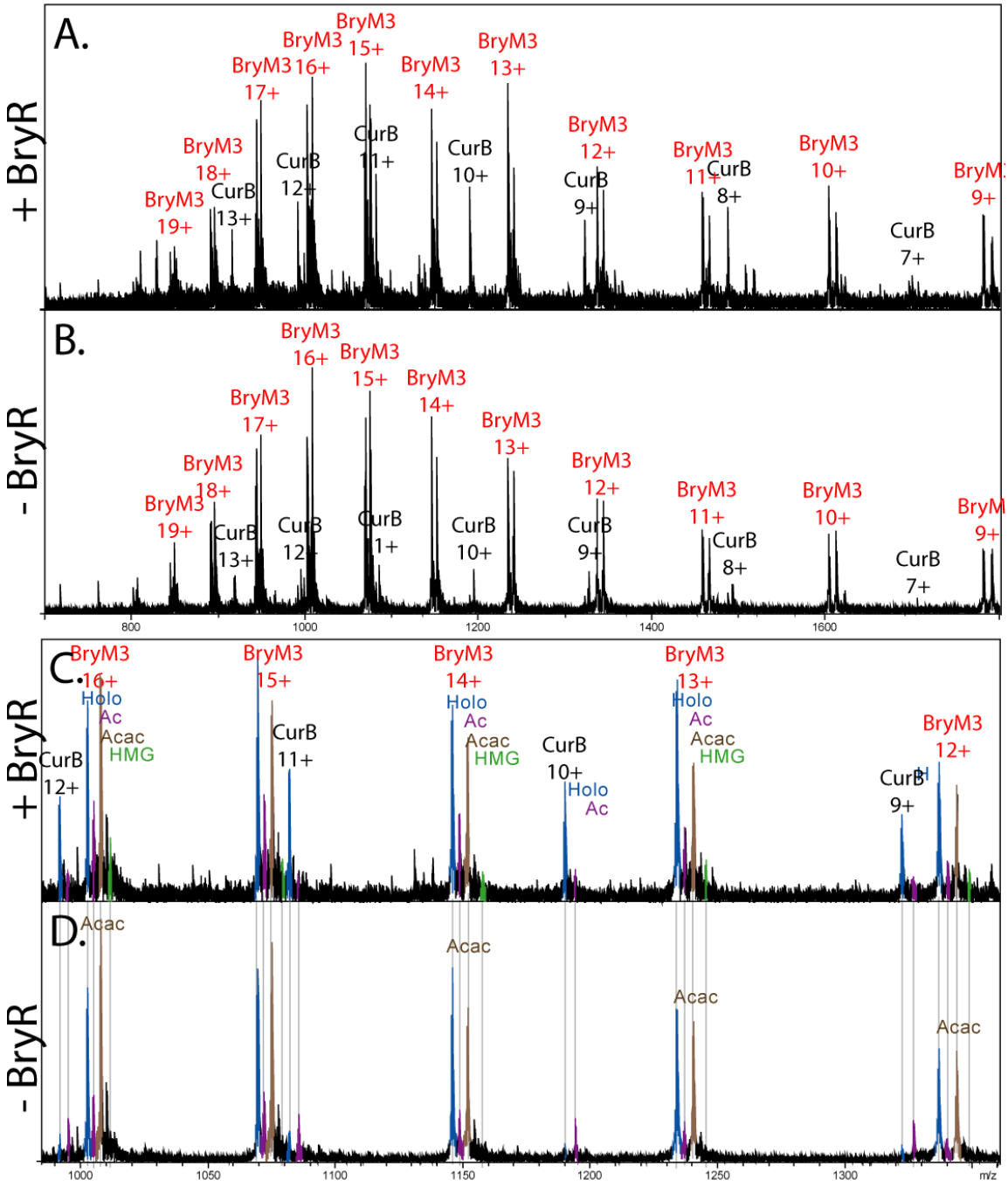


Figure 3-8. BryR catalyzed generation of HMG-BryM3 ACP from Ac-CurB and Acac-BryM3 as monitored by FTICR-MS. Intact donor and acceptor ACP data are illustrated. Data are presented as m/z versus abundance. (**A**) BryR + Ac-CurB + Acac-BryM3, Full spectrum. (**B**) Ac-CurB + Acac-BryM3, Full spectrum. (**C**) BryR + Ac-CurB + Acac-BryM3, zoom. (**D**) Ac-MacpC + Acac-BryM3, zoom. See **Figure 10** for Ppant ejection data and **3-11** and **3-12** for authentic standards.

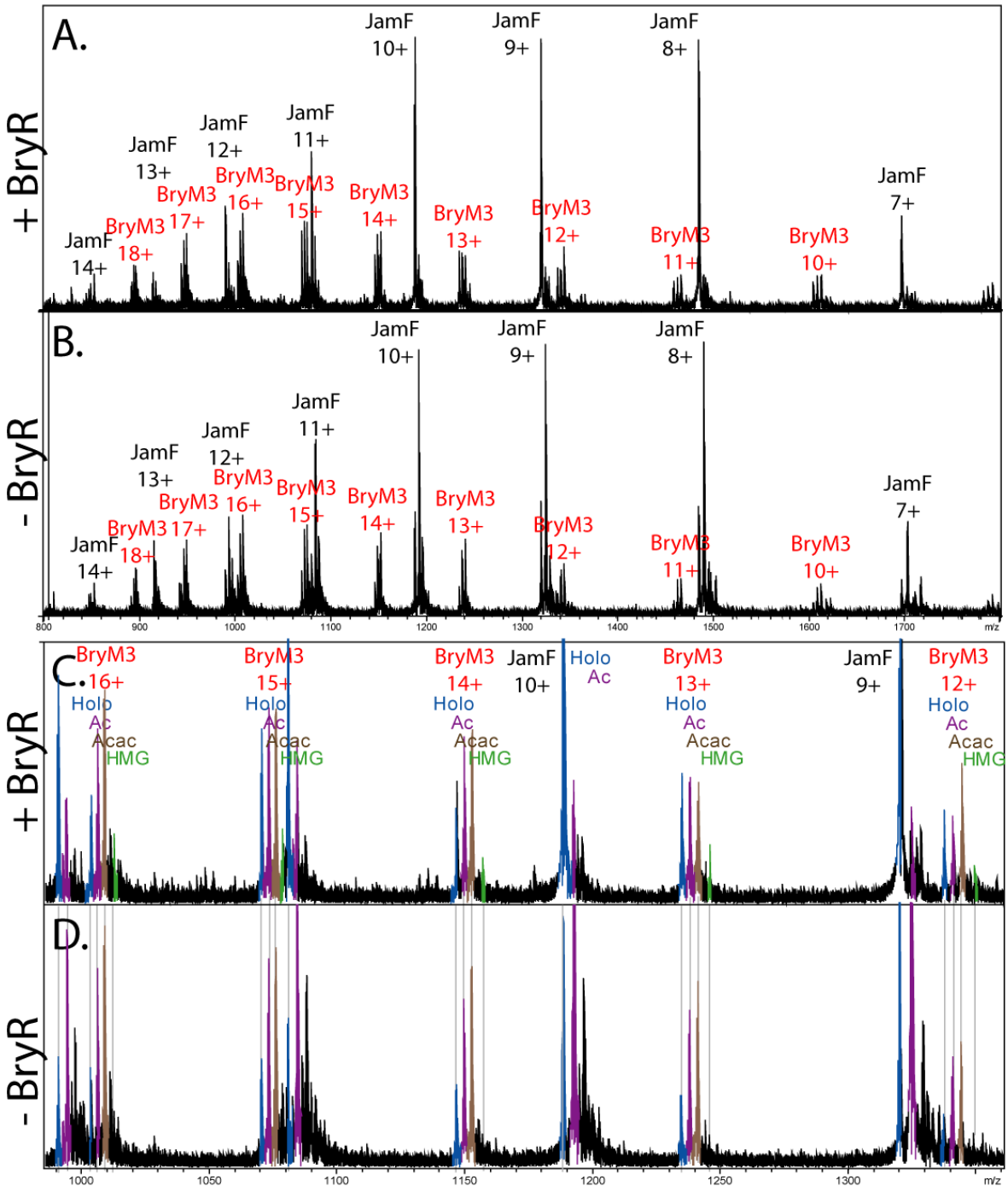


Figure 3-9. BryR catalyzed generation of HMG-BryM3 ACP from Ac-JamF and Acac-BryM3 as monitored by FTICR-MS. Intact donor and acceptor ACP data are illustrated. Data are presented as m/z versus abundance. (**A**) BryR + Ac-JamF + Acac-BryM3, Full spectrum. (**B**) Ac-JamF + Acac-BryM3, Full spectrum. (**C**) BryR + Ac-JamF + Acac-BryM3, zoom. (**D**) Ac-JamF + Acac-BryM3, zoom. See **Figures 3-10** for Ppant ejection data and **3-11** and **3-12** for authentic standards.

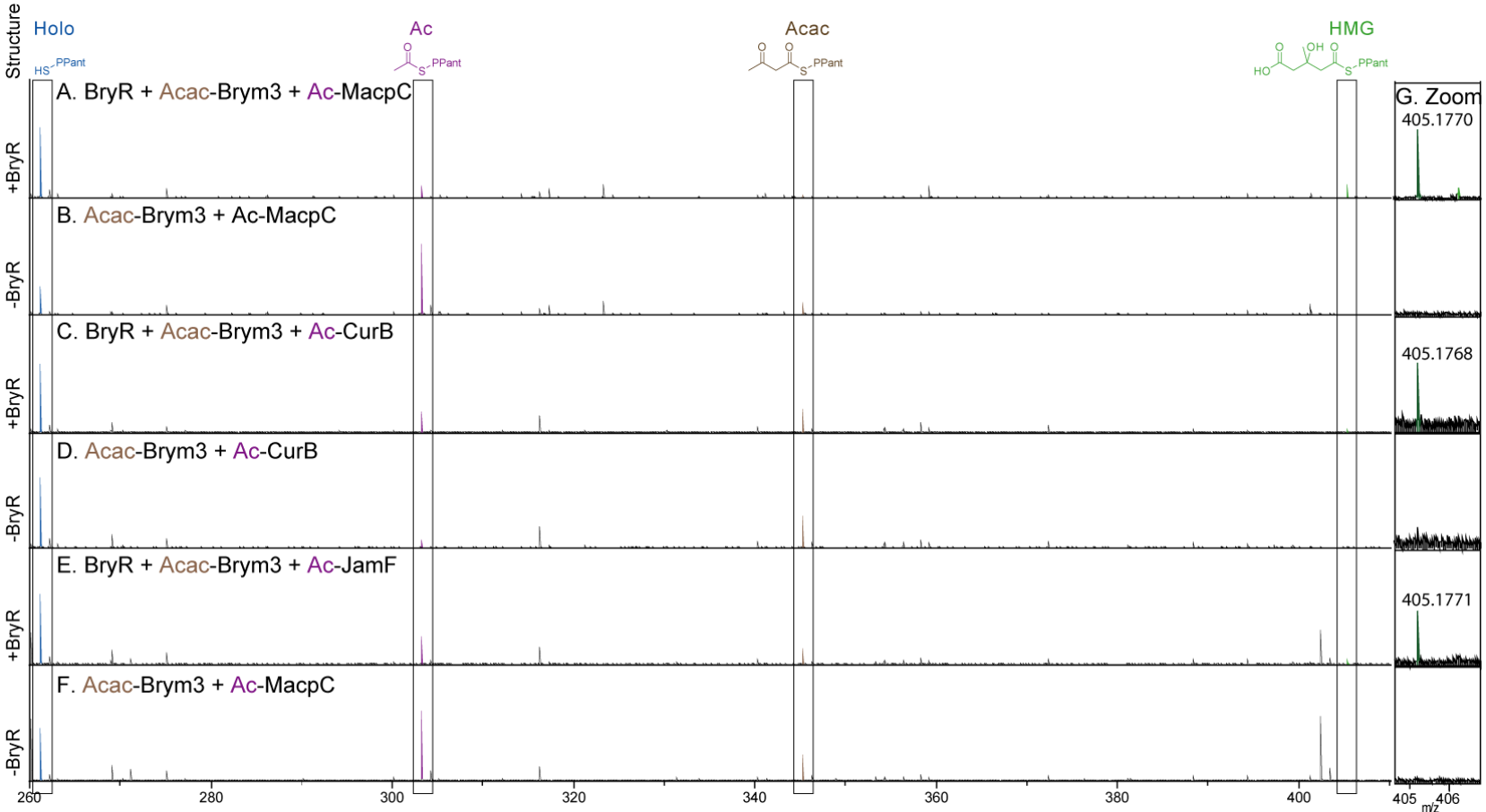


Figure 3-10. BryR catalyzed generation of HMG-BryM3 ACP. Ac-MacpC (**A-B**), Ac-CurB (**C-D**), and Ac-JamF (**E-F**), as monitored by FTICR-MS. Ppant ejection assay data from the entire charge state distribution are presented and are presented as m/z versus abundance. Ppant ejection peaks are in the +1 charge state, and the HMG-Ppant ion is only present when BryR is added to the reactions (**G**). Intact donor and acceptor ACP data are illustrated above: **3-6A**, **3-7, 3-8,** and **3-9**. See **Figures 3-11** and **3-12** for authentic standards.

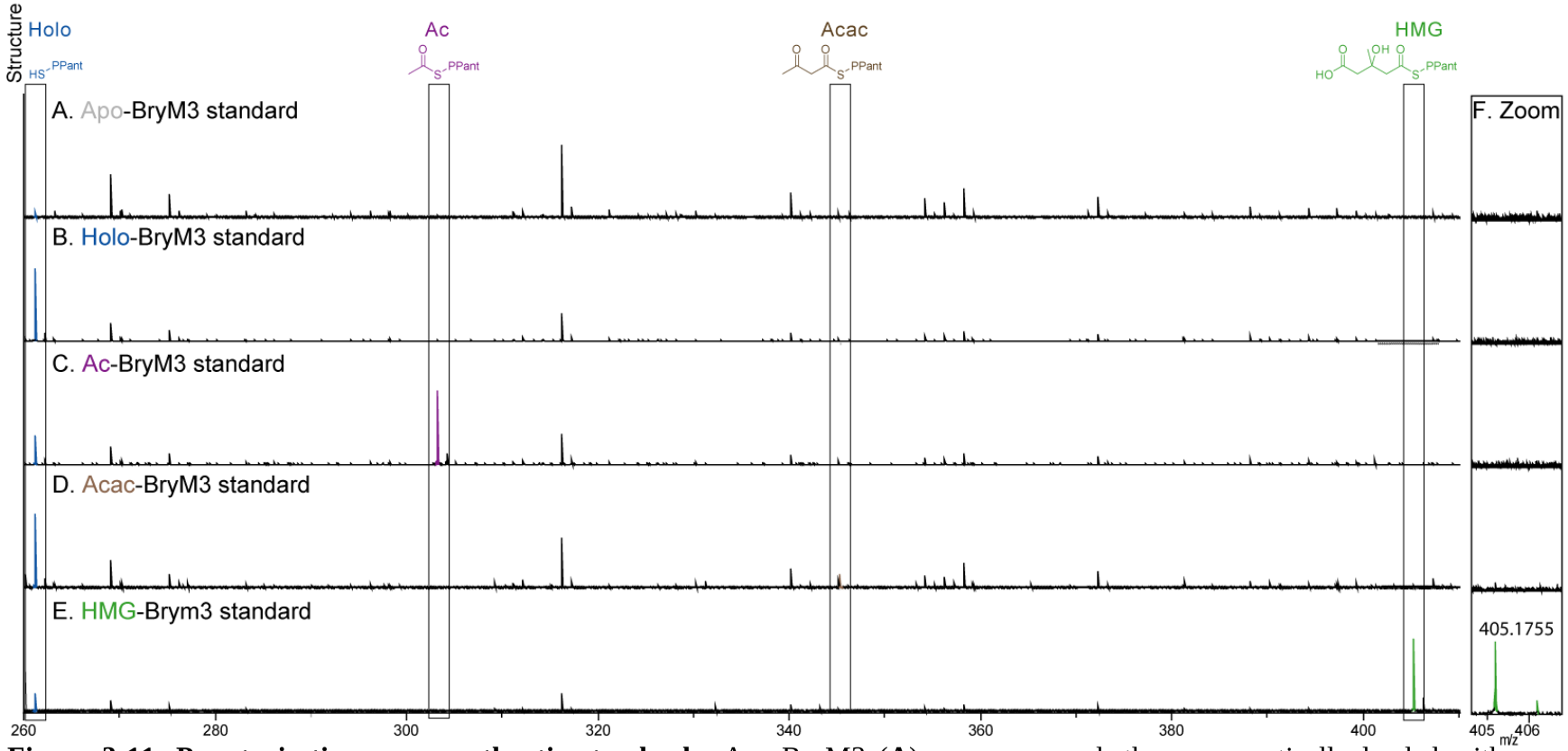


Figure 3-11. Ppant ejection assay authentic standards. Apo-BryM3 (**A**) was expressed, than enzymatically loaded with authentic coenzyme A (**B**), Ac-CoA (**C**), Acac-CoA (**D**), and HMG-CoA (**E**), and then subjected to Ppant ejection assay conditions. Data are presented as m/z versus abundance. Ppant ejection assay data from the entire charge state distribution are presented. The HMG-Ppant ion is only present when HMG-BryM3 is fragmented. All experimental HMG-Ppant ions are less than 5 ppm from the standard above.

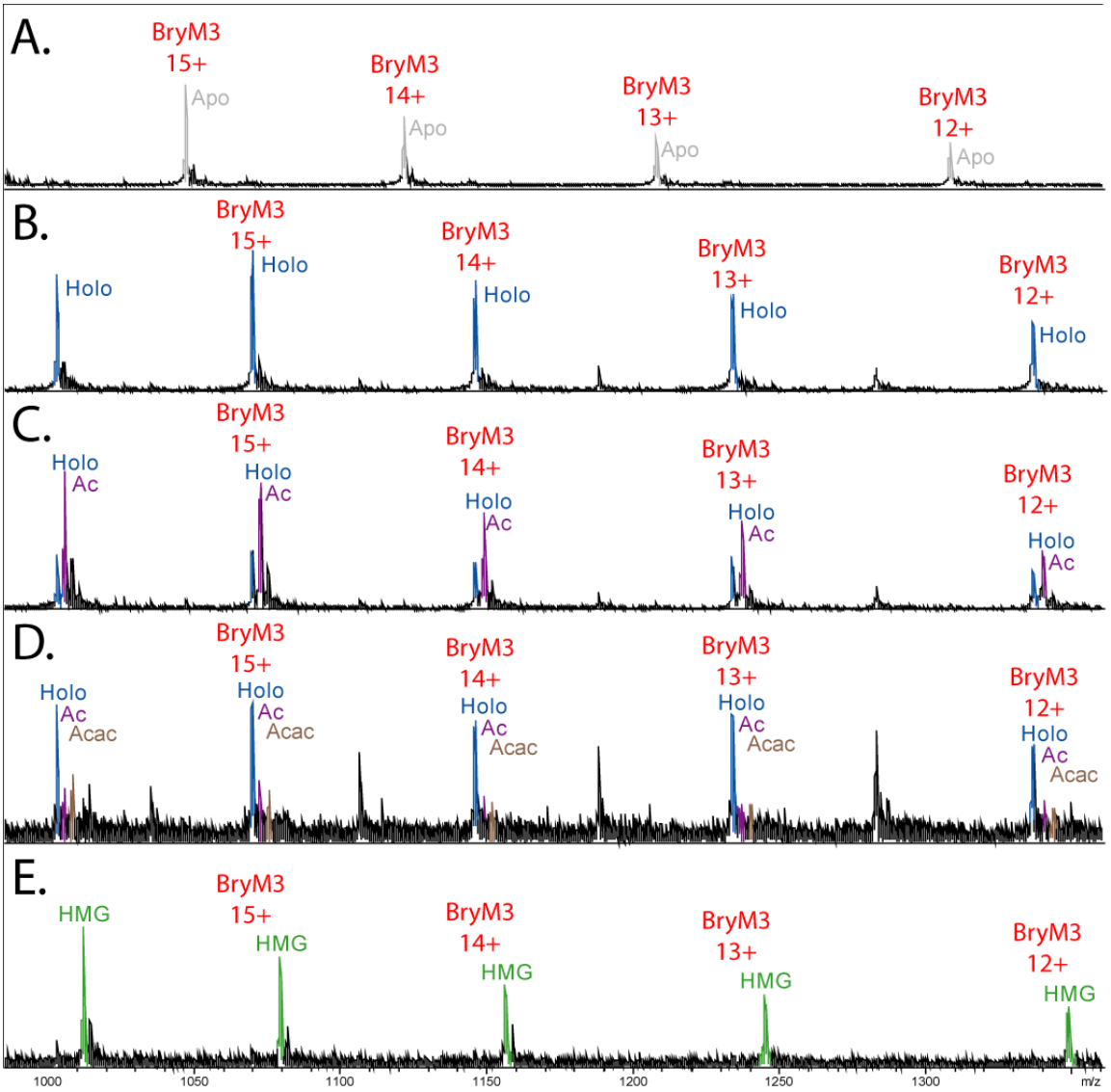


Figure 3-12. Intact accepter ACP authentic standards. Apo-BryM3 (**A**) was expressed, than enzymatically loaded with authentic coenzyme A (**B**), Ac-CoA (**C**), Acac-CoA (**D**), and HMG-CoA (**E**), and then intact ACP data was generated. Zoomed data are presented as m/z versus abundance. The Acac-BryM3 control reaction exhibited incomplete loading, possibly due to hydrolysis, compared to the other reactions present.

Generation of the Ac-BryR intermediate during the first half of the reaction can be visualized in the phosphorimage only when [1-^{14}C]-Ac-CurB (HMGS-cassette ACP_D), donates the acetyl group (**Figure 3-6B**). To confirm that the BryR reaction proceeds through the same enzyme intermediate as those observed in primary metabolism, we confirmed that the acetylation occurs on Cys114. BryR (10 μM) was reacted with Ac-MacpC (50 μM) in the absence of an ACP_A. After the sample was proteolyzed with

trypsin, peptides were separated by HPLC, and using LC-FTICR-MS and ion trap LC-MS/MS the BryR active site peptide was identified and acetylation of Cys114 was confirmed (**Table 3-1**). Additionally, a mutant form of the protein at this location (C114A) was enzymatically inactive (data not shown). These data represent the first direct demonstration of the Ac-Cys species in an HMGS homolog in polyketide biosynthesis.

Mass	**Intensity**	**ID***	**dPPM**
480.18	62	a_4	-28
463.13	30	a_4 - NH_3	-65
567.19	18	a_5	-67
550.24	28	a_5 - NH_3	79
708.28	76	a_7 - NH_3	19
753.19	51	b_7	-132
735.22	451	b_7 - H_2O	-76
736.23	145	b_7 - NH_3	-37
824.28	22	b_8	-53
806.54	16	b_8 - H_2O	276
807.32	91	b_8 - NH_3	28
732.35	17	y_6 - NH_3	-70
749.30	21	y_6 - NH_3	-171
845.34	20	y_7 - NH_3	-172
1,064.52	18	y_9	-63
1,047.46	35	y_9 - NH_3	-104

Table 3-1. **Ac-BryR active site peptide fragment ions observed in ion trap LC/MS/MS.** b- and y-type fragment ions originate from peptide backbone bond cleavage where b ions contain the peptide N terminus and y ions contain the peptide C terminus. The subscripted number indicates the number of amino acid residues in a particular fragment. a-type ions result from secondary fragmentation of b-ions via CO loss.

As evidence of the ability of BryR to catalyze the complete reaction (Ac-ACP_D + Acac-ACP_A → HMG-ACP_A), we observed a third radioactive band, consistent with modification of BryM3 ACP (a model acceptor substrate), an embedded ACP excised from the BryA tetramodule at one of the predicted HMGS modification sites (**Figures 3-1**). To identify the chemical modification on BryM3 ACP, the reaction mixtures were monitored by top-down FTICR-MS. The mass shift of +60.0 Da on the intact BryM3 ACP between the +/- BryR samples is consistent with conversion of the Acac-BryM3 ACP_A substrate to HMG-BryM3 ACP_A (**Figures 3-6A, and 3-7**). MS/MS analysis was performed using the Ppant ejection assay,[34,35] which confirmed that the mass shift between +/- BryR samples is due to modification of the Ppant prosthetic group (**Figure 3-10**). No product formation was observed when Ac-Bry FAS ACP was incubated with BryR and Acac-BryM3 ACP (data not shown). Other reactions without detectable product formation included an Ac-FAS ACP donor from *Streptomyces coelicolor* (SCO2389, *Sc* FAS ACP),[36] an Ac-CoA donor, using BryM3 ACP as both donor and acceptor, or Acac-FAS ACPs or Acac-MacpC as acceptors (data not shown).

Since the HMGS homologs found in secondary metabolism do exhibit a preference for acyl-ACPs, we sought to measure the affinity of BryR for these ACPs. The direct binding of BryR to a variety of potential Ac-ACP_Ds as well as the model acetoacetyl acceptor substrate, Acac-BryM3 ACP was assessed (**Figure 3-15**). After BryR immobilization to a BIAcore CM5 SPR chip (**Figure 3-13**), equilibrium binding analysis was performed using sequential injections of apo-, holo-, Ac-, or Acac-ACPs at varying concentrations (**Figure 3-14** and **Figure 3-15**). Active BryR (WT) and an enzymatically inactive (C114A) BryR mutant behaved similarly in our binding studies. BryR was able to bind to ACP_Ds from the curacin (CurB), jamaicamide (JamF), and mupirocin (MacpC) HMGS cassettes as well as to the excised native acceptor (BryM3 ACP) (**Figure 3-15**). Affinities (K_Ds) were in the middle to high micromolar range for ACP_Ds (40 – 110 μM) and the ACP_A (180 μM). The pattern of BryR binding affinities correlated well with that observed in our enzymatic activity assays. No significant binding was observed between Bry FAS ACP (up to 500 μM) or *Sc* FAS ACP (up to 650 μM) and BryR (**Figure 3-15**). No enhancement of affinity was observed between apo- and Ac-ACP_D or between apo-, holo-, Ac-, AcAc-, or HMG-ACP_A (**Figure 3-15**). Thus,

the affinity of BryR for the ACPs seems to be mediated mainly by protein-protein contacts (as opposed to protein-acyl chain or protein-Ppant contacts). These data suggest that specificity for a protein-bound acyl group is a distinguishing feature between HMGS homologs found in PKS or mixed PKS/nonribosomal peptide synthase (NRPS) biosynthetic pathways and those of primary metabolism.

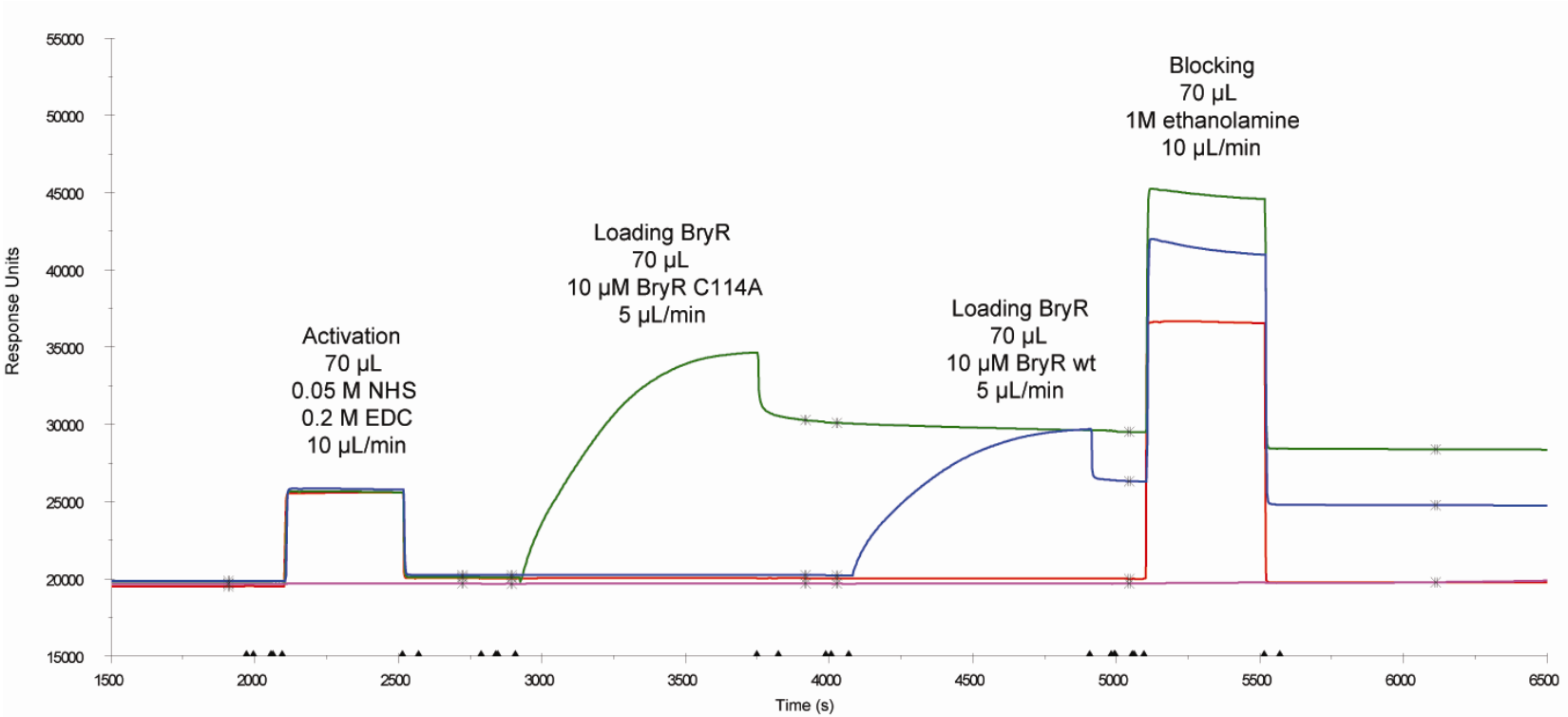


Figure 3-13. Raw sensorgram data from BIACORE 3000 Control software for immobilization of BryR and BryR C114A to the CM5 chip. FC1 is shown in red, FC2 in blue, FC3 in green and FC4 in pink.

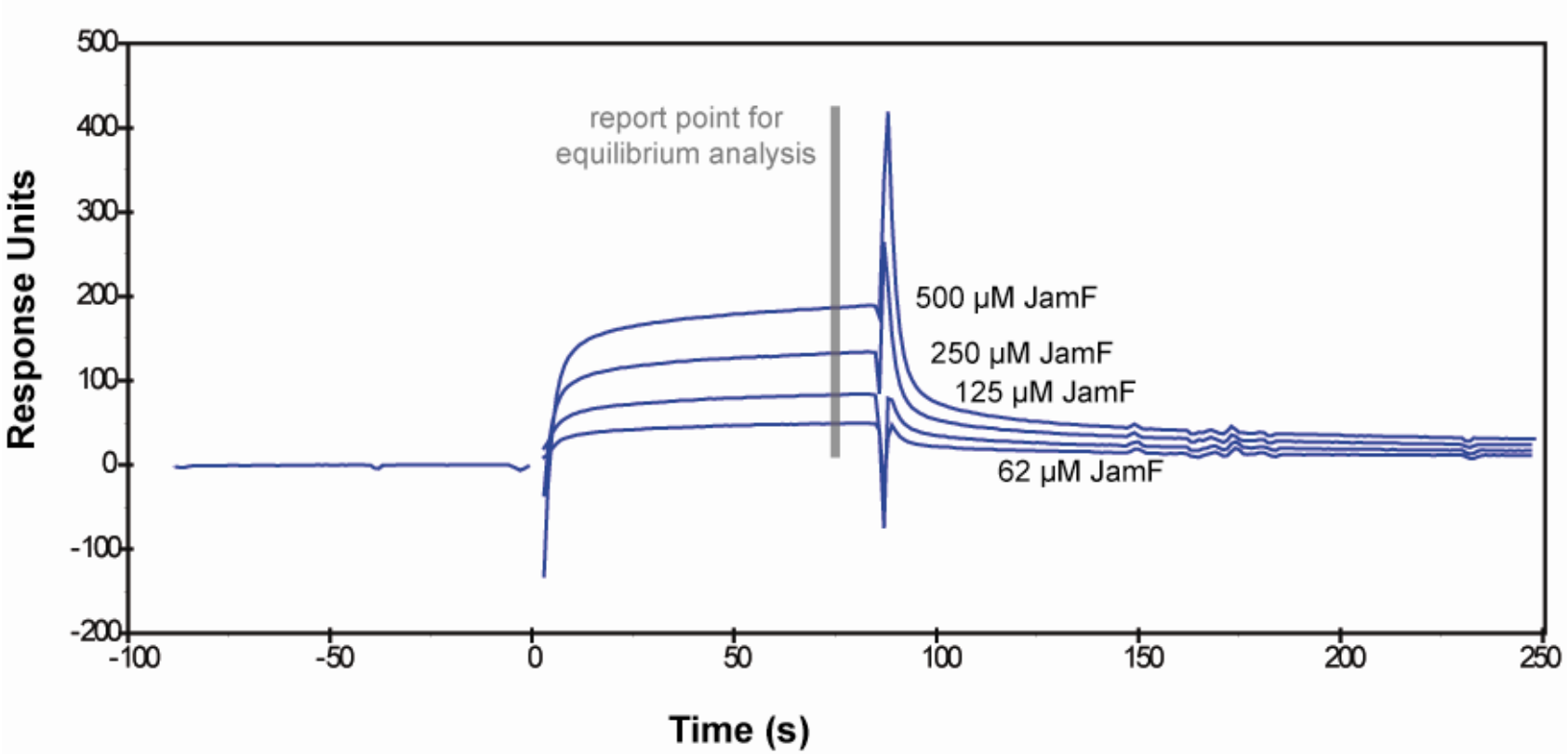
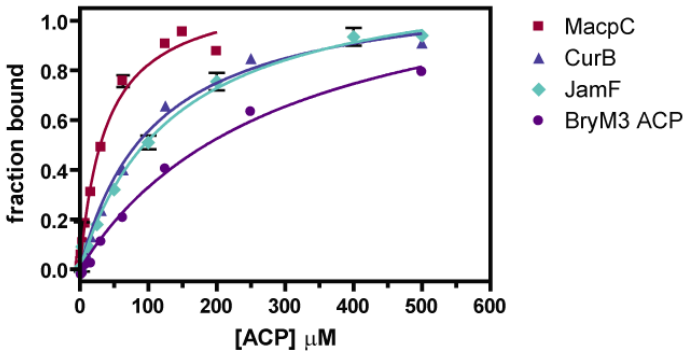


Figure 3-14. Subtracted BIAcore data for four concentrations of JamF:BryR binding. Data analyzed with BIAevaluation software. Report point was set at 75 seconds after the ACP injection.

A



B

		apo-	*holo*-	Ac-	Acac-	HMG-
ACP_D	MacpC	40 ± 5 μM		32 ± 9 μM		
	CurB	108 ± 7 μM		90 ± 8 μM		
	JamF	100 ± 20 μM		107 ± 7 μM		
ACP_A	BryM3 ACP	177 ± 7 μM	180 ± 24 μM	200 ± 12 μM	150 ± 51 μM	170 ± 13 μM
	Bry FAS ACP	> 500 μM				
	Sc FAS ACP	> 650 μM				

Figure 3-15. Binding of apo-ACPs to immobilized BryR, monitored by SPR. (**A**) Each data point is the average of triplicate measurements; error bars are standard deviation. The data were fit to a one-site binding model ($Y = B_{max}*X/(K_D + X)$). Y = fraction bound, B_{max} = maximal response, X = ACP concentration. Dissociation constants (K_Ds) are reported in (**B**) Forms of ACPs shaded grey in the table were not tested.

3.3 Discussion

We have investigated the enzymatic function of BryR (condensation of acetyl-ACP_D with acetoacetyl-ACP_A to form HMG-ACP_A) using two complementary methods, radio-SDS PAGE and FTICR-MS. The activity of BryR was dependent on pairing of the native Acac-BryM3 acceptor ACP with an appropriate surrogate Ac-ACP_D from a related HMGS cassette (CurB, JamF, or MacpC). In addition, the ability of BryR to discriminate between various ACPs was assessed using an SPR-based protein-protein binding assay. BryR bound selectively to ACPs obtained from a series of HMGS cassettes (MacpC, CurB, JamF, and BryM3 ACP). To date, no structural insights have been reported for the interaction of HMGS cassette enzymes with partner ACP_Ds. These future studies will be essential to determine the nature of BryR's ACP binding selectivity. Finally, this work, as

well as other recent studies[27,37] demonstrate further that natural product biosynthetic genes isolated from unculturable marine symbiotic bacteria can be manipulated *in vitro* in order to probe the functionalities of these enzymes from previously inaccessible sources.

3.4 Supplement

Expression and purification of proteins

Plasmids encoding N-terminal His_6- or His_6/MBP- fusion protein tags were transformed into *E. coli* BL21(DE3) and grown at 37 °C in TB medium to an OD_{600} of ~1.0 in 2 L flasks. The cultures were cooled to 18 °C, and isopropyl β-D-thiogalactopyranoside was added to a final concentration of 0.2 mM and grown 12-16 hr with shaking. The cells were harvested by centrifugation and frozen at -20 °C. Cell pellets were thawed to 4 °C and resuspended in 5X volume of lysis buffer (20 mM HEPES, pH 7.8, 300 mM NaCl, 20 mM imidazole, 1 mM $MgCl_2$, 0.7 mM Tris(2-carboxyethyl) phosphine (TCEP pH 7.5), ~100 mg CelLytic Express (Sigma-Aldrich)) before lysis via sonication. Centrifugation at 25,000 *x* g for 30 min provided clarified lysates. Proteins were purified using Ni-Sepharose affinity chromatography on an Akta FPLC. Briefly, after filtration of the supernatant through 0.45 μm membrane, the solution was loaded onto a 5 mL HisTrap nickel-nitrilotriacetic acid column. The column was washed with 10 column volumes of buffer A (20 mM HEPES, pH 7.8, 300 mM NaCl, 20 mM imidazole, 0.7 mM TCEP) and eluted with a linear gradient of buffer B (20 mM HEPES, pH 7.8, 300 mM NaCl, 400 mM imidazole, 0.7 mM TCEP pH 7.5). For ACP purifications, fractions were pooled, concentrated, and loaded onto a HiLoad 16/60 Superdex 75 (GE Healthcare Life Sciences) column equilibrated with storage buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 0.7 mM TCEP pH 7.5). Fractions were combined, concentrated, frozen, and stored at -80 °C. Because some of the acyl carrier proteins lack amino acids with appreciable absorbance at 280 nm, protein concentrations were determined via the bicinchoninic acid (BCA) method using BSA as a standard. BryR purificiations differed from ACP purifications in that all buffers contained 10% glycerol in addition to the components listed above. In addition, for SPR and FTICR-MS assays, His-MBP-tag removal was achieved by TEV protease incubation overnight at 4 °C in buffer A. TEV protease and the N-terminal His-MBP tag were removed by repassaging the solution over the HisTrap

column. Flow-through fractions were pooled, concentrated, and loaded onto a HiLoad 16/60 Superdex 200 column equilibrated with BryR storage buffer (10% glycerol, 20 mM HEPES, pH 7.4, 150 mM NaCl, 0.7 mM TCEP pH 7.5). Fractions were combined, concentrated, frozen, and stored at -80 °C. Protein concentrations were determined using absorbance at 280 nm and calculated extinction coefficients (1 A_{280} = 1.2 mg/mL). ACPs were greater than 95% pure following the above purification. Typical yields for BryR batches were ~ 3 mg/L of cell culture. TEV-cleaved BryR was approximately 85-90% pure. Purity estimates are based on SDS-PAGE (**Figure 3-5**).

Enzymatic analysis of BryR via radio-TLC

Radiolabeled and unlabeled acyl-CoA substrates were transferred onto the various ACPs using Svp, a phosphopantheine transferase from *Streptomyces verticillus*.[29] Acyl-CoAs (500 μM) were combined with 75 μM ACPs (CurB, BryM3 ACP) and 5 μM Svp in a Tris buffer (pH 7.4) containing $MgCl_2$ (10 mM) and DTT (1 mM), and the reaction proceeded for 1 hr at room temperature. The substrate-bound ACPs were desalted, and utilized for experiments with BryR. The purified acylated donor (15 μM) and acceptor ACPs (30 μM) were incubated with BryR (10 μM) in 25 mM Tris buffer (pH 7.4) with DTE (1 mM) at room temperature for 5 minutes. Reactions were quenched by the addition of SDS-PAGE gel loading buffer. Samples were separated on polyacrylamide gels by SDS-PAGE. The gels were first stained using SimplyBlue (Invitrogen), and were then exposed to Phosphoimager screens. The screens were scanned using a Typhoon Scanner (GE Healthcare), and analyzed using ImageQuant.

Enzymatic analysis of BryR via FTICR-MS

The preparation of acetyl-donor and acetoacetyl-acceptor ACPs was performed as above using Svp or Sfp PPANTases.[28] Acylated-ACPs were separated from CoA substrates via Zeba desalting columns (Pierce) or overnight dialysis in 3.5 kDa Slide-a-lyzer MINI dialysis units (Pierce) into 20 mM HEPES (pH 7), 150 mM NaCl. BryR (10 μM) was reacted with acetyl-donor ACP (50 μM) and acetoacetyl-acceptor ACP (80 μM) 75 mM HEPES (pH 7.5) buffer and 1 mM TCEP pH 7.5. After incubation for 60 min at room temperature, samples were acidified with 1% formic acid. Intact protein samples were

desalted with Handee Microspin columns (Pierce) packed with 20 µL of 300 Å polymeric C4 resin (Vydac). Samples were loaded onto the columns and washed with 30 column volumes of 0.1% formic acid prior to elution with 10 column volumes of 50% acetonitrile plus 0.1% formic acid. Intact protein samples were analyzed by an FTICR-MS (APEX-Q with Apollo II ion source and actively shielded 7T magnet; Bruker Daltonics). Data were gathered from m/z 200–2,000 utilizing direct infusion electrospray ionization in positive ion mode. Electrospray was conducted at 3,600 V with 24–60 scans per spectra utilizing 0.5 s external ion accumulation in a hexapole and 15 ICR cell fills prior to excitation and detection. Collision cell pressure was reduced to 2.5e-6 Torr for improved transmission of protein ions. Infrared multiphoton dissociation (IRMPD) MS/MS was performed in the FTICR cell. This approach is preferred over external collision induced dissociation, because time of flight effects during ion transport into the FTICR cell are avoided. The laser power was 10 W with an irradiation time of 0.05 to 0.25 s. The entire mass range was fragmented, without any prior mass selection. Data were processed in Data Analysis (Bruker Daltonics) and Midas (NHMFL). All mass shifts shown were confirmed across all charge states for each ACP present. An abundant charge state is used for **Figure 3-6A** and **Figures 3-7 to 3-9** illustrate all charge states for the intact ions. All identified species were accurate to 20 ppm with external calibration. All experiments were performed at least twice to verify the findings.

Identification of BryR active site acetylation by LC FTICR-MS and LC ion trap-MS

BryR (10 µM) was reacted with acetyl-donor ACP (50 µM) and no acceptor ACP in 75 mM HEPES (pH 7.5) buffer, 1 mM TCEP pH 7.5, then 1 mg/mL TPCK trypsin (Pierce) was added to a final 1:100 ratio. Samples were incubated at 37 °C overnight. 20 µL of samples was injected onto a Jupiter C18 1x150 mm 300 µm column (Phenomenex) using an Agilent 1100 LC system with a flow rate of 75 µL/min and a gradient of 2-98% acetonitrile over 85 minutes. 0.1% formic acid was added to the water and acetonitrile solvents. A divert valve was utilized for online desalting. The LC was coupled to an FTICR-MS (APEX-Q with Apollo II ion source and actively shielded 7T magnet; Bruker Daltonics). Data were gathered from m/z 200–2,000 in positive ion mode. Electrospray was conducted at 2,600 V with 4 scans per spectra utilizing 0.33 s external ion

accumulation in a hexapole and 4 ICR cell fills prior to excitation and detection. Data were analyzed using DECON2LC and VIPER (Pacific Northwest National Labs). The acetylated active site peptide QA**C**YSGTAGFQMAINFILSR (2219.050 Da expected) was observed at 2219.045 Da representing a mass error of -2 ppm with external calibration. The same LC conditions were coupled to an LTQ Deca XP iontrap MS (Thermo). Online MS identified the same modified peptide, and online MS^2 allowed for confirmation that the modification occurred on the active site cysteine (Cys114) with the following fragment ions identified (**Table 3-1**). Data were processed in Excaliber version 3.0 (Thermo).

Surface plasmon resonance assays of BryR, ACP_D and ACP_A

Sensor chips (CM-5) and HBS-P buffer were purchased from GE Healthcare Life Sciences. SPR experiments were performed on a BIAcore 3000 instrument. Running buffer for SPR was HBS-P+T (10 mM Hepes, pH 7.4, 0.15 M NaCl, 0.005% surfactant P20, 50 μM TCEP pH 7.5). The surface was prepared for immobilization of BryR by activating with 70 μL of a fresh mixture of 0.2 M 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) plus 0.05 M *N*-hydroxysuccinimide at 10 μL/min. BryR diluted to 20 μM in 10 mM phosphate/citrate buffer at pH 5.5 and loaded at 5 μL/min (**Figure 13**). Typically, 1000–8000 RU of BryR was immobilized. Activated carboxy groups were blocked with 1 M ethanolamine/HCl (70 μL at 10 μL/min). The surface was regenerated with 10 μL of 50 mM NaOH, 1 M NaCl after immobilization and between ACP binding cycles. To measure binding to BryR by SPR, solutions of ACPs in HBS-P+T were injected over the prepared surface as well as an ethanolamine treated control flow cell at a flow rate of 10 μL/min. Baseline subtraction was performed using a mock treated lane (activated with EDC/NHS and blocked with ethanolamine). Multiple injections (8–10 concentrations) were tested in duplicate or triplicate. Maximum testable concentrations for the ACPs were limited by their solubility. Data analysis was carried out using BIAevaluation software (GE Healthcare Life Sciences). Representative sensorgrams for four concentrations of apo-JamF binding to BryR are shown in **Figure 14**. Nonlinear curve fitting of the equilibrium binding response was carried out using GraphPad Prism software.

Design of expression constructs

Plasmids for the expression of CurB, JamF, *Sc* FAS ACP, Bry FAS ACP and BryR were generated by amplification using PCR with LIC overhangs and inserted into either the vector pMCSG7 (CurB, JamF and Bry FAS ACP) or pMCSG9.[41] CurB was amplified from plasmid pDHS2412 and JamF from Jamf:pET20. BryR was amplified from cosmid MM5 and Bry FAS ACP from cosmid MM7. PCR fragments were inserted into the vectors via ligation independent cloning. All mutants were generated according to the Quikchange® site-directed mutagenesis protocol (Stratagene/Agilent). All DNA sequences were confirmed by sequencing. The expression construct for MacpC (pGTB340) was a gift from Prof. Christopher M. Thomas. pDHS278 (BryM3 ACP in pMCSG7) was published previously.[37] The protein sequence for Bry FAS ACP in pMCSG7 is mhhhhhhssgvdlgtenlyfqsnaSNPSNTEERVKKIVAEQLGVKEDEVKMEAS FVEDLGADSLDTVELVMALEEEFETEIPDEDAEGITTVKLAIDYINAHLD. Further details in regards to construct generation are provided below (**Table 3-2**).

Primer name	**Primer sequence**	**Plasmid generated**
CurBLICFor	TACTTCCAATCCAATGCC atg agc aaa gaa caa gta cta	pDHS9780
CurBLICRev	TTATCCACTTCCAATGCTA caa ttt tgc tgc aaa taa atc	
JamFLICFor	TACTTCCAATCCAATGCC atg agc aaa gaa caa gta ctc aaa cta a	pDHS9781
JamFLICRev	TTATCCACTTCCAATGCTA taa ttt cgc cgc aaa taa atc agc	
BryFAS_ ACPLICFor	TACTTCCAATCCAATGCC agc aac cca agc aac act ga	pDHS9812
BryFAS_ ACPLICRev	TTATCCACTTCCAATGCTA atc tag gtg tgc gtt gat gta at	
*Sc*FAS_ ACPLICFor	TACTTCCAATCCAATGCC gcc act cag gaa ga	pDHS9758
*Sc*FAS_ ACPLICRev	TTATCCACTTCCAATGCTA ggc ctg gtg gtc gag gat gta	
BryRLICFor	TACTTCCAATCCAATGCC agg tat att ggt ata gaa tca at	pDHS279
BryRLICRev	TTATCCACTTCCAATGCTA att gat cca ctg ata ttc tct atg	
MacpC_ G53RRev	ggc atg cgc aac cta agg gcg ctc aaa gtc	

Table 3-2. Primers used for generation of expression plasmids via ligation independent cloning. All sequences are listed 5’ to 3’. Sequences in all capital letters represent the LIC overhangs necessary for insertion into the pMCSG7 and pMCSG9 vectors.

Portions of this chapter have been previously published in:

Tonia J. Buchholz, Christopher M. Rath, Nicole B. Lopanik, Noah P. Gardner, Kristina Håkansson, and David H. Sherman. Polyketide β-Branching in Bryostatin Biosynthesis: Identification of Surrogate Acetyl-ACP Donors for BryR, an HMG-ACP Synthase. Chemistry & Biology 17:1092-1100 (2010).

The authors thank Prof. Christopher M. Thomas for the MacpC expression construct (pGTB340). CMR received funding from the CBI training programs (T32 GM008597) at the University of Michigan. This work was supported by NIH grant GM076477 and the Hans W. Vahlteich Professorship (to DHS). Work in KH's laboratory is supported by an NSF Career Award (CHE-05-47699).

3.5 References

1. Sudek, S.; *et al*. *J Nat Prod* **2007**, *70*, 67.
2. Nelson, T.J.; Alkon, D.L. *Trends in Biochem Sci*, **2009**, *34*, 136.
3. Banarjee, S.; *et al*. *J Nat Prod*, **2008**, *71*, 492-496.
4. Sun, M.-K.; Hongpaisan, J.; Alkon, D.L. *Proc Natl Acad Sci USA 106*, **2009**, 14676.
5. Khan, T.K.; *et al*. *Neurobiol Dis* **2009**, *34*, 332.
6. Nelson, T.J.; Cui, C.; Luo, Y.; Alkon, D.L. *J Biol Chem*, **2009**, *284*, 34514.

7. Singh, R.; Sharma, M.; Joshi, P.; Rawat, D.S. *Anticancer Agents Med Chem,* **2008**, *8,* 603.

8. Staunton, J.; Weissman, K.J. *Nat Prod Rep*, **2001**, *18*, 380.

9. Weissman, K.J. *Meth Enzymol*, **2008**, *459*, 3.

10. Hertweck, C. *Angew Chem Int Ed,* **2009**, *48,* 4688.

11. Smith, S.; Tsai, S.C. *Nat Prod Rep,* **2007**, *24*, 1041.

12. Fischbach, M.A.; Walsh, C.T. *Chem Rev*, **2006**, *106*, 3468.

13. Calderone, C.T. *Nat Prod Rep,* **2008**, *25*, 845.

14. Geders, T.W.; *et al. J Biol Chem,* **2007**, *282*, 35954.

15. Gu, L.; *et al.* J *Am Chem Soc,* **2006,** *128,* 9014.

16. Simunovic, V.; Müller, R. *Chembiochem*, **2007**, *8*, 1273.

17. Simunovic, V.; Müller, R. *Chembiochem*, **2007**, *8*, 497.

18. Calderone, C.T.; *et al*. *Proc Natl Acad Sci USA*, **2006** *103*, 8977.

19. Wu, J.; *et al. Chem Commun*, **2007**, *20*, 2040.

20. Piel, J.; *et al. Proc Natl Acad Sci USA,* **2004** *101*, 16222.

21. Pulsawat, N.; Kitani, S.; Nihira, T. *Gene,* **2007**, *393*, 31.

22. Calderone, C.T.; *et al*. *Chem Biol*, **2007**, *14*, 835-846.

23. Liu, T.; Huang, Y.; Shen, B. *J Am Chem Soc*, **2009**, *131*, 6900.

24. Gu, L.; *et. al. Nature*, **2009**, *459*, 731.

25. Kusebauch, B.; *et al. Angew Chem Int Ed,* **2009**, *48,* 5001.

26. Chen, X.H.; *et al. J Bacteriol,* **2006**, *188*, 4024.

27. Fisch, K.M.; *et al. Nature Chem Biol,* **2009**, *5*, 450.

28. Lambalot, R.H.; *et al. Chem, Biol 3*, **1996**, 923.

29. Sanchez, C.; *et al*. *Chem Biol.* **2001**, *8*, 725-738.

30. Lange, B.M.; Rugan, T.; Martin, W.; Croteau, R. *Proc Natl Acad Sci USA*, **2000**. *97*, 13172.

31. Steussy, C.N.; *et al. Biochem*, **2006**, *45*, 14407.

32. Steussy, C.N.; *et al. Biochem*, **2005**, *44*, 14256.

33. Theisen, M.J.; *et al. Proc Natl Acad Sci USA*, **2004**, *101*, 16442.

34. Dorrestein, P.C.; *et al*. *Biochem*, **2006**, *45*, 1537.

35. Dorrestein, P.C.; *et al. Biochem*, **2006,** *45*, 12756.

36. Arthur, C.J.; *et al. ACS Chem Biol*, **2009**, *4*, 625.

37. Lopanik, N.B.; *et al. Chem Biol*, **2008**, *15*, 1175.

38. Wu, K.; Chung, L.; Revill, W.P.; Katz, L.; Reeves, C.D. *Gene*, **2000**, *251*, 81.

39. Clamp, M.; Cuff, J.; Searle, S.M.; Barton, G.J. *Bioinform*, **2004**, *20*, 426.

40. Alekseyev, V.Y; *et al*. Protein Sci, **2007**, *16*, 2093.

41. Stols, L.; *et al. Protein Expr Purif*, **2002**, *25*, 8.

Chapter 4

Chemoenzymatic Synthesis of Cryptophycin Anticancer Agents: Non-Amino Acid Incorporation Mediated by a NRPS Module

4.1 Introduction

Natural products have been widely applied to fight disease and offer chemical scaffolds for development of new analogs with improved/altered functions, achieved through semi-, total-, or chemoenzymatic synthesis efforts.[1-3] Cryptophycins are potent anticancer agents at picomolar concentrations and exert their cytotoxicities in both vinca alkaloid- and taxol-resistant cancer cells that contribute to the proliferation of drug-resistant tumors.[4] Their clinical potential as well as synthetic challenges have stimulated the development of alternative strategies to provide suitable amounts of material and new analogs with improved physiochemical properties for clinical studies. The cryptophycin gene cluster was recently elucidated and offers unique opportunities for assembly of the drug and new analogs using chemoenzymatic approaches.[5] The gene cluster is comprised of two type I polyketide synthase (PKS) genes, *crpA* and *crpB*, two nonribosomal peptide synthetase (NRPS) genes, *crpC* and *crpD*, and four tailoring enzyme genes including a key P450 epoxidase gene (*crpE*). Previous studies from this laboratory have demonstrated the feasibility and efficiency of biocatalysts from this gene cluster to properly macrocyclize and regio- and stereo-specifically epoxidize cryptophycin intermediates in generating the natural products and novel analogs.[5-7]

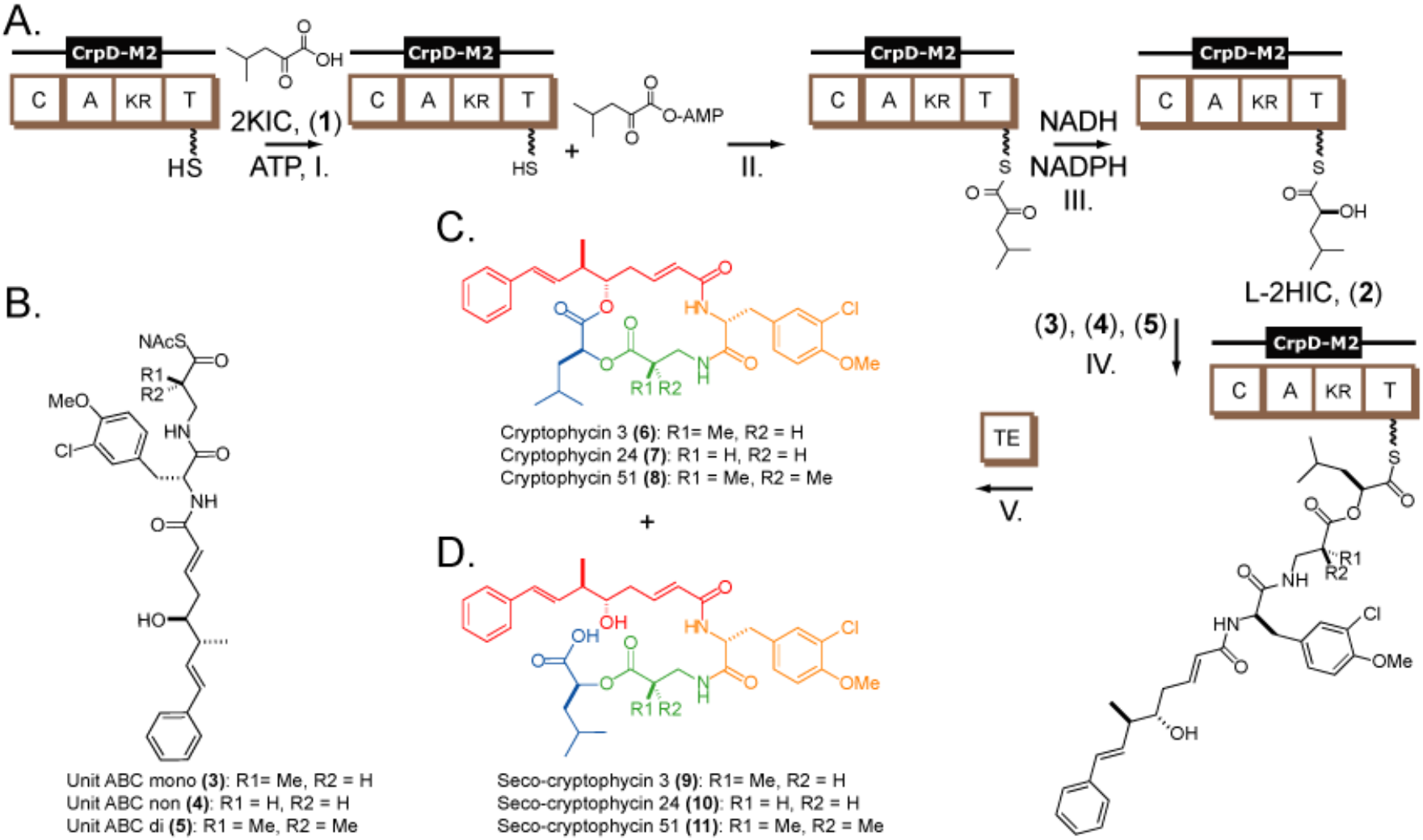


Figure 4-1. CrpD-M2 biosynthetic scheme. (**A**) Four sequential reactions catalyzed by CrpD-M2 and Crp TE mediated cyclization. Enzymatic domains within the CrpD-M2 polypeptide are noted with squares, and the phosphopantetheinyl arm is denoted by the linked SH group. (**B**) Three SNAC-ABC analogs used in this study (**3-5**).[7] (**C**) Cyclic cryptophycin products (**6-8**),[8-10] generated by CrpD-M2 and CrpTE.[7] (**D**) Linear seco-cryptophycin intermediates (**9-11**) produced through incomplete cyclization of (**6-8**). Units A, B, C, and D are labeled in red, orange, green, and blue, respectively. C: condensation domain, A: adenylation domain, KR: ketoreductase domain, T: thiolation domain, TE: thioesterase domain.

CrpD is a bimodular NRPS. Bioinformatic and a chemical feeding experiments suggest that the substrate of its first module is methyl-β-alanine converted from L-aspartic acid by CrpG, a β-methylaspartate-α-decarboxylase.[5,11] Chemical feeding experiment revealed that 2-ketoisocaproic acid (2KIC, (**1**)) instead of L-2-hydroxyisocaproic acid (L-2HIC, (**2**)) was the substrate of CrpD module 2 (CrpD-M2) and was able to be incorporated into cryptophycin as unit D (**Figure 4-1**).[5] Moreover, several natural cryptophycin analogs contain unit D variations as 3-methyl-2-hydroxyvalerate, 2-hydroxyvalerate, and 3-methyl-2-hydroxybutyrate (**Figure 4-2**).[5] Altered bioactivity of these analogs suggest the importance of this unit in cryptophycin anticancer action, but only limited synthetic efforts have been made to generate analogs carrying unnatural unit D structures.[12]

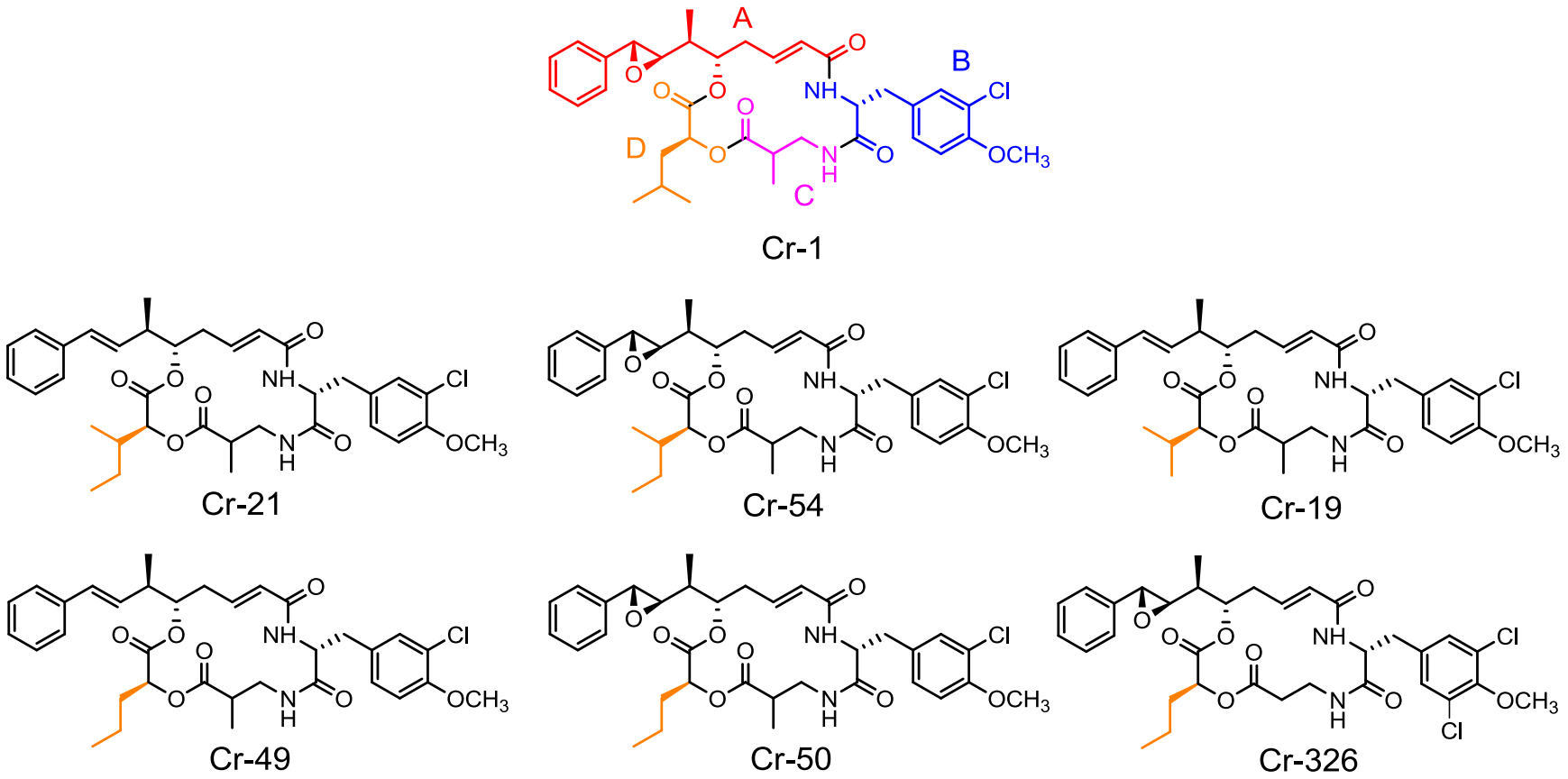


Figure 4-2. Chemical structures of natural cryptophycin analogs. Cryptopycin-1 is comprised of four units linked in clockwise order of δ-hydroxyoctenoic acid (unit A), 3-chloro *O*-methyl-D-tyrosine (unit B), methyl-β-alanine (unit C), and L-leucic acid (unit D). Several cryptophycin analogs carrying variable unit D moieties are also shown.

The cryptophycin biosynthetic catalyzed by CrpD-M2, is proposed to consist of four distinct steps (**Figure 4-1**). In step I, the free-acid extender unit (such as **1**) is activated by CrpD-M2 adenylation (A) domain to form the corresponding acyl-AMP. This intermediate is then loaded onto the thiolation (T) domain active site bound phosphopantetheine through a transthioesterification reaction to form the acyl enzyme intermediate in step II. In the presence of reducing equivalent the 2KIC enzyme intermediate is reduced stereoselectively to L-2HIC in step III by the unique 2-ketoreductase (KR) domain based upon analysis of known products. The nascent L-2-hydoxy group in unit D then accepts cryptophycin unit ABC biosynthetic intermediate transferred from CrpD module 1 T domain through formation of an atypical (in an NRPS module) ester linkage in step IV.

In this report, four sequential steps were biochemically validated to investigate the unique incorporation of 2-hydroxy acid in natural products, and to probe the intrinsic substrate flexibility and synthetic potential of CrpD-M2. Chemoenzymatic synthesis of three cryptophycins 3,[8] 24,[9] and 51[10] (**6-8**) with CrpD-M2 and CrpTE serves as a proof-of-principle for further efforts to generate cryptophycin analogs with unnatural structures of unit C and unit D through combined synthetic and biochemical methods.

4.2 Results

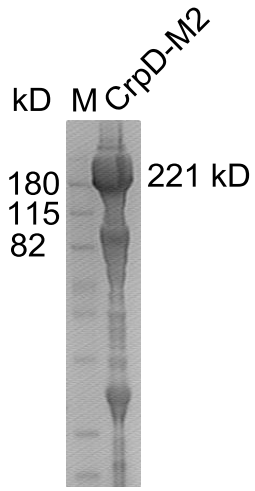


Figure 4-3. 4-12 % SDS-PAGE analysis of N-terminally His-tagged CrpD-M2 after Ni-NTA resin.

A CrpD-M2 expression construct was generated by amplifying a DNA fragment consisting of C, A, KR, and T domains by PCR, and cloning into the BamHI and XhoI sites of pET28a. This construct was heterologously overexpressed in *E. coli* BAP1 strain with for production of phosphopantetheinyled proteins.[13] The N-terminally His-tagged protein was purified with Ni-NTA resin to ~ 80 % purity (**Figure 4-3**). The integrity of the purified protein was verified by peptide map fingerprinting and FTICR-MS (**Figure 4-4A**, **Table 4-1**). The CrpD-M2 T domain active site was also identified and proper phosphopantetheinylation on the T domain active site was verified by the presence of a characteristic MS2 fragment (**Figure 4-3B**, **Table 4-2**).[14]

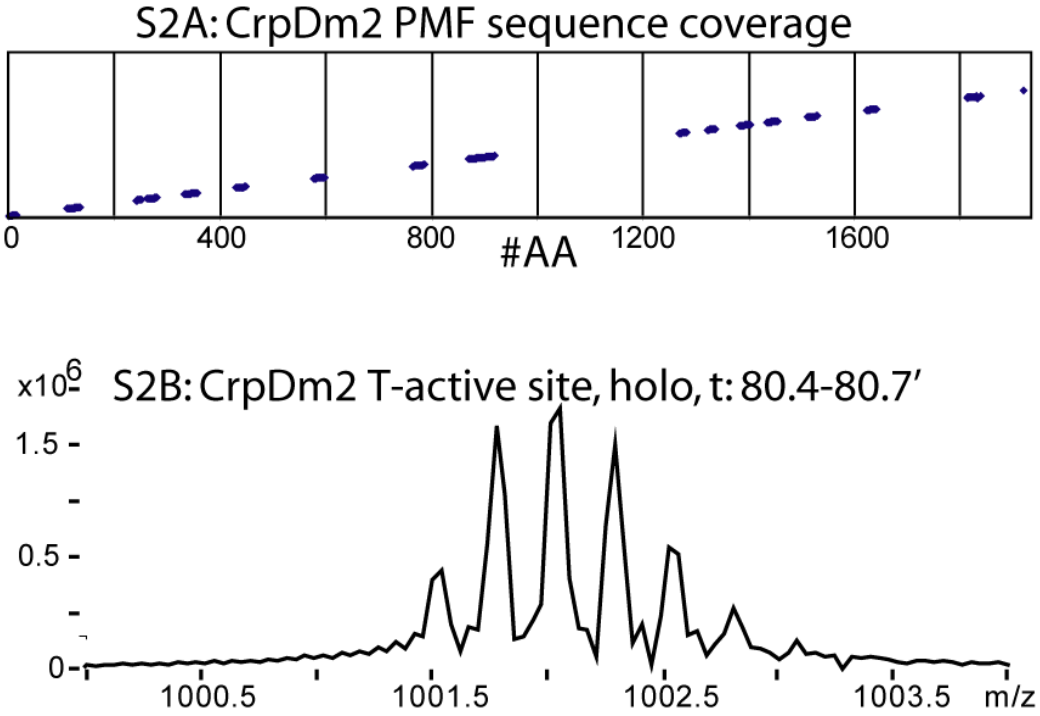


Figure 4-4. CrpD-M2 characterization. (**A**) Sequence coverage of purified CrpD-M2 by amino acid number as determined by direct inject FTICR-MS peptide fingerprint mapping. (**B**) Identification of CrpD-M2 T domain active site with LC FTICR-MS as m/z versus absolute intensity.

#	Observed	Expected	dPPM	Start	Stop	Sequence
1	1,493.82	1,493.82	1	3	16	TTNSALSLPPIQPR
2	2,760.52	2,760.48	14	112	136	ESVLHQQAQLAAITPFDLETAPLIR
3	1,410.72	1,410.70	14	242	254	GTTQSFSLNTDLK
4	2,227.18	2,227.15	15	263	282	NSGTTLFMTLHAAFATLLYR
5	2,983.41	2,983.46	-16	335	359	ETTLEAYEHQDVPFEQVVEVLQPQR
6	2,158.00	2,158.03	-13	432	450	MTAHFQNLCSAIVENPQQK
7	2,522.25	2,522.26	-4	578	599	IEQALQTAKGVEDCYVMVRNQK
8	2,784.30	2,784.32	-8	765	789	GITYINSDGSEQVQSYAQLLEDAQR
9	2,549.28	2,549.28	2	872	892	KWSQNNLNDDNFKLETIESLQK
10	3,216.50	3,216.46	11	893	921	FSTDKDKDYYNAQPEDLALFMLTSGSTGMSK
11	1,939.99	1,940.03	-20	1269	1284	SLLKQRFECGEFKSLR
12	2,034.06	2,034.02	19	1323	1339	TLTLIFTDNLGWQQDNR
13	2,571.28	2,571.27	2	1384	1404	QNSQVISQILHLWNYNEQTEK
14	2,399.29	2,399.30	-4	1436	1456	QQAVKLLWIANQSQLVHPTDK
15	1,862.01	1,862.01	0	1441	1456	LLWIANQSQLVHPTDK
16	2,205.17	2,205.12	21	1513	1531	NRERFVSGLEPVDMTAKEK
17	1,392.67	1,392.70	-21	1517	1529	FVSGLEPVDMTAK
18	2,226.15	2,226.14	4	1625	1644	TQLDGVFHMAGIIQETPIEK
19	2,354.28	2,354.28	1	1815	1834	FGIPNQINFVQLEQIPLTQR
20	2,075.06	2,075.07	-1	1840	1858	EQIAAIYGGLNTSEQTKPR
21	1,725.96	1,725.94	9	1908	1922	KNLPLATLFQNPTIER

Table 4-1. CrpD-M2 peptides identified by accurate mass peptide mass fingerprinting using direct injection FTICR-MS. All masses given are monoisotopic and deconvoluted in Da. Mass error in ppm, start and stop sites, as well as assigned sequence are all provided.

#	m/z mon	z	Obsd.	Theo.	dPPM	ID
1	1,248.3	3	3,741.98	3,741.97	2	Apo+80
2	1,215.7	3	3,644.04	3,643.97	19	Apo-18
3	1,335.0	3	4,001.94	4,002.06	-29	Holo 3^+
4	1,001.5	4	4,002.00	4,002.06	-16	Holo 4^+
5	261.1	1	260.12	260.12	9	$Ppant_1$
6	359.1	1	358.10	358.10	-8	$Ppant_2$
7	453.2	1	452.23	452.23	0	b_4-NH3
8	566.3	1	565.32	565.32	4	b_5-NH3
9	324.2	1	323.19	323.19	-15	b_3-NH3
10	1,409.8	1	1,408.78	1,408.76	12	b_{12}-NH3
11	788.1	3	2,361.28	2,361.24	18	b_{14}-NH3

Table 4-2. LC FTICR IRMPD MS/MS verification of CrpD-M2 T domain active site. The observed monoisotopic molecular weight, charge state, deconvoluted monoisotopic mass, and theoretical mass are all provided. Mass errors in ppm and peak identity are also provided.

A Domain	235	236	239	278	299	301	322	330	331	517	Specificity
CrpD-M2-A	V	A	I	F	L	G	S	S	G	K	2-KIC/2HIC
PksJ-A1	V	G	W	T	T	A	A	I	C	K	2-KIC
BarE-A	V	G	I	L	V	G	G	T	S	K	Trichloro-2-KIC
BSLS-A1	G	A	L	M	V	V	G	S	I	K	D-2HIV
BEAS-A1	G	A	L	M	I	V	G	S	I	K	D-2HIV
CseA-A1	V	G	V	W	V	G	T	S	G	K	2-KIC
CseB-A1	V	G	F	W	V	A	V	S	D	K	2-KIV
ENSYN-A1	G	A	L	H	V	V	G	I	C	K	D-2HIV
HctE-A1	V	G	V	W	L	A	L	F	C	K	2KIV
HctF-A1	V	G	V	W	L	A	L	F	C	K	2KIV
KtzG-A	V	T	Y	F	N	G	P	S	G	K	2-KIV
Vlm1-A1	A	A	L	W	I	A	V	S	G	K	2-KIV
Vlm2-A1	V	V	I	W	I	A	E	N	M	K	Pyruvate
BarD	**D**	A	I	L	L	G	G	A	A	K	L-Leucine

Table 4-3. CrpD-M2 A-domain predicted specificity.

Bioinformatics can be used to predict NRPS A domain specificity based upon binding pocket residue motifs. The conserved Asp235 involved in ionic interaction with the amino group of the substrate amino acid is replaced by Val235 in CrpD-M2 A domain (**Table 4-3**). Similar to unit D of cryptophycin, 2-hydroxy acid moiety is also found in nine other natural products including bacillaene,[15] barbamide,[16] bassianolide,[17] beauvericin,[18] cereulide,[19] enniatin,[20] hectochlorin,[21] kutzneride,[22] and valinomycin.[19] A similar replacement of Asp235 is conserved across all A domains responsible for incorporation of 2-hydroxy acid into these natural products (**Table 4-3**), predicting that the CrpD-M2 A domain prefers non α-amino acid substrates (ex. 2-hydroxy- or 2-keto-acids).

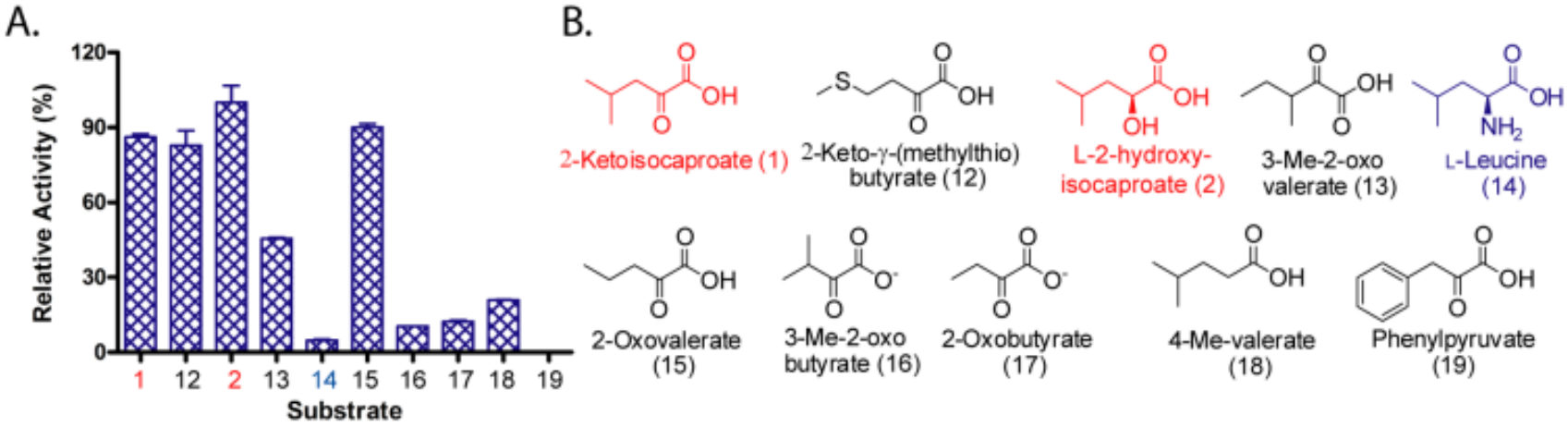


Figure 4-5. CrpD-M2 A-domain substrate specificity. (A) A-domain relative activity normalized to 2. The experiments were performed in duplicate. **(B) E**xtender units investigated in this assay.

The well-established ATP-PP_i exchange assay was then used to biochemically determine the substrate specificity of the CrpD-M2 A domain with ten acyl-acid substrates (**Figure 4-5**). CrpD-M2 activated 2-KIC (**1**) about 20 times more than its cognate amino acid l-leucine (**14**), consistent with bioinformatic prediction and previous feeding experiments.[5] L-2HIC (**2**), which was not incorporated into cryptophycin final structure in feeding experiment,[5]] was the best substrate in the assay. A similar level of selection for the natural substrate 2-oxovalerate (**15**) was observed. CrpD-M2 specificity to two other natural unit D fragments, 3-methyl-2-oxovalerate (**13**) and 3-methyl-2-oxobutyrate (**16**), was decreased about 50% and 90%. This result along with the observed weak activation of unnatural substrates 2-oxobutyrate (**17**) and phenyl pyruvate (**19**) suggests that the size and bulk of substrate side chains are important in CrpD-M2 A domain recognition. The linear substrate 2-keto-γ-(methylthio) butyrate (AKGB, **12**) was effectively activated. Activation of unnatural substrate AKGB demonstrates for the synthetic potential of native CrpD-M2 in producing novel cryptophycin analogs with altered unit D moiety. The weak activation of 4-methyl-valerate (**18**) by CrpD-M2 A domain reveals the importance of the α-position functional group recognition. CrpD-M2 A domain has relatively relaxed substrate specificity and exhibits a similar selectivity toward 2-keto and 2-hydroxy acids in step I shown in **Figure 4-1**. ATP-PP_i exchange assays have been previously applied to examine substrate preference of A domains in the biosynthesis of bacillaene,[15] barbamide,[16] cereulide,[19] enniatin,[20] hectochlorin,[21] kutzneride.[22] Only the cyanobacterial natural product hectochlorin displayed a similar selectivity toward both 2-keto and 2-hydroxy acid to CrpD-M2 A domain.[21]

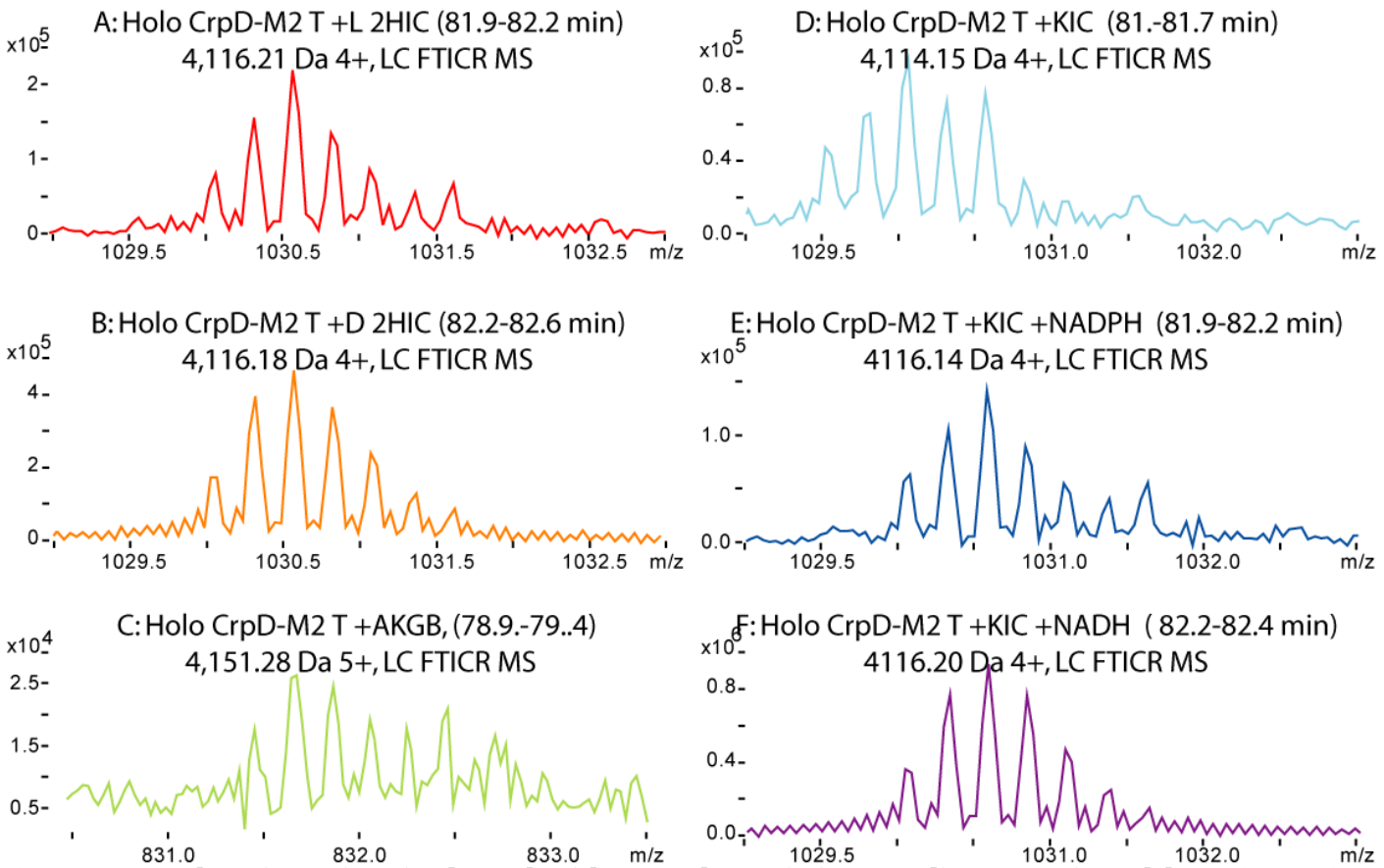


Figure 4-6. CrpD-M2 T domain active-site bound with extender unit intermediates monitored by LC FTICR-MS. A zoomed, mass spectrum averaged over the elution window is presented for each of the species observed. Intensity is presented in absolute signal. The active-site peptide bearing extender-unit loaded onto the phosphopantetheine arm, deconvoluted monoisotopic mass, and observed charge state are shown. The reactions of CrpD-M2 contained ATP and: (**A**) L-2HIC (2), (**B**) D-2HIC (20), (**C**) AKGB (12), (**D**) 2KIC (1), (**E**) 2KIC +NADPH, (**F**) 2KIC +NADH. Further LC FTICR-MS data (**Table 4-4**) and LC IT-MS/MS (**Table 4-5**)[23] confirmation are provided. Observed species were not present in no substrate control reactions.

FTICR-MS (**Figure 4-6**) was utilized to monitor substrate loading directly on the T-domain of CrpD-M2 (**Figure 4-1,** step II).[24-26] Enzyme reactions were terminated by proteolysis with trypsin, and the active-site peptide bound with extender units were separated and analyzed by LC-FTICR-MS and LC iontrap-MS/MS (LC-IT-MS/MS, **Figure 4-6**, **Tables 4-4 and 4-5**). As shown in **Figure 4-4A** and **4-4D**, T domain active site peptides bound with L-2HIC and 2-KIC showed masses of 4116.21 and 4114.15, respectively, at a charge state of 4+, are almost identical to theoretical values (**Table 4-4**). The substrate flexibility of CrpD-M2 A domain, with potential applications for combinatorial biosynthesis, is displayed by the loading of the unnatural substrate AKGB (**Figure 4-4C**). This is in agreement with the high degree of activity towards this substrate in the ATP-PP_i exchange assay (**Figure 4-5**). D-2HIC (**20**) was also loaded on CrpD-M2 T domain active site with a similar efficiency to L-2HIC as shown by the observed mass of 4116.18 and ion intensity (**Figure 4-4B**). Since only L-2HIC containing cryptophycin analogs have been isolated and characterized from cyanobacterium *Nostoc* sp., we suspect that other factors than A domain selectivity, such as substrate availability and/or downstream processing, determine the final outcome. It is well-known that 2-keto acids are indispensible intermediates in amino acid biosynthesis such as 2KIC, 3-methyl-2-oxovalerate (**13**) and 3-methyl-2-oxobutyrate (**16**) in the biosynthesis of leucine, isoleucine, and valine, respectively. The availability of free 2-hydroxy acid may ascribe to a pathway-specific enzyme. For example, A domains for the biosynthesis of bassianolide,[17] beauvericin,[18] enniatin,[20] are specific to D-2-hydroxy isovalerate (D-2HIV) and a pathway-specific NADPH-dependent reductase is found in their biosynthetic pathway to stereo-specifically reduce 2-keto isovalerate (2-KIV).[17,18,27] Since such a

reducing enzyme gene is not present in the cryptophycin gene cluster,[5] it is possible that 2KIC is the available native substrate of CrpD-M2 A domain.

#	Rxn	m obs	m theo	dppm	ID	Max Int.	FTICR-t
A	L-2HIC	4,116.21	4,116.12	22	2HIC	2.1E+05	82
B	D-2HIC	4,116.18	4,116.12	15	2HIC	4.3E+05	82
C	AKGB	4,151.28	4,151.16	30	AKGB	2.8E+04	79
D	KIC	4,114.15	4,114.12	9	KIC	8.8E+04	81
E	KIC+NADPH	4,116.14	4,116.12	5	2HIC	1.5E+05	82
F	KIC+NADH	4,116.20	4,116.12	20	2HIC	7.6E+05	82

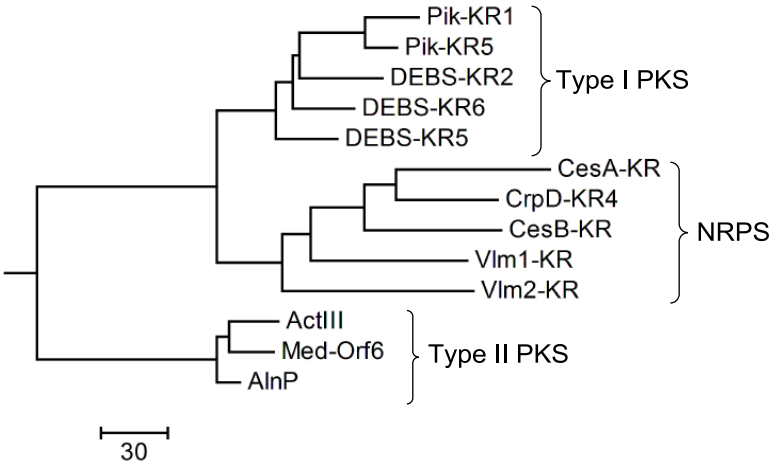
Table 4-4. CrpD-M2 active site bound intermediates identified by accurate mass using LC FTICR-MS. All masses given are monoisotopic, in Da. Mass error is provided in ppm as are peak IDs. Intensity is the average of peak intensity of the most abundant charge state averaged over the three most intense scans. Retention times are given in minutes, note that due to instrument configuration (dead-volume, actual performance, scan times) LC FTICR and LC LIT–MS retention times differ by approximately 8 minutes, although the order of eluted peaks and relative elution time in relation to total ion chromatogram is maintained.

#	Rxn	MSn	MSn t	MSn ID	MSn hits
A	L-2HIC	1030	74	2HIC	2HIC-Ppant$_2$, Apo-18
		1030->1215	74		Apo-18 -> (a$_{10}$, a$_{13}$-NH$_3$, a$_9$-NH$_3$, b$_{10}$-H$_2$O, b$_{13}$ -NH$_3$, b$_{14}$-H$_2$O,b$_{14}$-NH$_3$, b$_6$, b$_7$, b$_9$, b$_9$-NH$_3$, y$_{15}$-NH$_3$, y$_9$-H$_2$O)
B	D-2HIC	1030	74	2HIC	2HIC-Ppant$_1$, Apo-18, Apo+80
		1030->1215	75		Apo-18 -> (a$_{11}$-NH$_3$, a$_{16}$, a$_{16}$-NH$_3$, b$_{11}$-H$_2$O, b$_{11}$-H2O, b$_{12}$, b$_{12}$-NH$_3$, b$_8$-NH$_3$, y$_{13}$-NH$_3$, y$_{13}$-NH$_3$, y$_{15}$, y$_{16}$, y$_9$, y$_9$-H$_2$O)
		1373->1823	75		Apo-18 -> (a$_{12}$-NH$_3$, a$_{16}$-NH$_3$, b$_{12}$-H$_2$O, b$_{12}$-NH$_3$, b$_{14}$-H$_2$O, b$_{17}$-H$_2$O, b$_8$-H$_2$O, b$_8$-NH$_3$, y$_{10}$-NH$_3$, y$_{14}$, y$_{15}$, y$_{16}$, y$_{18}$)
C	AKGB	1385	70	AKGB	Apo-18, Apo+80
D	KIC	1030	74	KIC	Apo-18, Apo+80
E	KIC+NADPH	1030	75	2HIC	2HIC-Ppant$_1$
		1030->1373	75		Apo-18 -> (a$_{10}$, b$_{10}$-NH$_3$, b$_{10}$-NH$_3$, y$_{10}$-H$_2$O, y$_{13}$-H$_2$O, y$_{13}$-NH$_3$, y$_{14}$-NH$_3$, y$_{16}$, y$_7$)
		1373	75		Apo-18, Apo+80
F	KIC+NADH	1030	74	2HIC	2HIC-Ppant$_1$, Apo-18, Apo+80
		1373	75		Apo-18, Apo+80
		1373->1823	75		Apo -18 -> (a$_9$-NH$_3$, b$_{11}$H$_2$O, b$_{12}$H$_2$O, b$_{12}$-NH$_3$, b$_{15}$, b$_{16}$-H$_2$O, b$_8$-NH$_3$, b$_9$-H$_2$O)

Table 4-5. CrpD-M2 peptides identified by MS^2 and MS^3 LC LIT-MS. All masses given are average, in Da. Mass error of ±300 ppm was used for peak assignment. Retention times are given in minutes, note that due to instrument configuration (dead-volume, actual performance, scan times) LC-FTICR and LC-LIT –MS retention times differ by approximately 8 minutes, although the order of eluted peaks and relative elution time in relation to total ion chromatogram is maintained.[23]

As shown in the **Figure 4-1,** step III, the loaded 2KIC is proposed to be reduced into L-2HIC by the α-KR domain of CrpD-M2. This type of KR domain is also embedded in NRPS modules of cereulide,[19] hectochlorin,[21] kutzneride,[22] and valinomycin,[19] and its α-keto reduction activity and pure stereochemistry in its product were biochemically validated in the cereulide system.[19] Bioinformatic analysis indicates that this KR domain in CrpD-M2 is grouped with these NRPS KR domains, and is phylogenetically distinct from any type of PKS β-KR domains (**Figure 4-7A**). The stereochemistry outcome of PKS β-KR domains can be predicted based on conserved motifs.[28-30] A similar analysis was performed to predict 2-hydroxy chirality reduced by KR domains from CrpD-M2, CesA and CesB (**Figure 4-7B**). This domain in CesA and CesB produces L-2HIC and D-2HIV, respectively. However, none of these enzymes contains the conserved motifs necessary to group them into either type A or type B KR domains. Given phylogenetic distance and position difference of ketone group in their substrates between PKS β-KR domains and NRPS α-KR domains, this result is not unexpected. The first KR domain in the hybrid PK/NRPS PksJ involved in bacillaene biosynthesis catalyzes both β- and α-ketone reduction with 10-fold preference to the former reaction.[15] Its β-ketone reduction was determined to be A-type outcome, but similar to NRPS α-KR domains, this enzyme cannot be grouped in either type (**Figure 4-7B**). Therefore, bioinformatics analysis indicates that NRPS α-KR domains are a new group of ketoreductases and their product chirality is still not predictable (**Figure 4-8**). In CrpD-M2 KR reaction, the L-stereocenter is expected since all natural cryptophycins only contain this chirality.

A.



B.

```
Region 88-103
HMAGIIQETPIEKETP CrpDm2_KR
HAAATLDDGTVDTLTG DEBSKR1_B_type
HTAGALDDGIVDTLTA PikKR1_B_type
HAAGLPQQVAINDMDE DEBSKR2_A
HAAGVSTSTPLDDLTE DEBSKR5_A
HTAGVPESRPLHEIGE DEBSKR6_A
Region 134-149
FCSVNGFFGGTNVAAY CrpDm2_KR
FSSFASAFGAPGLGGY DEBSKR1_B_type
FSSVSSTLGIPGQGNY PikKR1_B_type
FSSGAGVWGSARQGAY DEBSKR2_A
FSSNAGVWGSPGLASY DEBSKR5_A
FSSGAGVWGSANLGAY DEBSKR6_A
```

Figure 4-7. (A) Phylogenetic analysis of CrpD-M2 KR domain and (B) multiple alignments showing specificity determining regions for PKS KR domains. CrpD-M2 KR domain was grouped with other NRPS KR domains, which was fallen into one different category with others from both type I and type II PKS KR domains. The tree scale was neighboring joining identity percentage. Alignment of sequences and the phylogenetic tree were constructed using MEGA 4.0.[31] Based on sequence analysis of multiple biosynthetic pathways and product analysis, guidelines were established for determing PKS-KR domain stereospecificity.[29] This findings were later verified with in vitro biochemistry for a series of free PKS enzyme domains with model substrates.[32] When this analysis is applied to CrpD-M2 we can determine that it is neither a A-type nor B-type PKS KR based upon its (unsurprising) ability to meet the criteria for either. For example in region 88-103 it does not contain the LDD motif typical of group B, with a strictly conserved L93, D94 or E94 and an invariant D95. In region 134-149 it does not contain the B-group P144 and N148, nor does it contain the A-group W141 (unless M142). Thus this NRPS KR domain appears to be in a separate class based upon primary amino acid sequence as well as product analysis (**Figure 4-8)**.

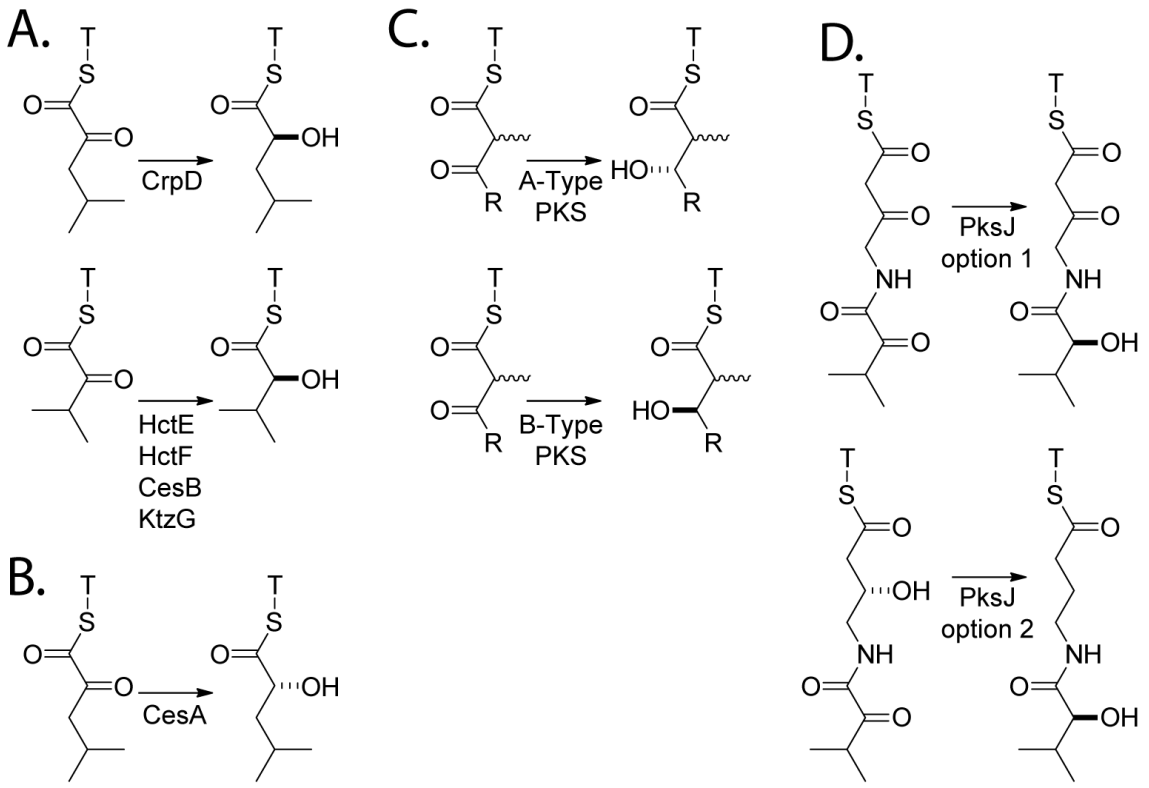


Figure 4-8. Known ketoreductase catalyzed reactions for T-domain bound substrates in NRP natural products.

After loading of 2KIC (**Figure 4-6D**), addition of reducing equivalent to CrpD-M2 reaction induced a mass shift as observed by FTICR-MS (**Figure 4-6E and 4-6F**). The increase of 0.5 m/z units (representing a 2-Da shift in the deconvoluted mass) is consistent with 2KIC conversion to 2HIC as the product of the α-ketoreduction reaction. Both NADH and NADPH appeared to operate within a similar (1-2 fold) efficiency as hydride donors based on peak intensity (**Figure 4-6E and 4-6F**). Further characterization of α-KR domain of CrpD-M2 (ex. crystallography) may contribute to our model for stereochemistry control in this enzyme subclass.

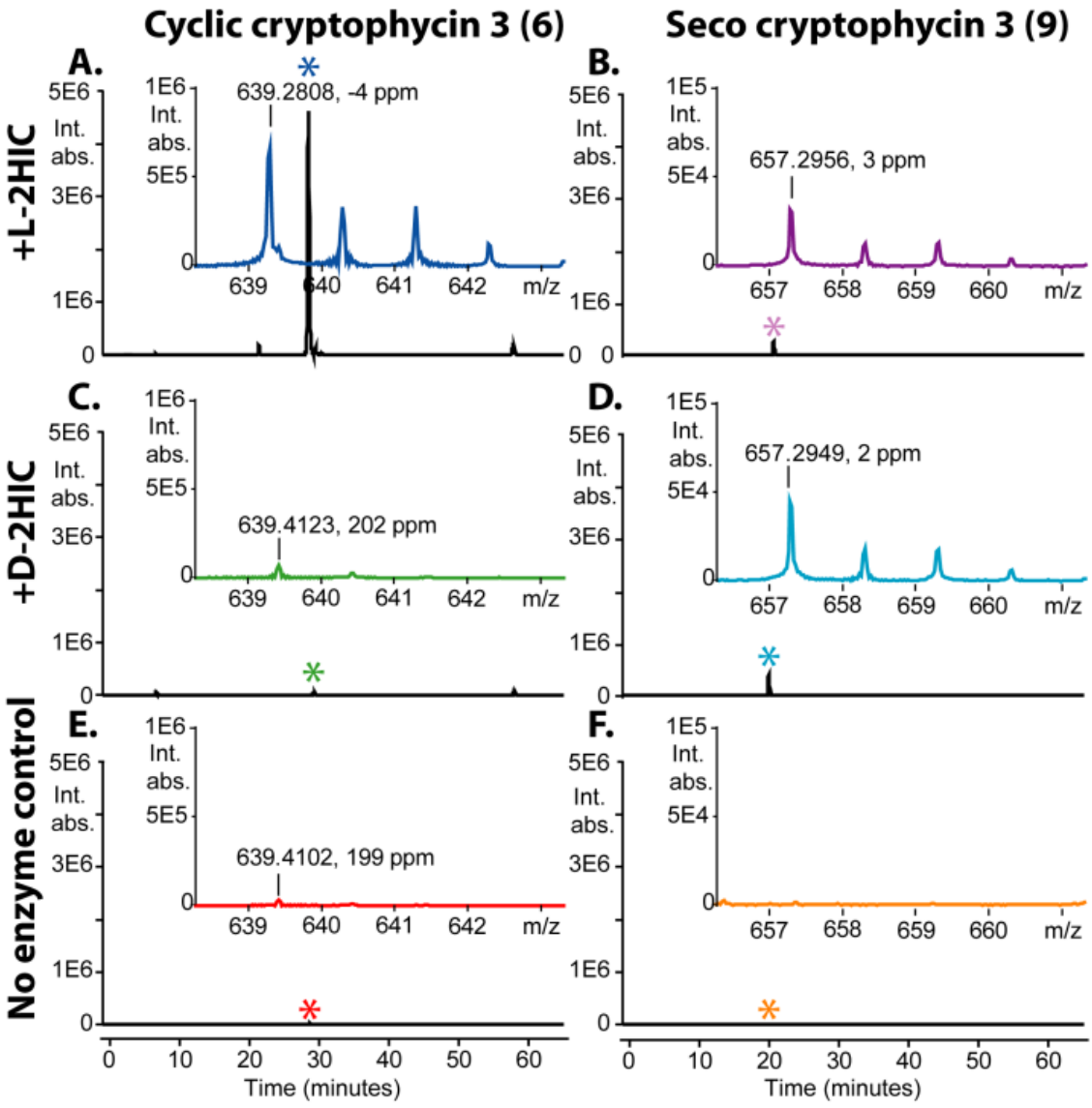


Figure 4-9. FTICR MS analysis of cryptophycin products from the reaction of unit C monomethyl chain elongation intermediate (3) with L/D-2HIC, ATP, CrpD-M2 and Crp TE. Extracted ion chromatograms are presented at ±15 ppm as time versus absolute signal. Inset mass spectra are time averaged over the 1 min elution window corresponding to the asterisk in the extracted ion chromatogram. Inset mass spectra are presented as m/z versus absolute signal. Monoisotopic mass $[MH]^+$ and the experimental mass error in ppm are also reported. Reactions with L-2HIC and D-2HIC are also included to monitor the formation of cyclic (**A** and **C**) and linear products (**B** and **D**). No enzyme control reactions monitoring cyclic (**E**) and linear product formation (**F**) are provided.

Next, the ability of CrpD-M2 to form *seco*-cryptophycin intermediate was investigated using synthetic SNAC-ABC chain elongation intermediates (**3-5**) as the starting point (**Figure 4-1B**). The synthetic scheme followed our previously established route[33] and the SNAC-ABC intermediates were confirmed with NMR and high resolution mass spectrometry. The intermediate with monomethylated unit C (3-amino-2(*R*)-methylpropionyl, **3**) was then combined with CrpD-M2 and L-2HIC (**2**). The C domain of CrpD-M2 is proposed to catalyze the formation of an ester bond with 2-hydroxy group of unit D as a nucleophilic acceptor (**Figure 4-1, step IV**). The formation of the ester bond was validated by successfully detecting the reaction products released from the CrpD-M2 T domain after addition of the excised Crp TE domain (**Figure 1, step V, Figure 4-9A and 4-9B**). Both cyclic cryptophycin 3 (**6**) (**Figure 4-9A**) and linear product (**9**) (**Figure 4-9B**) were observed in the extracted ion chromatograms (EIC). Previously, a bidomain NRPS (T-C), Fum14p, was shown to form a C-O bond in the biosynthesis of fungal mycotoxin fumonisin.[34] The only other proven C domain promoting C-O formation is a free-standing enzyme SgcC5 in C-1027 biosynthesis.[35] In both cases, donor substrates are tethered to T domains while the nucleophilic acceptors (-OH) is a small molecule. This study is first example of a C domain in a full NRPS module specifically incorporating non-amino acid moieties as an ester synthase rather than a common amide synthase. Similar to C domains catalyzing the amide bond formation, these ester synthases most likely require both substrates bound to T domains, indicating that Fum14p and SgcC5 may represent different evolutionary legacies.

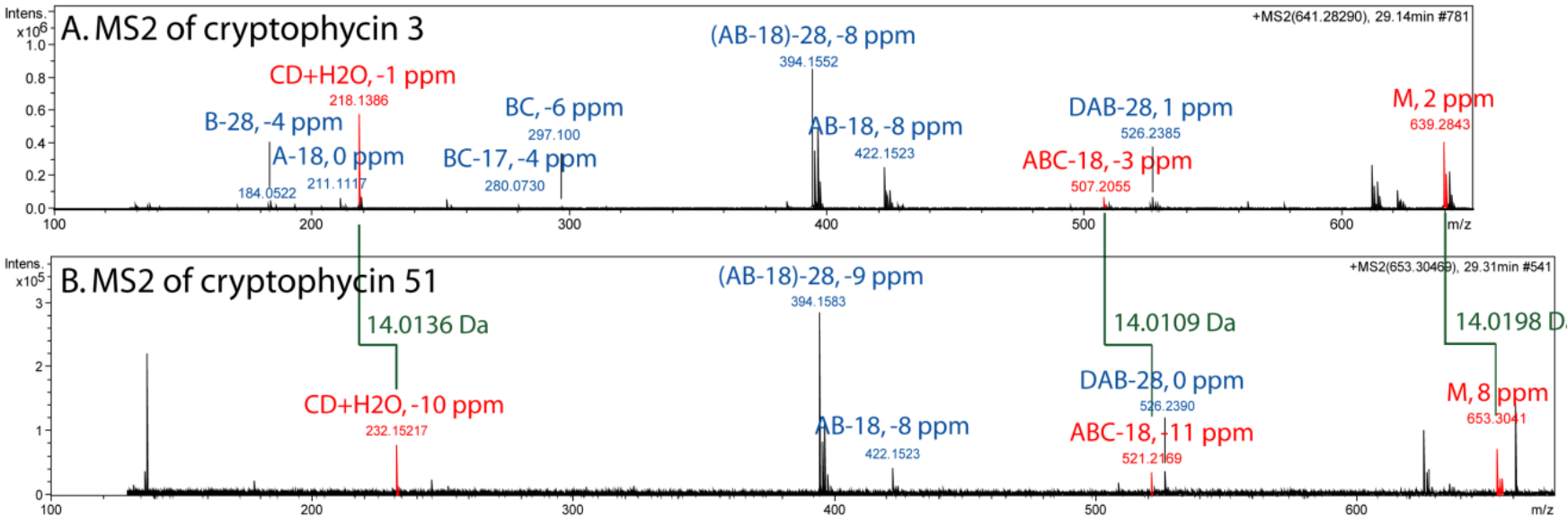


Figure 4-10. LC FTICR-MS/MS spectra of cryptophycins. Data are presented as m/z versus absolute signal. (**A**) cryptophycin 3 CID MS/MS fragmentation spectra. (**B**) cryptophycin 51 CID MS/MS fragmentation spectra. Assigned ions are donated in blue. Key ions are noted in red, with the mass shift in green. Mass errors and observed MH+ values are also provided. All assigned peaks were within 15 ppm of the theoretical value. All unit B containing peaks had the predicted +2 istope increase in abundance due to ^{37}Cl incorperation.

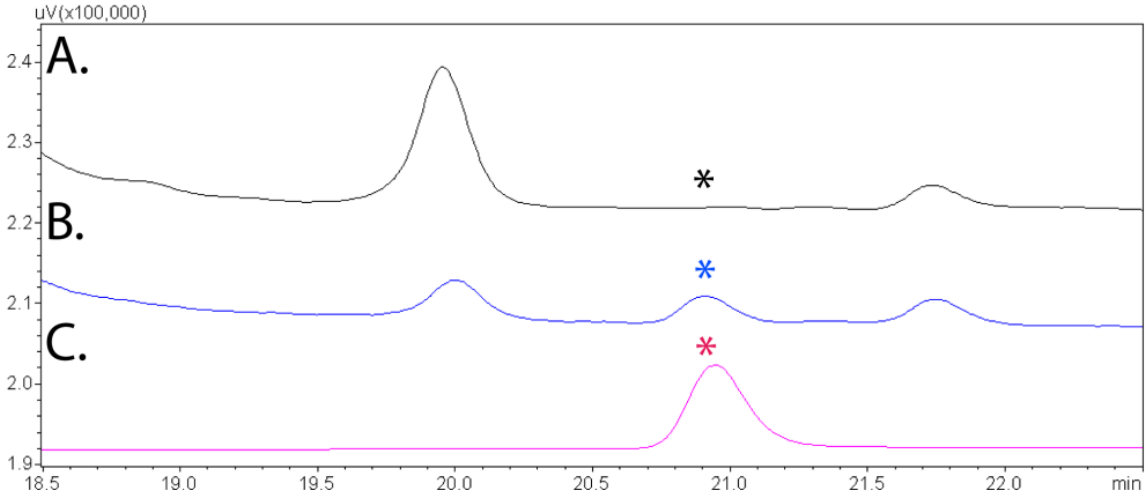


Figure 4-11. Crp 3 co-elution with authentic standard by HPLC. (**A**) A negative control reaction (ATP, L-2HIC extender unit, SNAC-ABC chain elongation intermediate (**3**), and boiled enzymes (CrpD-M2 and Crp TE). (**B**) A full reaction (ATP, L-2HIC extender unit, SNAC-ABC chain elongation intermediate (**3**), CrpD-M2, and Crp TE). (**C**) Authentic standard of cryptophycin 3. Reaction conditions were described above.

The cyclic cryptophycin 3 (**6**) was further characterized based on MS/MS (**Figure 4-10A**) and LC elution with an authentic standard (**Figure 4-11**). Observed isotope patterns of cyclic product and MS^2 fragments were consistent with expected +2 isotope abundance increase due to ^{37}Cl contributions from unit B. Assuming both cyclic and linear products share similar ionization efficiency in the positive ion mode, cyclic cryptophycin 3 (**6**) was dominantly formed over linear cryptophycin 3 (**9**). This result rigorously demonstrates chemoenzymatic synthesis of cryptophycin 3 (**6**) through five catalytic steps (**Figure 1, step I-V**).

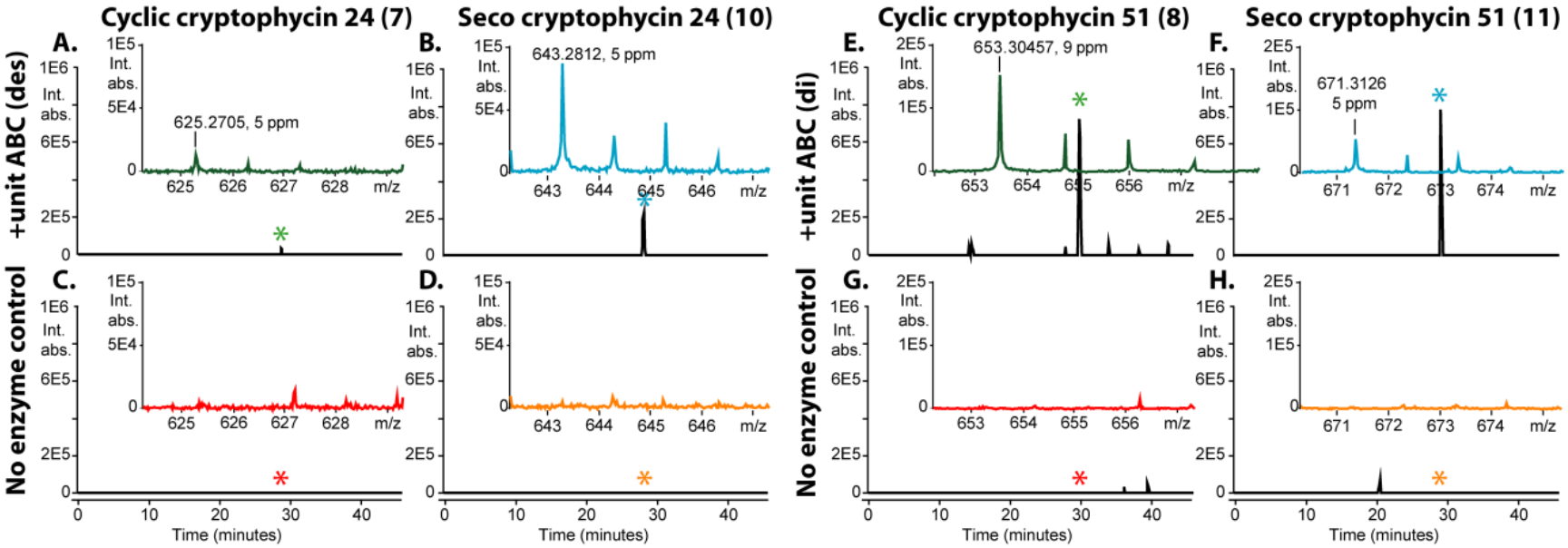


Figure 4-12 A-D. FTICR MS analysis of cryptophycin products from the reaction of unit C 3-amino-propionyl chain elongation intermediate (4) with L-2HIC, ATP, CrpD-M2 and Crp TE. Extracted ion chromatograms are presented at ± 15 ppm as time versus absolute intensity. Inset mass spectra are time averaged over the one minute elution window corresponding to the asterisk in the extracted ion chromatogram. Inset mass spectra are presented as m/z versus absolute intensity. Monoisotopic $[MH]^+$ mass and the experimental mass error in ppm are also reported. SNAC- ABC intermediate (**5**) reactions monitoring cyclic (**A**) and linear product formation (**B**) are provided. No enzyme control reactions monitoring cyclic (**C**) and linear product formation (**D**) are provided. **Figure 4-12 E-H. FTICR MS analysis of cryptophycin products from the reaction of unit C dimethyl chain elongation unit (5) with L-2HIC, ATP, CrpD-M2 and Crp TE.** Extracted ion chromatograms are presented at ± 15 ppm as time versus absolute intensity. Inset mass spectra are time averaged over the one minute elution window corresponding to the asterisk in the extracted ion chromatogram. Inset mass spectra are presented as m/z versus absolute intensity. Monoisotopic mass and the experimental mass error in ppm are also reported. +2-LHIC reactions monitoring cyclic (**E**) and linear product formation (**F**) are provided. No enzyme control reactions monitoring cyclic (**G**) and linear product formation (**H**) are provided.

Similarly, the synthetic chain elongation intermediate with desmethyl (**4**) or gem-dimethyl (**5**) unit C and L-2HIC (**2**) were used as substrates in CrpD-M2 along with Crp TE reaction. Both cryptophycin 24 (**7**) and cryptophycin 51 (**8**) were successfully generated by this chemoenzymatic route, indicating the versatility and robustness of CrpD-M2 C domain and CrpTE (**Figure 4-12A** and **4-12E**). The corresponding hydrolyzed linear products (**10**, **11**) were also observed (**Figure 4-12B** and **4-12F**). Assuming similar ionization intensity between linear and cyclic products, *seco*-cryptophycin 24 (**10**) was predominantly produced compared to cyclic cryptophycin 24 (**7**) (**Figure 4-12A** and **4-12B**), and cyclic cryptophycin 51 (**8**) was present in similar abundance compared to with its linear counterpart (**11**) (**Figure 4-12E** and **4-12F**). Using SNAC-ABCD analogs as native substrate mimics, Beck et al. investigated the effect extent of the unit C methylation degree on Crp TE mediated macrocylization.[7] The analog carrying the gem-dimethyl group was found to produce more cyclic product (6:1) than the one with desmethyl (5:1) but less than the one with monomethyl group (10:1). In this study, this order remained the same when the native T domain bound substrates generated by CrpD-M2 were supplied to CrpD TE.

When D-2HIC (**20**) was substituted for L-2HIC (**2**) in the chemoenzymatic reaction with chain elongation intermediate **3,** no cyclic depsipeptide product was detected (**Figure 4-9C**) but a small amount of linear product was formed (**Figure 4-9D**). This result suggests that the C domain of CrpD-M2 is able to recognize the stereoisomer of its acceptor substrate and then use it to form an ester bond with its donor substrate, albeit at a significantly lower level. It also indicates that the natural product of the KR reaction is L-2HIC dramatically favoring the following esterification and

macrocyclization reactions in cryptophycin biosynthesis—assuming similar substrate pools. The failure to macrocyclize ABC-D-2HIC substrates suggests that the "gate-keepers" for unit D stereochemical selection are CrpD-M2 α-KR domain and Crp TE rather than its C-domain or A-domain.

4.3 Conclusion

Non-amino acid moieties derived from NRPS enzymes have been found in a handful of natural products isolated from bacterial and fungal cells—however, a complete mechanism for understanding their incorporation has not been developed. In this study, CrpD-M2 coupled with Crp TE was used as a model system to fully understand non-amino acid selection, loading, reduction, elongation through an ester bond, and final production formation. With a powerful FTIC-MS toolkit, a scheme of five sequential biochemical reactions was verified for the first time to various degrees at the A-, C-, KR-, T-, and TE-domains. This is also the first study in a KR-NRPS to directly generate bioactive compounds from elaborate "natural" chain elongation intermediate precursors. Three cyclic cryptophycins 3 (**6**), 24 (**7**), and 51 (**8**) were chemoenzymatically synthesized. Thus, CrpD-M2 as a chemoenzymatic reagent or as part of a fermentation based production strategy has the potential to specifically generate novel cryptophycin analogs with altered physicochemical properties that may be beneficial to clinical application (ex. analogs with increased solubility and decreased peripheral neuropathy).

4.4 Supplement

Protein expression and purification

Proteins were cloned and expressed using standard molecular biology and biochemical techniques. Crp TE was cloned and expressed as previously described.[7,33] CrpD-M2 gene was amplified by PCR with a forward primer with a restriction site of *Bam*HI (CAAGGATCCTTACGTACTACTAATAGCGCA) and a reverse primer with a restriction site of *Xho*I (ATGCTCGAGTAGTTGTTGAATTGGTACTAATGG). The amplicons were purified and digested for cloning into pET28a. The plasmid encoding N-terminal His_6-CrpD-M2 was transformed into *E. coli* BAP1 and grown at 37 °C in TB medium to an OD_{600} of ~0.8 in 2 L flasks. The cultures were cooled to 18 °C, and isopropyl β-D-thiogalactopyranoside was added to a final concentration of 0.2 mM and grown for additional 12-16 hr with shaking. The cells were harvested by centrifugation and frozen at -20 °C. Cell pellets were thawed to 4 °C and resuspended in 5X volume of lysis buffer (20 mM HEPES, pH 7.8, 300 mM NaCl, 20 mM imidazole, 1 mM $MgCl_2$, 0.7 mM Tris(2-carboxyethyl) phosphine (TCEP PH 7.5), ~100 mg CelLytic Express (Sigma-Aldrich)) before lysis via sonication. Centrifugation at 25,000 *x* g for 60 min provided clarified lysates. Proteins were purified using Ni-Sepharose affinity chromatography with a gravity column. Briefly, after filtration of the supernatant through 0.45 μm membrane, the solution was loaded onto a 5 mL HisTrap nickel-nitrilotriacetic acid column. The column was washed with 10 column volumes of buffer A (20 mM HEPES, pH 7.8, 300 mM NaCl, 20 mM imidazole, 1 mM TCEP PH 7.5, 10% glycerol), 10 column volumes of buffer B (20 mM HEPES, pH 7.8, 300 mM NaCl, 50 mM imidazole, 1 mM TCEP PH 7.5, 10% glycerol), and then eluted with buffer C (20 mM

HEPES, pH 7.8, 300 mM NaCl, 400 mM imidazole, 1 mM TCEP PH 7.5, 10% glycerol). Protein containing fractions were pooled and desalted with pre-equilibrated PD-10 gravity flow columns in storage buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 1 mM TCEP PH 7.5, 10% glycerol). Fractions were combined, concentrated, frozen, and stored at -80 °C.

Integrity of purified CrpD-M2 as analyized by FTICR-MS

CrpD-M2 integrity was determined by peptide map fingerprinting. Briefly, CrpD-M2 was reduced and digested with trypsin (Pierce, TPCK modified). The sample was desalted with Handee Microspin columns (Pierce) packed with 20 µl of 300 Å polymeric C4 resin (Vydac). Samples were loaded onto the columns and washed with 30 column volumes of 0.1% formic acid prior to elution with 10 column volumes of 50% acetonitrile plus 0.1% formic acid. Peptides were then introduced into the FTICR-MS at a rate of 70 µL/hour with direct infusion. Peaks were identified with the thrash algorithm as implemented in MIDAS data analysis workstation (National High Magnetic Field Laboratory). Peaks were matched against a theoretical digest of CrpD-M2 (Protein Prospector) with a tolerance of ±20 ppm. Due to the extremely large size of the protein (221 kDa) the bottom-up approach was identified as more convenient. Note that the low sequence coverage observed is not surprising due to lack of LC separation, however, sequence coverage is obtained from residue 3-1,922 out of 1,964 suggesting that the purified protein is in full-length (**Figure 4-4**). Identified peptides are presented in **Table 4-1**.

ATP-[^{32}P] PP_i exchange assay

The exchange assay for determining A domain substrate specificity was conducted using a procedure modified from a previous protocol.[36] All acid substrates used in the assay were purchased from Sigma (St. Louis, MO). The reaction mixture (100 µL) contained 75 mM Tris-Cl, pH 7.5, 10 mM Mg Cl_2, 5 mM TCEP PH 7.5, 5 mM ATP, 1 mM tetrasodium pyrophosphate (PP_i), 5 mM free acid substrate, and 0.5 µCi tetrasodium [^{32}P] PP_i (Perkin Elmer, Boston, MA). The ATP-PP_i exchange was initiated by adding 1 µM of CrpD-M2 and allowed to proceed for 10 min at room temperature. The reaction was then terminated by the addition of cold charcoal solution (500 µL, 1.6% w/v activated charcoal, 0.1 M tetrasodium pyrophosphate, and 5% perchloric acid in water). Free [^{32}P]-PP_i was removed by centrifugation of the sample, and washing the charcoal pellet twice with wash buffer (0.1 M tetrasodium pyrophosphate, and 5% perchloric acid in water). The charcoal was finally resuspended in water (500 µL) and the bound radioactivity was determined by scintillation counting on a Beckman LS6500 (Fullerton, CA). All experiments were carried out in duplicated for each substrate with a negative control without enzyme.

Biochemical reactions and LC FTICR-MS/MS analysis

Enzymatic reaction conditions were as follows: 100 mM Tris-Cl, pH 7.5, 10 mM $MgCl_2$, 5 mM ATP, 1 mM TCEP PH 7.5, 500 µM free-acid extender unit, 100 µM SNAC-ABC chain elongation intermediate, 1 mM NADH, 1 mM NADPH, 1 µM CrpD-M2, and 1 µM Crp TE. Reaction components were excluded as appropriate (ex. no enzyme for no

enzyme control reactions, no chain elongation intermediate for PP_i adenylation domain assays).

For T-domain active site loading experiments samples were incubated for 60 mins at room temperature. Reactions were raised to pH 8 by the addition of concentrated Tris-base, after 4X dilution in 100 mM ammonium bicarbonate. Trypsin (TPCK, Pierce) at 1 mg/mL was added to a molar ratio of 10:1 (CrpD-M2:trypsin). Samples were incubated at 37 °C for 15 min, prior to addition of 10% formic acid to pH 4. Reactions were stored at -80 °C until analysis. The CrpD-M2 T domain active site (QLVEIFQEVLNLPSIGIHDNFFSLGGH**S**LLAVR) was first identified by accurate mass using LC FTICR-MS (**Figure 4-4B**). After the tryptic peptide's retention time (80-81 minutes) and the most abundant charge state (4+, 1001.5 m/z) were identified, online MS/MS was performed using external quadruple isolation and IRMPD in the FTICR cell for ion activation. The phosphopantetheine ejection ions specific markers for the post-translational modification, were observed at 261.1 m/z ($Ppant_1$) and 359.1 m/z ($Ppant_2$).[14,23] As well, the apo T-domain, charge-reduced, parent containing the residual phosphate (+80) or dehydroalanine (-18) was also observed. A short sequence of b-ions was observed, and added to the confirmation of peptide identity. Identified parent and product ions are shown in **Table 4-2**, and the parent ion mass spectrum is shown in **Figure 4-4B**.

Liquid chromatography of trypsin digested CrpD-M2 was performed on an Agilent 1100 with a Jupiter C18 300A 1x150 mm column (Phenomenex) at a rate of 75 µL/min using a

column heater at 50 °C. Twenty µL of sample was injected followed by an LC gradient: 2% solvent B 0-20 min, to 60% solvent B at 90 min, to 98% solvent B 105-108 min, back to 2% B at 110 min, and equilibrated at 2%B for 10 min. Mobile phase A was water with 0.1% formic acid, and mobile phase B consisted of acetonitrile with 0.1% formic acid. FTICR-MS (APEX-Q with Apollo II ion source and actively shielded 7T magnet; Bruker Daltonics) was conducted in positive ion mode from m/z 200–2,000. Electrospray was conducted at 3,000-4,000 V 16-32 scan per spectra utilizing 1 s external ion accumulation in the hexapole prior to analysis in the FTICR using a loop value of 4 for direct inject samples. For online LC FTICR-MS external ion accumulation time was set to 0.33 s with 1 scan per spectra and 128 K signal detected. Collision cell pressure was kept at 5.8e-6 torr and either CID or IRMPD was utilized for MS/MS. LC-FTICR MS data was processed in Data Analysis (Bruker Daltonics) and using DECON2LS and VIPER (Pacific National Labs) for online Thrash analysis. Protein Prospector (UCSF) was used to assist with manual assignment of MS^n data. All experiments were performed at least twice to verify the findings.

Determination of CrpD-M2 active site loading

The CrpD-M2 T domain active site peptide as identified above, was then loaded with the substrates: L-2HIC, D-2HIC, AKGB, and 2-KIC. Peaks were initially identified through LC FITCR-MS using accurate mass. For online confirmation of peak identity the same LC gradient (described above) was run using a Thermo LTQ linear ion trap (LIT) MS. Product peaks were subjected to MS^2 (phosphopantetheine ejection assay) and MS^3 (apo T-domain, charge-reduced, parent containing the residual dehydroalanine (-18) as

precursor) for further confirmation. Notably, the phosphopantetheine ejection assay performed well on the low resolution instrument, as previously reported.[23] See **Table 4-4** for LC FTICR-MS and **Table 4-5** for LC LIT-MS results. The loading and reaction of substrates on the CrpD-M2 T domain was monitored by looking for mass shifts to the loaded T-domain peptide by LC FITCR-MS. Identified peaks were verified by online MS^2 and MS^3 experiments using a LIT-MS.

Linear ion trap mass spectrometry was performed using a Thermo LTQ. A number of data independent scans were defined for each sample including: a survey scan, MS^2 fragmentation of the target T-domain peptide in the 3^+ and 4^+ charge state, and MS^3 of the charge reduced Apo-18 products from the MS^2 scans. An isolation window of 4 Da was utilized, with normalized collision energy of 40%. Xcaliber (Thermo) was used for initial data processing, followed by Protein Prospector (UCSF) for fragmentation assignments.

Cryptophycin MS and MS/MS

The cryptophycin products were generated by reacting CrpD-M2, SNAC-ABC chain elongation intermediates, unit D extender units, ATP, and CrpTE together as described above. After 60 minutes, the proteins were removed by precipitation with 3 volumes of methanol and pelleted by cetrifugation, and the reactions were injected into the LC FTICR-MS (described above). Online MS2 CID spectra were assigned to cryptophycin 3 and 51 (**Figure 4-10**). Product spectra were interpreted manually using cryptophycin 3 as a standard based on coelution with an authentic standard (**Figure 4-11**). Cyclic peptide

product ion assignment was aided by tools developed in the Dorrestien and Pevzner laboratory.[37,38] Cryptophycin MS/MS CID fragmentation spectra were assigned with the assumption that fragmentation would occur across the most labile bonds in the gas phase—amide and ester linkages.[37,38] Conveniently, this result breaks cryptophycin into the A, B, C, and D unit constituents with associated water (±18), ammonia (-17) and CO (-28) based mass shifts. Two key spectral features are present: the +14 Da shift of the CD+H2O, ABC-18, and MH+ ions between cryptophycin 3 (**Figure 4-10A**) and 51 (**Figure 4-10B**)—corresponding exactly to the one methyl group difference due to the different unit C moieties. This key feature helps to validate both spectral assignments.

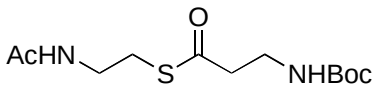
Cryptophycin 3 authentic standard elution

An authentic standard of cryptophycin 3 was used to verify the identity of chemoenzymatically generated cryptophycin 3 by HPLC co-elution analysis. A SHIMADZU LCMS-2010EV system was used for HPLC separation with a detection wavelength of 218 nm. The product separation was carried out with a Waters XBridge™ C_{18} (3.5 µm, 2.1×150 mm) column at a flow rate of 200 µL/min. Solvent B (acetonitrile with 0.1% formic acid) gradually increased from 50% to 99% and solvent A was water with 0.1% formic acid.

Synthesis of SNAC-ABC cryptophycin elongation intermeidates.

Reagents and General Procedures: All reactions were performed under nitrogen or argon atmosphere unless otherwise noted. Boc-β-Ala-OH was purchased from Adavnced ChemTech, (R)-3-(Boc-amino)-2-methylpropionic acid was purchased from Sigma-

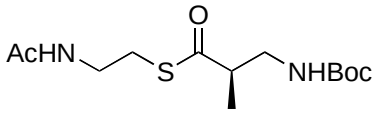
Aldrich (Fluka), and Boc-3-amino-2,2-dimethyl-propionic acid was purchased from PolyPeptide. Solvents were purchased as ACS Grade (CH_2Cl_2, DMF) from Sigma-Aldrich or Fisher Scientific and used as received. N-Acetylcysteamine (SNAC), 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC), 4-dimethylaminopyridine (DMAP), triethylamine (TEA), benzotriazole-1-yl-oxy-tris-pyrrolidino-phosphonium hexafluorophosphate (PyBOP), trifluoroacetic acid (TFA), and all other chemicals were obtained from Sigma-Aldrich or Advanced ChemTech and used directly. ^{1}H and ^{13}C NMR spectra were recorded on a Varian vnmrs 500 MHz or a Varian Performa IV 600 MHz. Proton chemical shifts are reported in ppm from an internal standard of residual chloroform (7.26 ppm) or residual methanol (3.31 ppm); carbon chemical shifts are reported in ppm using an internal standard of residual chloroform (77.16 ppm) or residual methanol (49.00 ppm). Proton chemical data are described as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, br = broad), coupling constant (in Hz), and integration. Mass spectra of chemical intermediates were recorded on a Micromass LCT time-of-flight mass spectrometer in electroscpray ionization (ESI) mode. Analytical thin-layer chromatography (TLC) was performed on silica gel 60 F_{254} TLC glass plates with a fluorescent indicator from EMD Chemicals. Visualization was accomplished with UV light (254 nm) and by dipping in a 1% solution of *p*-anisaldehyde in ethanol or a 0.05 M solution of $KMnO_4$ followed by heating.



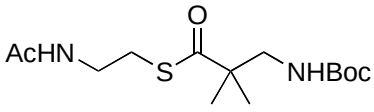
23

2-acetamidoethyl-3-(tert-butoxycarbonylamino)propanethioate (23). A solution of 3-(tert-butoxycarbonylamino)propanoic acid (**20**, 250 mg, 1.32 mmol), EDCI (506 mg, 2.64

mmol), and DMAP (16 mg, 0.13 mmol) in dichloromethane (4 mL) was prepared and allowed to stir at room temperature for 10 minutes. N-acetylcysteamine (169 µL, 1.59 mmol) was then added and reaction stirred at room temperature for 18 hours. Quenched by the addition of saturated aqueous NH_4Cl, the organic layer was removed and aqueous layer extracted with CH_2Cl_2 (2 x 10 mL), washed with saturated aqueous $NaHCO_3$ (10 mL), brine (10 mL), and water (10 mL), then combined organics dried with $MgSO_4$ and solvent removed under vacuum. Flash chromatography (100:0 to 95:5 DCM/methanol) afforded 230 mg of pure thioester **4** as a clear, colorless oil (60%). TLC R_f = 0.50 (5% MeOH/DCM); ^{1}H NMR ($CDCl_3$, 500 MHz) δ 6.15 (bs, 1H), 4.99 (bs, 1H), 3.41 (q, *J* = 5.0 Hz, 4H), 3.02 (t, *J* = 5.0 Hz, 2H), 2.76 (t, *J* = 5.0 Hz, 2H), 1.96 (s, 3H), 1.41 (s, 9H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 198.61, 170.55, 155.93, 79.69, 44.45, 36.97, 28.94, 28.48, 23.29; MS (ESI+) *m/z* 313.0 $[M+Na]^+$ ($C_{12}H_{22}N_2NaO_4S$ requires 313.1).



24

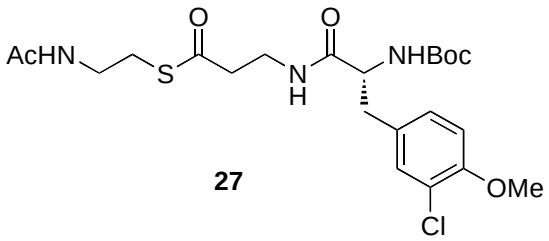


25

(R)-2-acetamidoethyl-3-(tert-butoxycarbonylamino)-2-methylpropanethioate (24) and **2-acetamidoethyl-3-(tert-butoxycarbonylamino)-2,2-dimethylpropanethioate (25)** were prepared in an identical manner to thioester **23**. Compound **24**: 82%; TLC R_f = 0.50 (5% MeOH/DCM); ^{1}H NMR ($CDCl_3$, 500 MHz) δ 6.17 (s, 1H), 4.95 (s, 1H), 3.37 (m, 4H), 2.98 (ddt, *J* = 6.0, 14.0, 36.5 Hz, 2H), 2.90 (m, 1H), 1.96 (s, 3H), 1.41 (s, 9H), 1.16 (d, *J* = 5.0 Hz, 3H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 202.85, 170.57, 156.09, 79.68, 49.01, 43.82, 39.10, 28.83, 28.48, 23.29, 15.26; MS (ESI+) *m/z* 327.0 $[M+Na]^+$ ($C_{13}H_{24}N_2O_4S$ requires 327.1). Compound **25**: 78%; TLC R_f = 0.50 (5% MeOH/DCM); ^{1}H NMR ($CDCl_3$, 500 MHz) δ 6.18 (s, 1H), 4.91 (s, 1H), 3.40 (q, *J* = 5.0 Hz, 2H), 3.27

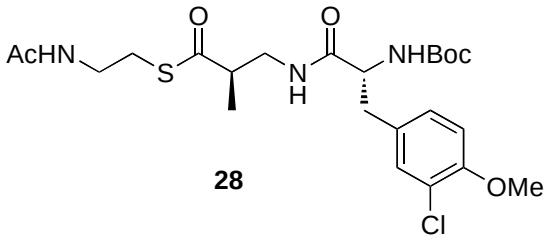
(d, *J* = 5.0 Hz, 2H), 3.00 (t, *J* = 5.0 Hz, 2H), 1.97 (s, 3H), 1.41 (s, 9H), 1.22 (s 6H); ^{13}C NMR ($CDCl_3$, 125 MHz) δ 205.81, 170.51, 156.26, 79.60, 51.24, 49.22, 38.92, 28.69, 28.47, 23.32; MS (ESI+) *m/z* 341.0 $[M+Na]^+$ ($C_{14}H_{26}N_2NaO_4S$ requires 341.2).

Boc-3-Cl-D-Tyr(Me)-OH (26). Boc-3-Cl-D-Tyr(Me)-OH was used from a preparation for a previous investigation.[33]

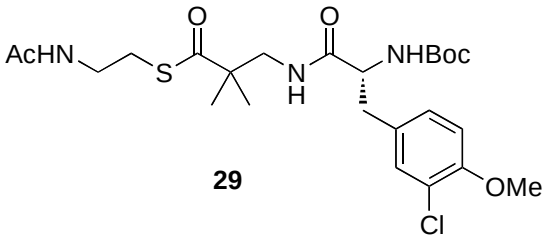


27

(R)-2-acetamidoethyl-3-((R)-2-(tert-butoxycarbonylamino)-3-(3-chloro-4-methoxyphenyl)propanamido)-2-methylpropanethioate (27). Thioester **23** (57 mg, 0.20 mmol) was dissolved in DCM (1 mL), charged with trifluoroacetic acid (500 μL), and stirred at room temperature for 1 hour. Volatiles removed under vacuum and residue dissolved in DMF (1 mL). To this solution was added a solution of Boc-3-Cl-D-Tyr(Me)-OH (**26**, 79 mg, 0.24 mmol), PyBOP (203 mg, 0.39 mmol), triethylamine (72 μL, 0.52 mmol), and DMAP (2 mg, 0.02 mmol). Reaction stirred for 18 hours at room temperature and then diluted with DCM (5 mL). Organic layer washed with saturated aqueous NH_4Cl (5 mL), brine (5 mL), and water (5 mL), then dried with Na_2SO_4. Solvent removed under vacuum and residue subjected to flash chromatography (100:0 to 95:5 DCM/methanol) to afford 63 mg of pure compound **27** as a clear, colorless oil (63%). TLC R_f = 0.45 (5% MeOH/DCM).



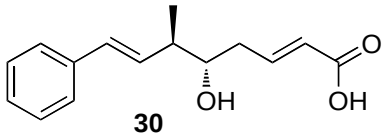
28



29

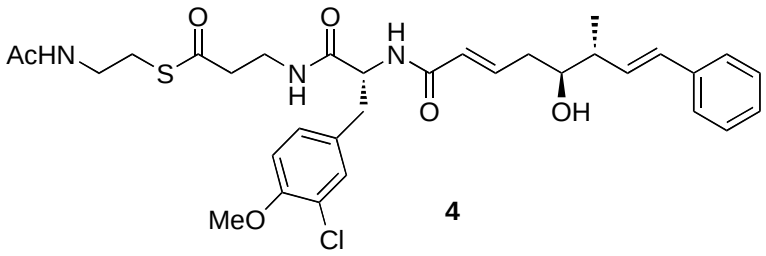
(*R*)-2-acetamidoethyl 3-((*R*)-2-(*tert*-butoxycarbonylamino)-3-(3-chloro-4-methoxyphenyl)propanamido)-2-methylpropanethioate (28) and **(*R*)-2-acetamidoethyl 3-(2-(*tert*-butoxycarbonylamino)-3-(3-chloro-4-methoxyphenyl)propanamido)-2,2-dimethylpropanethioate (29)** were prepared in a manner analogous to that used to prepare compound **27**. Compound **28**: (99%, ESI-MS *m/z* 524 [M+23].). Compound **29**: (124%, ESI-MS, etc).

Synthesis of cryptophycin unit A (30)



30

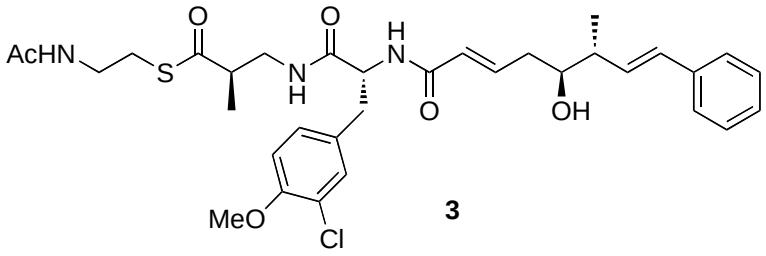
Cryptophycin unit A was synthesized previously according to the publication by Eggen, *et al*.[12]



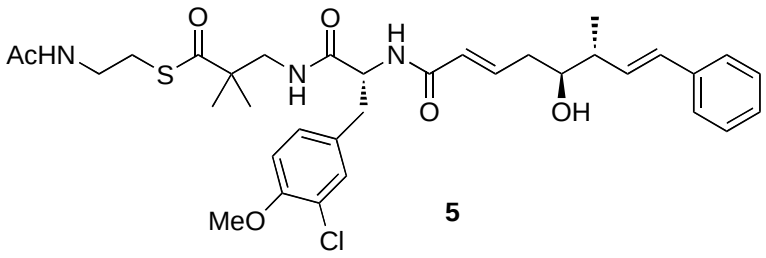
4

(R)-S-2-acetamidoethyl-3-((R)-3-(3-chloro-4-methoxyphenyl)-2-((2E,5S,6R,7E)-5-hydroxy-6-methyl-8-phenylocta-2,7-dienamido)propanamido)-2-methylpropanethioate (4). A solution of compound **27** (27 mg, 0.053 mmol) in DCM (500 μL) was treated with trifluoroacetic acid (250 μL) and stirred at room temperature

for 1 hour. Volatiles removed under vacuum and residue dissolved in DMF (1 mL). To this solution was added a solution of **30** (10 mg, 0.041 mmol), PyBOP (32 mg, 0.061 mmol), triethylamine (12 μL, 0.081 mmol), and DMAP (1 mg, 0.008 mmol). Reaction stirred for 18 hours at room temperature and then diluted with DCM (5 mL). Organic layer washed with saturated aqueous NH_4Cl (5 mL), brine (5 mL), and water (5 mL), then dried with Na_2SO_4. Solvent removed under vacuum, residue dissolved in methanol, and purified by HPLC (XBridge C18 Prep Column, 5 μm, 10 x 250 mm, 3 mL/min, 50:50 to 90:10 methanol/H_2O over 25 minutes, R_t = 20.5 minutes). Fractions lyophilized to recover 3.9 mg of pure **4** (12%). ^{1}H NMR (CD_3OD, 600 MHz); ^{13}C NMR (CD_3OD, 125 MHz) δ 198.43, 173.46, 173.41, 170.29, 168.14, 155.38, 143.53, 139.07, 132.55, 132.09, 131.86, 131.55, 129.47, 128.04, 127.13, 126.08, 123.21, 121.75, 113.36, 75.35, 56.58, 56.12, 44.24, 44.06, 40.00, 38.80, 37.96, 36.53, 29.37, 22.55, 17.50; MS, etc).



3



5

(*R*)-*S*-2-acetamidoethyl 3-((*R*)-3-(3-chloro-4-methoxyphenyl)-2-((2*E*,5*S*,6*R*,7*E*)-5-hydroxy-6-methyl-8-phenylocta-2,7-dienamido)propanamido)-2-methylpropanethioate (12) and ***S*-2-acetamidoethyl 3-((*R*)-3-(3-chloro-4-methoxyphenyl)-2-((2*E*,5*S*,6*R*,7*E*)-5-hydroxy-6-methyl-8-phenylocta-2,7-**

dienamido)propanamido)-2,2-dimethylpropanethioate (13) were prepared in a manner analogous to that used to prepare compound **11**. Compound **12** (13%, ESI-MS *m/z* 524 [M+23], R_t = 21.1 minutes). Compound **13** (8%, ESI-MS, R_t = 21.5 minutes).

Portions of this chapter have been previously published in:

Yousong Ding, Christopher M. Rath, Kyle L. Bolduc, Kristina Håkansson, David H. Sherman. Chemoenzymatic synthesis of cryptophycins 3, 24, and 51 through α-hydroxy-acid condensation and macrolactonization as monitored by FTICR-mass spectrometry. Drafting for Chemistry and Biology.

CMR received funding from the CBI training programs (T32 GM008597) at the University of Michigan. YD was supported by a University of Michigan Rackham Predoctoral Fellowship. This work was supported by NIH grant GM076477 and the Hans W. Vahlteich Professorship (to DHS). Work in KH's laboratory is supported by an NSF Career Award (CHE-05-47699).

4.5 References

1. Nicolaou, K.C.; *et al. J Am Chem Soc*, **2000**, *122*, 9939.
2. Wohlleben, W.; Pelzer, S. *Chem Biol*, **2002**, *9*, 1162.
3. Ran, N.; Rui, E.; Liu, J.; Tao, J. *Cur Pharm Design*, **2009**, *15*, 134.
4. Borst, P.; Evers, R.; Kool, M.; Winjholds, J. *J Nat Can Inst*, **2000**, *92*, 1295.
5. Magarvey, N.A.; *et al*. *ACS Chem Biol*, **2006**, *1*, 766.
6. Ding, Y.; Seufert, W.H.; Beck, Z.Q.; Sherman, D.H. *J Am Chem Soc*, **2008**, *130*, 5492.

7. Beck, Z.Q.; *et al*. *Biochem*, **2005**, *44*, 13457.

8. Trimurtulu, G.; *et al*. *J Am Chem Soc,* **1994**, *116*, 4729.

9. Kobayashi, M.; *et al*. *Tet Let*, **1994**, *35*, 7969.

10. Ghosh, A.K.; Swanson, L. *J Organic Chem,* **2003**, *68*, 9823.

11. Beck, Z.Q.; Burr, D.A.; Sherman, D.H *Chembiochem,* **2007**, *8*, 1373.

12. Eggen, M.; Georg, G.I. *Med Res Re,v* **2002**, *22,* 85.

13. Pfeifer, B.A.; *et al*. *Science,* **2001**, *291*, 1790.

14. Dorrestein, P.C.; *et al. Biochem,* **2006**, *45,* 12756.

15. Calderone, C.T.; *et al*. *Proc Nat Acad Sci USA*, **2008**, *105*, 12809.

16. Chang, Z.; *et al. Gene,* **2002**, *296*, 235.

17. Xu, Y.; *et al. Fungal Genet Biol*, **2009**, *46*, 353.

18. Xu, Y. *et al. Chem Biol*, **2008**, *15*, 898.

19. Magarvey, N.A.; Ehling-Schulz, M.; Walsh, C.T. *J Am Chem Soc,* **2006**, *128*, 10698.

20. Haese, A.; Schubert, M.; Herrmann, M.; Zocher, R. *Mol Microbiol*, **1993**, ***7***, 905.

21. Ramaswamy, A.V.; Sorrels, C.M.; Gerwick, W.H. *J Nat Prod*, **2007**, *70*, 1977.

22. Fujimori, D.G.; *et al. Proc Natl Acad Sci USA*, **2007**, *104*, 16498.

23. Meluzzi, D.; *et al*. *Bioorg Medicinal Chem Let*, **2008**, *18*, 3107.

24. Dorrestein, P.C.; Kelleher, N.L. *Nat Prod Rep*, **2006**, *23*, 893.

25. Gu, L.; *et al*. *Science,* **2007**, *318*, 970.

26. Gu, L.; *et al. Nature,* **2009**, *459*, 731.

27. Lee, C.; Gorisch, H.; Kleinkauf, H.; Zocher, R. *J Biol Chem,* **1992**, *267*, 11741.

28. Caffrey, P. *Chem Biol*, **2005**, *12*, 1060.

29. Caffrey, P. *Chembiochem,* **2003**, *4*, 654.

30. Siskos, A.P.; *et al. Chem Biol*, **2005**, *12*, 1145.

31. Tamura, K.; Dudley, J.; Nei, M.; Kumar, S. *Mol Biol Evol*, **2007**, *24*, 1596.

32. Weissman, K.J.; *et al. Biochem,* **1997**, *36*, 13849.

33. Seufert, W.; *et al*. *Ang Chem Int ed*, **2007**, *46*, 9298.

34. Zaleta-Rivera, K.; *et al. Biochem,* **2006**, *45*, 2561.

35. Lin, S.; Van Lanen, S.G.; Shen, B. *Proc Nat Acad Sci USA,* **2009**, *106*, 4183.

36. Chang, Z.; *et al. Gene,* **2002**, *296*, 235.

37. Liu, L.; *et al. Cancer Res*, **2009**, *69*, 6871.

38. Ng, J. *et al. Nature Meth,* **2009**, *6*, 596.

Chapter 5

Meta-omic analysis of a marine invertebrate microbial consortium provides a direct route to identify and characterize natural product biosynthetic systems

5.1 Introduction

ET-743 (Trabectedin, **1**) is a tetrahydroisoquinoline chemotherapeutic natural product isolated from the tunicate *Ecteinascidia turbinata*,[1] and is approved for use in Europe against ovarian neoplasms and sarcoma (**Figure 5-1**).[2] The drug operates by a unique mechanism of action as it alkylates within the minor groove of DNA,[3] which can lead to sequence-specific alterations in transcription[4] that trigger DNA cleavage.[5] Attempts to repair ET-743 DNA lesions may cause further double-stranded DNA breaks.[6] Obtaining sufficient amounts of ET-743 has been a significant challenge since it is isolated in extremely low yields from the natural source.[1] Aquaculture of the tunicate,[7] or total synthesis[8] cannot provide economical access to the drug.[9] Thus, ET-743 for clinical application is produced semi-synthetically from fermentation-derived cyanosafracin B in seventeen chemical steps.[10]

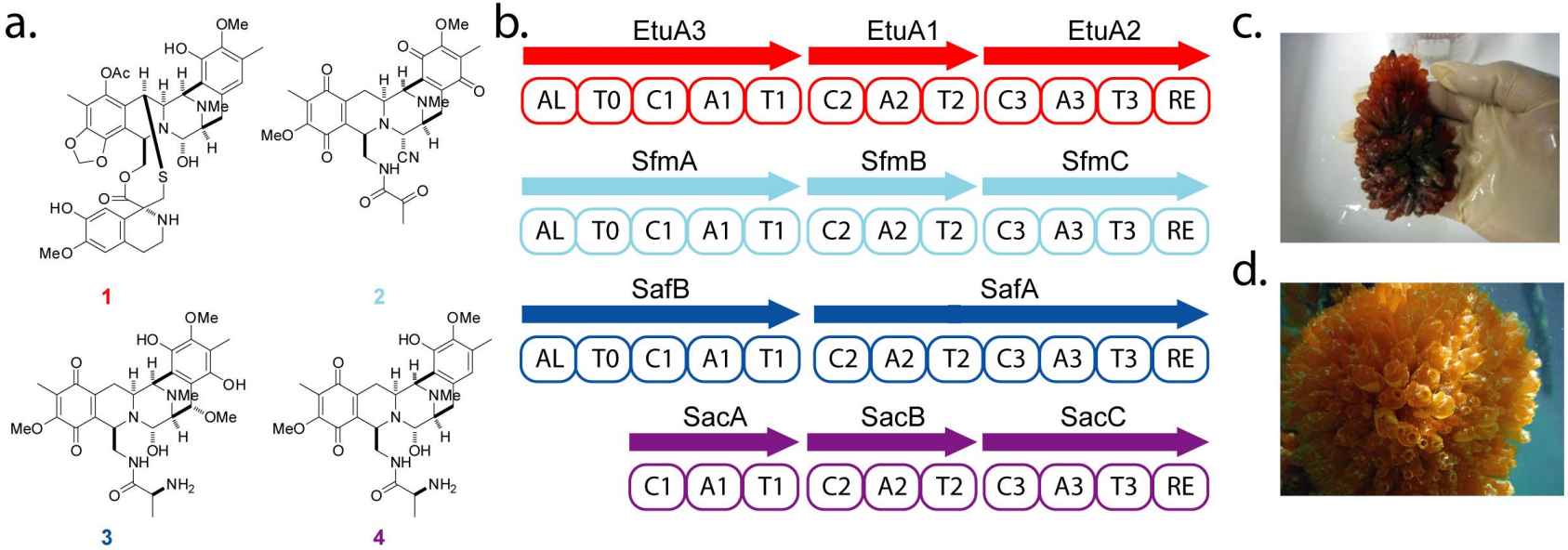


Figure 5-1. ET-743 (1) and tetrahydroisoquinoline natural products: saframycin A (2), saframycin Mx1 (3), and safracin (4). (**A**) ET-743 core modular NRPS proteins (EtuA1-3) and previously characterized Sfm, Saf, and Sac NRPS biosynthetic systems. (**B**) NRPS domains are: AL-acyl ligase, T-thiolation, C-condensation, A-adenylation, RE-reductive. (**C**) Collected *Ecteinascidia turbinata* samples (Erich Bartels, Mote Marine Laboratory). (**D**) *Ecteinascidia turbinata* in its natural environment (Cory Walter, Mote Marine Laboratory).

The similarity of ET-743 to three other bacterial derived natural products, including saframycin A (**2**) (*Streptomyces lavendulae*),[11] saframycin Mx1 (**3**) (*Myxococcus xanthus*),[12] and safracin B (**4**) (*Pseudomonas fluorescens*)[13] suggests that the drug is of prokaryotic origin (**Figure 5-1**),[14] Although the "symbiont hypothesis" has been supported for secondary metabolites isolated from invertebrate animals including bryostatin,[15,16] onnamide/pederin,[17,18] and psymberin,[19] the effort reported here is the first to apply combined meta-omic approaches to address this problem. The biosynthetic pathways for the tetrahydroisoquinoline natural products noted above have been previously characterized, thus providing a potential genetically conserved "biomarker" for the ET-743 system.[20-22] The tetrahydroisoquinoline pathways consist of three nonribosomal peptide synthase (NRPS) modules and a series of allied tailoring enzymes. Each module contains three domains: adenylation (A), condensation (C), and thiolation (T) that combine the amino acid building blocks. Two of these pathways are initiated by an acyl-ligase (AL) and a T didomain. All three NRPS trimodules are terminated by a signature reductase domain (RE) that utilizes NAD(P)H to release the enzyme bound intermediate as an aldehyde. The final C domain in the saframycin pathway serves as a "Pictet-Spenglerase" to cyclize the activated intermediate.[23] Koketsu and colleagues have shown that a fatty acid appended to the growing polypeptide on the NRPS T-domain is required to form the cyclic tri- and tetrapeptide derived from the biological Pictet-Spengler reaction.[23] In considering a meta-omics discovery strategy, we reasoned that the ET-743 pathway would likely be comprised of an AL-T for initiation, three NRPS modules for elongation, and termination by an RE domain (**Figure 5-1**, EtuA1-3).

Previous work directed toward identification of a producing organism and potential biosynthetic pathway assessed the phylogenetic diversity of bacterial species from *E. turbinata* as a source of ET-743 in the Mediterranean and Caribbean seas. A γ-proteobacterium *Candidatus Endoecteinascidia frumentensis* (AY054370) was identified as the most prevalent member from the tunicate at all collection sites,[24,25] providing indirect evidence for a potential bacterial producer of the ET-743 anticancer agent. We considered a cloning-independent approach that would avoid typical pitfalls encountered when handling environmental metagenomic DNA samples in order to gain direct access to the elusive gene cluster. Rapid advances in metagenomic and hologenomic sequencing technologies,[26] as well as bioinformatic tools for contig assembly, indicated that this direct approach would provide rapid access to the desired biosynthetic system derived from a host/symbiont community.

A key issue with metagenomic DNA derived from environmental samples, and unculturable microorganisms is the lack of an *in vivo* genetic system to establish the identity of the biosynthetic pathway. This limitation can be overcome by *in vitro* characterization of heterologously expressed gene products.[16] *In vitro* characterization provides a direct link between biosynthetic genes derived from field-collected samples and their corresponding metabolites, a key step toward understanding these complex systems. We also considered that proteomics would be an effective way to identify gene products in low abundance, particularly for samples consisting of multiple microbial species ("metaproteomics").[27] Direct amino acid sequence evidence for predicted biosynthetic proteins can effectively link gene-based bioinformatics to *in vitro* biochemical function in diverse microbial symbiont-host systems.

Herein, we describe the identification and initial biochemical characterization of the ET-743 biosynthetic pathway from the host/symbiont community derived from *E. turbinata*. After confirming the presence of the tetrahydroisoquinoline secondary metabolites from the animal, metagenomic sequencing was conducted to identify the target biosynthetic genes. High resolution mass spectrometry was then used to mine the metaproteome for the presence of ET-743 biosynthetic pathway enzymes predicted from the gene cluster sequence analysis. Finally, enzymatic activity for a key enzyme to form the tetrahydroisoquinoline core was verified *in vitro* with a model substrate to corroborate the identity of the metabolic pathway. This knowledge enables a clear path for accessing ET-743 and new analogs through heterologous expression technologies,[28,29] as well as provides a general strategy for identification and characterization of host/symbiont derived natural product systems.

5.2 Results

Secondary metabolite identification as a starting point for the "ET-743 bacterial symbiont producer" hypothesis. We confirmed that field-collected tunicate samples of *E. turbinata* from the Florida Keys contained ET-743 and related metabolites using high-resolution, high-mass accuracy, liquid chromatography-Fourier transform ion cyclotron resonance mass spectrometry (LC-FTICR-MS). Known biosynthetic precursors were identified from the tunicate by extracted ion chromatograms at ±20 ppm, including the $M + H^+$ and $(M - H_2O) + H^+$ for ET-743 (**1**), ET-597 (**19**), ET-594 (**21**) and ET-583 (**18**) (**Figure 5-2**). Confirmation by LC-MS/MS was performed on-line with FTICR-MS and an iontrap-mass spectrometer (IT-MS). Since all four compounds identified had

previously been characterized by MS/MS, assignment of product ions was straightforward (**Table 5-1**) as observed fragmentation was consistent between earlier studies using fast atom bombardment (FAB)-collision induced dissociation (CID),[30] and our work with electrospray ionization (ESI)-(CID) on FTICR and IT instruments. The presence of both ET-743 and presumed precursors strongly suggested that ET-743 biosynthesis occurred within the field-collected animal, and thus that the producing symbiont was present.

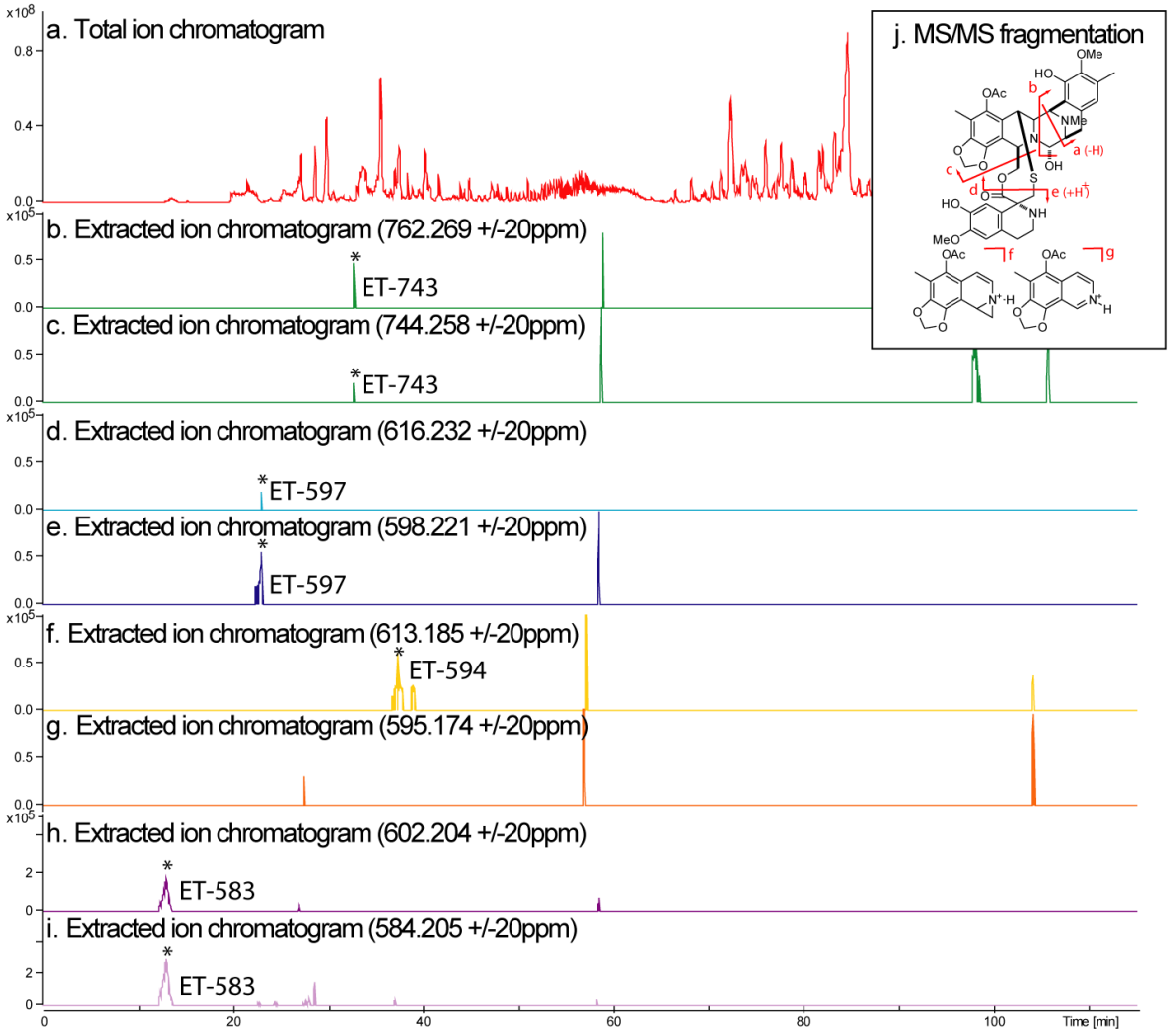


Figure 5-2. Liquid chromatography FTICR mass spectrometry (LC-FTICR-MS). Total ion chromatogram (**A**) and extracted ion chromatograms for M + H^+ (**B**, **D**, **F**, **H**), and (M – H_2O) + H^+ (**C**, **E**, **G**, **I**) for ET-743 (**1**), ET-597 (**19**), ET-594 (**21**), and ET-583 (**18**). Y axis is in arbitrary units. All identified compounds were verified by CID MS/MS (**Table 5-1**).

CID MS/MS fragmentation of ET-743 and related molecules. Expected fragments a-g for the ET-743 metabolite are from Sakai (**Figure 5-2J**).[30] The IT-MS provided greater signal and operated at a higher duty cycle for MS/MS, allowing assignment of more product ions compared to FTICR-MS. However, assigned IT-MS fragments are of low mass accuracy (±300 ppm) versus the high mass accuracy of FTICR-MS data (±20 ppm). For ET-743 (**1**), present at the lowest apparent abundance of

all analytes identified, FTICR-MS/MS provided only the $(M - H_2O) + H^+$ fragment, however iontrap-MS/MS provided the: b, c, d, e, f and $(M - H_2O) + H^+$ fragments. ET-597 (**19**) was confirmed by c, d and $(M - H_2O) + H^+$ ions from FTICR-MS/MS and a, b, c, d, g and $(M - H_2O) + H^+$ ions from IT-MS MS/MS. The $M + H^+$ of ET-594 (**21**) was not observed in FTICR-MS/MS spectra, however the d product ion was. In IT-MS all expected product ions were observed. ET-583 (**18**) provided the most complete structural information from CID MS/MS with c, d, f, and $(M - H_2O) + H^+$ in FTICR-MS/MS and all expected ions in IT-MS. In the future, the LC-FTICR-MS/MS metabolomic method applied herein will be expanded to identify novel predicted biosynthetic intermediates in field collected samples. Linking natural product abundance to spatial distribution may also further inform this system.[31]

		MH+	MH+ - H_2O	a	b	c	d	e	f	g
ET-743 (**1**)	Calculated m/z	762.269	744.258	204.102	218.118	463.187	493.197	224.075	260.092	246.077
	Observed m/z, FTICR-MS	762.276	744.268							
	Observed m/z, iontrap-MS	762.3	744.2		218.3	463.2	493.3	224.2	260.2	
ET-597 (**19**)	Calculated m/z	616.232	598.222	204.102	218.118	465.203	495.213		262.108	248.092
	Observed m/z, FTICR-MS	616.234	598.227			465.205	495.210			
	Observed m/z, iontrap-MS	616.2	598.2	204.1	218.2	465.2	495.2			248.2
ET-594 (**21**)	Calculated m/z	613.185	595.174	204.102	218.118	463.187	493.197			
	Observed m/z, FTICR-MS	613.188					493.191			
	Observed m/z, iontrap-MS	613.2	595.2	204.2	218.2	463.3	493.3			
ET-583 (**18**)	Calculated m/z	602.217	584.206	190.087	204.102	451.187	481.197		262.108	248.092
	Observed m/z, FTICR-MS	602.222	584.201			451.183	481.205		262.110	
	Observed m/z, iontrap-MS	602.2	584.2	190.2	204.2	451.2	4811.2		262.2	248.2

Table 5-1. CID-MS/MS confirmation of ET-743 and related metabolites. Expected product ions have been previously reported by Sakai from FAB-CID-MS/MS. Observed precursor and product ions are provided for ESI-FTICR-CID-MS/MS and ESI-iontrap-CID-MS/MS. See **Figure 5-2J** for product ion a-g definitions.

Metagenomic sequencing and phylogenetics. Based on identification of ET-743 from field-collected tunicates, we prepared total hologenomic DNA from *E. turbinata* samples. This DNA was used to prepare a 16S rRNA gene amplicon library and a random shotgun fragment library for 454 based FLX pyrosequencing. Raw reads from the first shotgun sequencing run, and an assembly of these data were filtered using relatedness of the translated protein sequences to the saframycin and safracin nonribosomal peptide synthetases (NRPSs) (MXU24657, DQ838002, AY061859) using BLASTx and tBLASTn. Linkage of these sequences was performed using a combination of traditional PCR and restriction-site PCR (RS-PCR) yielding six contigs of high interest containing NRPS domains for biosynthesis of ET-743.[32] A second sequencing run combined with the first generated another assembly of 839,923 reads with an average read length of 332 bp, bearing 77,754 total contigs, and 15,097 contigs larger than 500 bp. We identified a 22 kb contig that linked 4/6 of the high interest contigs from the first assembly and extended this putative NRPS-containing contig to > 35 kb using RS-PCR. This DNA fragment was PCR amplified and Sanger sequenced for confirmation. Twenty-five ET-743 biosynthetic genes were identified in this contig and annotated with proposed function using BLASTx against the NCBI NR database (**Figure 5-6, Table 5-2, Genbank HQ609499**). The individual genes appear to be of bacterial origin, suggesting that the cluster is not derived from the tunicate genome. In addition to the 35 kb putative NRPS contig, we identified sequences containing ribosomal RNA (rRNA) fragments. One of these rRNA sequences was located in a large contig (contig00422) that we extended to > 26 kb with RS-PCR. Contig00422 (**Table 5-3**, **Genbank HQ542106**)

contains a full 16S rRNA gene, which aligns (> 99% identity) to the 16S rRNA gene reported previously for *E. frumentensis* (AY054370) (DQ494516).[24,33]

Name	kb	ID	Sim	Taxonomy	BlastP ID	Function
EtuA1	1.9	29	54	proteobacteria	SafA (AAC44129)	NRPS module C-A-T
EtuA2	4.3	37	59	proteobacteria	SafA AAC44129	NRPS module C(PS)-A-T-RE
EtuA3	5.4	41	54	firmicutes	BarG (AAN32981)	NRPS dimodule FA-T-C-A-T
EtuD1	0.8	44	70	δ-proteobacteria -> Oceanospirillales	(YP_903342.1)	TatD Mg2+ dependant cytoplasmic DNAse
EtuD2	0.8	34	53	δ-proteobacteria	PRK05707 (ZP_01223804)	DNA polymerase III delta prime subunit
EtuD3	0.7	31	47	proteobacteria	(NP_221127)	DNA polymerase I 5'-3' exonuclease domain
EtuF1	1.3	64	81	δ-proteobacteria -> Vibrionaceae	(ZP_01161771)	Acetyl-CoA carboxylase biotin carboxylase subunit
EtuF2	0.5	43	67	δ-proteobacteria -> Enterobacteriaceae	(YP_002922996)	Acetyl-CoA carboxylase biotin carboxyl subunit
EtuF3	1.8	42	65	proteobacteria	CBUD_0858 (YP_001424243)	Penicillian acylase
EtuH	0.4	39	58	bacteria	SfmD (ABI22134)	Catechol hydroxylase
EtuM1	1.1	51	67	δ-proteobacteria	SfcF(AAL33761)	SAM dependant methyltransferase
EtuM2	0.7	46	72	proteobacteria	SafC (AAC44130)	SAM dependant O-methyltransferase
EtuN1	1.4	41	61	γ-proteobacteria	gatB (YP_344005)	Asp/Glu-tRNA amidotransferase subunit B
EtuN2	1.4	62	82	δ-proteobacteria	RICGR_0965 (ZP_02062061)	Asp/Glu-tRNA amidotransferase subunit A
EtuN3	0.2	33	62	firmicutes	BcellDRAFT_1794 (ZP_06363292)	Asp/Glu-tRNA amidotransferase subunit C
EtuO	1.5	34	56	Actinomycetes	SfmO2 (ABI22133)	FAD dependant monoxygenase
EtuP1	2.0	51	69	proteobacteria	(NP930029)	Pyruvate dehydrogenase E1 component
EtuP2	1.0	26	47		PRK11856 (ZP_06439425)	Pyruvate dehydrogenase E2 component
EtuR1	0.9	32	59	proteobacteria	S29x (AAB39275)	Bacterial symbiont gene for protein found in host
EtuR2	0.3	34	58	bacteria	HMPREF0446_00485 (ZP_05851657)	Transcriptional regulator MerR family
EtuR3	0.4	47	72	Neisseriaceae	dksA (NP_899844)	DNA K suppressor protein
EtuT	0.8	32	50	proteobacteria	FTM_0945 (YP_001891654)	Drug metabolite transporter superfamily protein
EtuU1	1.4	65	84	δ-proteobacteria	Patl_0190 (YP_659776)	EtuP peptidase U62 modulator of DNA gyrase
EtuU2	0.5	57	74	γ-proteobacteria	aroK (YP_094966)	Shikimate kinase I
EtuU3	0.3	55	70	γ-proteobacteria	VEA_003741 (YP_003286366)	Hypothetical protein

Table 5-2. ET-743 biosynthetic genes. Numerical order in the gene cluster is provided with gene size in kb. Identity, similarity, and protein ID (with accession numbers) are provided. Taxonomy data are shown at the level of agreement for >50% of the top 100 hits from BLASTx. Proposed gene product function is based on predicted and observed function for top BLASTx hits. Genes are named based on proposed function of the expressed protein: **EtuA**-NRPS modules, **EtuD**-DNA processing enzymes, **EtuF**-fatty-acid related enzymes, **EtuH**-hydroxylase, **EtuM**-methyltransferases, **EtuN**-amidotransferases, **EtuO**-monoxygenase, **EtuP**-pyruvate processing cassette, **EtuR**-regulatory enzymes, **EtuT**-drug metabolite transporter, **EtuU**-enzymes of unknown function.

Name	kb	ID	Sim	Tax	BlastP ID
etrA					16s rRNA gene
EtrB	2.6	52	72	gammaproteobacteria	preprotein translocase, SecA subunit (ZP_05729793)
EtrC	0.3	28	50	enterobacteria	hypothetical protein (ZP_06125969)
EtrD	1.2	62	77	gammaproteobacteria	cell division protein FtsZ (YP_114837)
EtrE	1.2	46	70	gammaproteobacteria	Cell division protein FtsA (ZP_01127086)
EtrF	1.2	51	72	gammaproteobacteria	fatty acid desaturase (YP_003525747)
EtrG	0.3	68	80	betaproteobacteria	4Fe-4S ferredoxin iron-sulfur binding domain protein (YP_001894230)
EtrH	0.7	46	64	proteobacteria	phosphopantetheine adenylyltransferase (YP_903601)
EtrI	0.3	31	49	bacteria	hypothetical protein (ZP 04754723)
EtrJ	0.5	46	65	gammaproteobacteria	predicted metal-sulfur cluster biosynthetic enzyme (ZP_01736718)
EtrK	0.3	37	57	proteobacteria	iron-sulfur cluster assembly accessory protein
EtrL	0.9	46	70	proteobacteria	cysteine desulphurases, SufS (YP_344460)
EtrM	1.3	27	50	proteobacteria	FeS assembly protein SufD (YP_003385038)
EtrN	0.8	59	79	gammaproteobacteria	FeS assembly ATPase SufC (YP_003526256)
EtrO	1.4	37	60	gammaproteobacteria	exodeoxyribonuclease I (YP_002303198)
EtrP	0.4	45	69	gammaproteobacteria	glycine cleavage system protein H (YP_437065)
EtrQ	1.4	34	58	proteobacteria	phospholipase D/Transphosphatidylase (ZP_01312794)
EtrR	3.3	28	50	gammaproteobacteria	UvrD/REP helicase (YP_003761202)
EtrS	1.3	23	46	gammaproteobacteria	hypothetical protein GPB2148_3550 (ZP_05093709)

Table 5-3. Etr 16S rRNA gene contig. Numerical order in the gene cluster is provided with gene size in kb. Identity, similarity and protein ID (with accession numbers) are provided. Taxonomy data are shown at the level of agreement for >50% of the top 100 hits from BLASTx. Proposed gene product function is based on predicted and observed function for top BLASTx hits. Genes are named based on order along the contig.

Taxonomic classification of the raw reads and of the total assembly was performed using the Metagenomic Rapid Annotations with Subsystems pipeline (MG-RAST).[34] Results from both sets were consistent, with ~40% of the classified sequences being of eukaryotic origin (mainly *Ciona* [sea squirt/tunicate]) and the remaining 60% being largely proteobacterial sequence (>90%) of which there were two major populations: α-proteobacterial (largely *Rhodobacteraceae*, 78-85%) and γ-proteobacterial (10-17%) (**Tables 5-4**). 16S rRNA gene amplicon sequencing runs identified 30 variants but only three significant ones (> 1% of the total reads) (**Tables 5-5**). The largest population of 16S rRNA gene reads was classified as *Rhodobacteraceae* (~78%), consistent with the classification of shotgun reads by MG-RAST. This 16S rRNA gene variant aligns to contig09113 from the shotgun sequence assembly, found previously in

two of the three tunicate sampling sites from the Caribbean (clone 2j, DQ494507).[25] The second most abundant 16S rRNA gene variant is an unclassified γ-proteobacterium (~19%) that aligns to contig00422 and represents *E. frumentensis* in the sample. A third small population of 16S rRNA gene reads was identified as unclassified bacteria and corresponded to one read from the shotgun sequencing runs (also identified previously).[24] These three variants account for > 97% of the 16S rRNA gene sequencing reads (**Figure 5-3**). None of these three strains form a close phylogenetic relationship with *S. lavendulae*, *M. xanthus*, or *P. fluorescens*, producers of the three tetrahydroisoquinoline antibiotics whose pathways have been previously characterized.

Classified by MG-RAST:	**Total Assembly**		**Raw Reads**	
	77,754		815,074	
Total	5,510	100.00%	65,267	100.00%
Eukaryota	2,390	43.38%	23,413	35.87%
Bacteria	**3,107**	**56.39%**	**41,651**	**63.82%**
Viruses	7	0.13%	60	0.09%
Archaea	6	0.11%	139	0.21%
Plasmids	0	0.00%	1	0.00%
broad host range plasmids	0	0.00%	3	0.00%
Phylum (in Bacteria)				
Actinobacteria	19	0.61%	241	0.58%
Aquificae	0	0.00%	16	0.04%
Bacteroidetes	1	0.03%	31	0.07%
Bacteroidetes/Chlorobi group	71	2.29%	944	2.27%
Chlamydiae/Verrucomicrobia	4	0.13%	92	0.22%
Chlorobi	0	0.00%	4	0.01%
Chloroflexi	13	0.42%	138	0.33%
Cyanobacteria	60	1.93%	684	1.64%
Deinococcus-Thermus	0	0.00%	31	0.07%
Fibrobacteres/Acidobacteria	11	0.35%	97	0.23%
Firmicutes	39	1.26%	642	1.54%
Fusobacteria	1	0.03%	7	0.02%
Planctomycetes	81	2.61%	942	2.26%
Proteobacteria	**2,798**	**90.05%**	**37,615**	**90.31%**
Spirochaetes	3	0.10%	48	0.12%
Synergistetes	2	0.06%	20	0.05%
Thermotogae	4	0.13%	90	0.22%
unclassified Bacteria	0	0.00%	9	0.02%
Total	3,107	100.00%	41,651	100.00%
Class (in Proteobacteria)				
α-proteobacteria	**2,376**	**84.92%**	**29,361**	**78.06%**
------>Rhodobacteraceae	**2,092**	**74.77%**	**24,873**	**66.13%**
β-proteobacteria	74	2.64%	1,179	3.13%
δ/ε-subdivisions	58	2.07%	716	1.90%
γ-proteobacteria	**287**	**10.26%**	**6,312**	**16.78%**
unclassified Proteobacteria	3	0.11%	47	0.12%
Total	2,798	100.00%	37,615	100.00%

Table 5-4. MG-RAST analysis of raw sequencing reads and assembly. In the MG-RAST pipeline, reads are classified by predicted protein homology to a manually curated protein database (the SEED). A cutoff of 1e-10 was used. No significant classified bacterial populations were observed beyond the Class level except in α-proteobacteria where Rhodobacteraceae comprised a significant portion of the reads. % values represent abundance in each taxonomic level.

16S contig	RDP Classification												# of 16S reads
contig00001	Bacteria	100%	Firmicutes	41%	Clostridia	29%	Clostridiales	26%	Veillonellaceae	5%	Anaerovibrio	3%	15
contig00002	Bacteria	95%	Bacteroidetes	37%	Sphingobacteria	32%	Sphingobacteriales	32%	Chitinophagaceae	27%	Segetibacter	19%	58
contig00003	Bacteria	100%	Proteobacteria	86%	Gammaproteobacteria	57%	Oceanospirillales	41%	Hahellaceae	8%	Endozoicomonas	8%	37
contig00004	Bacteria	100%	Proteobacteria	99%	Alphaproteobacteria	99%	Rhodobacterales	93%	Rhodobacteraceae	93%	Sagittula	30%	16
contig00005	Bacteria	100%	Bacteroidetes	99%	Sphingobacteria	96%	Sphingobacteriales	96%	Saprospiraceae	95%	Lewinella	56%	79
contig00006	Bacteria	96%	Verrucomicrobia	15%	Verrucomicrobiae	15%	Verrucomicrobiales	15%	Verrucomicrobiaceae	15%	Luteolibacter	13%	8
contig00007	Bacteria	98%	Bacteroidetes	62%	Sphingobacteria	28%	Sphingobacteriales	28%	Saprospiraceae	10%	Haliscomenobacter	6%	3
contig00008	Bacteria	100%	Proteobacteria	60%	Deltaproteobacteria	36%	Desulfobacterales	20%	Desulfobulbaceae	20%	Desulfopila	10%	12
contig00009	Bacteria	100%	Bacteroidetes	82%	Flavobacteria	54%	Flavobacteriales	54%	Flavobacteriaceae	40%	Wautersiella	1%	12
contig00010	Bacteria	100%	Bacteroidetes	100%	Flavobacteria	95%	Flavobacteriales	95%	Flavobacteriaceae	95%	Pseudozobellia	35%	47
contig00011	Bacteria	100%	Bacteroidetes	100%	Flavobacteria	100%	Flavobacteriales	100%	Flavobacteriaceae	100%	Muricauda	76%	8
contig00012	Bacteria	99%	Bacteroidetes	77%	Flavobacteria	71%	Flavobacteriales	71%	Flavobacteriaceae	67%	Flagellimonas	9%	3
contig00013	Bacteria	99%	OD1	89%	OD1_genera_incertae_sedis	89%							41
contig00014	Bacteria	100%	Proteobacteria	37%	Gammaproteobacteria	24%	Thiotrichales	15%	Thiotrichaceae	14%	Leucothrix	14%	187
contig00015	Bacteria	100%	Proteobacteria	33%	Gammaproteobacteria	24%	Thiotrichales	12%	Thiotrichaceae	10%	Leucothrix	10%	184
contig00016	Bacteria	100%	Tenericutes	11%	Mollicutes	11%	Haloplasmatales	11%	Haloplasmataceae	11%	Haloplasma	11%	753
contig00017	Bacteria	99%	Proteobacteria	85%	Deltaproteobacteria	85%	Syntrophobacterales	65%	Syntrophobacteraceae	65%	Desulforhabdus	55%	36
contig00018	Bacteria	100%	Proteobacteria	35%	Betaproteobacteria	17%	Burkholderiales	15%	Oxalobacteraceae	13%	Oxalicibacterium	11%	104
contig00019	Bacteria	100%	Proteobacteria	99%	Epsilonproteobacteria	98%	Campylobacterales	98%	Campylobacteraceae	97%	Arcobacter	96%	143
contig00020	Bacteria	100%	Proteobacteria	100%	Alphaproteobacteria	100%	Rhodobacterales	98%	Rhodobacteraceae	98%	Shimia	40%	78
contig00021	Bacteria	100%	Proteobacteria	74%	Gammaproteobacteria	38%	Thiotrichales	27%	Thiotrichaceae	23%	Leucothrix	23%	13264
contig00022	Bacteria	100%	Proteobacteria	91%	Alphaproteobacteria	76%	Rhodobacterales	37%	Rhodobacteraceae	37%	Pannonibacter	28%	16
contig00023	Bacteria	100%	Proteobacteria	100%	Alphaproteobacteria	99%	Rhodobacterales	91%	Rhodobacteraceae	91%	Roseovarius	41%	57
contig00024	Bacteria	100%	Proteobacteria	100%	Alphaproteobacteria	100%	Rhodobacterales	98%	Rhodobacteraceae	98%	Ruegeria	26%	55573
contig00025	Bacteria	99%	Proteobacteria	52%	Gammaproteobacteria	22%	Gammaproteobacter	12%	Methylohalomonas	10%			56
contig00026	Bacteria	98%	Proteobacteria	61%	Gammaproteobacteria	42%	Thiotrichales	21%	Thiotrichaceae	20%	Leucothrix	20%	158
contig00027	Bacteria	100%	Cyanobacteria	100%	Cyanobacteria	100%	Family II	100%	GpIIa	100%			64
contig00028	Bacteria	100%	Proteobacteria	100%	Alphaproteobacteria	100%	Sphingomonadales	100%	Erythrobacteraceae	100%	Erythrobacter	100%	60
contig00029	Bacteria	100%	Proteobacteria	100%	Gammaproteobacteria	100%	Thiotrichales	100%	Francisellaceae	99%	Francisella	99%	13
contig00030	Bacteria	100%	Proteobacteria	99%	Alphaproteobacteria	99%	Rhizobiales	93%	Hyphomicrobiaceae	71%	Cucumibacter	57%	71
													71156

Table 5-5. 16S rRNA gene identification. Reads from a 454 16S amplicon library were assembled using the 454 Newbler assembler at an identity of 95%. Each assembled contig was submitted to the Ribosomal Database Project (RDP) 16S Classifier. % values represent a bootstrap confidence estimate calculated by the RDP Classifier. The entire classification hierarchy is shown however it should be noted that the generally accepted confidence threshold is 80%. For example, contig 00001 is classified as an unknown bacterium and contig 00003 is classified as an unknown proteobacterium.

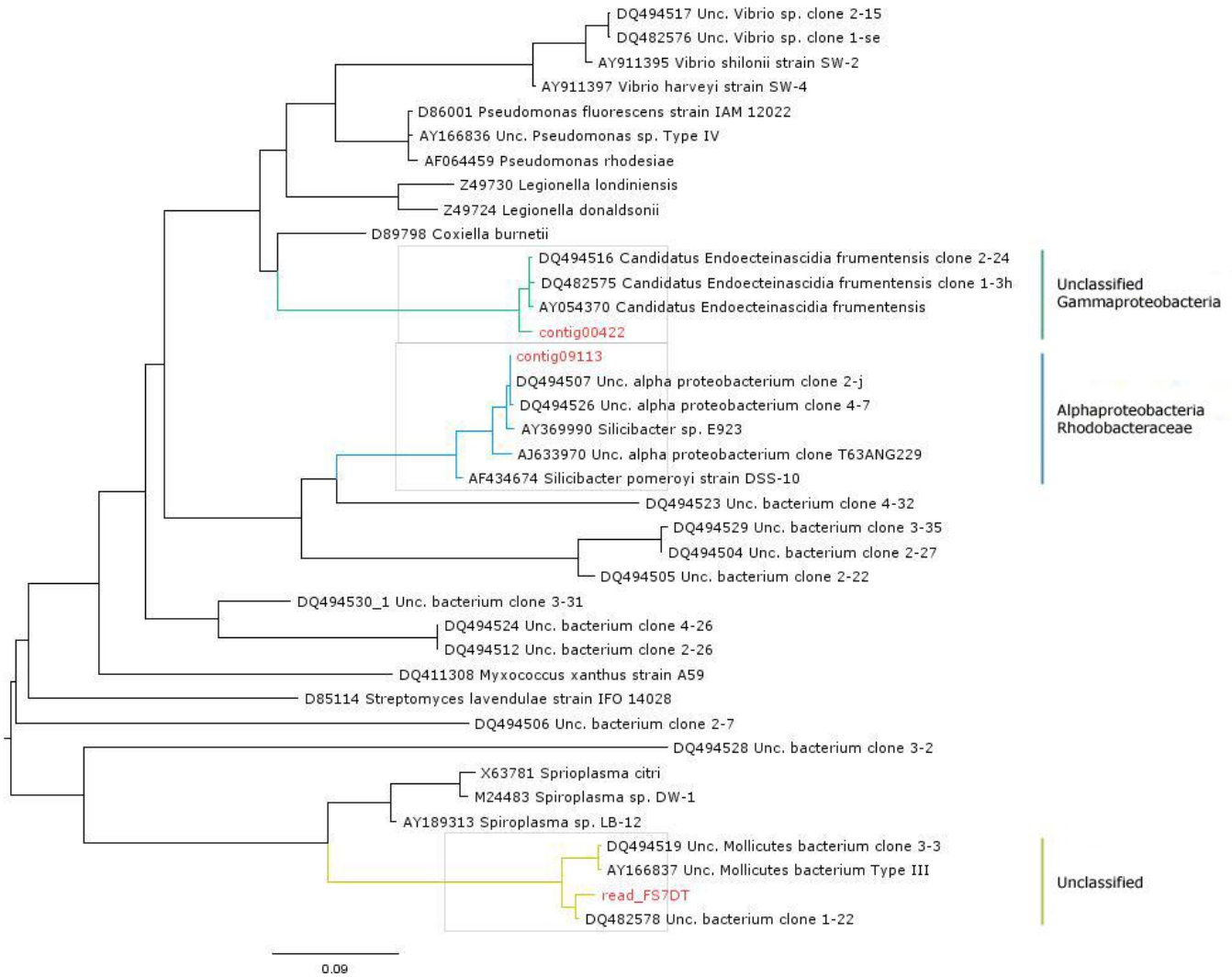


Figure 5-3. Multiple sequence alignment tree. 16S rRNA gene sequences reported in previous *E. turbinata* analyses[24,25] were aligned with 16S rRNA gene sequences representing the most abundant bacterial populations in our tunicate samples. A 16S rRNA gene-containing contig (00422) clusters with previously identified *E. frumentensis*.

We then sought to link the putative ET-743 35 kb biosynthetic gene cluster to the *E. frumentensis* 16S contig00422 by evaluating the codon usage bias. Bacteria typically do not employ synonymous codons equally and this can be exploited as a unique marker.[35] We performed a Relative Synonymous Codon Usage (RSCU) analysis using the annotated NRPS contig and contig00422 as well as ORFs identified in several contigs chosen at random. The RSCU score is the observed frequency of a codon divided by the frequency expected for equal usage of all synonymous codons, thereby making it a

measure of non-randomness.[35] RSCU scores for each codon are similar between the genes on the contig bearing the presumed NRPS biosynthetic genes and the *E. frumentensis* 16S rRNA gene-containing contig00422, but vary compared to RSCU scores from genes located in the random contigs from the total assembly (**Figure 5-4**). The extremely low GC content of the contig bearing the putative ET-743 NRPS genes (~23%) closely matches the GC content (26%) of the contig bearing the 16S rRNA gene corresponding to *E. frumentensis*, providing another strong marker of genetic linkage. On the other hand, Rhodobacteraceae appear to have uniformly high GC content (54% - 70%) according to current whole genome sequencing data, indicating that the contig containing NRPS genes is unlikely to be linked to this organism. The only fully sequenced and annotated tunicate genome, *Ciona intestenilis*, is 35% GC (NZ_AABS00000000). To account for GC bias in codon usage we included random genes from the low GC bacterium (~29%) *Clostridium botulinum* str. Okra. A comparison of the mean RSCU values for each codon revealed that only 12/60 values differed significantly ($p<.05$) between the putative ET-743 NRPS and contig00422 genes while 18/60 differed between the putative NRPS genes and random genes from *C. botulinum*. The significant differences between *C. botulinum* genes are most evident in the codons encoding isoleucine (AUU, AUC, AUA), lysine (AAA, AAG), aspartic acid (GAU, GAC), glutamic acid (GAA, GAG) and arginine (CGU, CGC, CGA, CGG, AGA, AGG). 49/60 codons differed significantly between the putative NRPS genes and random tunicate metagenome genes. In addition to RSCU analysis we used the contig containing the 25 predicted ET-743 pathway genes in a correspondence analysis using codonW to generate a codon adaptive index (CAI). This index was then used as a reference for

comparison with the same genes used in the RSCU analysis. Although all CAI scores differed significantly from the NRPS contig CAI score ($p<.05$), the *C. botulinum* CAI score and random gene CAI scores differed to a larger degree (**Figure 5-5).** We also analyzed the contig bearing the NRPS genes and contig00422 with the Naïve Bayesian Classifier (NBC) tool, a composition-based metagenome fragment classifier that uses N-mer frequency profiles.[36] NBC analysis based on 3- and 6-mer profiles results in high confidence classification of both contigs as γ-proteobacteria/Enterobacteriaceae. This same *E. frumentensis* 16S rRNA gene sequence has now been linked to *E. turbinata* collections from the Mediterranean, Caribbean, and Florida Keys. Taken together, these data suggest that the sequence contig bearing NRPS module genes are derived from the same organism as contig00422 (*E. frumentensis*).

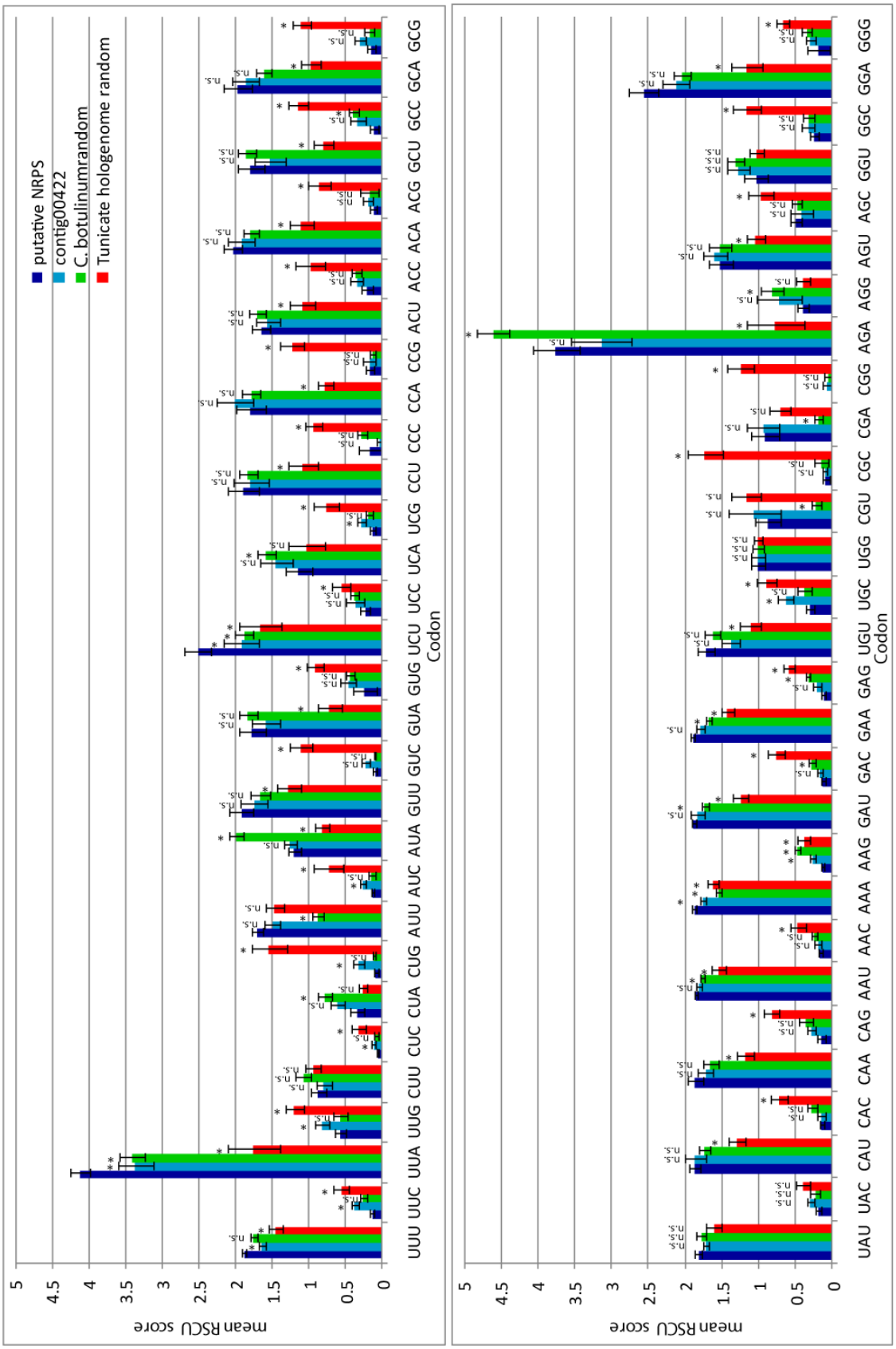

putative NRPS
contig00422
C. botulinumrandom
Tunicate hologenome random
mean RSCU score
Codon
UUU UUC UUA UUG CUU CUC CUA CUG AUU AUC AUA GUU GUC GUA GUG UCU UCC UCA UCG CCU CCC CCA CCG ACU ACC ACA ACG GCU GCC GCA GCG
UAU UAC CAU CAC CAA CAG AAU AAC AAA AAG GAU GAC GAA GAG UGU UGC UGG CGU CGC CGA CGG AGA AGG AGU AGC GGU GGC GGA GGG
n.s.
*
**

Figure 5-4. Relative Synonymous Codon Usage (RSCU) analysis. ORFs from the putative NRPS contig (26 ORFs), contig00422 (20 ORFs), random genes from *C. botulinum* Okra (34 ORFs) and from random contigs in the shotgun assembly (33 ORFs) were submitted to codonW for RSCU analysis. Codon preference is most similar between genes in contig00422 and genes in the putative NRPS (blue), whereas a random sampling of genes from the holo-metagenome shows a very different pattern of codon preference (red). Although RSCU values in the putative NRPS contig and those from *C. botulinum* (green) are also similar, several codons show large differences. Start and stop codons were omitted. Error bars display the standard error of the mean. A paired 2-sample T- test was used to compare means, *p<.05, n.s. (not significant).

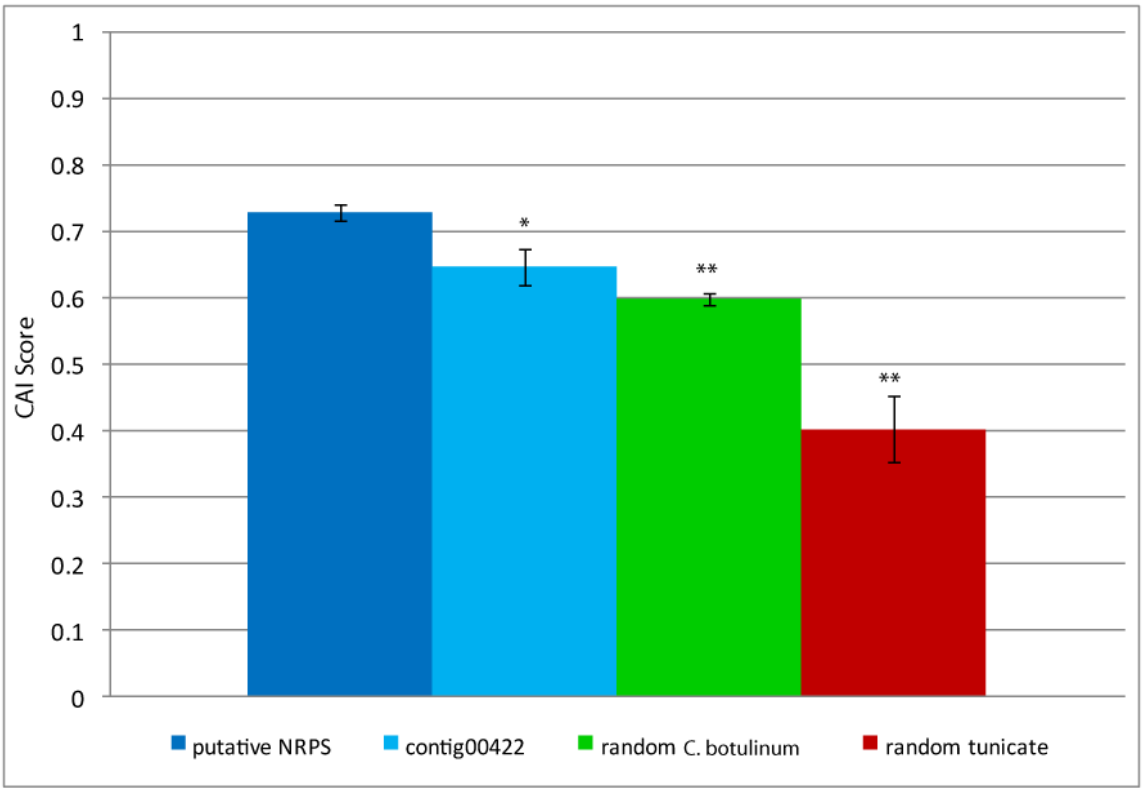


Figure 5-5. Codon Adaptive Index (CAI) scores. CodonW was used to perform a correspondence analysis on the 26 putative NRPS genes to generate a Codon Adaptive Index (CAI). This index was then used as a reference for calculation of a CAI score (value from 0 to 1) using the 26 NRPS genes, 20 genes on contig00422, 34 random genes from *C. botulinum* str. Okra and 33 random genes from the tunicate holo-metagenome. Error bars display the standard error of the mean. Mann-Whitney Test for significance, *p<.05 **p<.0001

EtuA1, EtuA2, EtuA3 are three predicted NRPSs with catalytic domains bearing predicted amino acid specificity motifs.[37] Sequence analysis and deep

annotation revealed that biosynthetic pathway architecture is non-collinear (as with SafA-B) and is represented by EtuA3→EtuA1→EtuA2. EtuA3 (AL-T-C-A-T) contains the AL-T starter module that is common to the saframycin, and saframycin Mx1 metabolic systems. The role of this module was elucidated for the saframycin biosynthetic pathway, where acylation of the precursor is required for further chain extension, cyclization and RE processing (Pictet-Spenglerase).[23] The NRPS A-domain, based upon the amino acid specificity motif,[37] was predicted to utilize cysteine (DLYNLSI, **Table 5-6**) with 100% sequence identity to the top three cysteine A domain sequence motifs. EtuA3 specificity is, therefore, unique to the Etu biosynthetic pathway and a key differentiator compared to other characterized tetrahydroisoquinoline systems, which all utilize alanine (DLFNNALT, **Table 5-6**). EtuA1 (C-A-T) has the greatest homology to SafA module 1 by BLASTx; however, the protein sequence identity and similarity are relatively low (29/54) compared to the other NRPSs in the pathway. An A-domain selectivity motif cannot be identified in EtuA1. Based on structural analysis of ET-743, a glycolic acid unit may be loaded and activated by the EtuA1 A-domain. Loading of hydroxy acids and formation of esters by NRPS modules have been characterized previously.[38,39] This extender unit represents another key difference relative to characterized tetrahydroisoquinoline antibiotics, for which a conserved core motif (7/8 amino acid identity) is both predicted and observed to select glycine (**Table 5-6**). EtuA2 (C-A-T-RE) contains the same A-domain specificity motif (DPWGLGLI, **Table 5-6**) for the final NRPS module as all known tetrahydroisoquinoline biosynthetic pathways. As verified in the saframycin biosynthetic system,[23] the EtuA2 homolog SfmC iteratively extends two 3H-4*O*-Me-5Me-Tyr residues. The terminal EtuR2 RE domain serves as a key marker of

the pathway and was examined biochemically to assess its activity in elaborating the tetrahydroisoquinoline core molecule.

Pathway	NRPS m1 A-domain	Predicted substrate	Expected substrate	NRPS m 2 A-domain	Predicted substrate	Expected substrate	NRPS m3 A-domain	Predicted substrate	Expected substrate
Etu (ET743, **1**)	DLYNLSLI	Cysteine	Cysteine	-	-	Glycolic acid	DPWGLGLI	3H-4-OMe-5Me-Tyr	3H-4-OMe-5Me-Tyr
Sfm (Saframycin **2**)	DLFNNALT	Glycine	Alanine	DILXLGLI	Glycine	Glycine	DPWGLGLI	3H-4-OMe-5Me-Tyr	3H-4-OMe-5Me-Tyr
Saf (saframycin Mx1, **3**)	DLFNNALT	Glycine	Alanine	DILXLGLV	Glycine	Glycine	DPWGLGLI	3H-4-OMe-5Me-Tyr	3H-4-OMe-5Me-Tyr
Sac (safracin, **4**)	DLFNNALT	Glycine	Alanine	DILQLGLI	Glycine	Glycine	DPWXLGLI	3H-4-OMe-5Me-Tyr	3H-4-OMe-5Me-Tyr

Table 5-6. A-domain specificity motifs for tetrahydroisoquinoline NRPS biosynthetic enzymes. The specificity determining motifs for the three key NRPS genes in previously described and ET-743 pathways are given. Bioinformatics derived NRPS A-domain specificity predictions are reported whereas expected substrate is based on analysis of the final natural product structure. The predicted amino acid incorporation (based upon natural product structure) for (**2-4**) is alanine-glycine-(3-hydroxy-4-*O*-methyl-5-methyltyrosine)$_2$, whereas for ET-743 (**1**) the proposed substrate incorporation, is cysteine-glycolic-acid-(3-hydroxy-4-*O*-methyl-5-methyltyrosine)$_2$.

DNA processing enzymes. Although unusual in natural prouct pathways, we hypothesize that EtuD1-3 may have a role in repairing damage induced by ET-743. EtuD1 appears to be a homolog of the TatD Mg^{2+} dependent DNAse[40] while EtuD2 shows similarity to a DNA polymerase III subunit δ', which has been characterized as part of the DNA-enzyme assembly complex. EtuD3 is a homolog of the 5'→3' exonuclease domain from DNA polymerase I.

Fatty acid processing enzymes. Pathway components that mediate production of essential cofactors or substrates are often encoded within biosynthetic gene clusters. EtuF1 and EtuF2 appear to represent subunits of an acetyl-CoA carboxylase. These enzymes transform acetyl-CoA to malonyl-CoA for fatty acid biosynthesis, and may supply substrate for synthesis of the fatty acid for EtuA3 AL. EtuF3 appears to be a penicillin acylase.[41] We propose that this key enzyme may act to release the predicted fatty acid modified intermediate of ET-743 after formation of the tetradepsipeptide and Pictet-Spengler cyclization (**Fig. 4)** prior to further processing into mature intermediates that are isolable from the tunicate.

Generation of 3-hydroxy-4-*O*-methyl-5-methyl-tyrosine (hydroxylase and methyltransferases). ET-743 is derived from at least two units of the unusual amino acid 3H-4*O*-Me-5Me-Tyr. The intermediate may be generated through 3-hydroxylation, 4-*O*-methylation, and 5-methylation of tyrosine. EtuH, an SfmD homolog, is predicted to hydroxylate tyrosine at the 3-position, whereas EtuM1, a SacF homolog, may be a SAM-dependent methyltransferase and a candidate for C-methylation at the 5-position. SafC, an EtuM2 homolog, has been characterized *in vitro* as a catechol 4-*O*-

methyltransferase.[42] Biochemical studies in the saframycin pathway revealed that SfmD (EtuH homolog), SfmM2 (EtuM1 homolog) and SmfM3 (EtuM2 homolog) form a minimal unit for 3H-4*O*-Me-5Me-Tyr production thus diverting tyrosine to secondary metabolism.[43]

EtuO is an FAD-dependent monooxygenase that shows high similarity to SfmO2 and SacJ. EtuO may catalyze modification of the tetrahydroisoquinoline to produce the hydroxylated species based on previous work involving *sacJ* gene disruption (**Fig. 4** 17-18).[20] *In vitro* biochemical characterization of this enzyme will require synthesis of an advanced intermediate to determine its precise activity.

Regulatory enzymes. EtuR1 has significant similarity (59%) to S29x, a protein previously shown to have a role in host-symbiont interactions between *Amoeba proteus* and the symbiotic Gram-negative X-bacteria.[44] This fascinating protein is excreted from the bacterium, and localized to the *A. proteus* nucleus.[45] The role of S29x in host-symbiont interactions is unclear with no other homologs characterized. The presence of a homolog to a characterized symbiont-derived gene in the Etu cluster suggests that regulated host-symbiont interactions may be involved in ET-743 biosynthesis. BLASTx analysis of EtuR2 shows (34/58%) identity/similarity to a MerR family transcriptional regulator. This class of regulators has been found in diverse classes of bacteria and responds to toxic effectors including heavy metals and antibiotics.[46] EtuR3 resembles the TraR/DksA transcriptional regulator that functions as a DNAK suppressor protein.

EtuT appears to be a drug transporter protein. Members of this superfamily are commonly represented in natural product biosynthetic pathways and could serve as part of a resistance/export mechanism for ET-743.[47]

Gene products of unknown function. EtuU1 is related to a putative EtuP peptidase modulator of DNA gyrase, whereas EtuU2 appears to be a shikimate kinase I. EtuU3 is an unknown hypothetical protein. EtuN1, EtuN2, and EtuN3 appear to encode the three subunits of a Glu-tRNAGln amidotransferase.[48] This enzyme forms correctly acylated Gln-tRNA^{Gln} by transamidation of aberrant Glu-tRNA^{Gln}. The role of these genes in the ET-743 pathway is unknown. EtuP1 and EtuP2 form two components (E1 and E2) of a possible pyruvate dehydrogenase complex. Pyruvate dehydrogenase catalyzes the transformation of pyruvate into acetate, but its function remains unclear in the Etu pathway.

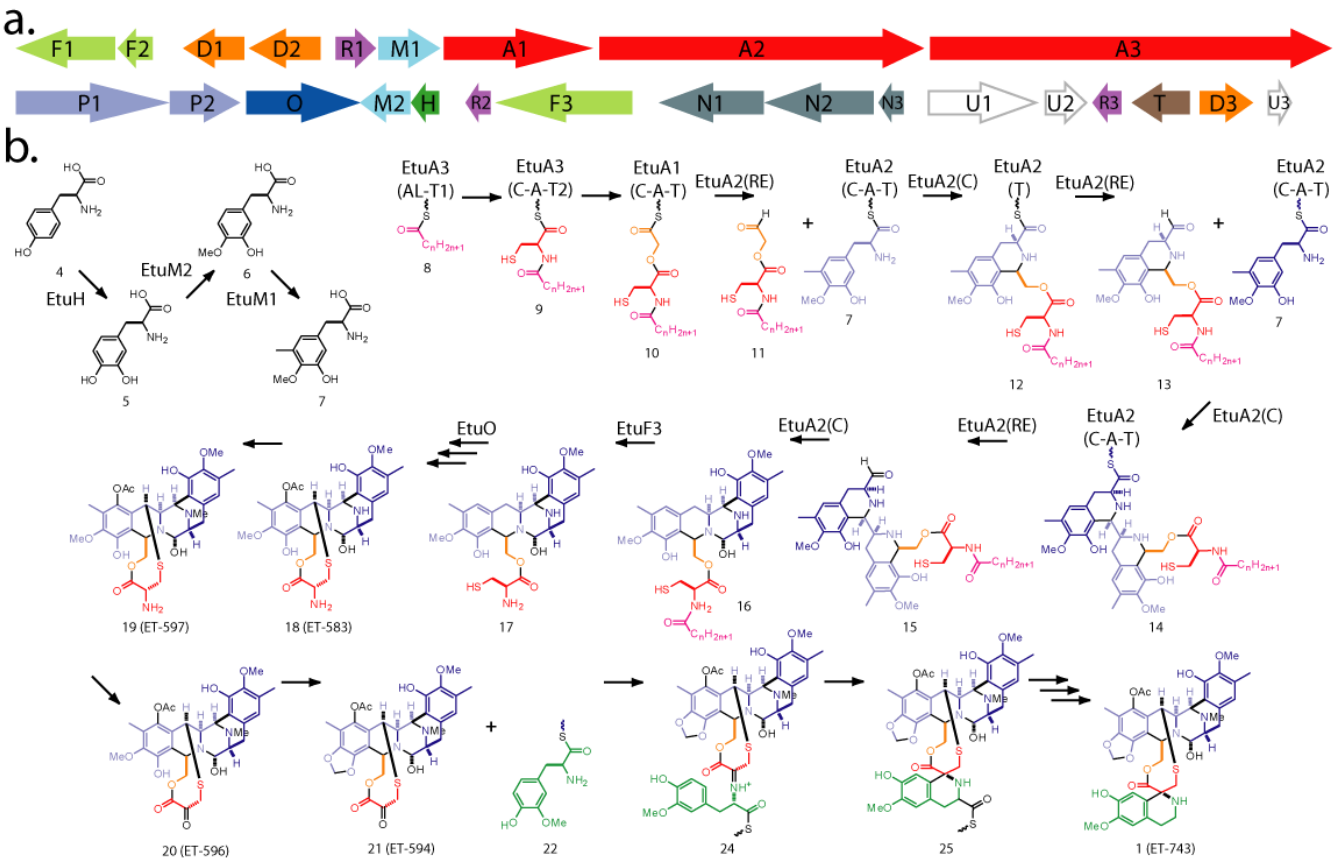


Figure 5-6. ET-743 biosynthetic gene cluster. (A) Gene names relate to proposed function for each protein: **EtuA**-NRPS, **EtuD**-DNA processing, **EtuF**-fatty-acid enzymes, **EtuH**-hydroxylase, **EtuM**-methyltransferases, **EtuN**-amidotransferases, **EtuO**-monoxygenase, **EtuP**-pyruvate cassette, **EtuR**-regulatory enzymes, **EtuT**-drug transporter, **EtuU**-unknown function. Proposed biosynthetic pathway for ET-743. (**B**) Named intermediates (characterized), enzymes (if assigned) and enzyme intermediates (thioester-bound) are shown.

Proposed scheme for ET-743 biosynthesis. Our scheme begins with assembly of the key subunit 3H-4*O*-Me-5Me-Tyr (**7**) (**Figure 5-6**). This non-proteinogenic amino acid is formed by 3-hydroxylation of (**4**), 4-*O* methylation of (**5**), and 5-methylation (**6**) catalyzed by EtuH, EtuM2, and EtuM1,[42] respectively. Next, the fatty acid CoA ligase of EtuA3 loads a fatty acid (**8**) onto the T domain. This fatty acid is condensed with cysteine that is activated and loaded by the C-A-T module of EtuA3 (**9**). Cysteine condenses with a T-loaded glycolate on EtuA1 to form the acylated-depsipeptide (**10**). Based on Koketsu's model, (**10**) is reductively released by the EtuA2 RE-domain as an

aldehydo-depsipeptide (**11**) from the EtuA1 T-domain. Such a "reach-around" model is not unprecedented in natural product biosynthesis.[49] EtuA2 loaded with 3H-4*O*-Me-5Me-Tyr (**7**) is then condensed with (**11**) to form the cyclic aldehydo-tridepsipeptide (**12**) through the presumed Pictet-Spenglerase activity of the EtuA2 C domain. Intermediate (**12**) is released from the EtuA2 T by the RE-domain activity as an aldehyde (**13**). Based on Koketsu's model it is proposed that EtuA2 catalyzes a second Pictet-Spengler reaction between another unit of 3H-4*O*-Me-5Me-Tyr (**7**) and (**13**). The protein-bound tetradepsipeptide (**14**) is then reductively released to form aldehyde (**15**) that may undergo a further enzyme-catalyzed Pictet-Spengler reaction to form the fatty-acid bound carbinolamine pre-ET-743 (**16**). The penicillin acylase EtuF3 (**16**) is then proposed to cleave the fatty acid unit, which may serve to sequester substrate in the EtuA2 active site during repeated loading/release, forming pre-ET-743 (**17**). Proposed intermediates ET-583 (**18**), ET-597 (**19**), ET-596 (**20**), and ET-594 (**21**) have all been isolated, characterized,[30] and all except ET-596 (**20**) have been confirmed by our secondary metabolite analysis (**Fig. S3**). We propose that pre-ET-743 (**17**) is hydroxylated by EtuO, acetylation and formation of the thioether ring are both catalyzed by unknown enzymes/mechanisms and intermediates to form ET-583 (**18**). An unidentified *N*-methyltransferase acts on ET-583 (**18**) to generate ET-597 (**19**). In accordance with Sakai, we propose that a transamination reaction proceeds on (**19**) to produce ET-596 (**20**). Another unknown protein catalyzes formation of a methylene dioxybridge in the A ring to generate ET-594 (**21**). Since compounds (**18 - 21**) are isolable, and tryptophan analogs of ET-743 have also been observed,[30] it is reasonable to propose that the final subunit to complete biosynthesis of the drug is added at a late stage, perhaps by formation

of an imine to the β-carbonyl and the new tyrosine analog. In both ET-743 total synthesis and semi-synthesis schemes, the α-ketone (**21**) is transformed to the final tetrahydroisoquinoline ring system by addition of 4-*O*-methyl-tyrosine under mild conditions.[8,10] Further processing steps are hypothetical, with neither enzyme nor intermediate identified. We propose that another tyrosine analog, 4-*O*-methyl-tyrosine (**22**) is condensed with ET-594 (**21**). The proposed imine intermediate (**24**) may then undergo another Pictet-Spengler-type reaction to form the final ring system (**25**). It is unknown if this unusual cyclization reaction and reduction is catalyzed by EtuA2 or an additional enzyme. The mechanism by which the proposed thioester of ET-743 is released from the proposed enzyme as ET-743 (**1**) remains to be established. Full validation of this proposed pathway will require synthesis of the predicted enzyme substrates, and direct biochemical analysis.

Biochemical confirmation of a key enzymatic activity. The transformation of thioester-bound acylated-depsipeptide (**10**) to the aldehydo-didepsipeptide (**11**) is a key enzymatic step thought to be catalyzed by the EtuA2 RE domain. We, therefore, cloned and overexpressed the excised EtuA2 RE domain to test this activity. Koketsu and coworkers had shown the same activity for the matching saframycin substrate analogs (**25** → **26**) with the SfmC A-T-RE-tridomain.[23] Therefore, the known saframycin substrate analog (**25**) was synthesized and transformed to the previously characterized saframycin aldehydo-dipeptide (**26**) by the EtuA2 RE domain (**Fig. 5-7**). As a positive control, substrate (**25**) was converted with high efficiency to (**26**) by purified apo-SfmC (C-A-T-RE). The differential activity was expected as (**25**), while clearly a well-tolerated substrate, is missing the cysteine-derived thiol and has a glycine in place of the glycolic

acid compared to the native substrate (**10**). Experimental data was in agreement with a synthetic authentic standard aldehydo-dipeptide (**27**). Confirmation of RE enzyme activity links our predicted biochemical scheme to demonstrated function in the ET-743 biosynthetic pathway.

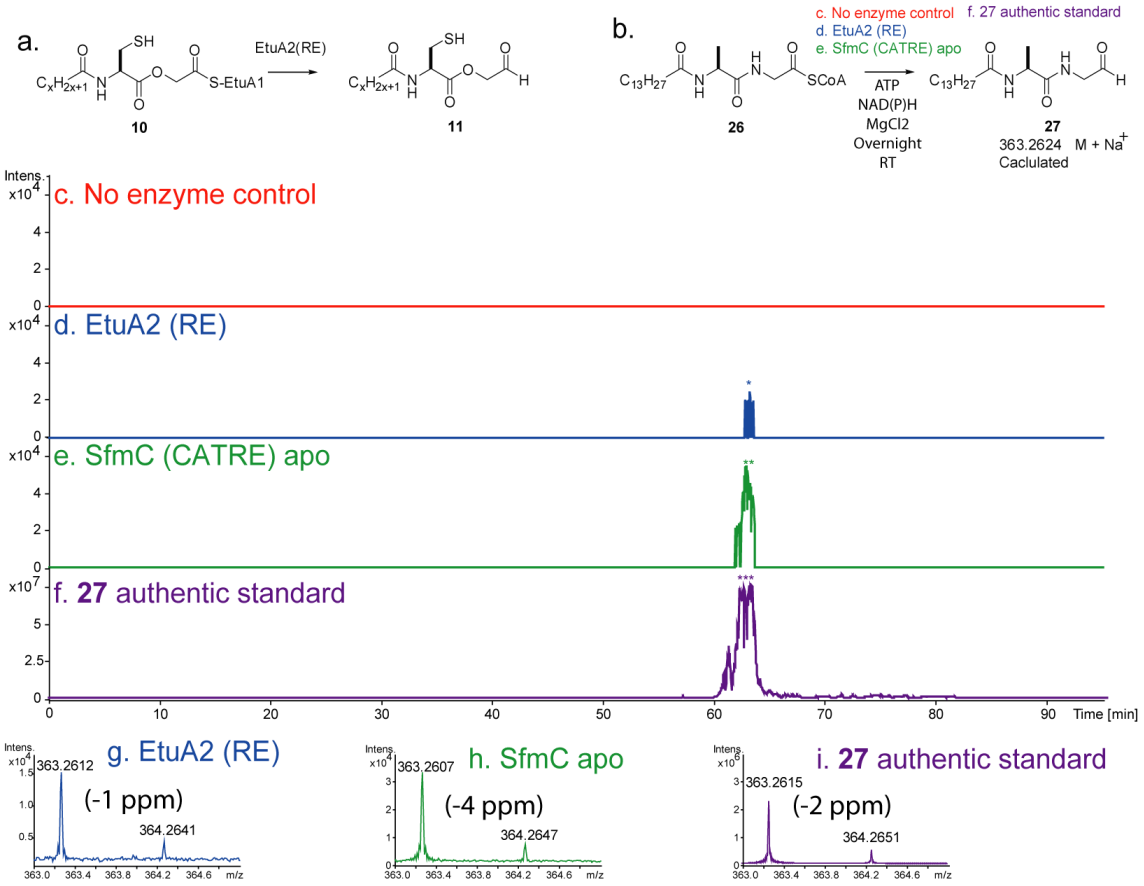


Figure 5-7. EtuA2 RE and SfmC reactions with (26). (**A**) The proposed biochemical activity of EtuA2 RE-domain in transforming activated didepsipeptide acyl-thioester (**10**) to the aldehyde (**11**). (**B**) The analogous reaction for SfmC is the transformation of (**26**) to (**27**) as reported by Koketsu.[23] The reaction of (**26**) to (**27**) was investigated with no enzyme control (**C**), EtuA2 RE-domain (**D,G**), SfmC (**E,H**), and an authentic standard of (**27**) (**F,I**). The aldehydo-dipeptide product (**27**) was monitored as the Na^+ adduct in positive ion mode by LC-FTICR-MS with an EIC at +/-20 ppm.

Metaproteomics to identify ET-743 biosynthetic proteins. Total tryptic peptides from the field-collected *E. turbinata* sample were fractionated by strong-cation-exchange chromatography, desalted, and then analyzed by reverse phase nano-LC MS/MS. Datasets were collected on LTQ-Orbitrap and 12T Q-FTICR mass spectrometers, with high-resolution/mass-accuracy MS1 spectra (and MS2 for FTICR only). Data were processed in Trans Proteomic Pipeline[50] with four distinct search engines (X!tandem, OMSSA, Inspect, and Spectrast) and the Peptide and Protein Prophet probability models with false discovery rates at the protein level of 0.6-0.9%. The database searched consisted of a six-frame translation of the total metagenome assembly filtered to contain all possible polypeptides >60 amino acids in length (SI). Sequence length-based cutoffs were utilized rather than ORF prediction due to the short length of many metagenomic contigs derived from the 454 metagenomic sequencing. Filtering resulted in a six-fold reduction in total sequence length versus the unfiltered six-frame translation. A 60 amino acid cut-off represents a 0.2% chance of any random sequence producing a translation without a stop codon appearing. Based upon 23S/16S analysis the closest fully sequenced organisms to the four principle constituents of the assemblage were included to assign homologous proteins derived from genes that may have been incompletely sequenced in the metagenomic analysis (tunicate: *Ciona intestinalis NZ_AABS00000000*, α-proteobacteria: *Ruegeria pomeroy*i DSS-3 NC_003911, γ-proteobacteria: *Coxiella burnetii* RSA 331 NC_010115, unknown bacteria: *Mycoplasma mycoides* subsp. mycoides SC str. PG1 NC_005364). Reversed sequences for all proteins were included as decoys in the search database.

A total of 289 proteins were identified at a probability >95% from Inter Prophet pooled analysis of all four search engines prior to protein prophet analysis **(Table 5-7** and **5-8**). Three of the proteins identified were derived from the Etu pathway with two identified by Orbitrap and one by FTICR and Orbitrap MS. The penicillin acylase EtuF3 was identified with two unique peptides, 3+ TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (N_{115}=deamidated) and 2+ RPIELR, and the protein was identified in 3/4 search engines providing a total protein probability of 99.99%. The bacterial symbiont protein EtuR1 was identified with two unique peptides, 2+ GSNIHYDLENDHNDYEK and 3+ GSNIHYDLENDHNDYEK, identified by 3/4 search engines at the protein level with a combined protein probability of 100.00% The EtuM1 SAM dependent methyltransferase was identified by one unique peptide, 2+ LLDVGGGTAINAIALAK and 2/4 search engines at the protein level with a probability of 99.16%.

	Iprophet	**X!tandem**	**OMSSA**	**Inspect**	**Spectrast**
Correct	289	230	142	176	2
Incorrect	2	2	1	1	0
FDR (protein)	0.7%	0.9%	0.7%	0.6%	0.0%

Table 5-7. Metaproteomics protein identification. Data are presented for each of the four search engines utilized, as well as the pooled Inter prophet data at a protein probability of >95% as calculated in Prophet Prophet. The calculated number of true and false postives as well as false discovery rates are reported as output by the appropriate probability models.

Protein:	Iprophet	X!tandem	OMSSA	Inspect	Spectrast
EtuF3	99.99%	99.99%	66.32%		99.99%
EtuM1	99.16%	98.99%		98.98%	
EtuR1	100.00%	100.00%	99.06%		100.00%

Table 5-8. Assignment of Etu proteins. Assigned Protein Prophet probabilities are provided for each of the three Etu proteins identified both from the combined Inter Prophet analysis, as well as for each of the individual search engines.

All identified Etu peptides were validated by comparison with synthetic peptide standards by LC elution time (±2 minutes on the same nano-LC system), and MS/MS fragmentation spectra (**Figures 5-8** and **5-9**). Detailed spectral information is provided (**Tables 5-20 through 5-25**, **Figure 5-13 through 5-31**). These three biosynthetic proteins identified with high probabilities and confidence levels by multiple search algorithms and comparison with authentic standards strongly suggests that ET-743 biosynthetic genes are expressed in the tunicate microbial symbiont assemblage.

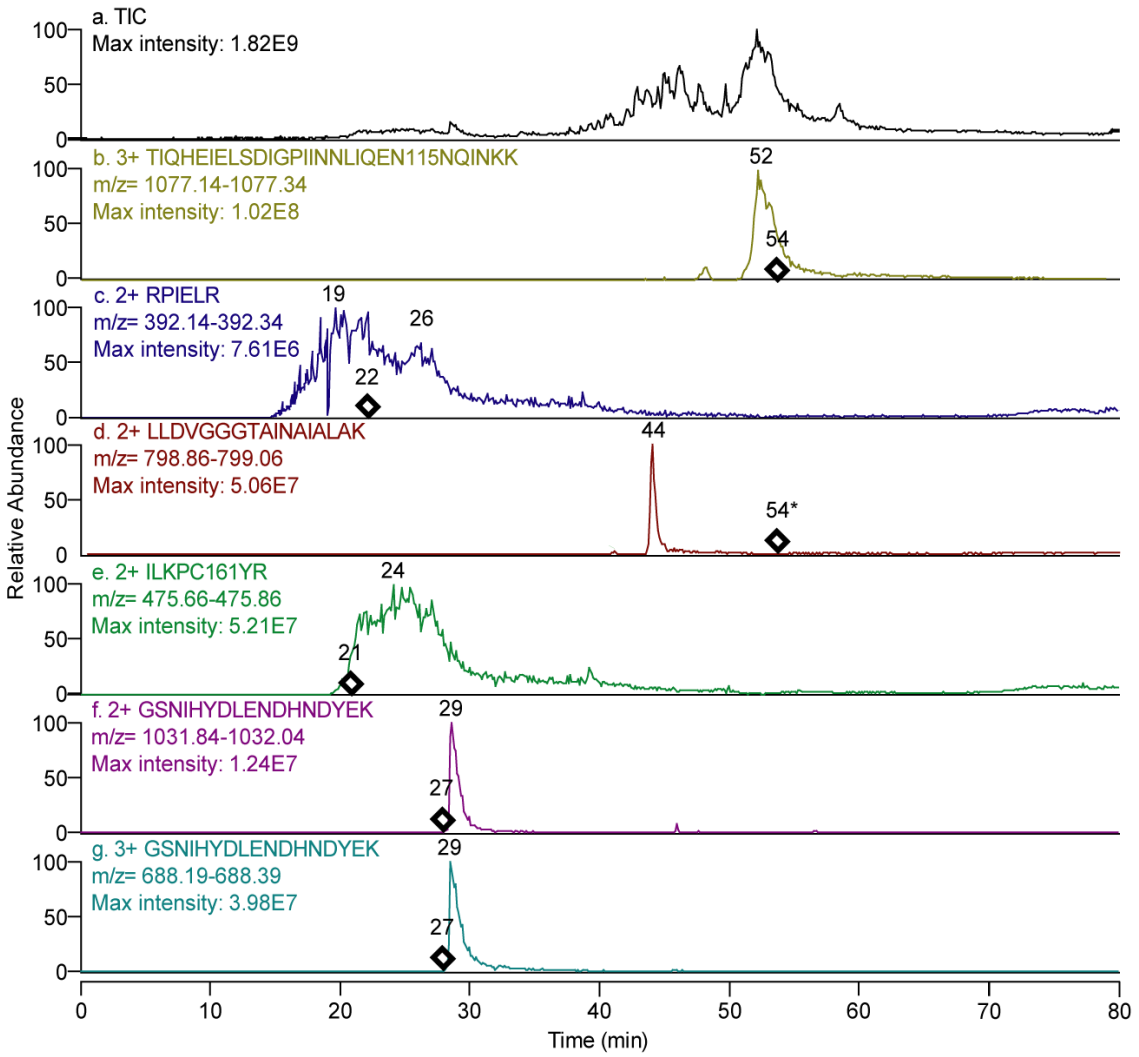


Figure 5-8. Synthetic peptides as authentic standards to verify metaproteomics peptide assignments. (**A**) Total ion chromatogram for the standard peptide mixture on the LTQ-orbitrap. (**B-G**) extracted ion chromatograms generated at +/- 0.1 m/z for each of the synthetic peptides in the mixture. Chromatograms are presented as time versus normalized intensity. Maximum intensity in each normalized total or extracted ion chromatogram is noted. *Denotes that the experimental retention time for doubly protonated tryptic LLDVGGGTAINAIALAK was obtained on a different LC system with a different gradient and column compared to the authentic standard. In the case of all other synthetic standard versus experimental identifications the LC system and gradient were identical, although a different column was used. ◊ denotes the elution time of the experimental MS2 spectra assigned to each of the peptides.

The mixture of six synthetic peptides was analyzed on the LTQ-Orbitrap. Extracted ion chromatograms were generated at +/- 0.1 Da. For triply protonated

TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (**Figure 5-8B**), doubly protonated RPIELR (**Figure 5-8C**), and doubly protonated ILKPC161YR (**Figure 5-8E**) the metaproteomics spectra fall within the elution window of the authentic standard synthetic peptides. For doubly protonated GSNIHYDLENDHNDYEK (**Figure 5-8F**) and triply protonated GSNIHYDLENDHNDYEK (**Figure 5-8G**) the metaproteomics and standard retention times fall within two minutes of each other (within the range of experimental error). Doubly protonated LLDVGGGTAINAIALAK (**Figure 5-8D**) appears to show a 10 minute difference in retention time. However, this peptide was originally identified on a completely different LC system with a completely different gradient (Q-FTICR-MS in Ballerica MA versus LTQ-Orbitrap Ann Arbor MI). Unfortunately, the standard peptide could not be analyzed on the same instrumentation as originally identified on. These data still provide weak confirmation as both metaproteomics and authentic standard peptides did display late elution profiles in the two different LC systems and methods. Further data regarding the interpreted spectra for the metaproteomics versus authentic standard peptides are given below in **Figure 5-9**.

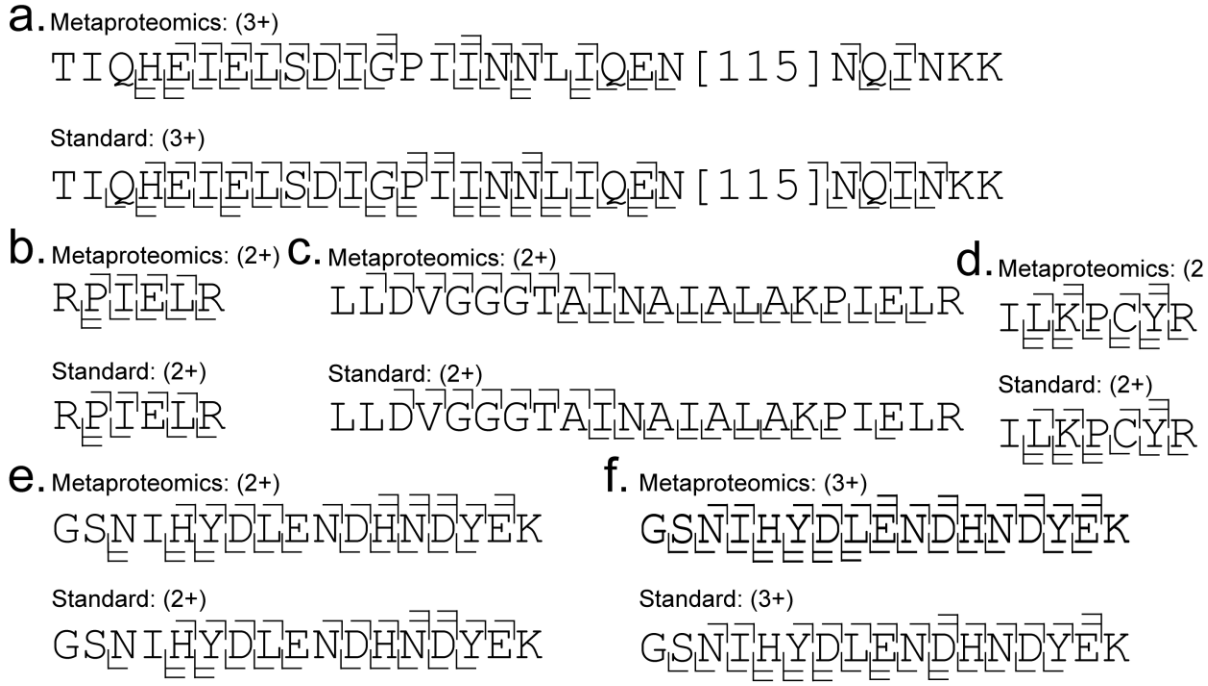


Figure 5-9. Peptide MS2 sequence coverage for metaproteomics versus authentic standard synthetic peptides. Only b and y ion assignments are shown although other ions (e.g., a, b - H_2O, b - NH_3, y - H_2O, and y - NH_3) could also be assigned. Multiple bars indicate that a given fragment can be assigned to multiple charge states

For all assigned peptides (**Figure 5-9**) the agreement between data obtained from metaproteomics and authentic standard peptides provides outstanding correlative evidence. This excellent agreement is further corroborated upon detailed inspection of each assignment below (**MS2 spectra for Etu peptides p215**). Those data reflect manual assignments of individual spectra. The authentic standard synthetic peptide data strongly support the metaproteomics peptide assignments in terms of elution profile (**Figure 5-8**), and MS2 spectral assignment (**Figure 5-9**) for each of the peptides investigated.

BlastP was used to search the three biosynthetic proteins identified from metaproteomics analysis against the NR database (limited to *Ciona* and bacteria) in an effort to identify the likely parental organism. A prokaryotic origin is suggested with bacteria, γ-proteobacteria, and proteobacteria assigned to EtuM1, EtuP3, and EtuR1,

respectively (**Table 5-9)** The contigs containing the encoded proteins were analyzed by MG-RAST and assigned specifically to γ-proteobacteria for all three proteins (**Table 5-10**). The same exercise was performed for the total proteins identified and taxonomy was assigned by BlastP (**Table 5-11**).

	E value	Protein	Organism	Full taxonomy	Ascension
EtuF3	0E+00	penicillin acylase	γ-proteobacteria	*Coxiella burnetii Dugway 5J108-111*	YP_001424243
EtuM1	9E-90	methyltransferase	bacteria	*Streptomyces anulatus*	ADG27364
EtuR1	6E-22	29kDa protein	proteobacteria	*Candidatus Legionella jeonii*	AAB39275

Table 5-9. Matched Etu peptides and proteins—BlastP derived protein taxonomy. Identified Etu proteins were searched against the NCBI NR database with BlastP restricted to *Ciona* and bacteria. For each protein identified expectation values, assigned protein function, organism class, full organism taxonomy, and accession numbers are provided.

	E value	%ID	start	stop	total	organism	full taxonomy
EtuF3	8E-158	53	447	16148	17479	γ-proteobacteria	*Photorhabdus luminescens subsp. laumondii TTO1*
EtuM3	8E-158	53	447	16148	17479	γ-proteobacteria	*Photorhabdus luminescens subsp. laumondii TTO1*
EtuR1	8E-158	53	447	16148	17479	γ-proteobacteria	*Photorhabdus luminescens subsp. laumondii TTO1*

Table 5-10. Matched Etu peptides and proteins—MGRAST of contig containing identified protein. Identified Etu proteins were matched to their parent contigs and MG-RAST results were tabulated. For each protein identified expectation value, % ID, match start, match stop, and total DNA contig length, organism, and full taxonomy are provided. MGRAST E values.

	# Assigned
Total	**391**
tunicate	283
bacteria	91
n/a	17
Bacteria	**91**
proteobacterial	71
n/a	20
Proteobacterial	**71**
α-proteobacterial	24
γ-proteobacterial	26
n/a	21

Table 5-11. Matched total proteins—BlastP derived protein taxonomy. Identified total proteins were searched against the NCBI NR database with BlastP restricted to Ciona and bacteria. Only BlastP results at the order/family level of taxonomy are provided.

The 289 proteins identified represent the minimum number of distinct DNA sequences that fully represent the assigned dataset, however as many as 391 distinct DNA sequences can be assigned to this same set of peptides (e.g. multiple copy genes, homologous genes, shared peptides). Of the maximum possible 391 proteins, 283 were tunicate derived, 91 were bacterial derived, and 17 produced no significant similarity (**Table 5-11**). Of the 91 bacterial proteins identified, 71 appeared to be of proteobacterial origin and 20 could not be assigned at the family level. Of the 71 proteobacterial proteins identified, 24 were α-proteobacterial, 26 were γ-proteobacterial origin, and 21 could not be assigned beyond proteobacterial. Apparent distribution of predominant species (tunicate, α-proteobacterial, γ-proteobacterial) roughly correlates with metagenomics/16S

rRNA genes with higher amounts of tunicate proteins observed than would be expected from a direct correlation of DNA and protein levels. More γ-proteobacterial proteins were observed compared to α-proteobacterial assuming DNA is proportional to expression level.

5.3 Discussion

The work described in this study was motivated by the outstanding opportunity to identify and characterize an enormous range of host-symbiont derived natural product systems that have remained refractory to analysis. The inability to culture the vast majority of bacterial and fungal symbionts (outside of their natural host or environmental niche) that produce secondary metabolites have limited our access to a huge genetic diversity relating to untapped chemical resources for therapeutic and other industrial applications. This includes complex marine (e.g., sponge, tunicates, dinoflagellates) and terrestrial (e.g., plant-microbe, biofilm, insect-gut, human-gut) microbial consortia where the presence of large populations of diverse microorganisms and their corresponding genomes that bear natural product gene clusters remain unexplored. This unique source of metabolic and chemical diversity will lead to important new basic knowledge, and also contribute to on-going drug discovery efforts against many disease indications. In order to initiate this meta-omic analysis, ET-743 was chosen as a model system due to the predicted genetic composition of core components of its biosynthetic pathway. This was based on the assumption of a highly conserved overall architecture from previously characterized pure culture bacterial-derived metabolic pathways for related

tetrahydroisoquinoline natural product scaffolds (e.g., safracin, saframycin C, saframycin MX1). Moreover, recent advances in next-generation sequencing and bioinformatic tools to assemble contigs from large metagenomic datasets, and analysis of proteomic data to identify low-abundance proteins enabled the approaches described in this report.

In these studies, several steps were taken to obtain evidence for identification of the ET-743 biosynthetic pathway and the corresponding producing microbial symbiont. First**,** the presence of the ET-743 natural product and intermediates were used as markers for the producing bacterium in the tunicate/microbial consortium. Second, codon usage similarity between the biosynthetic gene cluster and a contig containing a 16S rRNA gene sequence support *E. frumentensis* as the bacterial producer of ET-743. Direct functional analysis of a key biosynthetic enzyme confirmed its predicted catalytic assignment in the pathway. Finally, symbiont-derived expression of three ET-743 biosynthetic enzymes was confirmed by metaproteomic and bioinformatic analysis, enabling the direct correlation between natural product, the Etu gene cluster, and predicted biosynthetic proteins. This tiered strategy provides a general approach for future efforts to characterize orphan and target natural product biosynthetic systems from complex marine and terrestrial microbial assemblages including animal-microbe symbiont consortia, and dinoflagellates. Moreover, this initial characterization of 25 putative ET-743 biosynthetic proteins will enable future efforts to confirm the function of individual enzymes by direct biochemical analysis. This work also provides the first key step toward supplying ET-743 and analogs through heterologous expression in an amenable production host.

5.4 Supplement

Data Deposition

Sequence data have been deposited in Genbank, http://www.ncbi.nlm.nih.gov/genbank (Accession numbers HQ542106 and HQ609499). The proteomics dataset has been deposited in Tranche Proteomic commons: (https://proteomecommons.org/dataset.jsp?i=FOIwaTzxhqbiEK1DShCVR4shblJ4c%2BAR%2BKAKY3c5fBd7uFYC6Ti6pdjvPxSPK2VgaSHDTEzDPeu%2FyshMZLe9qMe2gooAAAAAAACkEg%3D%3D)with the password: tunicate_et

Materials and methods

E. turbinata sample collection

Specimens were collected in the Florida Keys (24^{0}39'31.9", -81^{0},25'20.1"), frozen on dry ice, and shipped. Samples were prepared by grinding 25 g of tunicate on N_2(l) in lysis buffer (50 mM HEPES, 300 mM NaCl, 10 mM imidazole 10% glycerol, 1 mM TCEP PH 7.5 pH 8).

Secondary metabolite identification by LC-FTICR-MS and confirmation by LC-MS/MS

Tunicate samples were deproteinized with MeOH (2:1 ratio,1 hr at -20 ^{o}C), the protein was removed by centrifugation (14,000 RCF x G), and the supernatant was concentrated 5-fold. 50 µL of this sample was analyzed on a Luna C18 100 Å 2x250 mm 5 µm column (Phenomenex). The following gradient was generated on an Agilent 1100 HPLC: 0 (98,2), 10 (98,2), 95 (2,98), 100 (2,98) and 105 (98,2). Values are given as Time (%A,

%B). Time is given in minutes with the total run time being 120 minutes at a flow rate of 0.2 mL/min. A column heater was operated at 50 ^{o}C. The flow was diverted for the first 10 minutes of the run. Buffer A consisted of 0.1% formic acid in DDI water and Buffer B consisted of 0.1% formic acid in acetonitrile.

FTICR-MS was performed on an APEX-Q (Apollo II ion source, 7T magnet, Bruker Daltonics). Data were gathered by ESI in positive ion mode (2,400 V, m/z 150–1,000, transient 128 K, 1 scan/spectrum) with external ion accumulation (0.33 s), dynamic trapping, and 1 ICR cell fill per spectrum. External calibration utilized HP-mix (Agilent). For FTICR-MS/MS experiments auto-MS/MS was selected with Q-isolation (10 m/z, 5 precursor ions, collision energy of -16 to -21 V). A peak list of possible ET-743 related metabolites was used for precursor ion selection. Data were processed in Data Analysis (Bruker Daltonics) and MS/MS spectra were interpreted manually. Metabolite peaks were detected over multiple samples and runs. Iontrap-MS/MS was performed with HPLC conditions as above, except with a Surveyor HPLC (ThermoFisher). An LTQ Deca XP Ion trap MS (ThermoFisher) was employed for data-dependent MS/MS (1 precursor ion scan, 400-1800 m/z, 7 MS/MS events, isolation width 3 m/z, normalized collision energy 35%). Data analysis was performed in Excalibur version 3.0 (Thermo) and MS/MS spectra were interpreted manually.

454 and 16S rRNA gene library construction and sequencing methods

Metagenomic DNA was extracted from frozen *E. turbinata* samples using a DNeasy Tissue kit (Qiagen). DNA was used to prepare a 16S rRNA gene targeted amplicon

library using primers (TGCTGCCTCCCGTAGGAGT and AGAGTTTGATCCTGGCTCAG) and a random shotgun 454 FLX library. Sequencing was performed on a Roche/454 Life Sciences FLX Sequencer. Later, a second shotgun library was prepared using the 454 Titanium upgrade. Tunicate raw sequencing reads from the first FLX run were assembled using the 454 Newbler assembler (v2.0.00.20). The second 454 Titanium sequencing run was assembled together with the first sequencing run data producing a second assembly (Newbler v2.0.01.14).

NRPS module identification

Reads/contigs were filtered by protein homology to the saframycin, saframycin Mx1, and safracin NRPS genes characterized in *S. lavendulae* (DQ838002), *M. xanthus* (U24657), and *P. fluorescens* (AY061859) using BLASTx/tBLASTn searches. Primers were designed from the ends of filtered sequences (VectorNTI 9, Informax) and PCR reactions were designed based on the location of the BLAST hit on the reference sequences. Sequencing of positive reactions with linked sequences was performed with Sequencher 4.9, Gene Codes. Flanking sequence from high interest contigs was obtained by restriction-site PCR (RS-PCR) including two rounds of PCR using a semi-degenerate primer in conjunction with nested primers of known sequence.[32]

Analysis of the metagenomic population

Classification of the raw reads and total assembly was performed with MG-RAST.[34] Sequences were classified by protein homology to a manually curated database (the SEED). The 16S rRNA gene amplicon sequencing run was analyzed by assembling the raw reads with an identity threshold of 95%. The assembled contigs were submitted to

the Ribosomal Database Project (RDP) for classification.[51] Multiple sequence alignment (MSA) for 16S rRNA gene sequences was performed by http://greengenes.lbl.gov[52,53] Default parameters were used except for minimum length (300) and minimum %identity (50%). Formatting was changed to "remove common alignment gap characters" to provide an equal length MSA. The correct tree-building model was selected,[54] and assembled using the maximum likelihood method with the HYK nucleotide substitution matrix and additional parameters selected by Modelgenerator (Phyml v2.4.4).[55] The cladeorgram was displayed (FigTree v1.3.1) using midpoint rooting and colored with clade annotation. Gene-finding was performed on the NRPS contig, *E. frumentensis* 16S contig (contig00422), and random contigs from the total shotgun assembly (AMIgene with manual curation).[56] Relative Synonymous Codon Usage (RSCU) analysis and CAI analysis was performed with codonW.[57] Further phylogenetic classification of the NRPS contig and contig00422 was performed using the Naïve Bayesian Classification Tool using the 3- and 6-mer setting.[58]

EtuA2 and SfmC cloning and expression

To make the SfmC over-expression construct, the *sfmC* gene was amplified using genomic DNA from *Streptomyces lavendulae* NRRL 11002 as template and SfmC_F (5'- GCAGAATTCCATATGGTGACCCGGCACGAGCC -3', *Nde*I site underlined) and SfmC_R (5'- TTTGGATCCAAGCTTTCATCGCTCCTCCTCCAGCGTGC -3' , *Hin*dIII site underlined) as primers. The PCR product was digested with *Nde*I and *Hind*III and cloned to the same sites of pET-28a to generate pET28a-sfmC. *PfuTurbo*® DNA Polymerase (Stratagene) was used in *sfmC* cloning.

To make the overexpression construct for the RE domain of EtuA3, the RE coding sequence (1,251 bp) was amplified via PCR using the metagenomic DNA mixture as template and EtA3RE_F (5'-GCAGAATTCCATATGACCTTGCAAAAAGAAGGAATTG-3', *Nde*I site underlined) and EtA3RE_R (5'- CGCGGATCCTCGAGTTATATTTTTTTCGGATGAGGAAAG-3', *Xho*I site underlined) as primers, digested with *Nde*I and *Xho*I, and further cloned to the same sites on pET28a to generate pET28a-RE. KOD DNA Polymerase (Novagen) was used in the cloning of the EtuA3 RE domain.

The *N*-His_6-tagged RE domain protein and the *N*-His_6-tagged SfmC protein expression constructs were separately transformed into *E. coli* BL21 (DE3) +pRare. The two strains were grown at 37 °C in 0.5 L TB medium to an OD_{600} of ~0.8 in 2 L flasks. The cultures were cooled to 18 °C, and isopropyl β-D-thiogalactopyranoside was added to a final concentration of 0.2 mM and grown 12-16 hr with shaking. The cells were harvested by centrifugation and frozen at -80 °C. Cell pellets were thawed to 4 °C and resuspended in 5X volume of lysis buffer (20 mM HEPES, pH 7.8, 300 mM NaCl, 20 mM imidazole, 1 mM Tris(2-carboxyethyl) phosphine (TCEP PH 7.5), ~20 mg CelLytic Express (Sigma-Aldrich)) before lysis via sonication. Centrifugation at 40,000*x*g for 30 min provided clear lysates. Proteins were purified using affinity chromatography with Nickel-NTA resin (Qiagen). Briefly, after filtration of the supernatant through 0.45 μm membrane, the solution was loaded onto a 1 mL gravity flow column. The column was washed with 10 column volumes of wash buffer (20 mM HEPES, pH 7.8, 300 mM NaCl, 50 mM imidazole, 1.0 mM TCEP PH 7.5, 10% glycerol) and eluted with 20 mM HEPES, pH 7.8,

300 mM NaCl, 400 mM imidazole, 1.0 mM TCEP PH 7.5, 10% glycerol. Fractions were pooled, concentrated, and loaded onto a PD10-desalting column (GE Healthcare Life Sciences) equilibrated with storage buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 1.0 mM TCEP PH 7.5, 20% glycerol). Fractions were combined, concentrated, frozen, and stored at -80 °C. Protein concentrations were calculated using A280 and predicted protein extinction coefficients. Proteins were approximately 80% and 95% pure by SDS-PAGE with yields of 1 mg/L for EtuA2 RE and 3 mg/L for SfmC (**Figure 5-10**).

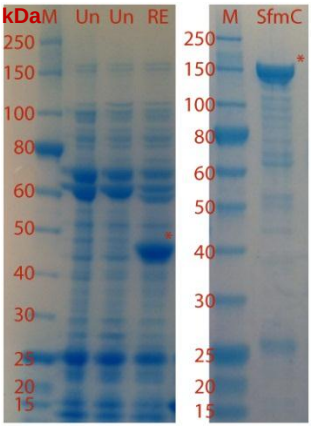
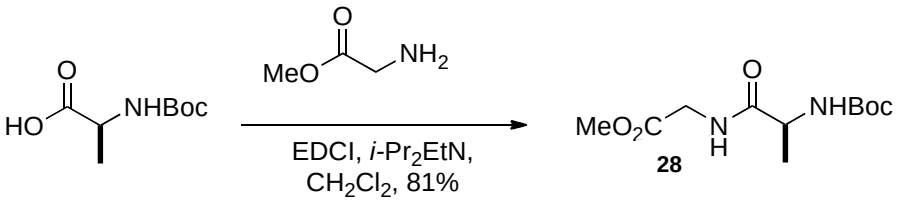
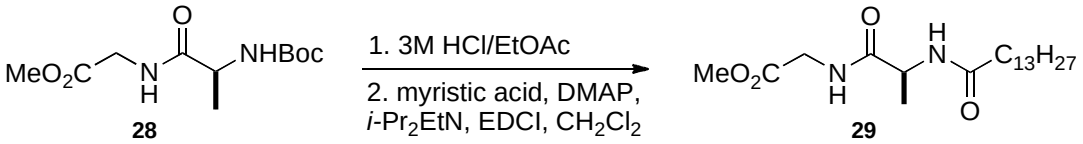


Figure 5-10. **4-12% NuPage gels stained with Simply Blue Safe Stain.** Samples include marker (M), control *E. coli* Ni-NTA elution without pET28a-RE (Un), EtuA2 RE (RE), and SfmC apo (SfmC).

Synthesis of substrate (26) for EtuA2 RE reactions.

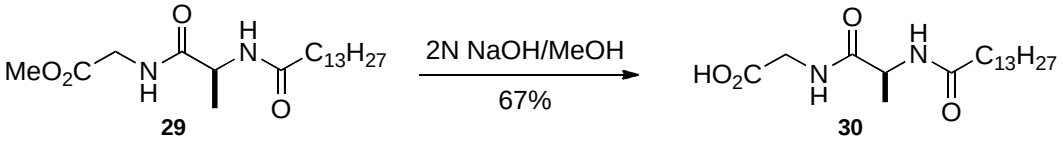


(*S*)-methyl 2-(2-((*tert*-butoxycarbonyl)amino)propanamido)acetate (28).[23] To a stirred solution of *N*-t-Boc alanine (3.15 g, 16.63 mmol, 1.2 eq) and glycine methyl ester•HCl salt (1.74 g, 13.83 mmol, 1.0 eq) in CH_2Cl_2 (46 mL) at 0 °C were added *N*,*N*-diisopropylethylamine (5.8 mL, 33.25 mmol, 2.4 eq) and EDCI (3.19 g, 16.63 mmol, 1.2 eq). The resulting suspension was stirred vigorously for 15 hrs, warming gradually to room temperature. The reaction was then quenched by addition of EtOAc (500 mL) and 0.5 M aqueous HCl (100 mL). The layers were separated, and the organic layer was washed sequentially with saturated aqueous $NaHCO_3$ (100 mL) and brine (100 mL). The organic layer was then dried (Na_2SO_4), concentrated, and purified by passage through a silica plug (30% EtOAc/hexanes, then 50% EtOAc/hexanes as eluent). *En vacuo* concentration then furnished 2.92 g (81%) of product (**28**) as a dark yellow oil. Data are: ^{1}H NMR ($CDCl_3$, 300 MHz) δ 7.23 (bs, 1H), 5.50 (d, *J*=3.3 Hz, 1H), 4.20 (m, 1H), 3.93, (d, *J*=2.85 Hz, 2H), 3.63 (s, 3H), 1.33 (s, 9H), 1.28 (d, *J*=3.6 Hz, 3H). ^{13}C NMR ($CDCl_3$, 75.5 MHz) δ 173.7, 170.4, 155.7, 80.0, 52.4, 50.1, 41.2, 28.4, 18.6.



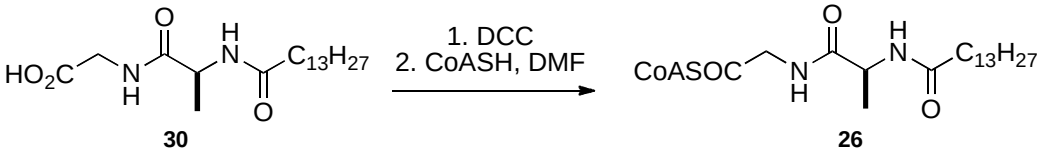
(*S*)-methyl 2-(2-tetradecanamidopropanamido)acetate (29).[23] Methyl ester (**28)** (973 mg, 3.74 mmol, 1.0 eq) was stirred vigorously in 3M HCl in EtOAc (22 mL) at RT for 30 minutes. It was then made basic by addition of Et_3N and was subsequently concentrated. The crude product was diluted with Et_2O and re-concentrated, then diluted again with Et_2O and re-concentrated to give the deprotected HCl salt intermediate. This was suspended in CH_2Cl_2 (15 mL) and cooled to 0 °C with vigorous stirring. To this

suspension were added myristic acid (939 mg, 4.11 mmol, 1.1 eq), DMAP (114 mg, 0.93 mmol, 0.25 eq), EDCI (1.43 g, 7.48 mmol, 2.0 eq), and *i*-Pr_2EtN (1.3 mL, 7.48 mmol, 2.0 eq). The resulting solution was then stirred overnight for 18 hours, warming gradually to room temperature. The reaction was quenched by addition of H_2O (10 mL), 3N H_3PO_4 (20 mL), and $CHCl_3$ (50 mL), and the layers were separated. The aqueous layer was extracted with $CHCl_3$ (3 x 50 mL). The combined organic layers were then washed with saturated $NaHCO_3$ and brine; dried ($MgSO_4$), concentrated, and purified by passage through a silica plug (EtOAc as eluent). The resulting yellow solid was then concentrated and purified by overnight recrystallization from hot MeOH/EtOAc to afford 460 mg (33%) of product **(29)** as an off-white solid. Data are: ^1H NMR ($CDCl_3$, 400 MHz) δ 6.82 (bs, 1H), 6.13 (bs, 1H), 4.61-4.52 (m, 1H), 4.01 (t, *J*=7.2 Hz, 2H), 3.75 (s, 3H), 2.20 (t, *J*=10.4 Hz, Hz, 2H), 1.64-1.56 (m, 2H), 1.39 (d, *J*=4.8 Hz, 3H), 1.32-1.25 (m, 20H), 0.87 (t, *J*=9.2 Hz, 3H); ^{13}C NMR ($CDCl_3$, 100 MHz) δ 174.2, 172.3, 169.7, 61.7, 48.7, 41.5, 36.7, 32.1, 29.9-29.6, 25.8, 22.9, 18.5, 14.3. HRMS calcd. for $C_{20}H_{39}N_2O_4$ $[M+H]^+$ 371.29, found 371.29.



(*S*)-2-(2-tetradecanamidopropanamido)acetic acid (30). Ester (**29)** (460 mg, 1.24 mmol, 1.0 eq) was stirred vigorously in 2N NaOH in anhydrous MeOH (6.2 mL) at room temperature overnight (14.5 hrs). Once the starting material was consumed (TLC, 100% EtOAc as eluent, bromocresol green stain), the reaction was concentrated, rediluted in H_2O (10 mL), and washed with Et_2O (2 x 5 mL). The aqueous layer was then carefully acidified to pH 2.0 with 1 N aqueous HCl and was extracted with EtOAc (3 x 20 mL).

The combined organic layers were dried (Na_2SO_4) and concentrated en vacuo to afford 296 mg (67%) of acid (**30)** as a white solid. Data are: ^{1}H NMR (MeOH-d_4, 400 MHz) δ 4.38 (q, J_1=3.6 Hz, 1H), 3.88 (dd, J_1=7.6 Hz, J_2=8.8 Hz, 2H), 2.21 (t, J=7.6 Hz, 2H), 1.61-1.55 (m, 2H), 1.33 (d, J=3.6 Hz, 3H), 1.34-1.26 (m, 20H), 0.87 (t, J=6.8 Hz, 3H); ^{13}C NMR (MeOH-d_4, 100 MHz) δ 175.0, 174.3, 171.6, 40.6, 35.6, 31.9, 29.6-29.3, 25.6, 22.6, 16.9, 13.3. HRMS calcd. for $C_{19}H_{37}N_2O_4$ $[M+H]^+$ 357.27, found 357.27.



30 **26**

CoA dipeptide fatty acid (26). To a stirred solution of acid (**30)** (5.9 mg, 16.55 μmol, 1.0 eq) in CH_2Cl_2 (0.565 mL) and THF (0.147 mL) under Ar was added DCC (4.1 mg, 19.86 μmol, 1.2 eq), and the resulting solution was stirred at room temperature overnight (17 hrs). The crude material was concentrated, rediluted in DMF (0.165 mL), and cooled to 0 °C. Coenzyme A sodium salt hydrate (2.5 mg, 3.3 μmol, 0.2 eq) was then added, followed by Et_3N (1.4 μL, 9.92 μmol, 0.6 eq), and the resulting suspension was stirred vigorously for 30 min. The reaction was then brought to neutral pH by dropwise addition of 0.1 N aqueous HCl, concentrated to remove volatile impurities, and remaining (**26**) was suspended in DMF. (**26**) was characterized by FTICR-MS in both negative and positive ion mode (**Figure 5-11**) and was detected with a 0.6 ppm error in negative mode and 1.6 ppm in positive mode.

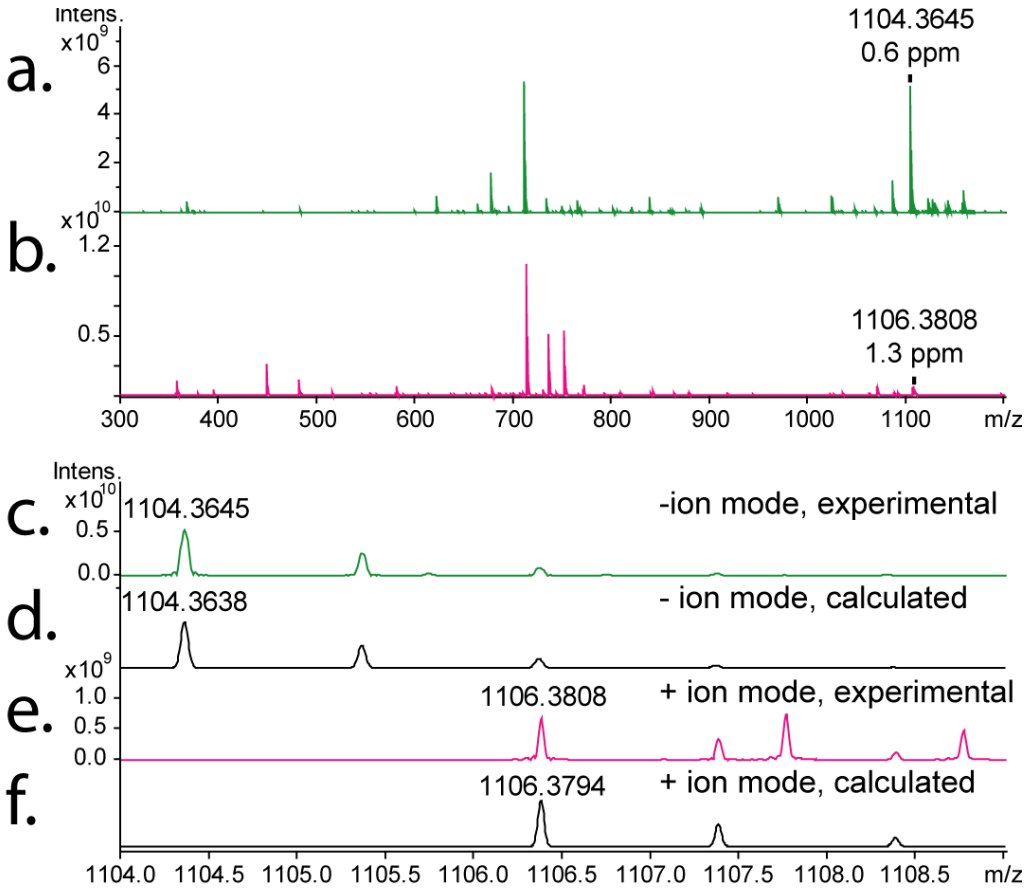
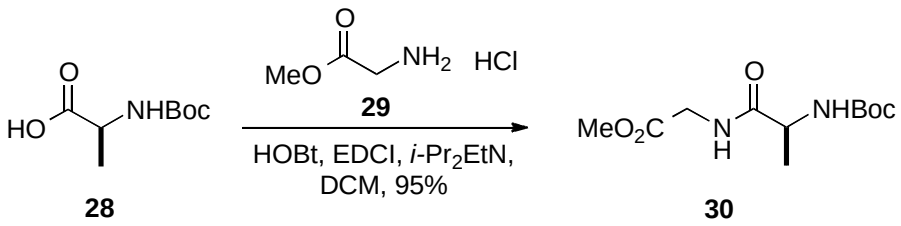
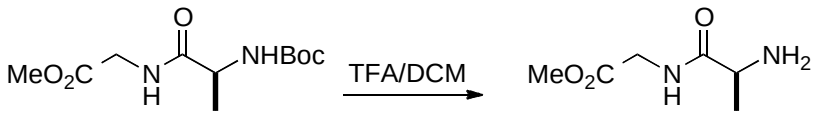


Figure 5-11. FTICR-MS characterization of (26). Positive ion mode (**Figure 5-11A,C**), and negative ion mode (**Figure** 5**-11B,E**) spectra of (**26**). Simulated positive and negative mode spectra (Bruker Daltonics Data analysis) are also provided (**Figure 5-11D,F**).



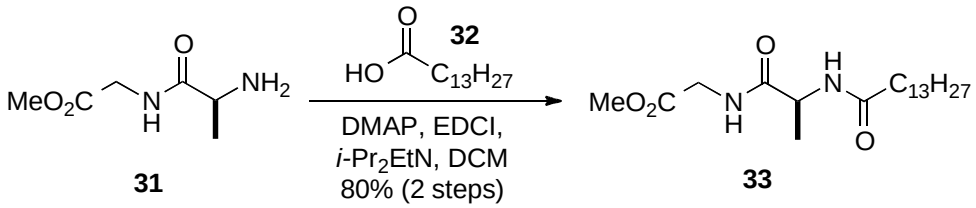
(*S*)-methyl 2-(2-((*tert*-butoxycarbonyl)amino)propanamido)acetate (30). By following the procedure detailed in Zhang,[59] 1.5 g (11.95 mmol, 1.0 eq) of glycine methyl ester•HCl salt **28** yielded 2.97 g (95%) of product **30** as a yellow oil. Data are: ^{1}H NMR ($CDCl_3$, 300 MHz) δ 7.23 (bs, 1H), 5.50 (d, *J*=3.3 Hz, 1H), 4.20 (m, 1H), 3.93,

(d, *J*=2.85 Hz, 2H), 3.63 (s, 3H), 1.33 (s, 9H), 1.28 (d, *J*=3.6 Hz, 3H). ^{13}C NMR ($CDCl_3$, 75.5 MHz) δ 173.7, 170.4, 155.7, 80.0, 52.4, 50.1, 41.2, 28.4, 18.6.



30 **31**

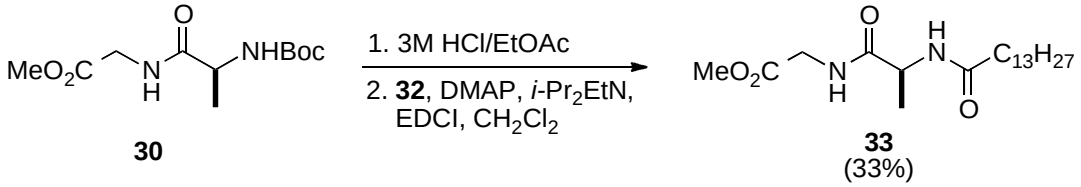
(*S*)-methyl 2-(2-aminopropanamido)acetate (31). To a stirred solution of **30** (2.94 g, 11.28 mmol, 1.0 eq) in CH_2Cl_2 (38 mL) at 0 °C was added fresh trifluoroacetic acid (12.26 mL). The resulting solution was stirred 2 hrs. It was then concentrated to give an orange oil, which was rinsed three times with Et_2O. This caused immediate product precipitation. The orange Et_2O supernatant was discarded with each rinse, ultimately affording 2.95 g of **31** as an off-white solid. Data are: ^{1}H NMR (DMSO-d_6, 300 MHz) δ 8.88 (d, *J*=2.25 Hz, 1H), 8.17 (bs, 2H), 3.91-3.88 (m, 3H), 3.60, (s, 3H), 1.33 (d, *J*=2.55 Hz, 3H). ^{13}C NMR (DMSO-d_6, 75.5 MHz) δ 170.8, 162.3, 52.5, 48.6, 41.2, 17.7.



31 **33**

Method 1: (*S*)-methyl 2-(2-tetradecanamidopropanamido)acetate (33). To a stirred solution of **31** (2.73 g, 18.68 mmol, 1.0 eq) in CH_2Cl_2 (75 mL) at 0 °C were added myristic acid **32** (5.12 g, 22.42 mmol, 1.2 eq), DMAP (571 mg, 4.67 mmol, 0.25 eq), EDCI (7.16 g, 191.70 mmol, 2.0 eq), and *i*-Pr_2EtN (6.5 mL, 37.36 mmol, 2.0 eq). The resulting solution was then stirred overnight, warming to room temperature, for 23 hrs. The reaction was then quenched by addition of H_2O (100 mL) and $CHCl_3$ (75 mL), and the layers were separated. The aqueous layer was extracted with $CHCl_3$ (3 x 75 mL).

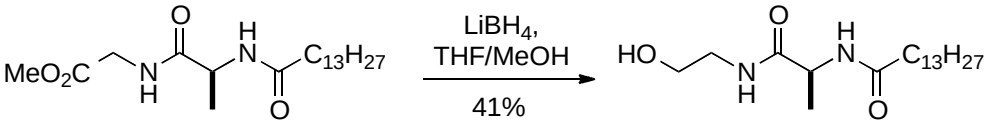
The combined organic layers were then washed sequentially with 3N H_3PO_4, saturated $NaHCO_3$, and brine; dried ($MgSO_4$); concentrated, and flushed through a silica plug (2% MeOH/CH_2Cl_2 as eluent). The resulting yellow solid was then purified by three successive overnight recrystallizations from hot MeOH/EtOAc, to give a combined total of 3.39 g (49%, 80% over two steps from **53**) of **57** as a white-yellow solid. Data are: ^{1}H NMR ($CDCl_3$, 400 MHz) δ 6.68 (bs, 1H), 6.02 (bs, 1H), 4.57-4.49 (m, 1H), 4.00 (d, *J*=2.6 Hz, 2H), 3.73 (s, 3H), 2.18 (t, *J*=7.6 Hz, Hz, 2H), 1.65-1.53 (m, 2H), 1.37 (d, *J*=3.6 Hz, 3H), 1.32-1.23 (m, 20H), 0.85 (t, *J*=7.2 Hz, 3H); ^{13}C NMR ($CDCl_3$, 100 MHz) δ 174.2, 172.3, 169.7, 52.6, 48.7, 41.4, 36.7, 32.1, 29.8-29.3, 25.8, 22.9, 18.3, 14.2.



30 **33** (33%)

Method 2: (*S*)-methyl 2-(2-tetradecanamidopropanamido)acetate (33). Substrate **30** (973 mg, 3.74 mmol, 1.0 eq) was stirred vigorously in 3M HCl in EtOAc (22 mL) at room temperature for 30 min. It was then made basic by addition of Et_3N and was subsequently concentrated. The crude product was next diluted with Et_2O and re-concentrated (repeat 1x) to give the deprotected HCl salt intermediate. This was then re-suspended in CH_2Cl_2 (15 mL) and cooled to 0 °C with vigorous stirring. To this suspension were added myristic acid **32** (939 mg, 4.11 mmol, 1.1 eq), DMAP (114 mg, 0.93 mmol, 0.25 eq), EDCI (1.43 g, 7.48 mmol, 2.0 eq), and *i*-Pr_2EtN (1.3 mL, 7.48 mmol, 2.0 eq). The resulting solution was then stirred overnight, warming to room temperature, for 18 hrs. The reaction was then quenched by addition of H_2O, 3N H_3PO_4, and $CHCl_3$, and the layers were separated. The aqueous layer was extracted with $CHCl_3$

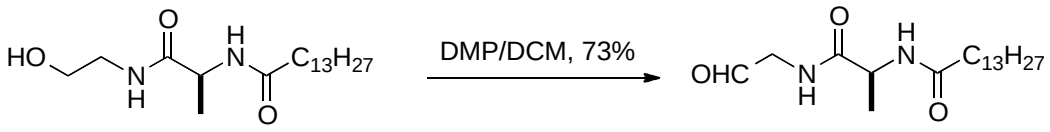
(3 x). The combined organic layers were then washed with saturated $NaHCO_3$ and brine; dried ($MgSO_4$); concentrated, and flushed through a silica plug (EtOAc as eluent). The resulting yellow solid was then concentrated and purified by overnight recrystallization from hot MeOH/EtOAc to give 460 mg (33%) of **33** as an off-white solid. Data are: ^{1}H NMR ($CDCl_3$, 400 MHz) δ 6.82 (bs, 1H), 6.13 (bs, 1H), 4.61-4.52 (m, 1H), 4.01 (t, *J*=7.2 Hz, 2H), 3.75 (s, 3H), 2.20 (t, *J*=10.4 Hz, Hz, 2H), 1.64-1.56 (m, 2H), 1.39 (d, *J*=4.8 Hz, 3H), 1.32-1.25 (m, 20H), 0.87 (t, *J*=9.2 Hz, 3H); ^{13}C NMR ($CDCl_3$, 100 MHz) δ 174.2, 172.3, 169.7, 61.7, 48.7, 41.5, 36.7, 32.1, 29.9-29.6, 25.8, 22.9, 18.5, 14.3. HRMS calcd. for $C_{20}H_{39}N_2O_4$ $[M+H]^+$ 371.29, found 371.29. Note: Despite a lower yield (compared with conditions above to **33**), this procedure gave a purer batch of **33** necessary for probe **27**.



33 **34**

(*S*)-*N*-(1-((2-hydroxyethyl)amino)-1-oxopropan-2-yl)tetradecanamide (34). By following the procedure detailed in Koketsu,[23] 163 mg (0.439 mmol, 1.0 eq) of methyl ester **33** yielded 62 mg (41%) of product **34** as a white solid. Data are: $[\alpha]_D^{25}$ = +7.0 (*c* 1.63, $CHCl_3$); ^{1}H NMR ($CDCl_3$, 300 MHz) δ 6.87 (bs, 1H), 6.26 (m, 1H), 4.50 (m, 1H), 3.71 (t, *J*=5.0 Hz), 3.50-3.45 (m, 2H), 2.20 (t, *J*=9Hz, 2H), 1.65-1.57 (m, 2H), 1.38 (d, *J*=4.6 Hz, 3H), 1.37-1.25 (m, 20H), 0.88 (t, *J*=6.6 Hz, 3H); ^{13}C NMR ($CDCl_3$, 75.5 MHz) δ 173.9, 173.5, 62.1, 49.2, 42.3, 42.7, 36.8, 31.2, 29.9-29.4, 25.8, 22.9, 18.4, 14.4. Note: Two heavily staining upper spots were seen by TLC (CH_2Cl_2, R_{f1} = 0.56, R_{f2} = 0.28), but were not the product. The product was obtained after pushing these spots off

the column ($CHCl_3$ as eluent) and then flushing the column with MeOH. The product spot is nearly baseline by TLC (CH_2Cl_2). Also, Koketsu reports this alcohol's optical rotation as –33.4. Two separate syntheses in our lab, however, both confirmed a positive optical rotation.



34 **27**

(*S*)-*N*-(1-oxo-1-((2-oxoethyl)amino)propan-2-yl)tetradecanamide (27). By following the procedure detailed in Koketsu,[23] 61 mg (0.178 mmol, 1.0 eq) of substrate **34** yielded 44 mg (73%) of product **27** as an off-white solid. Data are: TLC R_f = 0.34 (5% MeOH/CH_2Cl_2, 2,4-dinitrophenyl hydrazine stain); ^{1}H NMR ($CDCl_3$, 400 MHz) δ 9.61 (s, 1H), 7.04 (bs, 1H), 6.09 (d, *J*=3.6 Hz, 1H), 4.60-4.53 (m, 1H), 4.14 (d, *J*=2.6 Hz, 2H), 2.20-2.15 (m, 2H), 1.61-1.63 (m, 2H), 1.38 (d, *J*=3.6 Hz, 3H), 1.31-1.22 (m, 20H), 0.85 (t, *J*=6.8 Hz, 3H); ^{13}C NMR ($CDCl_3$, 100 MHz) δ 196.5, 174.0, 50.3, 48.7, 36.7, 32.1, 29.8-29.5, 25.8, 22.9, 18.2, 14.3. HRMS calcd. for $C_{19}H_{37}N_2O_3$ $[M+H]^+$ 341.28, found 341.28. Dess-Martin periodinane was prepared freshly before use over two steps.[60]

Biochemical reaction of EtuA2 RE-domain and SfmC with the CoA dipeptide fatty acid (26)

The biochemical reaction of compound (**26**) to (**27**) was performed as described by Koketsu et al.[23] Reactions were made in reaction buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 1.0 mM TCEP PH 7.5, 20% glycerol) with either no enzyme, 10 μM EtuA RE-domain, or 10 μM SfmC. Cofactors including 2 mM ATP, 10 mM $MgCl_2$, 10 μM

$MnCl_2$, 3 mM NADPH, and 3 mM NADH were then added from concentrated stocks followed by addition of compound (**26**) to 200 μM in DMF. The reaction was incubated overnight at room temperature and monitored by LC-FTICR-MS as described above in the **Secondary metabolite identification by LC-FTICR-MS and confirmation by LC-MS/MS**.

Metaproteomic analysis of biosynthetic gene expression

Tunicate protein samples were precipitated with acetone (4:1 at -20 ^{o}C for 60 minutes) followed by centrifugation (4 ^{o}C for 10 minutes at 14,000 RCFxG) and resolubilization (500 μL 8 M urea in 10 mM HEPES, pH 8). The sample was reduced (DTT for 60 minutes at room temperature), alkylated (iodoaceteic acid 5.5 mM at room temperature in the dark), diluted (4X in DI), and digested (20 μg trypsin, 16 hr at 37 ^{o}C). TFA was added to pH 2.7 and the sample was fractionated over 40 min into 20 fractions on a Luna SCX 300 Å 50x4.6 mm 5 μm column as previously described.[61] Each fraction was then desalted with a C18 spin column, dried, and reconstituted in 0.1% formic acid.

The 20 peptide fractions were analyzed once on an LTQ-Orbitrap XL (Thermo-Fisher Scientific) interfaced with a nanoLC 2D system (Eksigent Technologies). Peptides were separated on a column (75 μm × 15 cm) in-house packed with 3 μm C18 resin (Sepax HP-C18) after loading on a C18 trap column over a 90 min gradient of 10-50% solvent B (90% acetonitrile with 0.1% formic acid) at a flow rate of 250 nL/min and sprayed into the mass spectrometer via a chip-based nanoelectrospray source (Advion Triversa Nanomate) in positive ion mode. The LTQ-orbitrap was operated in data-dependent mode by alternating single MS scan (300-1700 m/z) in the orbitrap analyzer and sequential

MS/MS scans in the LTQ for the seven most intense ions from each MS survey scan. MS scans were acquired with a resolution set at 60,000 at m/z 400 and an automatic gain control (AGC) target of 1×10^6. MS/MS scans were triggered on ions with signal intensities above 500. Recurring precursor ions were dynamically excluded for 30 s. By applying charge-state monitoring, ions with 1+ or unassigned charge states were rejected. Full scans were obtained in profile mode and MS/MS scans were saved as line spectra. Raw data files generated from LC-MS/MS experiments were converted into sets of DTA peaklists by using BioWorks Browser 3.3.1 (Thermo) with the default parameters for LTQ Orbitrap. DTA peaklists were merged in an mgf file for further analysis. .mzXML files were generated for submission to Tranche Proteome commons.

The 20 peptide fractions were also analyzed in duplicate on a Solarix 12T hybrid Q-FTICR (Bruker Daltonics) interfaced with a U3000 nanoLC system (Dionex). Peptides were separated on a column (75 µm × 20 cm) packed in-house with 3 µm C18 resin (Alltech) after loading on a C18 trap column over 90 minutes of 5-80% solvent B (acetonitrile with 0.1% formic acid). The FTICR operated in data-dependent mode with each parent mass scan followed by up to six product ion scans in the FTICR-MS cell. The instrument was operated with sidekick instead of gated trapping in the cell, and calibrated with sodium TFA clusters. Ions in the $2\text{-}6^+$ charge states with an abundance greater than $5x10^6$ were subjected to MS/MS and excluded for one minute after 4 scans. All scans were collected in the profile mode with a transient of 256 K and a calculated resolving power of 24,000 at m/z 400. Peak picking was performed from the profile mode spectra in Data Analysis 4.0 (Bruker) using the FTMS algorithm and Protein Analysis function to export an .mgf file for further processing. FTMS parameters were S/N threshold 4,

relative intensity threshold 0.01%, and absolute intensity threshold 100. Find AutoMS(n) parameters were: Intensity threshold positive 0, maximum number of compounds 10000, retention time window 0, profile spectra only. The .mgf files were reformatted from the Bruker output to the Thermo .mgf format, transformed to .dta files, which were used to generate .mzXML files for downstream analysis (Dr. Damian Fermin, Nesvizhskii laboratory).

The sixty raw data files, in the appropriate .mgf (OMSSA,[62] Inspect,[63]) or mzXML (X!tandem,[64] Spectrast[65]) format were then searched in each of the four search engines utilized in this study. For OMSSA searching the following command line parameters were utilized: omssacl -e 0 -i 1,4 -mf 2 -mv 1,4 -tem 0 -tom 0 -te 0.1 -to 0.3 -tez 1 -he 100000 -zcc 1 -hl 1 -v 2 -zh 4 -zoh 3 -op c:\OUTPUTFILE.pep.xml -d c:\DATABASE.fasta -fm c:\INPUTFILE.mgf. These parameters correspond to: trypsin, b/y ions, fixed modification carboxymethyl cysteine, variable modification deamidation of N and Q and M oxidation, monoisotopic precursor, monoisotopic fragments, precursor tolerance 0.1 Da, product tolerance 0.3 Da, charge dependency of precursor mass, mass expect value 100000, use input file charge, retain top hit, maximum 2 missed cleavages, maximum charge 4, maximum product ion charge 3, output type .pep.xml. For Inspect the following parameters were utilized: spectra,c:\INPUTFILE.mgf, db,c:\DATABASE.trie, protease,trypsin, mod,+58.005479,C,fix, mod,+15.994915,M,opt, mod,+0.984016,N,opt, mod,+0.984016,Q,opt, mods,2 ParentPPM,25 IonTolerance,0.3, Instrument,FT-Hybrid, TagCount,1, RequireTermini,2. The output text files were converted to the .pep.xml format with the script inspecttopepxml.py. X!tandem search parameters were: parent monoisotopic mass error 25 ppm, monoisotopic mass error

allowed, static modification 58.005479@C, potential modifications 15.994915@M,0.984@N,0.984@Q, no semi-tryptic cleavages, 2 maximum missed cleavages allowed, no refinement mode. Output .tandem files were transformed to pep.xml files in the TPP GUI. Spectrast searching used all default parameters. A Spectrast library was constructed from Orbitrap and Q-FTICR-MS runs of synthetic peptide standards of all Etu peptides identified.

All .pep.xml output files generated by the four search engines were then processed in the Trans-Proteomic Pipeline (TPP)[50] version TPP v4.4 VUVUZELA rev 1, Build 201010121551 (MinGW) running under Windows 7 on a four-core Intel Core 7 PC with 4GB of RAM.

OMSSA Peptide Prophet command line search:

1. interactparser c:\OUTPUTFILE.pep.xml INPUTFILE1.pep.xml … INPUTFILEN.pep.xml –L6 –Etrypsin –C –P

2. refreshparser c:\OUTPUTFILE.pep.xml DATABASE.fasta

3. peptideprophetparser c:\OUTPUTFILE.pep.xml DECOY=###REV### MINPROB=0 NONPARAM

(This search corresponds to a minimum peptide length of 6, trypsin digest, OMSSA input, and a decoy based non-parametric model)

Inspect Peptide Prophet command line search:

1. interactparser c:\OUTPUTFILE.pep.xml INPUTFILE1.pep.xml … INPUTFILEN.pep.xml –L6 –Etrypsin –P

2. peptideprophetparser c:\OUTPUTFILE.pep.xml DECOY=###REV### MINPROB=0 NONPARAM

3. refreshparser c:\OUTPUTFILE.pep.xml DATABASE.fasta

(This search corresponds to a minimum peptide length of 6, trypsin digest, OMSSA input, and a decoy based non-parametric model)

X!tandem Peptide Prophet searches were performed from the GUI with the parametric model, accurate mass binning, expect values for scoring, and a minimum peptide length of 6. Spectrast Peptide prophet searches were performed from the GUI with the parametric model and default search options. Search results were generated from all four search engine specific pep.xml files. Protein Prophet results were generated with the default search options for each of the search engine specific pep.xml files, as well as the combined interprophet.pep.xml files. All reported data from the prot.xml files were filtered at probability > 95%. All peptides/proteins discussed in the text were manually inspected and verified. Synthetic peptide standards of identified Etu peptides were ordered from GenScript. Proteins were reduced and alkylated then analyzed on the Thermo LTQ-Orbitrap MS as described above in the proteomics method section (with a different capillary LC column) or on a Bruker Apex Q-FTICR-MS as described in the metabolite section using a 2 mm x 150 mm 5 µM Jupiter C4 column.

Metaproteomics analysis of the ET-743 assemblage

The tunicate metaproteome sample was reduced, alkylated, digested, and then separated into 20 SCX fractions. These fractions were monitored by nLC MS/MS with FTICR and Orbitrap instruments. Mass lists were generated based on high resolution/accurate mass

MS1 data (and MS2 for FTICR) and subjected to X!Tandem,[64] OMSSA,[62] Inspect,[63] and Spectrast[65] search algorithms in the TPP. The target database was the six-frame translation of the total metagenomic assembly filtered for polypeptides of 60 amino acids or greater, plus the closest sequenced organisms to each of the three principle constituents of the assemblage. Reverse sequences were also included. The proteins identified from the Etu and Etr databases were also added to the sequence database. Identified proteins were manually inspected for all cases where only one peptide was identified. Total performance characteristics are presented in **Table 5-12.** Automated results for all peptides composing these proteins are given below in **Tables S14-S19**. Further database searching results are given in **Table 5-20**. More detailed manual spectral interpretation is given in **MS2 spectra for Etu peptides** in **Figures 5-12 through 5-31** and **Tables 5-20 through 5-25, p215**.

	X!tandem	OMSSA	Inspect	Spectrast
Total Spectra	259,749	174,833	23,673	5
Assigned Spectra	7,543	1,787	853	5
Unique Peptides	866	425	285	4
Unique Proteins	326	203	227	2
Single Hits	142	104	135	0

Table 5-12. Total metaproteomics performance characteristics. Performance characteristics are provided for each of the four search engines used in the metaproteomics analysis. Total spectra represent total spectra as reported from the .pep.xml output files. Assigned spectra represent all peptides assigned with a greater than 90% probability in the .pep.xml files. Unique peptides, unique proteins, and single hits represent the maximum number of possible assignable species in the .prot.xml output files. Despite a relatively low number of assigned spectra, (3%, 1%, 4%, 100%) we have obtained excellent ratios of correct to incorrect protein identifications based on the probability models in Peptide-, Inter-, and Protein- Prophet. Possible contributing factors include: chemical noise, poor quality spectra, chimeric spectra, PTMs, eukaryotic splicing events, and peptides not present in our search database. It should be noted that environmental and metaproteomics datasets often have much lower numbers of total assigned spectra than for similarly sized studies with well characterized model organisms.

The number of total proteins identified is modest for each of the four search engines, with the most number of proteins identified correlating with the higher false discovery rates as expected. Calculated protein false discovery rates are excellent (<1%), and represent a high quality dataset. Of the total 289 proteins identified, three can be assigned with very high probability (>99%) to genes derived from the ET-743 biosynthetic gene cluster by at least two of the four search engines.

EtuF3 was assigned a probability of 99.99% based on pooled results. This high value is due to identification of two peptides TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (N_{115} denotes deamidation) in the +3 charge state (**Table 5-13**) and RPIELR in the +2 charge state (**Table 5-14**). Both peptides were identified by LTQ-Orbitrap MS.

Etu_F3: 3+ TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK	Iprophet	X!tandem	OMSSA	Inspect	Spectrast
Score	99.88%	99.90%	66.32%		99.83%
Coverage		40/108	23/54		N/A
Total	3	1	1		1

Table 5-13. Peptide assignment for EtuF3: 3+ TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK. The peptide probabilities contributing to the total protein probabilities are given four each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided. N_{115} denotes deamidation

The peptide 3+ TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK was assigned with 3/4 search engines, although the OMSSA result is less than confident at 66.32% (**Table 5-13**). In each case only one spectrum was assigned. It should be noted that it is possible that the exact site of deamidation (N_{115} denotes deamidation) could be incorrect due to the multiple modifiable residues in close proximity. In addition, the actual ion observed could represent a population with the deamidated residue at multiple sites.

EtuF3: 2+ RPIELR	Iprophet	X!tandem	OMSSA	Inspect	Spectrast
Score	93.43%	87.48%			94.69%
Coverage		11/10			N/A
Total	3	1			2

Table 5-14. Peptide assignment for EtuF3: 2+ RPIELR. The peptide probabilities contributing to the total protein probabilities are given for each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided.

The peptide 2+ RPIELR was assigned with 2/4 search engines (**Table 5-14**). One spectrum was assigned by X!tandem, and two were assigned by Spectrast. More caution should be taken in assigning this peptide due to the small size, although the manual spectral interpretation and standard peptide data (shown below) strongly support this assignment. The peptide RPIELR contains an apparent missed cleavage site, but further investigation reveals that trypsin almost never cleaves if a proline follows the basic amino acid.[66]

EtuM1 was assigned a probability of 99.16% based on pooled Inter Prophet results. This high value is due to identification of two peptides LLDVGGGTAINAIALAK in the +3 charge state by Q-FTICR-MS2 (**Table 5-15**) and LKPC_{161}YR in the +2 charge state by LTQ-Orbitrap MS2 (**Table 5-16**). C_{161} denotes the expected cysteine modification from iodoaceteic acid. These two unique peptides represent 7% total sequence coverage of EtuM1.

EtuM1:

2+ LLDVGGGTAINAIALAK

	Iprophet	**X!tandem**	**OMSSA**	**Inspect**	**Spectrast**
Score	99.16%	98.99%		98.98%	
Coverage		24/32		N/A	
Total	4	2		2	

Table 5-15. Peptide assignment for EtuM1: 2+ LLDVGGGTAINAIALAK. The peptide probabilities contributing to the total protein probabilities are given for each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided.

The peptide LLDVGGGTAINAIALAK in its 2+ charge state was assigned with 2/4 search engines (**Table 5-15**) with two spectra assigned by X!tandem and Inspect. Manual interpretation increases this total to four (spectra number 4505, 4507, 4509, and 4512 in the LC run). This peptide assignment is highly confident due to the use of Q-FTICR-MS, providing high mass accuracy MS1 and MS2 data. It should be noted that, in our experience, Spectrast does not seem to perform well with Q-FTICR-MS data and the failure to identify this peak is thus not surprising. The unmodified, fully tryptic peptide LLDVGGGTAINAIALAK is observed with a mass error less than 10 ppm for the parent masses and very high quality MS2 data with almost all peaks in the spectrum identified with very tight mass errors from 0 to +25 ppm. This peptide was observed in SCX fraction 7 in both FTICR-MS runs with elution times of 53.8-54.5 minutes (spectra 4490-4450) and 54.5-55.5 minutes (spectra 4520-4564), respectively. However, this peak was only selected for MS/MS in the second run with four total MS/MS spectra: 4512 and 4505, 4507, 4509.

	EtuM1:				
	2+ ILKPC$_{161}$YR				
	Iprophet	**X!tandem**	**OMSSA**	**Inspect**	**Spectrast**
Score	15.62%	15.62%			
Coverage		13/12			
Total	1	1			

Table 5-16. Peptide assignment for EtuM1: 2+ ILKPC$_{161}$YR. The peptide probabilities contributing to the total protein probabilities are given for each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided.

The peptide ILKPC_{161}YR in its 2+ charge state was assigned with 1/4 search engines (**Table 5-16**: one spectrum assigned by X!tandem). It has a very low probability as it is a hexamer peptide, and the assignment does not contribute positively in the Protein Prophet model. It is noted in this paper for completeness and due to the favorable standard peptide and manual inspection results. For example, this peptide may actually be present, but still cannot be confidently assigned by the appropriate software packages due to probability based issues. ILKPC_{161}YR, appears to contain one missed cleavage site, however, as cleavage N-terminal to proline is rarely observed, it can be considered to contain zero expected missed cleavage sites.[66] This peptide, which represents the most N-terminal EtuM1 peptide that is likely observable, is alkylated at the cysteine as expected. C_{161} denotes the expected cysteine modification from iodoaceteic acid.

EtuR1 was assigned a probability of 100.00% based on pooled results. This high value is due to identification of the same peptide ion GSNIHYDLENDHNDYEK by LTQ-Orbitrap-MS2 in the +2 charge state (**Table 5-17**) and the +3 charge state (**Table 5-18**).

EtuR1: 2+ GSNIHYDLENDHNDYEK	Iprophet	X!tandem	OMSSA	Inspect	Spectrast
Score	99.91%	99.92%	99.06%		99.93%
Coverage		26/32	19/32		N/A
Total	3	1	1		1

Table 5-17. Peptide assignment for EtuR1: 2+ GSNIHYDLENDHNDYEK. The peptide probabilities contributing to the total protein probabilities are given for each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided.

EtuR1: 3+ GSNIHYDLENDHNDYEK	Iprophet	X!tandem	OMSSA	Inspect	Spectrast
Score	99.89%	99.87%			99.91%
Coverage		28/64			N/A
Total	2	1			1

Table 5-18. Peptide assignment for EtuR1: 3+ GSNIHYDLENDHNDYEK. The peptide probabilities contributing to the total protein probabilities are given for each of the four search engines and the combined analysis. The calculated sequence coverage is provided except for Inspect and Spectrast for which it is not calculated. The number of total spectra identified by each search engine is also provided.

The peptide GSNIHYDLENDHNDYEK in its 2+ charge state was assigned with 3/4 search engines (**Table 5-17**), with one spectrum each from LTQ-Orbitrap MS2. The same peptide in its 3+ charge state was assigned with 2/4 search engines (**Table 5-18**), with one spectrum each from LTQ-Orbitrap MS2. Identification of the same peptide in different charge states in subsequent scans further corroborates this assignment and are considered unique peptides in the –Prophet statistical models.

Rationalizing the observed peptides **(Table 5-13 through 5-18)** is speculative, however it should be noted that the relatively basic peptides observed were limited to late eluting SCX fractions (10/20 and 15-19/20), which should contain the more basic peptides in the mixture. Having multiple basic sites may lead to relatively high ionization efficiency in positive mode ESI, thereby resulting in peptides observed even for low abundant parent peptides as compared to less basic peptides. Abundance of the proteins may be an important factor for the success of this method. 60% of the total metagenomic sequencing reads are bacterial in origin, however, this number may not correspond to relative total levels of bacterial versus tunicate derived proteins present in the cell free protein extract of the tunicate specimen. Assuming the suggested *E. frumentensis* species is the correct producer of ET-743, and that 16S rRNA gene reads present a relatively accurate view of bacterial consortium, approximately 14% of the bacterial protein present may be from the correct species. The expression level of ET-743 biosynthetic proteins within this 8% of the total protein is difficult to predict. Biosynthetic enzymes have previously been detected from native producers by 1D-SDS-PAGE at wild-type expression levels, which suggests expression ratios >1% of total protein.[67] Therefore, if these assumptions are within one to two orders of magnitude of this estimation, we hypothesized that it would be possible to directly detect ET-743 biosynthetic proteins from the total collected tunicate-microbial consortium metaproteome.

After assigning the biosynthetic proteins and peptides above and comparing with authentic standard synthetic peptides, we sought to address whether the assignments were correct through other methods. This task was conducted with a series of bioinformatics

tools as noted below (**Table 5-19**). Using the program Mascot with mass errors as described above, identified Etu peptide spectra were searched against the Non-redundant database, resulting in no significant peptide hits. This result suggests that the assigned peptides are not false positives that better match *any* known peptide sequences. A similar search was performed with the online search engine OMSSA with all species. None of the spectra scored significant hits (expectation value, e <0.1) except for the peptide RPIELR, which was assigned to the peptide RPLELR (which is identical for the purpose of CID-MS/MS assignment) from the frog *Xenopus*. These data further suggest that the assignment of this peptide is correct as it is unlikely that freshwater frogs were present in the marine tunicate collection, or due to laboratory contamination thus it seems likely that this peptide is derived from the assigned biosynthetic protein EtuF3. The assigned peptide spectra were also searched against the online version of Global Proteome Machine X!tandem[64] with similar parameters and all species selected. No significant peptide hits were observed.

Next, the assigned Etu biosynthetic peptides were subjected to automated *de novo* sequencing with Inspect Pep-Novo (**Table 5-19**).[68] Unlike the previous statistical methods of database (Mascot, OMSSA, GPM) searching, in theory, *de novo* will assign sequences without any bias derived from the target database. This task is very challenging in practice, however useful sequence tags can often be obtained. The EtuF3 peptides had substantial sequence tags correctly assigned as: RPIELR and TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK. In the case of the FTICR-MS detected EtuM1 peptide the two spectra provided the sequence tags LN and VGGGTAL, matching

to LLDVGGGTAINAIALAK. The orbitrap detected EtuM1 peptide was assigned to the N- and C- terminal sequences (LLK and YR) matching to ILKPC$_{161}$YR. The first EtuR1 peptide had different sequence tags assigned to each charge state, for high combined coverage of GSNIHYDLENDHNDYEK. The program MSGF attempts to generate a scoring function for MS/MS peptide assignments independent of the total database searched (**Table 5-19**).[69] Significant log(p) values of less than 3E-8 were obtained for all peptide assignments. The identified peptide sequences were also analyzed with the BlastP program against the NR database to determine if the identified peptide sequences were unique, assuming the spectra were assigned to the correct peptide sequences. Except for the two smallest peptides, ILKPC$_{161}$YR and RPIELR, no peptide sequences were found in the NR database. (**Table 5-19**)

Sequence	Mascot NR:	OMSSA all species:	GPM all species:	De-novo	P-value MSGF	BlastP peptide
3_TIQHEIELSDIGPIINNLIQEnNQINKK	No hits	No hits	No hits	HELELTTLGT	4.50E-17	Unique
2_RPIELR	No hits	RPLELR (3ppm, -1, Xenopus)	No hits	VGPLELR	3.20E-08	Not unique
2_LLDVGGGTAINAIALAK	No hits	No hits	No hits	LPSNLNR, VGGGTAL	8.20E-05	Unique
2_ILKPcYR	No hits	No hits	No hits	LLKEEYR	7.60E-09	Not unique
2_GSNIHYDLENDHNDYEK	No hits	No hits	No hits	HLDLNTK	1.00E-16	Unique
3_GSNIHYDLENDHNDYEK	No hits	No hits	No hits	LHYDQDNDHLDYEK	1.10E-16	Unique

Table 5-19. Additional database searching. Results from searching assigned Etu peptide spectra against the full protein databases available online for Mascot, OMSSA, and GPM. The total and correct sequence tags from *De novo* searching of the same spectra as well as the log(p) generated by the program MS GF are provided. The identified peptide sequences were also analyzed with the BlastP program against the NR database.

MS2 spectra for Etu peptides

Detailed MS spectral data are provided for each of the six peptides identified. Data include tables of manually assigned CID spectra for metaproteomics and authentic standard peptides (**Tables 5-20** through **5-25**). Other data presented include plots of mass error versus m/z for metaproteomics and authentic standard peptides (**Figure 5-12, 5-15, 5-18, 5-23, 5-26, 5-29**). These data illustrate that mass errors fall within a similar range, within expected instrument tolerances, for metaproteomics and authentic standard peptides. Automatically assigned MS2 spectra are shown as derived from X!tandem (the only search engine to identify all six peptides) to illustrate the spectral quality and b and y ion assignments used in the metaproteomics experiments (**Figure 5-13, 5-16, 5-19, 5-20, 5-24, 5-27, 5-30**). Finally, direct visual comparison between metaproteomics and authentic standard assigned spectra is provided (**Figure 5-14, 5-17, 5-21, 5-25, 5-28, 5-31**).

For triply protonated TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table 5-20**), a graph illustrating mass error in manually assigned b and y ions for metaproteomics versus standard peptides (**Figure 5-12**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-13**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-14**) are provided.

For doubly protonated RPIELR a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table 5-21**), a graph illustrating mass error

in manually assigned b and y ions for metaproteomics versus standard peptides (**Figure 5-15**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-16**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-17**) are provided.

For doubly protonated LLDVGGGTAINAIALAK a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table 5-22**), a graph illustrating mass error in manually assigned b and y ions for metaproteomics versus standard peptides (**Figure 5-18**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-19** and **5-20**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-21**) are provided. Inspect assignment is provided in **Figure 5-22**.

For doubly protonated ILKPC_{161}YR a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table S24**), a graph illustrating mass error in manually assigned b and y ions for metaproteomics versus standard peptides (**Figure 5-23**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-24**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-25**) are provided.

For doubly protonated GSNIHYDLENDHNDYEK a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table 5-24**), a graph illustrating mass error in manually assigned b and y ions for metaproteomics versus

standard peptides (**Figure 5-26**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-27**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-28**) are provided.

For triply protonated GSNIHYDLENDHNDYEK a table of metaproteomics versus authentic standard peptide manual assignments of b and y ions (**Table 5-25**), a graph illustrating mass error in manually assigned b and y ions for metaproteomics versus standard peptides (**Figure 5-29**), a figure showing the X!tandem assigned product and parent ions (**Figure 5-30**), and a figure comparing the spectra from the metaproteomics and authentic standard peptides (**Figure 5-31**) are provided.

3+ TIQHEIELSDIGPIINNLIQEN[115]NQINKK, 3228.6987 Da									
3228.7117 Da, 1077.2445, 4 ppm					3228.6557 Da, 1077.2259 m/z, -13 ppm				
O_16.4412					Standard				
Exptl	I	d(m/z)	Calc	Assign	Exptl	I	d(m/z)	Calc	Assign
					480.3	4	0.1	480.3	b4
609.4	23	0.1	609.3	b5	609.3	23	0.0	609.3	b5
722.5	19	0.1	722.4	b6	722.4	25	0.0	722.4	b6
851.3	33	-0.1	851.4	b7	851.4	25	0.0	851.4	b7
964.6	41	0.1	964.5	b8	964.5	36	0.0	964.5	b8
1051.6	20	0.1	1051.5	b9	1051.4	16	-0.1	1051.5	b9
1166.6	28	0.0	1166.6	b10	1166.5	40	-0.1	1166.6	b10
1279.5	39	-0.2	1279.7	b11	1279.5	66	-0.1	1279.7	b11
1336.5	25	-0.2	1336.7	b12	1336.5	26	-0.2	1336.7	b12
669.1	14	0.2	668.8	b12+2					
					1433.4	1	-0.3	1433.7	b13
					717.5	3	0.1	717.4	b13+2
					1546.7	8	-0.1	1546.8	b14
773.8	7	-0.2	773.9	b14+2	773.9	6	0.0	773.9	b14+2
1659.8	5	-0.1	1659.9	b15	1659.8	9	-0.1	1659.9	b15
830.6	9	0.2	830.5	b15+2					
1773.9	5	0.0	1773.9	b16	1773.6	4	-0.3	1773.9	b16
					1888.0	2	0.0	1888.0	b17
944.7	7	0.2	944.5	b17+2	944.3	1	-0.2	944.5	b17+2
					1001.5	10	0.5	1001.0	b18+2
1057.2	31	-0.4	1057.6	b19+2	1057.4	7	-0.1	1057.6	b19+2
					1186.6	2	0.5	1186.1	b21+2
					1243.8	2	0.1	1243.6	b22+2
1301.1	14	0.4	1300.7	b23+2	1300.7	5	0.1	1300.7	b23+2
					1364.7	1	0.0	1364.7	b24+2
1421.5	20	0.3	1421.2	b25+2	1421.3	2	0.0	1421.2	b25+2
					1478.3	3	0.0	1478.3	b26+2
					389.4	2	0.1	389.3	y3
502.2	12	-0.1	502.3	y4	502.4	2	0.0	502.3	y4
630.8	13	0.4	630.4	y5	630.5	7	0.1	630.4	y5
					744.5	10	0.0	744.4	y6
859.6	21	0.1	859.5	y7	859.4	12	0.0	859.5	y7
988.1	30	-0.4	988.5	y8	988.4	13	-0.1	988.5	y8
					494.4	1	-0.4	494.8	y8+2
1116.7	22	0.1	1116.6	y9	1116.5	16	-0.1	1116.6	y9
1229.7	23	0.1	1229.6	y10	1229.6	18	0.0	1229.6	y10
615.5	7	0.1	615.3	y10+2	615.7	4	0.4	615.3	y10+2
					1342.7	4	0.0	1342.7	y11
					671.9	4	0.1	671.9	y11+2
1456.8	5	0.1	1456.8	y12	1456.8	6	0.0	1456.8	y12
729.3	12	0.4	728.9	y12+2	729.1	2	0.2	728.9	y12+2
					1570.6	14	-0.2	1570.8	y13
786.2	11	0.3	785.9	y13+2	786.1	9	0.2	785.9	y13+2
1683.7	16	-0.2	1683.9	y14	1684.1	3	0.2	1683.9	y14
					842.6	12	0.1	842.5	y14+2
					1893.7	4	-0.3	1894.0	y16
					947.5	31	0.0	947.5	y16+2
					1950.8	3	-0.2	1951.1	y17
976.0	91	0.0	976.0	y17+2	976.0	72	0.0	976.0	y17+2
1032.7	84	0.1	1032.6	y18+2	1032.8	99	0.2	1032.6	y18+2
1090.0	12	-0.1	1090.1	y19+2	1089.8	4	-0.3	1090.1	y19+2
1133.9	99	0.2	1133.6	y20+2	1133.6	59	-0.1	1133.6	y20+2
1190.1	46	0.0	1190.1	y21+2	1190.1	33	-0.1	1190.1	y21+2
793.6	8	-0.2	793.8	y21+3					
1254.8	57	0.2	1254.7	y22+2	1254.9	88	0.2	1254.7	y22+2
					836.3	1	-0.4	836.8	y22+3
1311.4	100	0.2	1311.2	y23+2	1311.1	58	-0.1	1311.2	y23+2
1376.1	66	0.4	1375.7	y24+2	1376.0	58	0.3	1375.7	y24+2
917.9	6	0.4	917.5	y24+3					
1444.5	92	0.2	1444.3	y25+2	1444.5	58	0.2	1444.3	y25+2
963.2	29	0.0	963.2	y25+3	963.4	70	0.3	963.2	y25+3
					1508.6	4	0.3	1508.3	y26+2

Table 5-20. TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in m/z, calculated m/z values, and ion assignments are provided

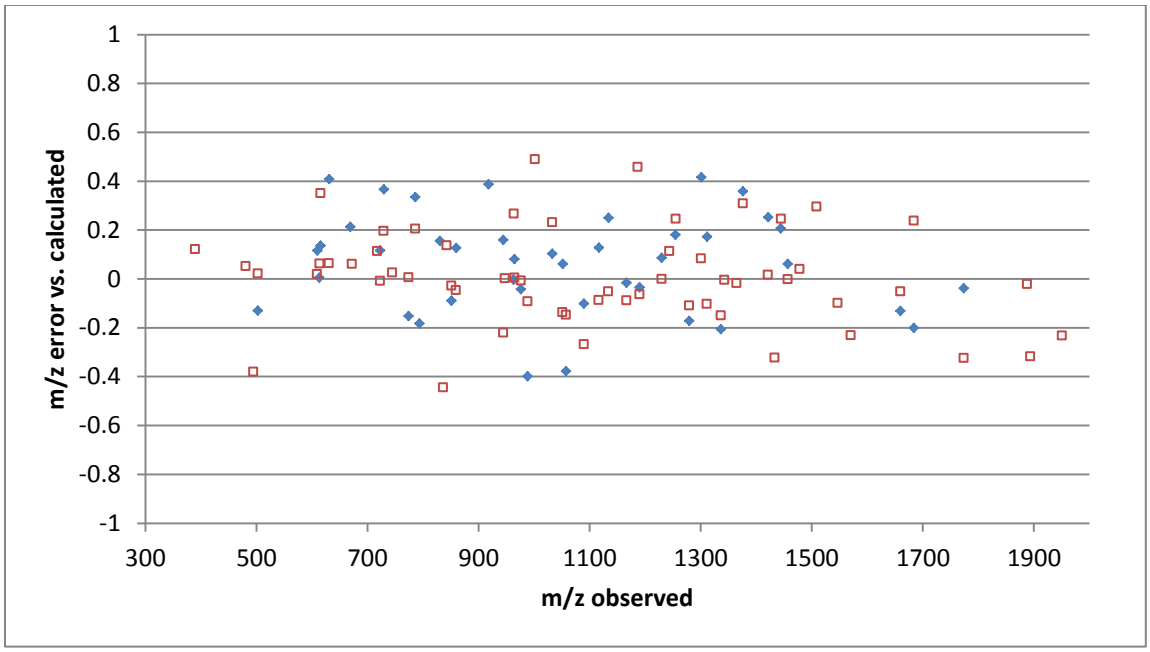
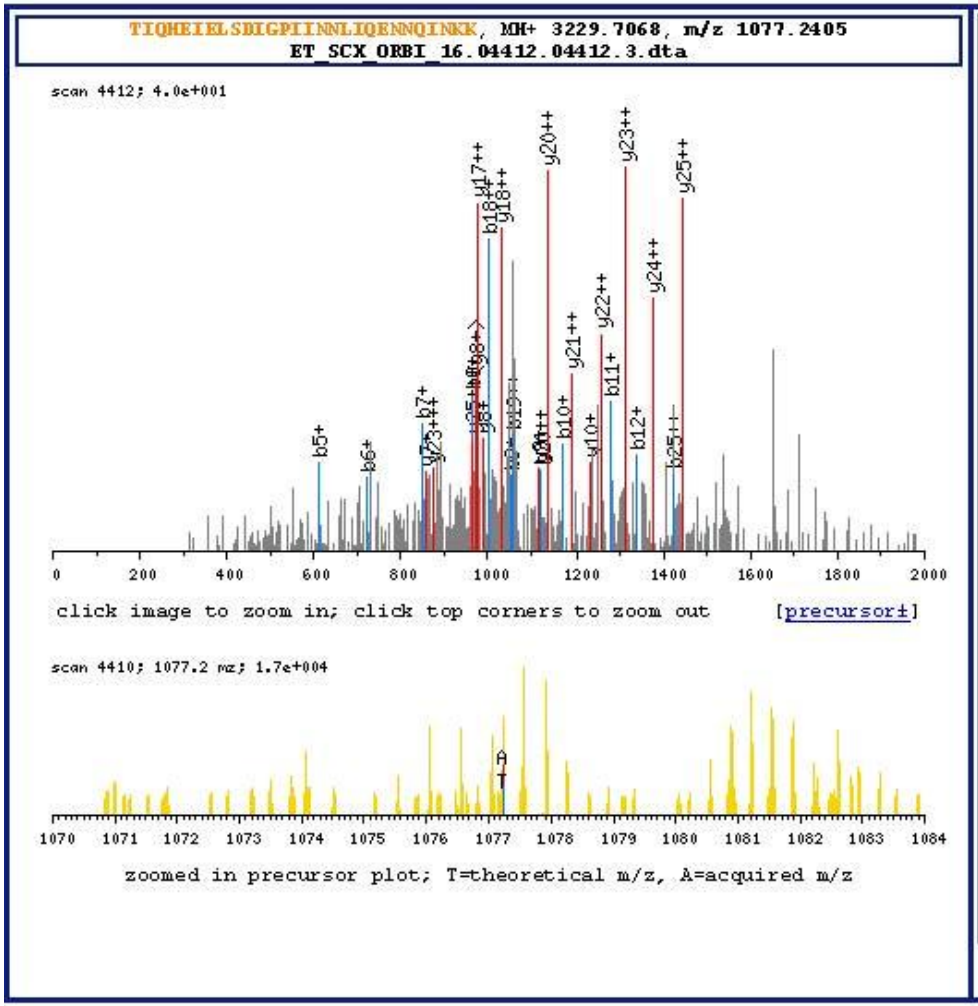


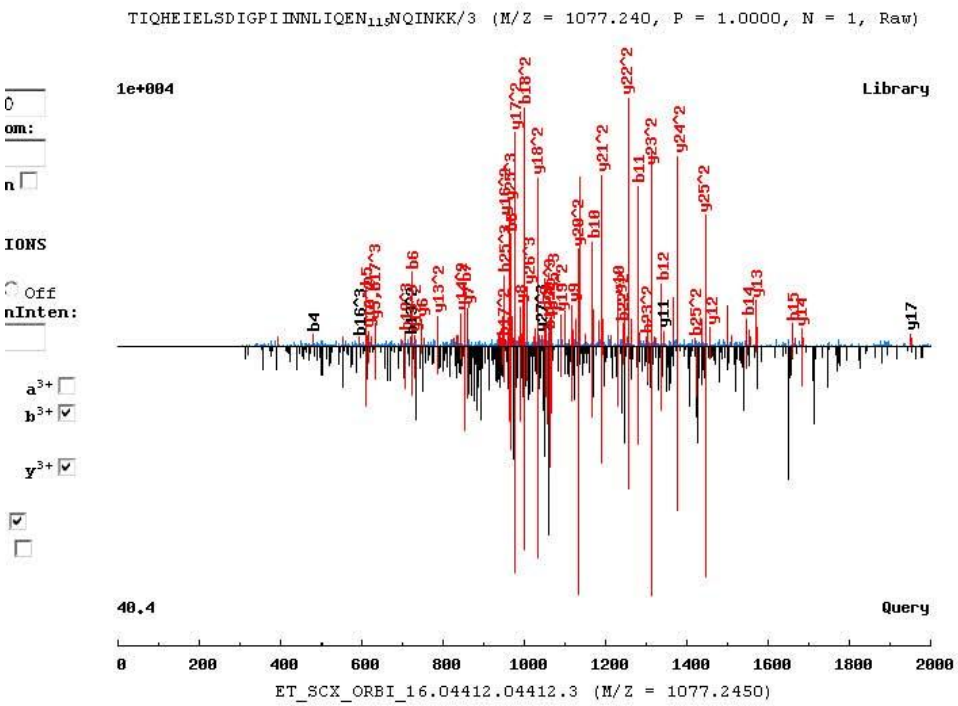
Figure 5-12. TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) mass error in manually assigned b and y ions for metaproteomics versus standard peptides. Y-axis and X-axis values are given in m/z. Metaproteomics assignments are provided with a closed blue diamond (♦), and standard peptide data with an open red square (□).



b^+	b^{2+}	#	AA	#	y^+	y^{2+}	y^{3+}
102.0555	51.5317	1	T	28			
215.1396	108.0737	2	I	27	3128.6591	1564.8335	1043.5583
343.1981	172.1030	3	Q	26	3015.5751	1508.2914	1005.8636
480.2571	240.6324	4	H	25	2887.5165	**1444.2622**	**963.1774**
609.2997	305.1537	5	E	24	2750.4576	**1375.7327**	917.4911
722.3837	361.6958	6	I	23	2621.4150	**1311.2114**	**874.4769**
851.4263	426.2171	7	E	22	2508.3309	**1254.6694**	836.7822
964.5104	482.7591	8	L	21	2379.2883	**1190.1481**	793.7680
1051.5424	526.2751	9	S	20	2266.2043	**1133.6060**	756.0733
1166.5693	583.7886	10	D	19	2179.1722	1090.0900	727.0626
1279.6534	640.3306	11	I	18	2064.1453	**1032.5766**	688.7203
1336.6749	668.8413	12	G	17	1951.0612	**976.0345**	651.0256
1433.7276	717.3677	13	P	16	1894.0398	947.5238	632.0185
1546.8117	773.9098	14	I	15	1796.9870	898.9974	599.6675
1659.8958	830.4518	15	I	14	1683.9029	842.4554	561.9729
1773.9387	887.4733	16	N	13	1570.8189	785.9133	524.2782
1887.9816	944.4947	17	N	12	1456.7759	728.8919	486.2639
2001.0657	**1001.0368**	18	L	11	1342.7330	671.8704	448.2496
2114.1497	**1057.5788**	19	I	10	**1229.6489**	615.3284	410.5549
2242.2083	**1121.6081**	20	Q	9	**1116.5649**	558.7864	372.8602
2371.2509	1186.1294	21	E	8	**988.5063**	494.7571	330.1740
2486.2778	1243.6428	22	N	7	**859.4637**	430.2358	287.1598
2600.3207	1300.6643	23	N	6	744.4368	372.7223	248.8175
2728.3793	1364.6936	24	Q	5	630.3939	315.7009	210.8032
2841.4634	**1421.2356**	25	I	4	502.3353	251.6716	168.1170
2955.5063	1478.2571	26	N	3	389.2512	195.1295	130.4223
3083.6013	1542.3046	27	K	2	275.2083	138.1081	92.4080
		28	K	1	147.1134	74.0606	49.7097

N(22):+115.03

Figure 5-13. TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) automatically assigned spectrum from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectra as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b^{1+}	b^{2+}	b^{3+}	#	AA	#	y^{1+}	y^{2+}	y^{3+}
102.0550	51.5311	34.6898	1	T	28			
215.1390	108.0731	72.3845	2	I	27	3128.6586	1564.8329	**1043.5577**
343.1976	172.1024	115.0707	3	Q	26	3015.5745	1508.2909	**1005.8630**
480.2565	240.6319	160.7570	4	H	25	2887.5159	**1444.2616**	**963.1768**
609.2991	305.1532	203.7712	5	E	24	2750.4570	**1375.7322**	917.4905
722.3832	361.6952	241.4659	6	I	23	2621.4144	**1311.2109**	874.4763
851.4258	426.2165	284.4801	7	E	22	2508.3304	**1254.6688**	836.7816
964.5098	482.7585	322.1748	8	L	21	2379.2878	**1190.1475**	793.7674
1051.5418	526.2746	351.1855	9	S	20	2266.2037	**1133.6055**	756.0728
1166.5688	583.7880	389.5278	10	D	19	2179.1717	**1090.0895**	727.0621
1279.6528	640.3301	427.2225	11	I	18	2064.1448	**1032.5760**	688.7198
1336.6743	668.8408	446.2296	12	G	17	**1951.0607**	**976.0340**	651.0251
1433.7271	**717.3672**	478.5805	13	P	16	1894.0392	**947.5233**	632.0179
1546.8111	773.9092	516.2752	14	I	15	1796.9865	898.9969	599.6670
1659.8952	830.4512	553.9699	15	I	14	**1683.9024**	**842.4548**	561.9723
1773.9381	887.4727	**591.9842**	16	N	13	**1570.8184**	**785.9128**	524.2776
1887.9810	**944.4942**	**629.9985**	17	N	12	**1456.7754**	**728.8914**	486.2633
2001.0651	**1001.0362**	667.6932	18	L	11	**1342.7325**	671.8699	448.2490
2114.1492	**1057.5782**	**705.3879**	19	I	10	**1229.6484**	**615.3279**	410.5543
2242.2077	1121.6075	748.0741	20	Q	9	**1116.5644**	558.7858	372.8596
2371.2503	1186.1288	791.0883	21	E	8	**988.5058**	494.7565	330.1735
2486.2773	**1243.6423**	829.4306	22	N[115]	7	**859.4632**	430.2352	287.1593
2600.3202	**1300.6637**	867.4449	23	N	6	**744.4363**	372.7218	248.8169
2728.3788	1364.6930	910.1311	24	Q	5	**630.3933**	315.7003	210.8026
2841.4628	**1421.2351**	**947.8258**	25	I	4	502.3347	251.6710	168.1164

Figure 5-14. TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) comparison between the authentic standard peptide and the metaproteomics spectrum. Spectra are plotted with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is on top and the metaproteomics spectrum is on the bottom and inverted. Assigned b and y product ions present in both spectra are noted in tabular format in bold.

For the N-deamidated peptide TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) a long series of b and y ions in both the 1+ and 2+ charge state can be assigned to both the metaproteomics and authentic standard peptides (**Table 5-20**). As illustrated in **Figure 5-13** the error in mass assignment follows a similar distribution for both metaproteomics and standard peptides. In **Figure 5-13** with the assigned product and parent ion spectra, long series of sequential b and y ions are assigned. As well, the parent ion spectrum (**Figure 5-13**) clearly illustrates that the correct monoisotopic parent ion was assigned. Near identical intensity distributions of the long b and y ion series are noted in **Figure 5-14**. Taken together these data strongly suggest that the assignment of the peptide sequence TIQHEIELSDIGPIINNLIQEN$_{115}$NQINKK (3+) to the metaproteomics data is an excellent match.

2+ RPIELR, 782.4762 Da									
782.4787 Da, 392.2466 m/z , 3 ppm					**782.47683, 392.2457 m/z, 1 ppm**				
O_19.1621					**Standard**				
Exptl	**I**	**d(m/z)**	**Calc**	**Assign**	**Exptl**	**I**	**d(m/z)**	**Calc**	**Assign**
254.3	2	0.1	254.1612	b2	254.2	3	0.1	254.16	b2
367.5	3	0.2	367.2452	b3	367.3	3	0.0	367.25	b3
496.3	100	0.0	496.2878	b4	496.3	100	0.1	496.29	b4
609.5	11	0.1	609.3719	b5	609.4	10	0.0	609.37	b5
175.1	5	0.0	175.119	y1	175.1	7	0.0	175.12	y1
288.2	89	0.0	288.203	y2	288.3	100	0.1	288.2	y2
417.0	1	-0.3	417.2456	y3					
530.4	2	0.0	530.3297	y4	530.4	3	0.1	530.33	y4
627.5	17	0.1	627.3824	y5	627.4	20	0.0	627.38	y5
314.3	2	0.1	314.1949	y5+2	314.4	3	0.2	314.19	y5+2

Table 5-21. RPIELR (2+) metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in m/z, calculated m/z values, and ion assignments are provided.

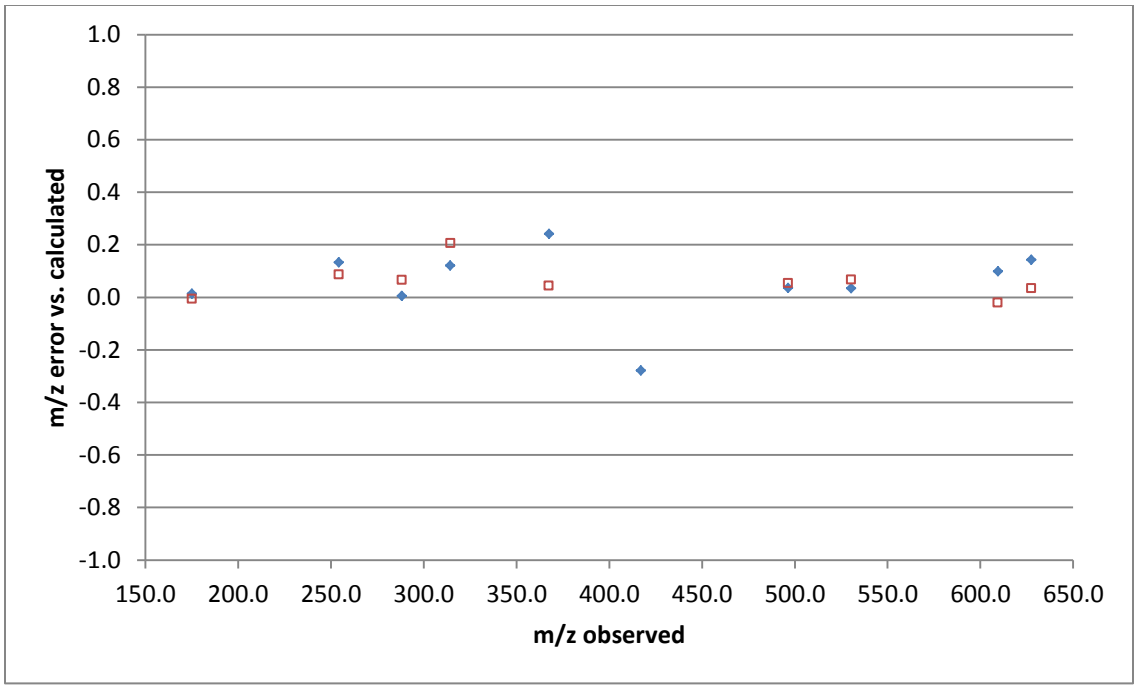
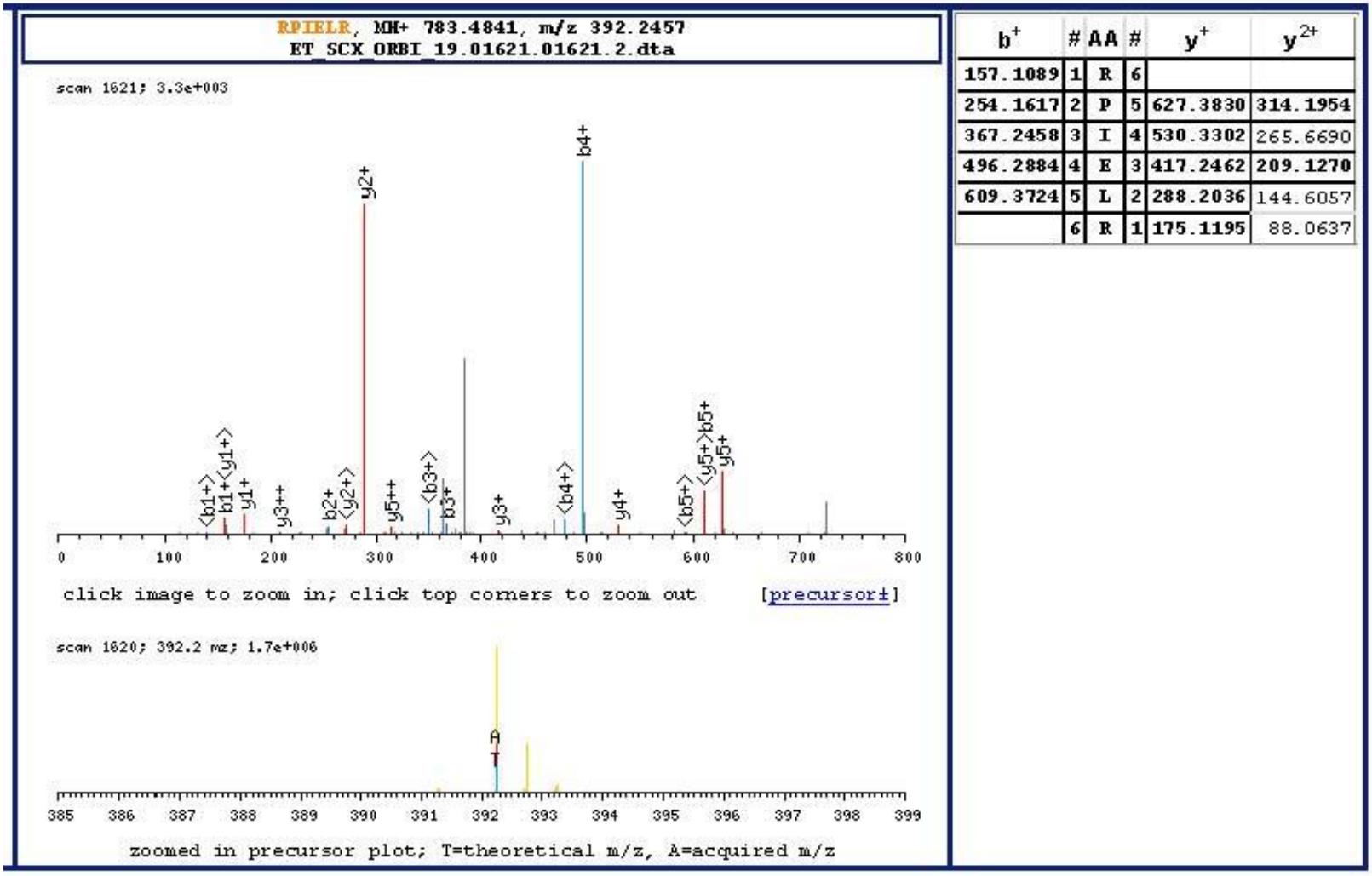
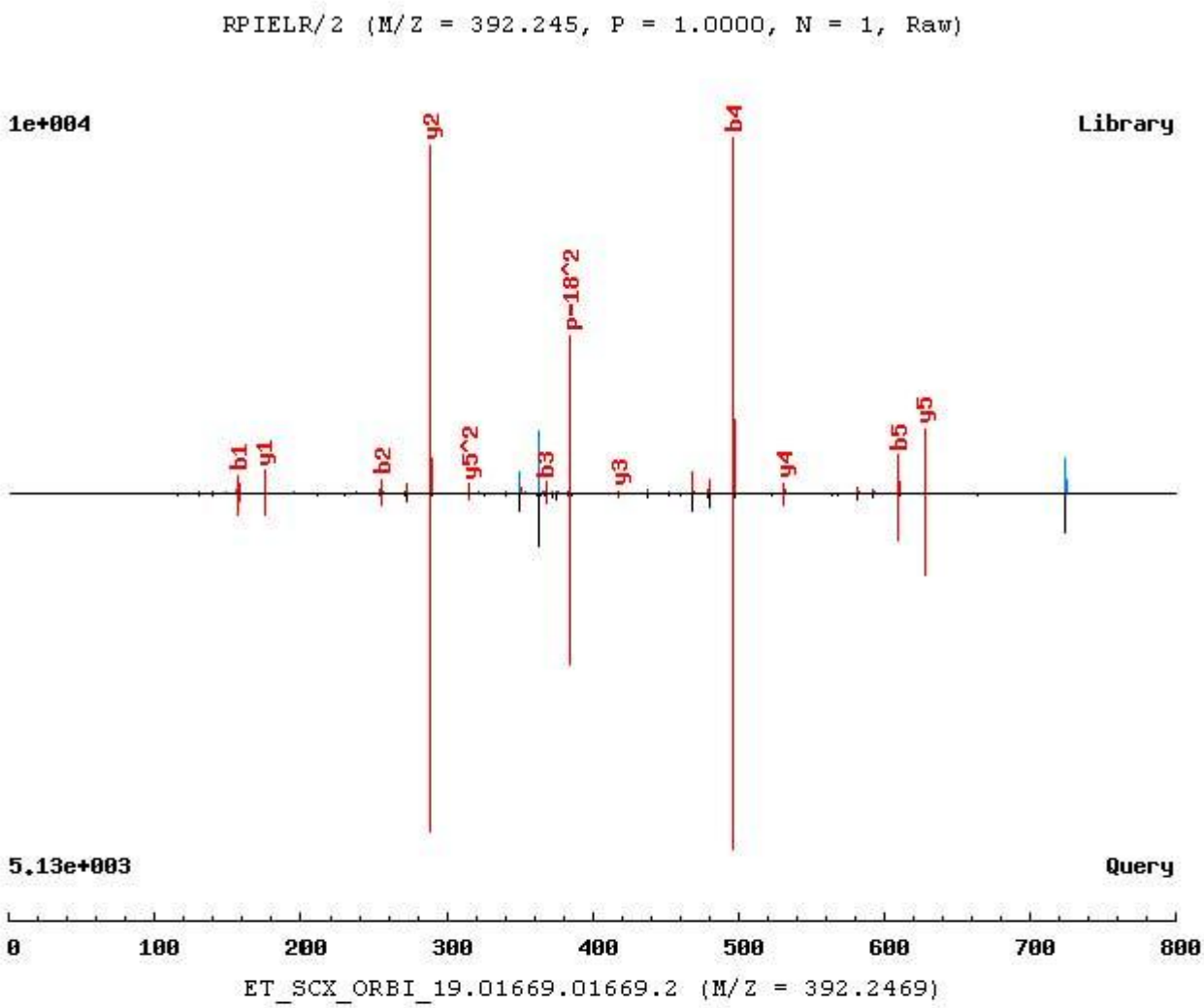


Figure 5-15. RPLIER (2+) mass error in manually assigned b and y ions for metaproteomics versus standard peptides. Y-axis and X-axis values are given in m/z. Metaproteomics assignments are provided with a closed blue diamond (♦), and standard peptide data with an open red square (□).



b^+	#	AA	#	y^+	y^{2+}
157.1089	1	R	6		
254.1617	2	P	5	627.3830	314.1954
367.2458	3	I	4	530.3302	265.6690
496.2884	4	E	3	417.2462	209.1270
609.3724	5	L	2	288.2036	144.6057
	6	R	1	175.1195	88.0637

Figure 5-16. RPIELR (2+) automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectrum as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
157.1084	79.0578	**1**	**R**	**6**		
254.1612	127.5842	**2**	**P**	**5**	**627.3824**	**314.1948**
367.2452	184.1262	**3**	**I**	**4**	**530.3297**	265.6685
496.2878	248.6475	**4**	**E**	**3**	**417.2456**	209.1264
609.3719	305.1896	**5**	**L**	**2**	**288.2030**	144.6051
		6	**R**	**1**	**175.1190**	88.0631

Figure 5-17. RPIELR (2+) comparison between the authentic standard peptide and metaproteomics spectrum. Spectra are illustrated with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is on top, and the metaproteomics spectrum is on the bottom and inverted. Assigned b and y product ions present in both spectra are noted in tabular format in bold.

For RPIELR (2+), complete b and y ion series are observed (b_1 ions are not typically observed) (**Table 5-21**). As illustrated in **Figure 5-15** the error in mass assignment follows a similar distribution for both metaproteomics and standard peptides with individual pairs often being closer than 0.1 Da. In **Figure 5-16** complete b and y ion series are assigned. In addition, the parent ion spectrum illustrates that the correct monoisotopic parent ion was assigned. Near identical intensity distributions between the complete b and y ion series are noted in **Figure 5-17**. Taken together these data strongly suggest that the assignment of the peptide sequence RPIELR (2+) to the metaproteomics data is an excellent match—it is a "text-book" CID spectrum with abundant cleavage after E and before P.[66]

2_LLDVGGGTAINAIALAK, 1595.9247 Da																		
798.97772, 10 ppm				798.97766, 10 ppm				798.97748, 10 ppm				798.97766, 10 ppm				798.96161, -10 ppm			
Scan 4505				Scan 4507				Scan 4509				Scan 4512				Standard			
m/z	I	ion	dPPM	m/z	I	ion	dPPM	m/z	I	ion	dPPM	m/z	I	ion	dPPM	m/z	I	ion	dPPM
227.1761	7	b2	3.0	227.1750	5	b2	-1.8	227.1753	4	b2									
342.2030	45	b3	2.0	342.2028	46	b3	1.2	342.2029	48	b3	0.3	342.2029	45	b3	1.6	342.2025	31	b3	0.5
441.2724	29	b4	3.7	441.2713	24	b4	1.2	441.2718	27	b4	2.3	441.2723	25	b4	3.5	441.2717	28	b4	2.1
498.2945	19	b5	4.6	498.2939	17	b5	3.4	498.2938	16	b5	-0.1	498.2930	18	b5	1.5	498.2927	26	b5	0.9
555.3150	10	b6	2.3	555.3176	8	b6	7.0	555.3149	8	b6	4.2	555.2529	6	b6	0.3	555.3146	12	b6	1.7
612.3411	6	b7	9.7	612.3348	8	b7	-0.5	612.3302	6	b7	18.0	612.3419	7	b7	11.0	612.3379	12	b7	4.6
713.3881	8	b8	7.4	713.3853	9	b8	3.5	713.3883	8	b8	4.4	713.3828	8	b8	0.0	713.3859	9	b8	4.3
784.4230	10	b9	3.8	784.4258	10	b9	7.4	784.4210	16	b9	6.8	784.4219	8	b9	2.5	784.4181	4	b9	-2.3
				897.5172	9	b10	15.0	897.5145	6	b10	11.0					897.5051	9	b10	1.2
218.0069	3	y2	0.1	218.1499	8	y2	-0.3	218.1500	6	y2	0.5	218.1499	6	y2	0.1				
331.2345	19	y3	1.4	331.2343	13	y3	0.9	331.2341	18	y3	-0.3	331.2341	13	y3	0.3	331.2340	6	y3	0.0
402.2718	45	y4	1.8	402.2722	47	y4	2.8	402.2723	47	y4	1.6	402.2722	49	y4	2.8				
515.3573	14	y5	4.2	515.3570	10	y5	3.5					515.3555	11	y5	0.6	515.3568	15	y5	3.3
586.3941	9	y6	3.1	586.3934	11	y6	1.8	586.3229	10	y6	2.1	586.3264	7	y6	4.4	586.3943	13	y6	3.5
700.4393	42	y7	5.8	700.4400	43	y7	6.9	700.4383	45	y7	1.7	700.4382	40	y7	4.3	700.4386	41	y7	4.8
813.5251	65	y8	7.1	813.5269	63	y8	9.3	813.5265	60	y8	1.4	813.5240	58	y8	5.8	813.5109	100	y8	-10.0
884.5675	47	y9	13.0	884.5666	41	y9	12.0	884.5659	40	y9	16.0	884.5648	38	y9	9.5	884.5583	56	y9	2.1
985.6155	15	y10	12.0	985.6160	16	y10	12.0	985.6018	14	y10	12.0	985.6152	14	y10	11.0	985.6117	14	y10	7.8
1042.6390	14	y11	13.0	1042.6404	19	y11	14.0	1042.6459	16	y11	-2.3	1042.6374	17	y11	11.0	1042.6316	9	y11	5.9
1099.6610	29	y12	1.9	1099.6636	32	y12	15.0	1099.6619	22	y12	8.8	1099.6676	22	y12	19.0	1099.6593	7	y12	11.0
1156.6856	52	y13	20.0	1156.6822	53	y13	12.0	1156.6824	49	y13	24.0	1156.6844	48	y13	14.0	1156.6753	15	y13	6.0
1255.7642	6	y14	24.0	1255.7432	8	y14	5.1	1255.7427	10	y14	15.0								

Table 5-22. LLDVGGGTAINAIALAK (2+) metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in ppm, calculated m/z values, and ion assignments are provided

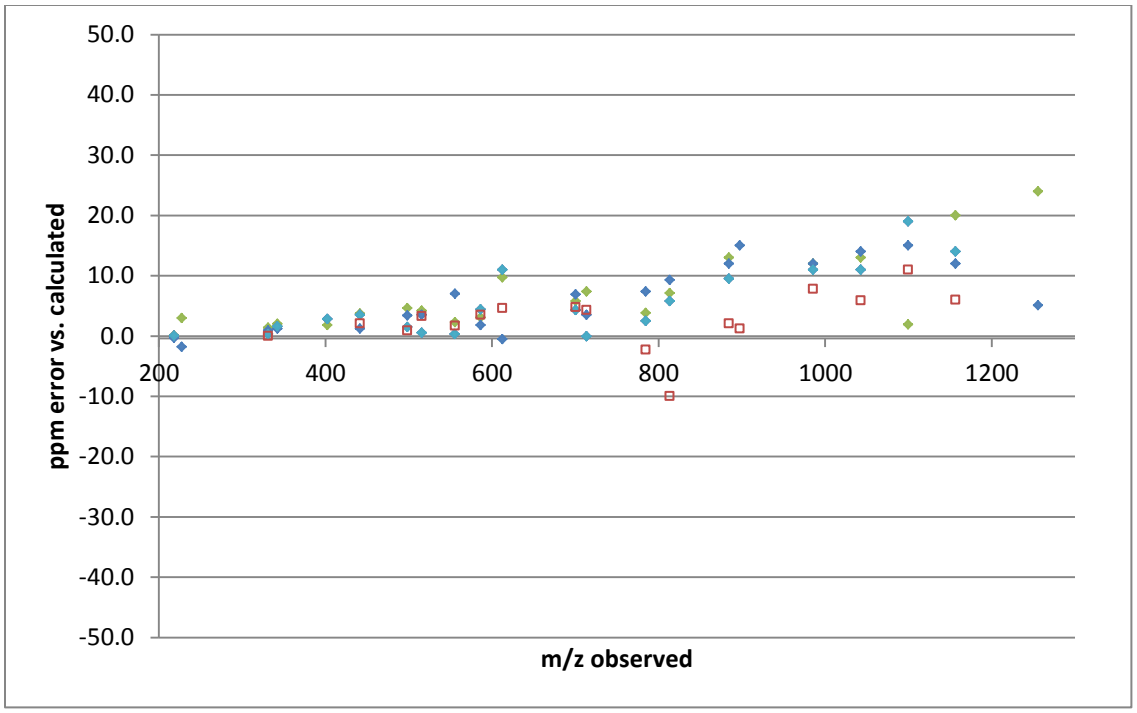
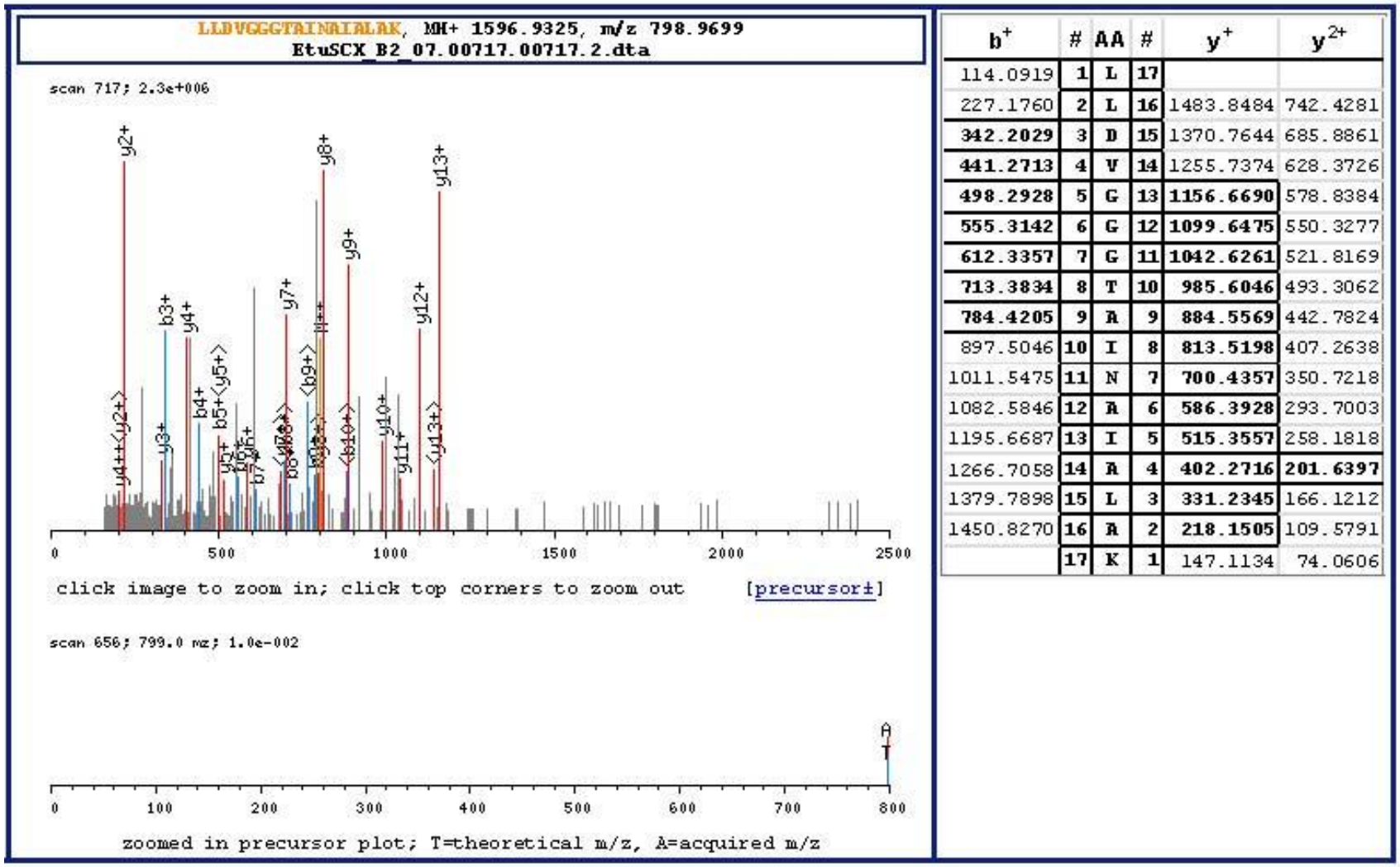
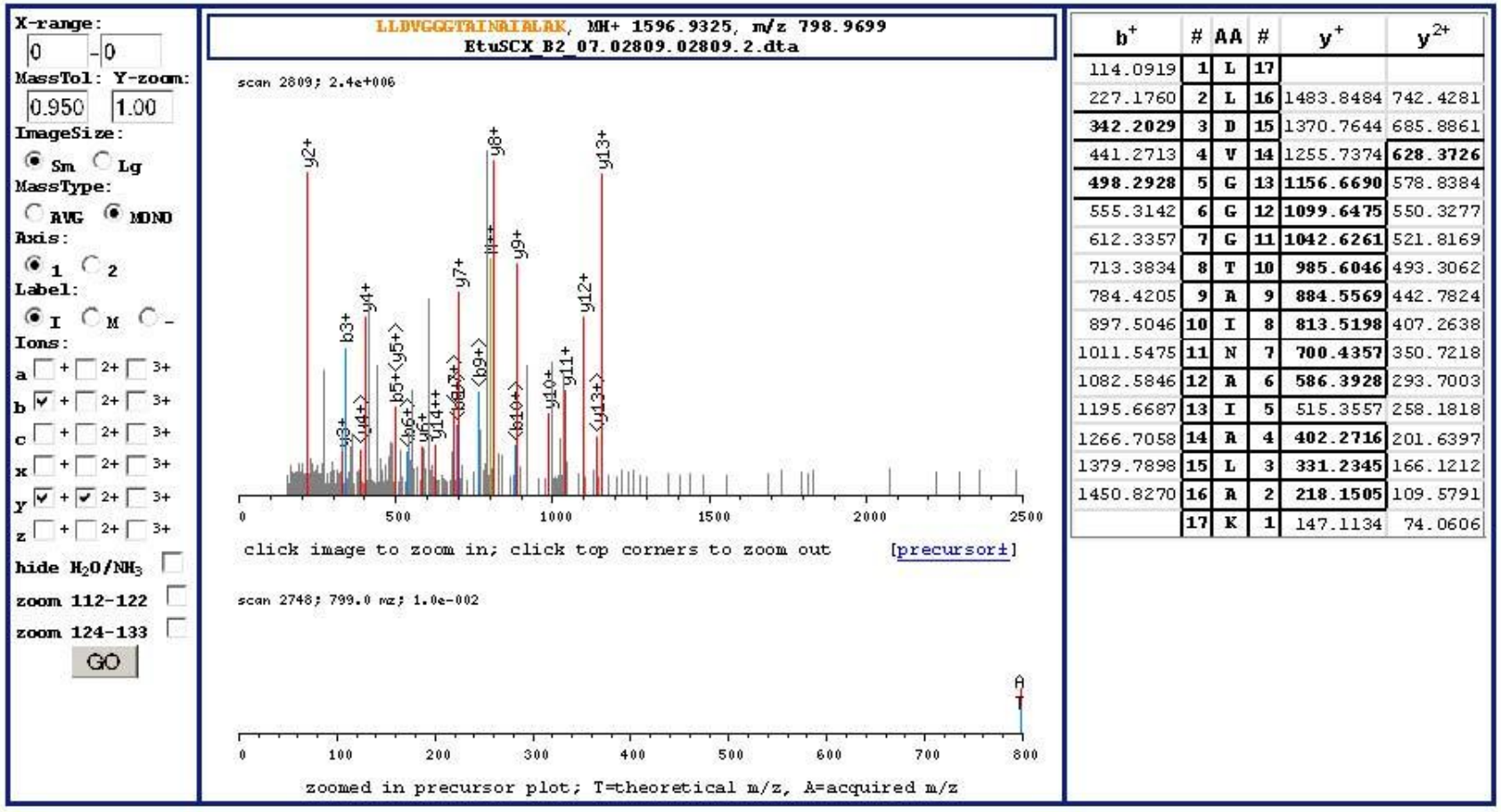


Figure 5-18. LLDVGGGTAINAIALAK (2+) mass error in manually assigned b and y ions for metaproteomics versus standard peptides. Y-axis are in PPM and X-axis values are given in m/z. Metaproteomics assignments are provided with closed blue diamonds (Scan 4505: ♦, Scan 4507: ♦, Scan 4509: ♦, Scan 4512: ♦,), and standard peptide data with an open red square (□).



b^+	#	AA	#	y^+	y^{2+}
114.0919	1	L	17		
227.1760	2	L	16	1483.8484	742.4281
342.2029	3	D	15	1370.7644	685.8861
441.2713	4	V	14	1255.7374	628.3726
498.2928	5	G	13	**1156.6690**	578.8384
555.3142	6	G	12	**1099.6475**	550.3277
612.3357	7	G	11	**1042.6261**	521.8169
713.3834	8	T	10	**985.6046**	493.3062
784.4205	9	A	9	**884.5569**	442.7824
897.5046	10	I	8	**813.5198**	407.2638
1011.5475	11	N	7	**700.4357**	350.7218
1082.5846	12	A	6	**586.3928**	293.7003
1195.6687	13	I	5	**515.3557**	258.1818
1266.7058	14	A	4	**402.2716**	**201.6397**
1379.7898	15	L	3	**331.2345**	166.1212
1450.8270	16	A	2	**218.1505**	109.5791
	17	K	1	147.1134	74.0606

Figure 5-19. LLDVGGGTAINAIALAK (2+) automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectra as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b+	#	AA	#	y+	y2+
114.0919	1	L	17		
227.1760	2	L	16	1483.8484	742.4281
342.2029	3	D	15	1370.7644	685.8861
441.2713	4	V	14	1255.7374	**628.3726**
498.2928	5	G	13	**1156.6690**	578.8384
555.3142	6	G	12	**1099.6475**	550.3277
612.3357	7	G	11	**1042.6261**	521.8169
713.3834	8	T	10	**985.6046**	493.3062
784.4205	9	A	9	**884.5569**	442.7824
897.5046	10	I	8	**813.5198**	407.2638
1011.5475	11	N	7	**700.4357**	350.7218
1082.5846	12	A	6	**586.3928**	293.7003
1195.6687	13	I	5	515.3557	258.1818
1266.7058	14	A	4	**402.2716**	201.6397
1379.7898	15	L	3	**331.2345**	166.1212
1450.8270	16	A	2	**218.1505**	109.5791
	17	K	1	147.1134	74.0606

Figure 5-20. LLDVGGGTAINAIALAK (2+) automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectra as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.

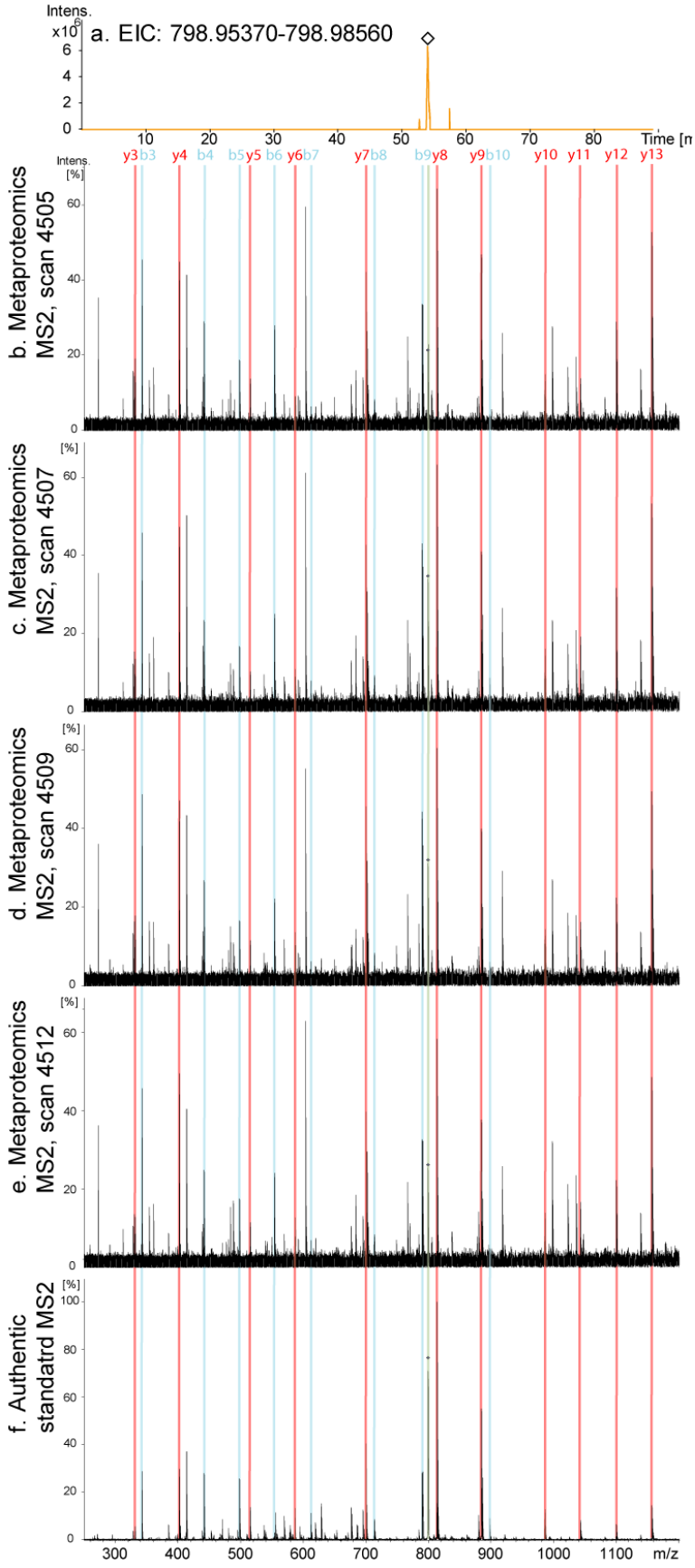


Figure 5-21. LLDVGGGTAINAIALAK (2+) comparison between the authentic standard peptide and the metaproteomics spectra. A parent ion extracted ion chromatogram is illustrated on top at +/-20ppm. Spectra are illustrated with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is at the bottom, and the four metaproteomics spectra are on top. y-ions are illustrated in red, and b-ion series are illustrated in blue. The parent ion is illustrated in green.

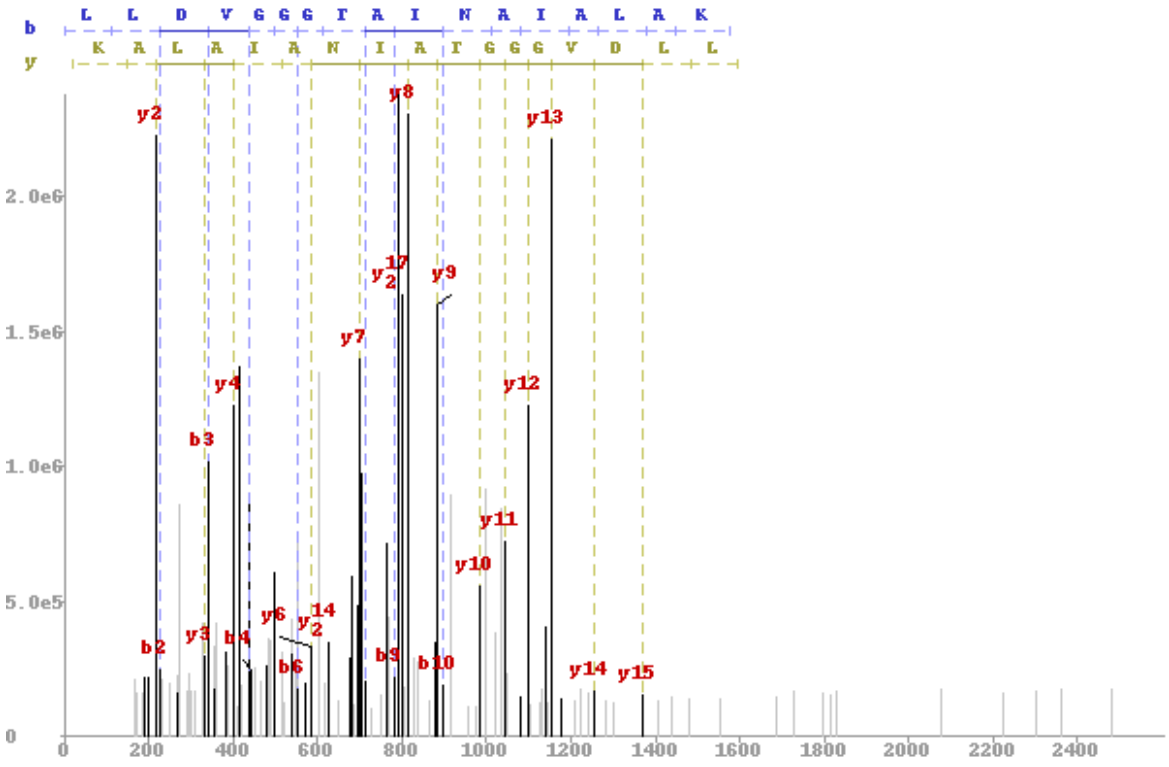


Figure 5-22. LLDVGGGTAINAIALAK (2+) assignment with the online implementation of Inspect. This spectrum is presented as m/z on the x-axis and intensity on the y-axis with the b and y ion ladders noted.

For LLDVGGGTAINAIALAK (2+) two additional spectra could be manually assigned to the peptide. All four spectra correspond to sequential elution times (Scan # 4505, 4507, 4509, and 4512) and two long series of product ions can be observed: b2-10 and y2-14. The sequence coverage for the four metaproteomics spectra is slightly better than that of the authentic standard peptides, all at a mass error of less than 25 ppm (**Table 5-22**). As illustrated in **Figure 5-18** the error in mass assignment follows a similar distribution for both metaproteomics and standard peptides. The typical performance of FTICR-MS is observed with increasing mass error correlating with increasing m/z values. The slight difference in calibration between the metaproteomics and standard peptide data at >800 m/z likely represents differences in instrument calibration (Solarix 12T Q-

FTICR-MS was calibrated with Sodium TFA clusters whereas Apex 7T Q-FTICR-MS was calibrated with HP-mix). In **Figures 5-19** and **5-20** the assigned product ion spectrum display long series of sequential b and y ions assigned. In addition, the parent ion spectrum clearly illustrates that the correct monoisotopic parent ion was assigned. In **Figure 5-21** the near identical distributions of intensity between the long series of b and y ions are also illustrated for all four experimental spectra as well as the authentic standard peptide. In **Figure 5-22**, as spectrum assigned with the program Inspect is shown, with an excellent match. Interestingly, this match could only be observed with the online "live-search" implementation of the program and thus a score could not be assigned for processing in Peptide- and Protein- Prophet. Taken together these data strongly suggest that the assignment of the peptide sequence 2+ LLDVGGGTAINAIALAK to the experimental data is an excellent match.

2+ ILKPCYR,949.5055 Da								
949.5087 Da, 475.7616 m/z , 3 ppm					949.5061 Da, 475.7603 m/z, 1 ppm			
O_15.01581					Standard			
Exptl	d(m/z	Calc	Assign		Exptl	d(m/z)	Calc	Assign
227.1	-0.1	227.2	b2		227.0	-0.2	227.2	b2
355.3	0.0	355.3	b3		355.1	-0.2	355.3	b3
178.1	0.0	178.1	b3+2					
613.2	-0.2	613.3	b5		613.3	0.0	613.3	b5
776.2	-0.2	776.4	b6		776.4	0.0	776.4	b6
389.2	0.4	388.7	b6+2		388.5	-0.3	388.7	b6+2
475.5	-0.2	475.8	MH+2		475.8	0.0	475.8	MH+2
175.0	-0.1	175.1	y1		175.1	0.0	175.1	y1
338.4	0.2	338.2	y2		338.4	0.3	338.2	y2
169.3	-0.3	169.6	y2+2					
499.3	0.1	499.2	y3		499.3	0.1	499.2	y3
249.9	-0.2	250.1	y3+2					
596.3	0.0	596.2	y4		596.3	0.1	596.2	y4
					298.7	0.1	298.6	y4+2
724.4	0.0	724.3	y5		724.4	0.0	724.3	y5
362.7	0.0	362.7	y5+2		362.9	0.2	362.7	y5+2
837.5	0.1	837.4	y6		837.4	0.0	837.4	y6
419.0	-0.2	419.2	y6+2		419.4	0.2	419.2	y6+2

Table 5-23. 2+ ILKPC$_{161}$YR metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in m/z, calculated m/z values, and ion assignments are provided.

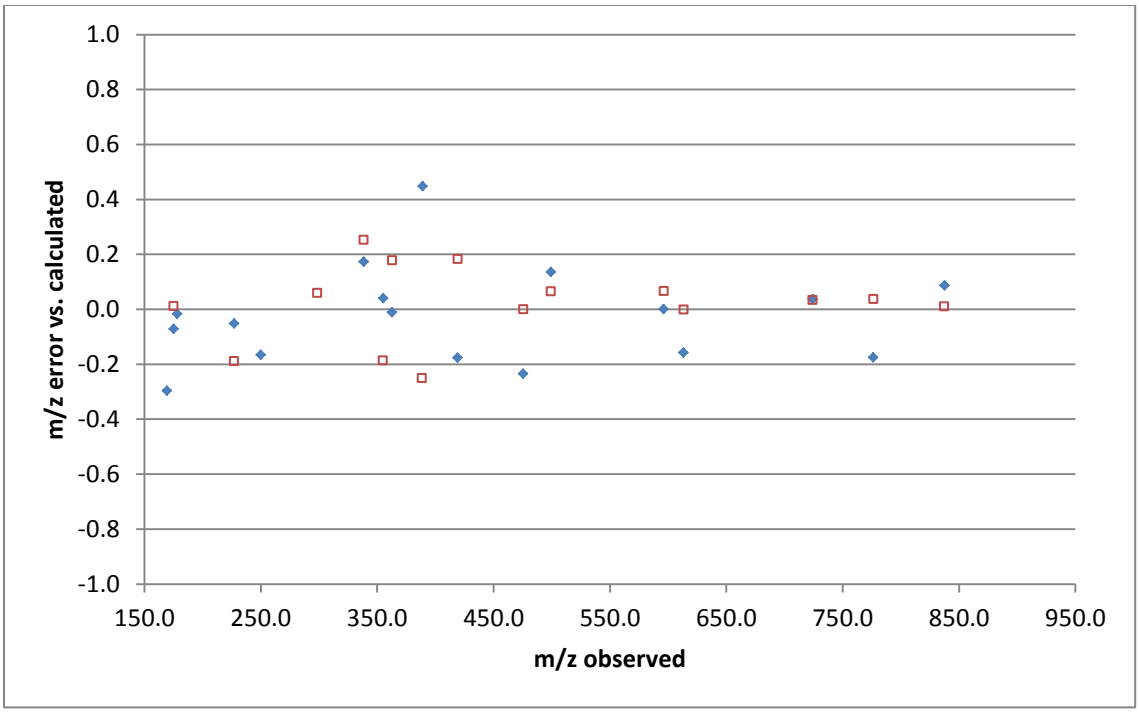
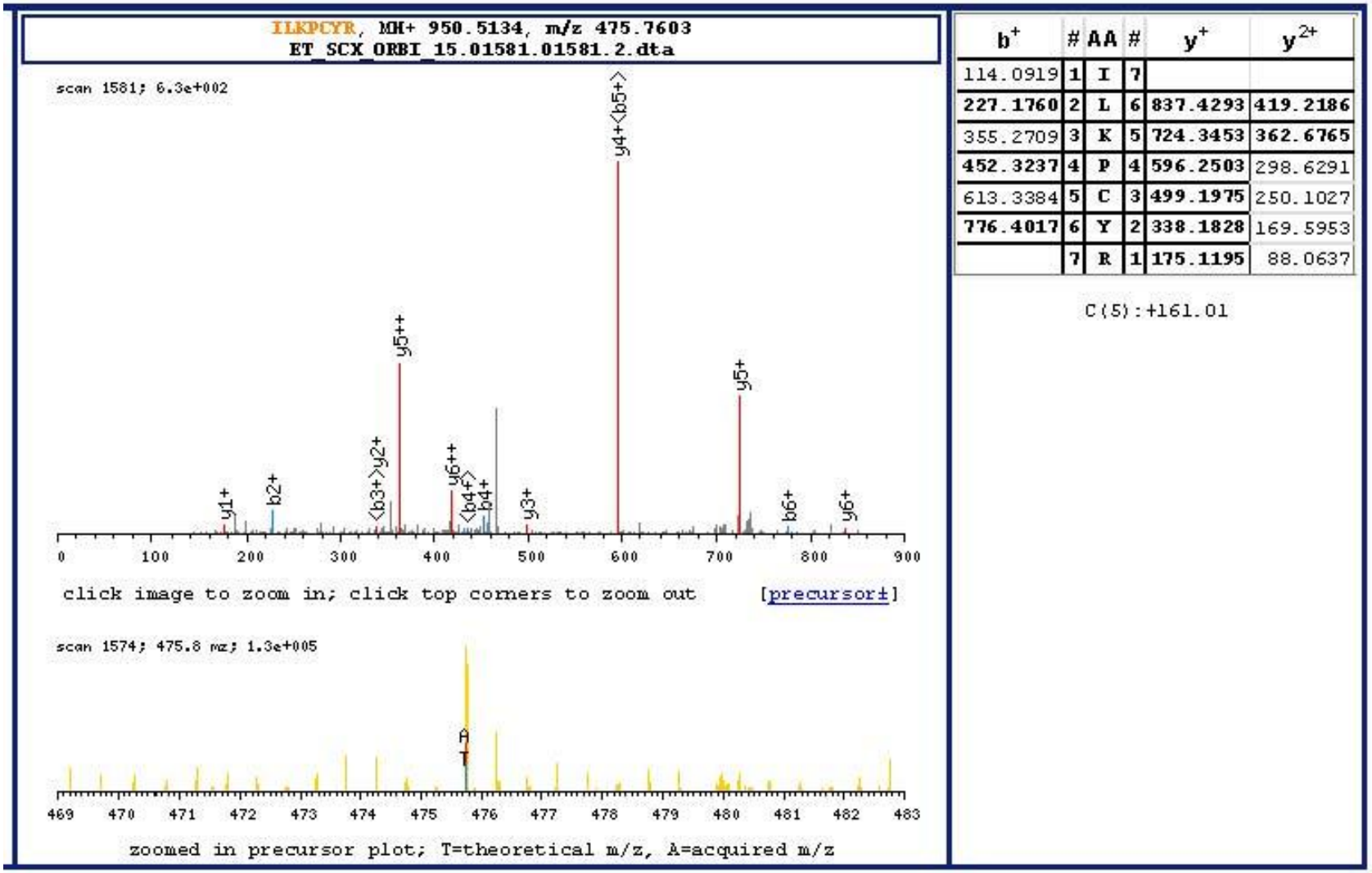
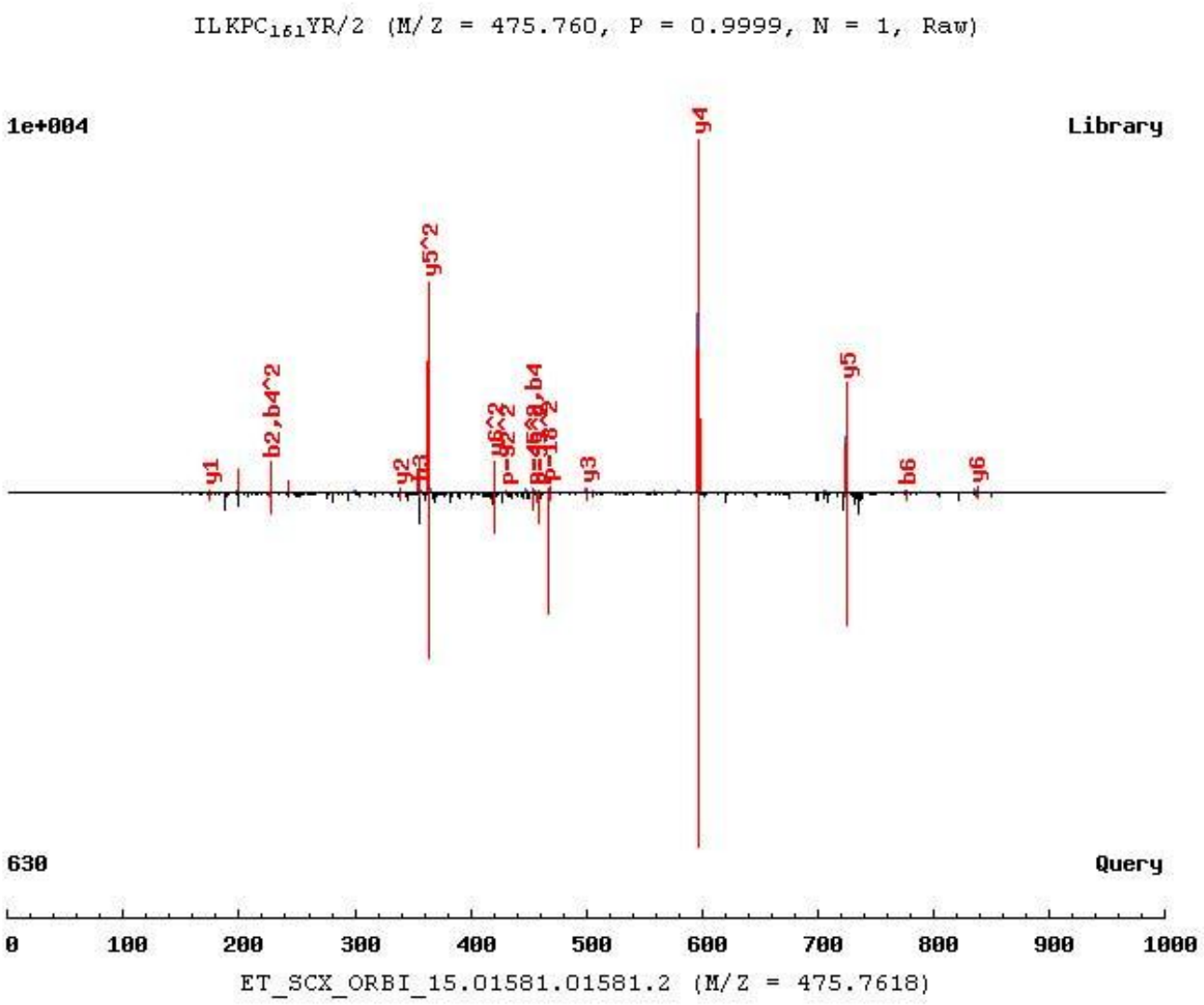


Figure 5-23. ILKPC$_{161}$YR (2+) **mass error in manually assigned b and y ions for metaproteomics versus standard peptides.** Y-axis and X-axis values are given in m/z. Metaproteomics assignments are provided with a closed blue diamond (♦), and standard peptide data with an open red square (□).



b+	#	AA	#	y+	y2+
114.0919	1	I	7		
227.1760	2	L	6	**837.4293**	**419.2186**
355.2709	3	K	5	**724.3453**	**362.6765**
452.3237	4	P	4	**596.2503**	298.6291
613.3384	5	C	3	**499.1975**	250.1027
776.4017	6	Y	2	**338.1828**	169.5953
	7	R	1	**175.1195**	88.0637

Figure 5-24. 2+ ILKPC$_{161}$YR automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectrum as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
114.0913	57.5493	1	I	7		
227.1754	114.0913	2	L	6	**837.4287**	**419.2180**
355.2704	178.1388	3	K	5	**724.3447**	**362.6760**
452.3231	**226.6652**	4	P	4	**596.2497**	298.6285
613.3378	307.1725	5	C[161]	3	**499.1969**	250.1021
776.4011	388.7042	6	Y	2	**338.1823**	169.5948
		7	R	1	**175.1190**	88.0631

Figure 5-25. ILKPCYR (2+) comparison between the authentic standard peptide and metaproteomics spectrum. Spectra are illustrated with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is on top, and the metaproteomics spectrum is on the bottom and inverted. Assigned b and y product ions present in both spectra are noted in tabular format in bold.

For 2+ ILKPCYR 4/6 b ions and all 6/6 y ions are observed with an excellent match as compared to the authentic standard peptide (**Table 5-23**). As illustrated in **Figure 5-23** the error in mass assignment follows a similar distribution for both experimental and standard peptides, indeed, individual pairs are often closer than 0.1 m/z. In **Figure 5-24** the assigned product and parent ion spectra, the complete 3/6 b and 6/6 y ions are assigned. As well, the parent ion spectrum clearly illustrates that the correct monoisotopic parent ion was assigned. In **Figure 5-25** the near identical distributions of intensity between the complete b and y ion series are also illustrated. This low mass peptide has a complete y-ion series and multiple b-ions as well. Indeed the ion assigned at 596.3 could be either the y4 or b5 (as automatically assigned). The y4 ion may actually be correct as it would correspond to a favored cleavage N-terminal to proline.

2+ GSNIHYDLENDHNDYEK, 2061.8667 Da									
2061.8767 Da, 1031.9456 m/z, 5 ppm					2061.8670 Da, 1031.9408 m/z, 0 ppm				
O_17.2123					Standard				
Exptl	I	d(m/z)	Calc	Assign	Exptl	I	d(m/z)	Calc	Assign
509.4	14	0.2	509.2	b5	509.3	9	0.0	509.2	b5
672.3	19	-0.1	672.3	b6	672.3	15	0.0	672.3	b6
787.3	27	0.0	787.3	b7	787.3	20	0.0	787.3	b7
900.4	16	0.0	900.4	b8	900.4	12	0.0	900.4	b8
1143.4	3	-0.1	1143.5	b10	1143.3	4	-0.2	1143.5	b10
1258.4	40	-0.2	1258.5	b11	1258.5	31	-0.1	1258.5	b11
1395.4	10	-0.1	1395.6	b12	1396.5	3	0.9	1395.6	b12
698.2	1	-0.1	698.3	b12+2					
1509.5	9	-0.1	1509.6	b13	1509.6	4	0.0	1509.6	b13
755.6	3	0.3	755.3	b13+2	755.6	2	0.3	755.3	b13+2
1624.6	15	-0.1	1624.7	b14	1624.5	16	-0.1	1624.7	b14
813.0	9	0.1	812.8	b14+2	813.1	4	0.3	812.8	b14+2
1787.3	3	-0.4	1787.7	b15	1787.7	3	-0.1	1787.7	b15
1916.7	2	-0.1	1916.8	b16					
959.1	100	0.3	958.9	b16+2	958.9	38	0.0	958.9	b16+2
439.3	7	0.0	439.2	y3	439.3	7	0.1	439.2	y3
554.5	1	0.2	554.2	y4	554.2	3	0.0	554.2	y4
668.4	4	0.2	668.3	y5	668.3	6	0.0	668.3	y5
805.4	49	0.1	805.3	y6	805.3	33	0.0	805.3	y6
920.7	4	0.3	920.4	y7	920.4	6	0.0	920.4	y7
1163.7	9	0.2	1163.5	y9	1163.4	8	0.0	1163.5	y9
1276.5	33	-0.1	1276.5	y10	1276.5	22	-0.1	1276.5	y10
1391.5	23	-0.1	1391.6	y11	1391.5	23	-0.1	1391.6	y11
1554.6	29	-0.1	1554.6	y12	1554.5	25	-0.1	1554.6	y12
778.0	2	0.1	777.8	y12+2	778.0	1	0.2	777.8	y12+2
1692.0	7	0.3	1691.7	y13	1691.6	10	-0.1	1691.7	y13
846.6	33	0.3	846.4	y13+2	846.7	25	0.3	846.4	y13+2
1918.8	2	0.0	1918.8	y15					
960.3	16	0.4	959.9	y15+2	959.5	39	-0.4	959.9	y15+2

Table 5-24. GSNIHYDLENDHNDYEK (2+) metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in m/z, calculated m/z values, and ion assignments are provided

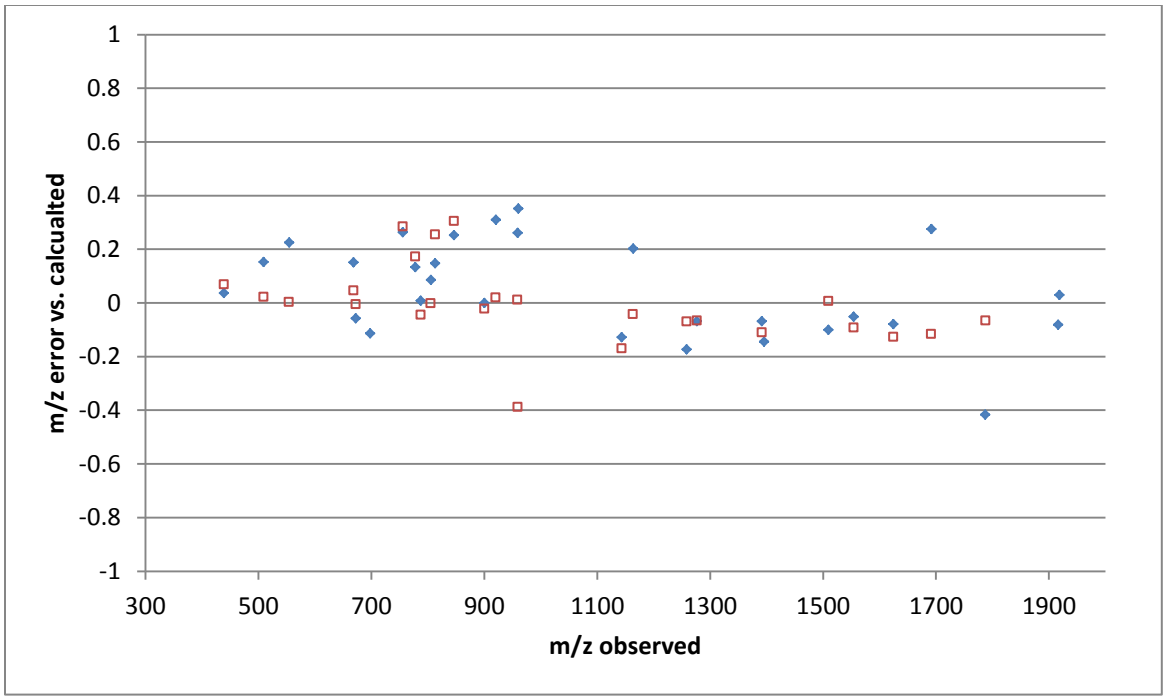
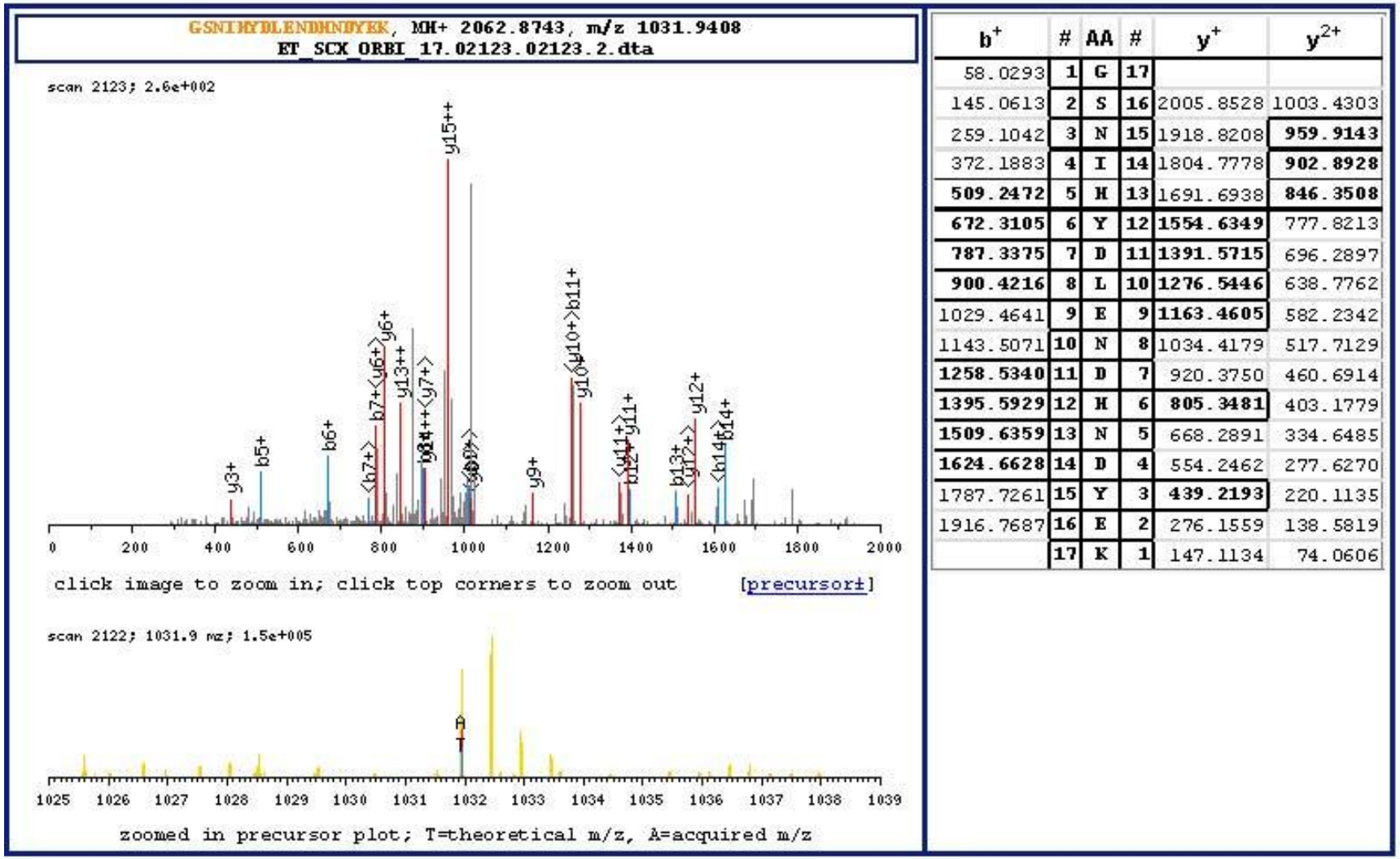
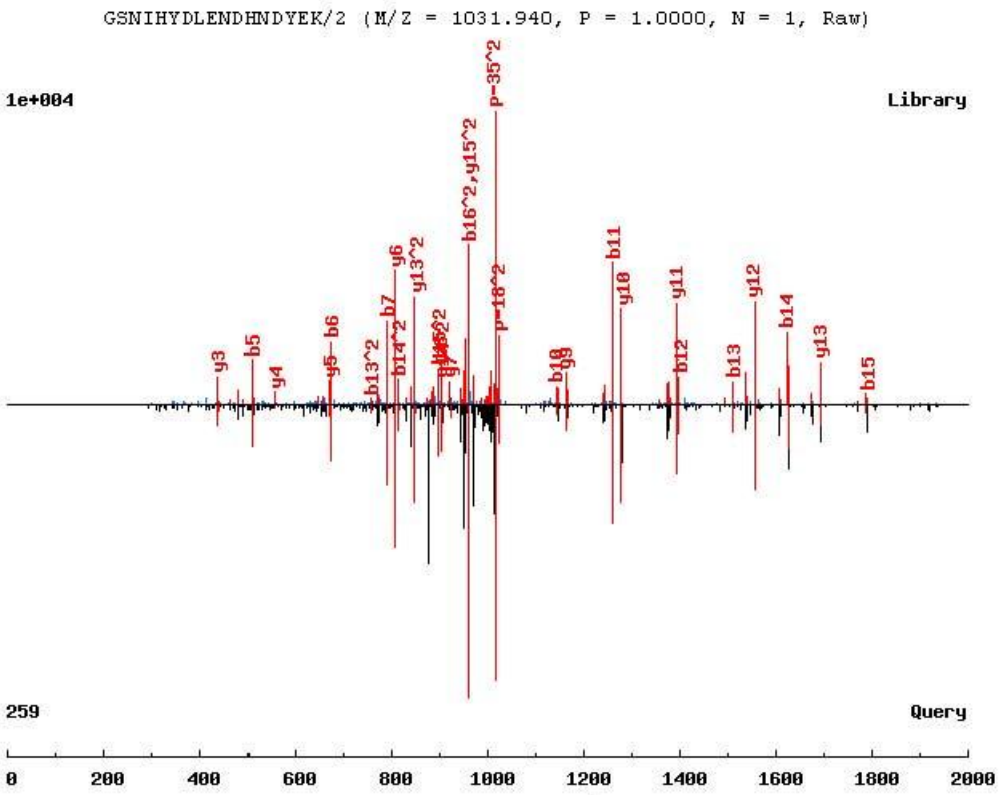


Figure 5-26. GSNIHYDLENDHNDYEK (2+) mass error in manually assigned b and y ions for metaproteomics versus standard peptides. Y-axis and X-axis values are given in m/z. Metaproteomics assignments are provided with a closed blue diamond (♦), and standard peptide data with an open red square (□).



b^+	#	AA	#	y^+	y^{2+}
58.0293	1	G	17		
145.0613	2	S	16	2005.8528	1003.4303
259.1042	3	N	15	1918.8208	**959.9143**
372.1883	4	I	14	1804.7778	**902.8928**
509.2472	5	H	13	1691.6938	**846.3508**
672.3105	6	Y	12	**1554.6349**	777.8213
787.3375	7	D	11	**1391.5715**	696.2897
900.4216	8	L	10	**1276.5446**	638.7762
1029.4641	9	E	9	**1163.4605**	582.2342
1143.5071	10	N	8	1034.4179	517.7129
1258.5340	11	D	7	920.3750	460.6914
1395.5929	12	H	6	**805.3481**	403.1779
1509.6359	13	N	5	668.2891	334.6485
1624.6628	14	D	4	554.2462	277.6270
1787.7261	15	Y	3	**439.2193**	220.1135
1916.7687	16	E	2	276.1559	138.5819
	17	K	1	147.1134	74.0606

Figure 5-27. GSNIHYDLENDHNDYEK (2+) automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectrum as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
58.0287	29.5180	1	G	17		
145.0608	73.0340	2	S	16	2005.8522	1003.4297
259.1037	130.0555	3	N	15	1918.8202	**959.9137**
372.1878	186.5975	4	I	14	1804.7773	**902.8923**
509.2467	255.1270	5	H	13	**1691.6932**	**846.3502**
672.3100	336.6586	6	Y	12	**1554.6343**	777.8208
787.3369	394.1721	7	D	11	**1391.5710**	696.2891
900.4210	450.7141	8	L	10	**1276.5440**	638.7756
1029.4636	515.2354	9	E	9	**1163.4600**	582.2336
1143.5065	572.2569	10	N	8	1034.4174	517.7123
1258.5335	629.7704	11	D	7	**920.3744**	460.6909
1395.5924	698.2998	12	H	6	**805.3475**	403.1774
1509.6353	**755.3213**	13	N	5	**668.2886**	334.6479
1624.6622	**812.8348**	14	D	4	**554.2457**	277.6265
1787.7256	**894.3664**	15	Y	3	**439.2187**	220.1130
1916.7682	**958.8877**	16	E	2	276.1554	138.5813
		17	K	1	147.1128	74.0600

Figure 5-28. GSNIHYDLENDHNDYEK (2+) comparison between the authentic standard peptide and metaproteomics spectrum. Spectra are illustrated with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is on top, and the metaproteomics spectrum is on the bottom and inverted. Assigned b and y product ions present in both spectra are noted in tabular format in bold.

For 2+ GSNIHYDLENDHNDYEK several long series of ions (b5-b8, b10-b16, y3-15) can be assigned both the metaproteomics and authentic standard peptides (**Table 5-24**). As illustrated in **Figure 5-26** the error in mass assignment follows a similar distribution for both experimental and standard peptides, with many experimental/standard ion pairs falling closely together. In **Figure 5-27** with the assigned product and parent ion spectra, numerous b and y ions are assigned. As well, the parent ion spectrum clearly illustrates that the correct monoisotopic parent ion was assigned. In **Figure 5-28** the near identical distributions of intensity between the long series of b and y ions are also illustrated. Taken together these data strongly suggest that the assignment of the peptide sequence **2_GSNIHYDLENDHNDYEK** to the experimental data is an excellent match, especially when identification of **3_GSNIHYDLENDHNDYEK** is also taken into account.

3+ GSNIHYDLENDHNDYEK, 2061.8667 Da									
2061.8713 Da, 688.2977 m/z, 2 ppm					2061.8670 Da, 688.2963 m/z, 0 ppm				
ORBI_17.2124					Standard				
Exptl	I	d(m/z)	Calc	Assign	Exptl	I	d(m/z)	Calc	Assign
258.7	5	-0.4	259.1	b3	259.1	3	0.0	259.1	b3
372.3	3	0.1	372.2	b4	372.1	2	-0.1	372.2	b4
672.6	2	0.3	672.3	b6	672.5	6	0.2	672.3	b6
787.5	27	0.1	787.3	b7	787.3	19	0.0	787.3	b7
900.7	2	0.3	900.4	b8	900.5	2	0.0	900.4	b8
1029.7	4	0.2	1029.5	b9	1029.4	10	0.0	1029.5	b9
515.1	2	-0.2	515.2	b9+2					
1143.7	2	0.2	1143.5	b10	1143.3	1	-0.2	1143.5	b10
1258.5	2	0.0	1258.5	b11	1258.4	6	-0.1	1258.5	b11
629.9	5	0.1	629.8	b11+2	630.1	7	0.3	629.8	b11+2
1396.1	5	0.5	1395.6	b12					
					698.3	1	0.0	698.3	b12+2
504.4	2	0.5	503.9	b13+3	755.5	3	0.2	755.3	b13+2
812.9	25	0.0	812.8	b14+2	812.9	25	0.1	812.8	b14+2
542.4	4	0.2	542.2	b14+3					
894.5	17	0.1	894.4	b15+2	894.2	6	-0.1	894.4	b15+2
959.3	50	0.4	958.9	b16+2	959.2	51	0.3	958.9	b16+2
639.6	11	0.0	639.6	b16+3	639.5	27	-0.1	639.6	b16+3
276.4	3	0.3	276.2	y2					
439.5	14	0.3	439.2	y3	439.3	18	0.1	439.2	y3
					554.3	2	0.1	554.2	y4
668.5	8	0.2	668.3	y5	668.3	10	0.0	668.3	y5
805.2	7	-0.2	805.3	y6	805.3	10	-0.1	805.3	y6
920.4	7	0.0	920.4	y7	920.4	3	0.0	920.4	y7
					460.9	2	0.2	460.7	y7+2
517.7	26	0.0	517.7	y8+2	518.0	17	0.2	517.7	y8+2
					1163.4	2	0.0	1163.5	y9
582.0	2	-0.2	582.2	y9+2	582.3	2	0.0	582.2	y9+2
638.9	38	0.1	638.8	y10+2	638.9	44	0.1	638.8	y10+2
426.4	1	0.2	426.2	y10+3					
1392.1	3	0.5	1391.6	y11	1391.5	2	-0.1	1391.6	y11
696.6	2	0.3	696.3	y11+2	696.5	3	0.2	696.3	y11+2
778.3	49	0.5	777.8	y12+2	778.1	45	0.3	777.8	y12+2
518.5	2	-0.4	518.9	y12+3	519.0	1	0.1	518.9	y12+3
846.7	100	0.4	846.4	y13+2	846.6	100	0.3	846.4	y13+2
565.1	8	0.5	564.6	y13+3	564.9	7	0.4	564.6	y13+3
602.5	6	0.2	602.3	y14+3	602.5	6	0.2	602.3	y14+3
					960.1	15	0.2	959.9	y15+2
640.5	16	0.3	640.3	y15+3	640.6	18	0.3	640.3	y15+3
669.5	21	0.2	669.3	y16+3	669.7	7	0.4	669.3	y16+3

Table 5-25. 3+ GSNIHYDLENDHNDYEK metaproteomics versus authentic standard peptide manual assignments of b and y ions. For the metaproteomics and authentic standard peptide parent ion monoisotopic masses, observed m/z, and observed mass error in ppm are given. Assigned product ion m/z values (>1% normalized intensity), normalized intensity, mass errors in m/z, calculated m/z values, and ion assignments are provided

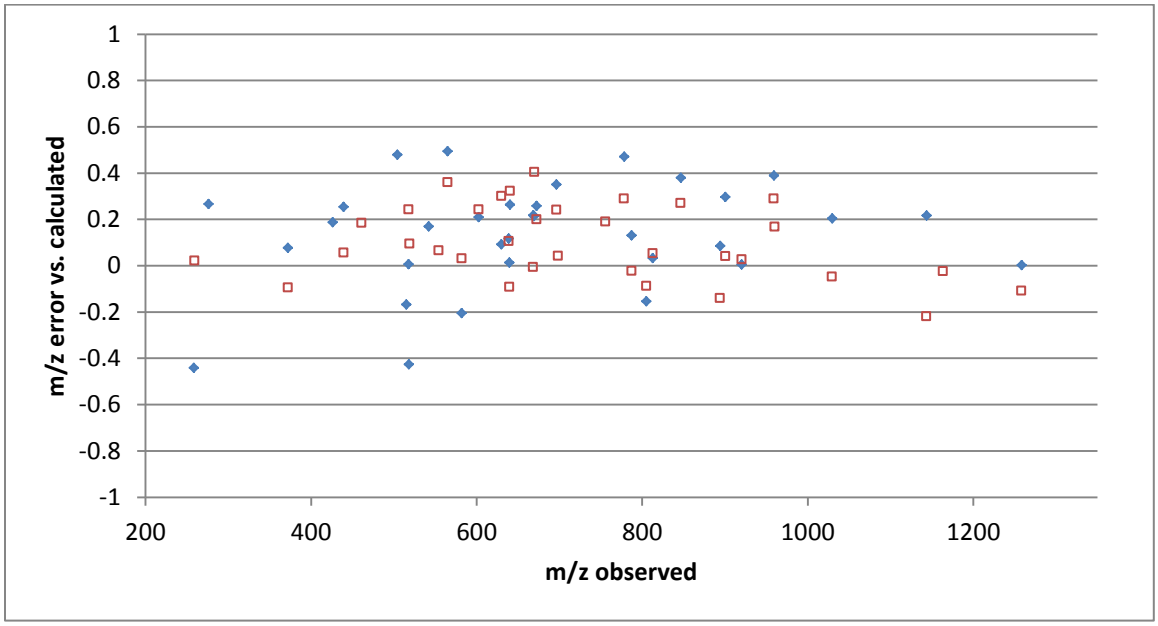
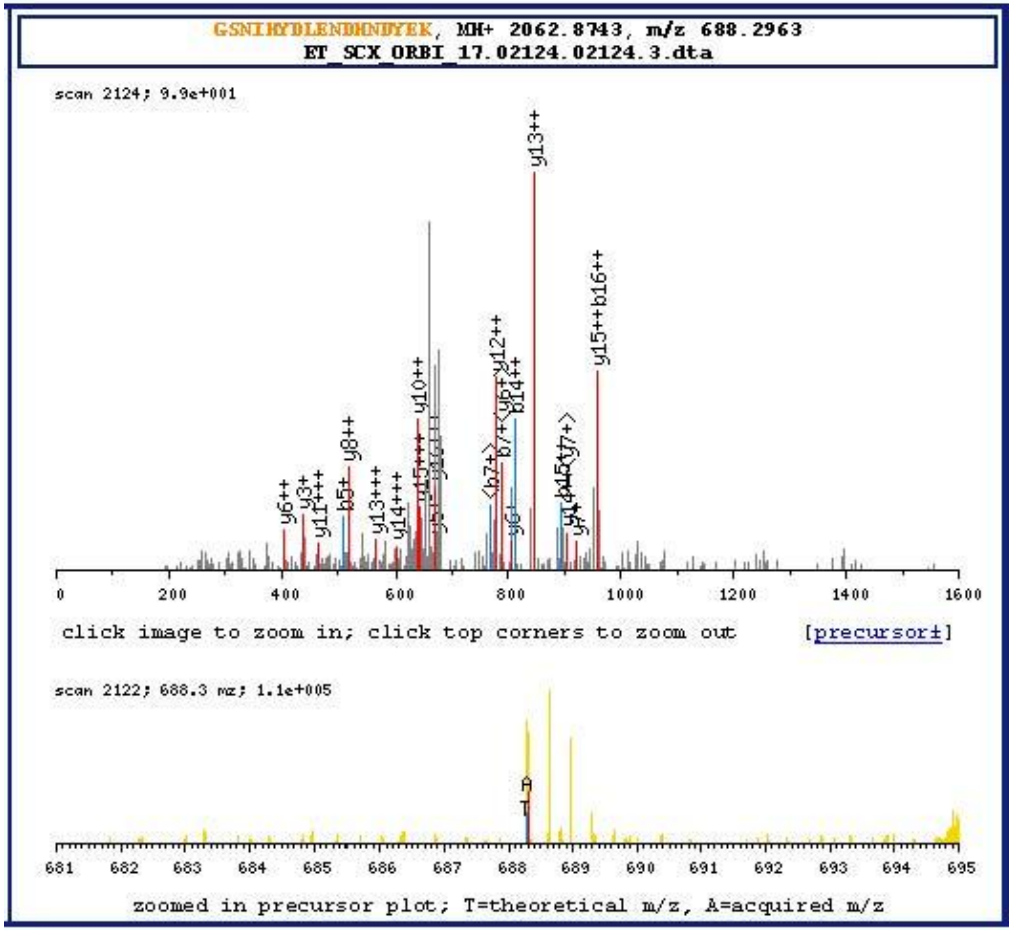
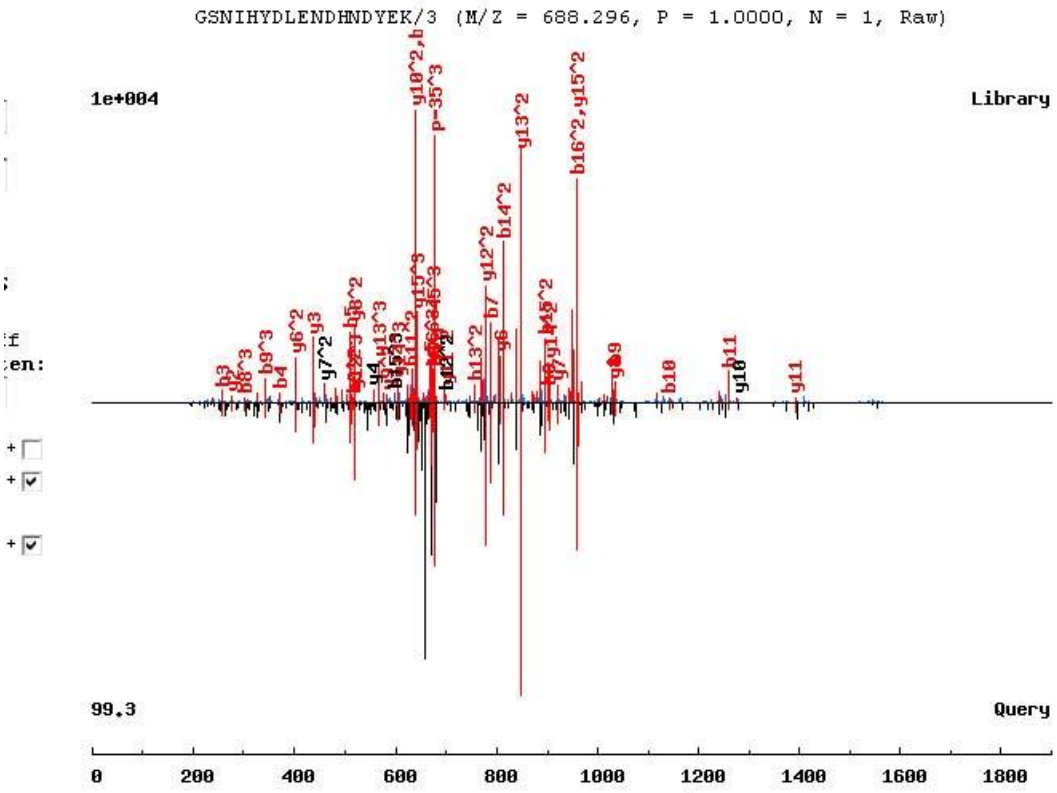


Figure 5-29. GSNIHYDLENDHNDYEK (3+) mass error in manually assigned b and y ions for metaproteomics versus standard peptides. Y-axis and X-axis values are given in m/z. Metaproteomics assignments are provided with a closed blue diamond (♦), and standard peptide data with an open red square (□).



b^{+}	b^{2+}	#	AA	#	y^{+}	y^{2+}	y^{3+}
58.0293	29.5186	1	G	17			
145.0613	73.0346	2	S	16	2005.8528	1003.4303	**669.2895**
259.1042	130.0560	3	N	15	1918.8208	**959.9143**	**640.2788**
372.1883	186.5981	4	I	14	1804.7778	**902.8928**	**602.2645**
509.2472	255.1275	5	H	13	1691.6938	**846.3508**	**564.5698**
672.3105	336.6592	6	Y	12	1554.6349	**777.8213**	518.8835
787.3375	394.1727	7	D	11	1391.5715	696.2897	**464.5291**
900.4216	450.7147	8	L	10	1276.5446	**638.7762**	426.1867
1029.4641	515.2360	9	E	9	1163.4605	582.2342	388.4921
1143.5071	572.2575	10	N	8	1034.4179	**517.7129**	345.4779
1258.5340	629.7709	11	D	7	**920.3750**	460.6914	307.4636
1395.5929	698.3004	12	H	6	**805.3481**	**403.1779**	269.1212
1509.6359	755.3218	13	N	5	**668.2891**	334.6485	223.4349
1624.6628	**812.8353**	14	D	4	554.2462	277.6270	185.4206
1787.7261	**894.3670**	15	Y	3	**439.2193**	220.1135	147.0783
1916.7687	**958.8883**	16	E	2	276.1559	138.5819	92.7239
		17	K	1	147.1134	74.0606	49.7097

Figure 5-30. GSNIHYDLENDHNDYEK (3+) automatically assigned spectra from X!tandem. Various experimental details are noted in this figure. Key panels include the product and parent ion mass spectrum as m/z versus intensity. The automatic assignment of b and y ions is noted by bold boxes in tabular form.



b^{1+}	b^{2+}	b^{3+}	#	AA	#	y^{1+}	y^{2+}	y^{3+}
58.0287	29.5180	20.0144	1	G	17			
145.0608	73.0340	49.0251	2	S	16	2005.8522	1003.4297	**669.2889**
259.1037	130.0555	87.0394	3	N	15	1918.8202	**959.9137**	**640.2782**
372.1878	186.5975	124.7341	4	I	14	1804.7773	**902.8923**	**602.2639**
509.2467	255.1270	170.4204	5	H	13	1691.6932	**846.3502**	**564.5693**
672.3100	336.6586	224.7749	6	Y	12	1554.6343	**777.8208**	**518.8829**
787.3369	394.1721	263.1172	7	D	11	**1391.5710**	**696.2891**	464.5285
900.4210	450.7141	**300.8119**	8	L	10	**1276.5440**	**638.7756**	426.1862
1029.4636	**515.2354**	**343.8260**	9	E	9	1163.4600	**582.2336**	388.4915
1143.5065	572.2569	381.8404	10	N	8	**1034.4174**	**517.7123**	345.4773
1258.5335	**629.7704**	420.1827	11	D	7	**920.3744**	**460.6909**	307.4630
1395.5924	**698.2998**	465.8690	12	H	6	**805.3475**	**403.1774**	269.1207
1509.6353	**755.3213**	503.8833	13	N	5	**668.2886**	334.6479	223.4344
1624.6622	**812.8348**	542.2256	14	D	4	**554.2457**	277.6265	185.4201
1787.7256	**894.3664**	**596.5800**	15	Y	3	**439.2187**	220.1130	147.0778
1916.7682	**958.8877**	**639.5942**	16	E	2	**276.1554**	138.5813	92.7233
			17	K	1	147.1128	74.0600	49.7091

Figure 5-31. GSNIHYDLENDHNDYEK (3+) comparison between the authentic standard peptide and metaproteomics spectrum. Spectra are illustrated with m/z on the x-axis and intensity on the y-axis. The authentic standard spectrum is on top, and the metaproteomics spectrum is on the bottom and inverted. Assigned b and y product ions present in both spectra are noted in tabular format in bold.

For 3+ GSNIHYDLENDHNDYEK several long series of ions (b7-b10, y5-14) can be assigned both the experimental and authentic standard peptides (**Table 5-25**). As illustrated in **Figure 5-29** the error in mass assignment follows a similar distribution for both experimental and standard peptides, with many experimental/standard ion pairs falling closely together. In **Figure 5-30** with the assigned product and parent ion spectra, numerous b and y ions are assigned, with almost a complete y-ion ladder observed when all charge states are taken into account. As well, the parent ion spectrum clearly illustrates that the correct monoisotopic parent ion was assigned. In **Figure 5-31** the near identical distributions of intensity between the long series of b and y ions are also illustrated, although it is clear that the experimental spectra is of lower intensity and thus has a worse signal to noise ratio. Taken together these data strongly suggest that the assignment of the peptide sequence **2_GSNIHYDLENDHNDYEK** to the experimental data is an excellent match. The fully-tryptic, unmodified peptide GSNIHYDLENDHNDYEK is identified in both the 2+ and 3+ charge states in subsequent spectra. Mass errors are less than 5 ppm for both parent spectra, with very strong y- and b- ions series for the 2+ ion, and a weaker series for the 3+ ion (likely due to lower parent ion intensity).

Portions of this chapter have been previously published in:

Christopher M. Rath*, Benjamin Janto*, Josh Earl, Azad Ahmed, Fen Z. Hu, Luisa Hiller, Meg Dahlgren, Rachael Kreft, Fengan Yu, Jeremy J. Wolff, Hye Kyong Kweon, Michael A. Christiansen, Kristina Håkansson, Robert M. Williams, Garth D. Ehrlich,

David H. Sherman. Meta-omic analysis of a marine invertebrate microbial consortium provides a direct route to identify and characterize natural product biosynthetic systems. Manuscript submitted PLOSone.

We thank Erich Bartels, Vicki Woodbridge, and Mote Marine Laboratories for assistance with sample collection, and Kate Noon at the UM Pharmacology Mass Spectrometry Facility for assistance with IT-MS. Bruker Daltonics is gratefully acknowledged for access to the 12T FTICR-MS, and Philip Andrews for access to the Orbitrap-MS (supported by NCRR-P41) used in this study. We thank Dr. Damian Fermin of the Nesvizhskii laboratory for help with the TPP. We thank Dr. George Chlipala for assistance with Perl scripts. This work was supported by NIH Grant CA070375 (R.M.W. and D.H.S), the H. W. Vahlteich Professorship (D.H.S), a Microfluidics in Biomedical Sciences Training Grant fellowship (C.M.R.), and the Allegheny Singer Research Foundation and DHHS/HRSA C76HF00659 (G.D.E). This work was also inspired by NIH grant U01 TW007404 as part of the International Cooperative Biodiversity Group initiative at the Fogarty International Center.

5.5 References

1. Rinehart, K.L.; *et al*. *J Org Chem,* **1990**, *55*, 4512.
2. www.emea.europa.eu/humandocs/PDFs/EPAR/yondelis/H-773-en6.pdf
3. Izbicka, E.; *et al*. *Annals Oncol,* **1998**, *9*, 981.
4. Minuzzo, M.; *et al*. *Proc Nat Acad Sci USA,* **2000**, *97*, 6780.
5. Pommier, Y.; *et al*. *Biochemistry,* **1996**, *35*, 13303.

6. Takebayashi, Y.; *et al*. *Nat Med* **2001,** *7*, 961.

7. Carballo, J.L.; Naranho, S.; Kukurtzu, B.; De La Calle, F.; Hernandez-Zanuy A. *J World Aquaculture Soc,* **2000,** *31*, 481.

8. Corey, E.J.; Gin, D.Y.; Kania, R.S. *J Am Chem Soc,* **1996,** *118*, 9202.

9. Cuevas, C.; Francesch, A. *Nat Prod Rep,* **2009,** *26*, 322.

10. Cuevas, C.; *et al*. *Org Letters,* **2000,** *2*, 2545.

11. Arai, T.; Takahashi, K.; Nakahara, S; Kubo, A. *Cell Mol Life Sci*, **1980**, *36*, 1025.

12. Irschik H.; Trowitzschi-Kienast W.; Gerth K.; Hofle G.; Reichenbach H. *J Antibiotics,* **1988**, *41*, 993.

13. Ikeda Y; Shimada Y; Honjo K; Okumoto T; Munakata T. *J Antibiotics,* **1983**, *36,* 1290.

14. Piel, J.; *Curr Med Chem,* **2006**, *13,* 39.

15. Sudek, S.; *et al*. *J Nat Prod,* **2007,** *70*, 67.

16. Lopanik, N.B.; *et al. Chem Biol,* **2008,** *15*, 1175.

17. Piel, J.; *Proc Nat Acad Sci USA,* **2002**, *99*, 14002.

18. Piel, J.; et al. *Proc Nat Acad Sci USA,* **2004,** *101*, 16222.

19. Fisch, K.M.; *et al*. *Nat Chem Biol,* **2009,** *5*, 494.

20. Velasco, A.; *et al*. *Mol Microbiol,* **2005,** *56*, 144.

21. Li, L.; *et al*. *J Bacteriol* **2008,** *190*, 251.

22. Pospiech, A.C.; Bietenhader, J.; Schupp, T. *Microbiol*, **1995,** *141*, 1793.

23. Koketsu, K.; Watanabe, K.; Suda, H.; Oguri, H.; Oikawa, H. *Nat Chem Biol,* **2010**, *6*, 408.

24. Moss, C.; *et al*. *Mar Biol,* **2003,** *143*, 99.

25. Parez-Matos, A.E.; Rosado W.; Govind N.S. *Antonie van Leeuwenhoek,* **2007,** *92*, 155.

26. Ehrlich, G.; Hiller, N.L; Hu, F. *GenomeBiology.com,* **2008,** *9*, 225.

27. Wilmes, P.; Bond, P.L. *Trends in Microbiol,* **2006,** *14*, 92.

28. Ro, D.-K.; *et al*. *Nature,* **2006,** *440*, 940.

29. Wenzel, S.C.; Muller, R.; *Cur Op Biotech,* **2005,** *16*, 594.

30. Sakai, R.; Jares-Erijman, E.A.; Manzanares, I.; Silva Elipe, M.V.; Rinehart, K.L. *J. Am, Chem, Soc,* **1996,** *118*, 9017.

31. Yang, Y.-L.; Xu, Y.; Straight, P.; Dorrestein, P.C. *Nat Chem Biol,* **2009,** *5*, 885.

32. Ragin, C.C.R.; Reshmi, S.C.; Gollin, S.M. *Int J Cancer,* **2004,** *110*, 701.

33. D'Agostino, G.; *et al*. *Int. J. Gynecol. Cancer,* **2006,** *16*, 71.

34. Meyer, F.; *et al*. *BMC Bioinform,* **2008,** *9*, 386.

35. Sharp, K.H.; Davidson, S.K.; Haygood M.G. *ISME J,* **2007,** *1*, 693.

36. Yamamoto, S.; He, Y.; Arakawa, K.; Kinashi, H. *J Bacteriol,* **2008,** *190*, 1308.

37. Bachmann, B.O.; Ravel, J.; David, A.H. in Methods in Enzymology, Vol. Volume 458 181-217 (Academic Press, **2009**).

38. Magarvey, N.A.; Ehling-Schulz, M.; Walsh, C.T. *J Am Chem Soc,* **2006,** *128*, 10698.

39. Calderone, C.T.; Bumpus, S.B.; Kelleher, N.L.; Walsh, C.T.; Magarvey, N.A. *Proc Nat Acad Sci USA,* **2008,** *105*, 12809.

40. Wexler, M.; *et al*. *J Biol. Chem.* **2000,** *275*, 16717.

41. Arroyo, M.; Mata, I.; Acebal, C.; Castillion M. P. *Ap Microbiol Biotech,* **2003,** *60*, 507.

42. Nelson, J.T.; Lee, J.; Sims, J.W.; Schmidt, E.W. *Ap Environ Microbiol*, **2007,** *73*, 3575.

43. Fu, C.-Y.; *et al*. *J. Microbiol Biotech,* **2009,** *19*, 439.

44. Pak, J.; Jeon K. W. *Gene*, **1996,** *171*, 89.

45. Pak, J.; Jeon K. W. *J Eurk Microbiol,* **1997,** *44*, 614.

46. Brown, N.L.; Stoyanov, J.V.; Kidd, S.P; Hobman, J.L. *FEMS Microbiol Rev,* **2003,** *27*, 145.

47. Jack, D.L.; Yang, N.M.; H. Saier, M. *Eur. J Biochem,* **2001,** *268*, 3620.

48. Curnow, A.W.; *et al. Proc. Nat Acad Sci USA,* **1997,** *94*, 11819.

49. Kittendorf, J.D.; Beck, B.J.; Buchholz, T.J.; Seufert, W.; Sherman, D.H. *Chem Biol,* **2007,** *14*, 944.

50. Deutsch, E.W.; *et al*. *Proteomics,* **2010**, *10*, 1150.

51. Wang, J.; *et al*. *Proc Nat Acad Sci USA,* **2007,** *104*, 7612.

52. DeSantis, T.Z.; *et al*. *Ap. Environ. Microbiol* **2006,** 72, 5069.

53. DeSantis, T.Z.; *et al*. *Nuc Acids Res,* **2006,** *34*, W394.

54. Keane, T.M.; Creevey, C.J.; Pentony, M.M.; Naughton, T.J.; McLnerney, J.O. *BMC Evol Biol,* **2006,** *6*, 29.

55. Guindon, S.; Gascuel, O. *Sys Biol,* **2003,** *52*, 696.

56. Bocs, S.; Cruveiller, V.D.; Nuel, G.; Medigue, C. *Nucleic Acids Res,* **2003,** *31*, 3723.

57. Wang, J.; *et al*. *Proc Nat Acad Sci USA,* **2007,** *104*, 7612.

58. Hicks, L.M.; Moffitt, M.C.; Beer, L.L., Moore, B.S.; Kelleher, N.L. *ACS Chem Biol,* **2006,** *1*, 93.

59. Zhang, W.; Sun, T.-T.; Li, Y.-X. *J Pep Sci,* **2009,** *15*, 366.

60. Frigerio, M.; Santagostino, M.; Sputore, S. *J Org Chem,* **1999,** *64*, 4537.

61. Staeva-Vieira, T.; von Herrath, M. *Clin Exper Immun,* **2007,** *148*, 17.

62. Geer, L.Y.; *et al. J. Prot Res,* **2004,** *3*, 958.

63. Tsur, D.; *et al. Nat Biotech,* **2004,** *23*, 1562.

64. Craig, R.; Beavis R.C. *Bioinformatics,* **2004,** *20,* 1466.

65. Lam, H. *et al. Nat Meth,* **2008,** *5*, 873.

66. Breci, L.A.; Tabb, D.L.; Yates, J.R.; Wysocki, V.H. *Anal Chem,* **2003**, *75*, 1963.

67. Caffrey, P.; Bevitt, D.J.; Staunton, J.; Leadlay, P.F.; *et al*. *FEBS Let,* **1992**, *304*, 225.

68. Tanner, S.; *et al*. *Anal Chem,* **2005**, *77*, 4626.

69. Kim, S.; Gupta, N.; Pevzner, P.A. *J Prot Res,* **2008,** *7*, 3354.

Chapter 6

Future directions

6.1 Summary

The previous four chapters have illustrated different paradigms for dissecting natural product pathways with an FTICR-MS-centric analytical approach. Results, in terms of fundamental mechanistic details, chemoenzymatic generation of natural products, and characterization/identification of novel symbiont derived pathways have all been generated—each raising additional hypotheses. Herein, further investigations into Type I PKS biosynthetic pathways (pikromycin, tylosin, and erythromycin) are discussed, building upon the developed analytical expertise. Two projects utilizing mass spectrometry to probe substrate flexibility in chemoenzymatic synthesis within the pikromycin and cryptophycin biosynthetic pathways are explored. Further investigations into the ET-743 biosynthesis that will inform future efforts at chemoenzymatic synthesis and heterologous expression are also reviewed.

6.2 Introduction

The three major themes explored in this work have been: detailed exploration of enzymes in biosynthetic pathways, chemoenzymatic synthesis of natural products, and exploring new symbiont derived natural product pathways. In particular, the

development and application of FTICR-MS tools has been the focus of these investigations. In chapter 2, the mechanism of CoA extender unit selection was probed using steady state kinetics and FTICR-MS analysis of active-site occupancy. Follow-up investigations to this project in terms of our fundamental understanding of PKS mechanisms including the role/fate of chain elongation intermediates, the inter-protein substrate transfer, and the role of docking are detailed below. Investigations into mechanisms and selectivity within the pikromycin, tylosin, and erythromycin biosynthetic pathways will further inform our attempts at chemoenzymatic production of macrolides—particularly an ongoing investigation into role of the thiol containing moiety on chain elongation intermediates. Developed methods and findings could also be applied to novel symbiont derived pathways such as cryptophycin and bryostatin.

Two projects focusing on exploring substrate flexibility in chemoenzymatic synthesis in the cryptophycin and pikromycin biosynthetic pathways are also reported. As shown through chemoenzymatic methods and natural diversity, there is a great deal of substrate flexibility in the cryptophycin biosynthetic pathway.[1,2] A digital microfluidic device is proposed for generating an artificial, reconfigurable biosynthetic pathway. This project will depend on advanced mass spectrometry methodologies for product characterization and will also inform our fundamental understanding of programmable flexibility in these systems. The second chemoenzymatic project proposed focuses on exploring substrate flexibility of the chemoenzymatic reagent RhFRED-PikC.[3] Libraries of artificial and natural compounds will be simultaneously hydroxylated with ^{18}O and ^{16}O, and then product generation will be monitored by LC-FTICR MS/MS. The heavy oxygen label will give a unique isotope pattern—enhancing the ability of the high-

performance mass spectrometry to identify new products. Furthermore, the altered isotopic pattern will allow for the site of hydroxylation to be localized. In addition to identifying substrates for chemoenzymatic C-H activation, this study will also help develop methods for monitoring biochemical reactions in either the study of fundamental biosynthetic mechanisms of in novel symbiont derived pathways.

Further investigations into the ET-743 biosynthetic pathway are proposed. First we aim to characterize the 25 proteins identified to date, through cloning, overexpression and *in vitro* biochemistry. Key enzymes such as the novel Pictet-Spenglerase will be targets for structural biology as methods to fully explain their catalytic capabilities.[4] Next, we will seek to better understand the role of the biosynthetic pathway in the native biological context. Chemical probes will be applied to purify key enzymes based on catalytic activity—thus validating biological function assigned from bioinformatics.[5] These probes can also be applied for imaging analysis in conjunction with traditional optical imaging.[6] One fascinating opportunity would be to directly apply mass spectrometry based imaging techniques to localize small molecules and proteins within the tunicate bacterial assemblage.[7] Finally, more modern sequencing technologies could be applied to further probe the metagenome of our organism and allow us to complete the assembly.[8] All of these efforts together which will allow us to assemble and characterize a minimal module for ET-743 biosynthesis which can be used *in vitro* or engineered into a strain for heterologus expression of the pathway and production of ET-743 and related analogs. This project represents the pinnacle of all the work to date as it directly applies (and extends) our knowledge of natural product biosynthesis to chemoenzymatically synthesize an approved drug for which there is no efficient way to produce it. This

project also served as template for new MS-driven technologies to characterize host-symbiont biosynthetic systems.

6.3 *In vitro* biochemistry of type I PKS biosynthetic enzymes by FTICR-MS

Further investigations into biosynthetic mechanisms are proposed in the pikromycin, tylosin and erythromycin biosynthetic systems. These three pathways are similar enough to allow for easy exchange of components while retaining enough key differences to allow the role of specific factors to be elucidated (**Figure 6-1**).

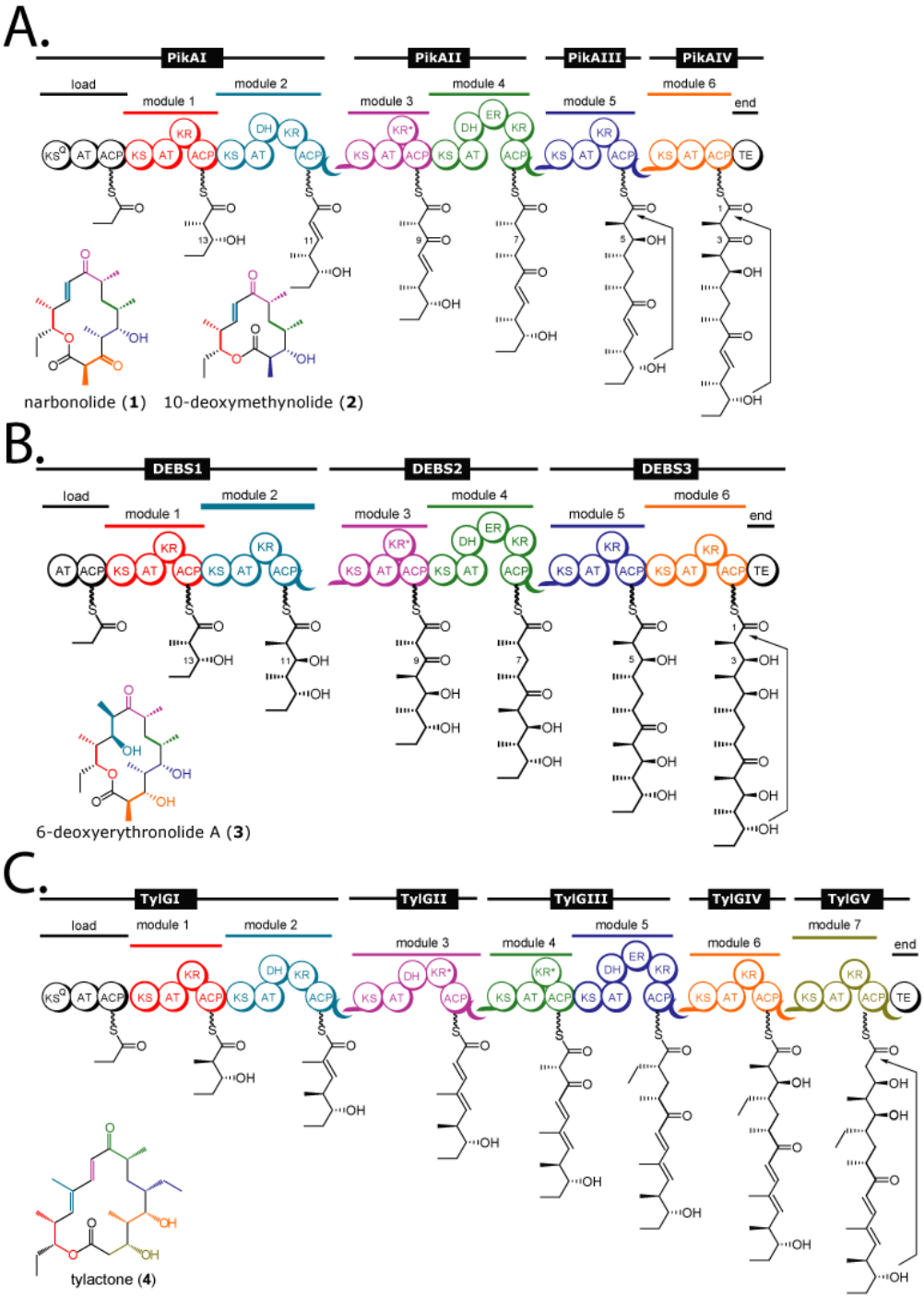


Figure 6-1. The pikromycin, erythromycin, and tylosin biosynthetic pathways. (**A**) The pikromycin biosynthetic pathway is illustrated with the 14- and 12- membered ring products narbonolide (**1**) and 10-DML (**2**). (**B**) The erythromycin biosynthetic pathway is illustrated with the 14-membered macrolide product 6-deoxyerythronolide A. (**C**) The tylosin biosynthetic pathway with the 16-membered macrolide product tylactone (**4**). Polypeptides are noted with each protein name. Individual modules are color coded with the contribution shown in the final natural product.

While the Sherman lab has a made the pikromycin biosynthetic pathway (**Figure 6-1A**) a key model system since 1998[9], the erythromycin biosynthetic pathway[10] has only recently emerged as a model system (**Figure 6-1B**).[11] *In vitro* biochemical investigations in the tylosin biosynthetic pathway[12] (**Figure 6-1C**), are just beginning to produce results. Expression of the expression of DEBS3 modules 5/6 and synthesis of the DEBS pentaketide natural substrate analogs[11] will allow for a series of mix/match experiments to be performed. Likewise, soluble, active expression of tylosin modules 6 and 7 (TylGIV, V) will expand our repertoire.

Key factors to be explored include macrolide product ring size as Pik produces 12- or 14-membered products, Debs 14-membered products, and Tyl 16-membered products. In addition, the terminal modules for TylG and PikA are monomodules versus a dimodule for DEBS3 (**Figure 6-1**). Potentially, differential analysis may elucidate the role of inter-protein transfers. The number of product double bonds present (0 for DEBS, 1 for PikA, and 2 for TylG) also varies. As this factor has been proposed to be important in TE-catalyzed cyclization,[13] the differential analysis may prove quite informative. Further potential challenges to working with the TylGIV/V and DEBS3 systems include the identification of key active site residues for active site occupancy experiments. A series of these planned and ongoing experiments are detailed below (**6.3.1-6.3.4**).

6.3.1 PikAIII pentaketide leaving group analogs

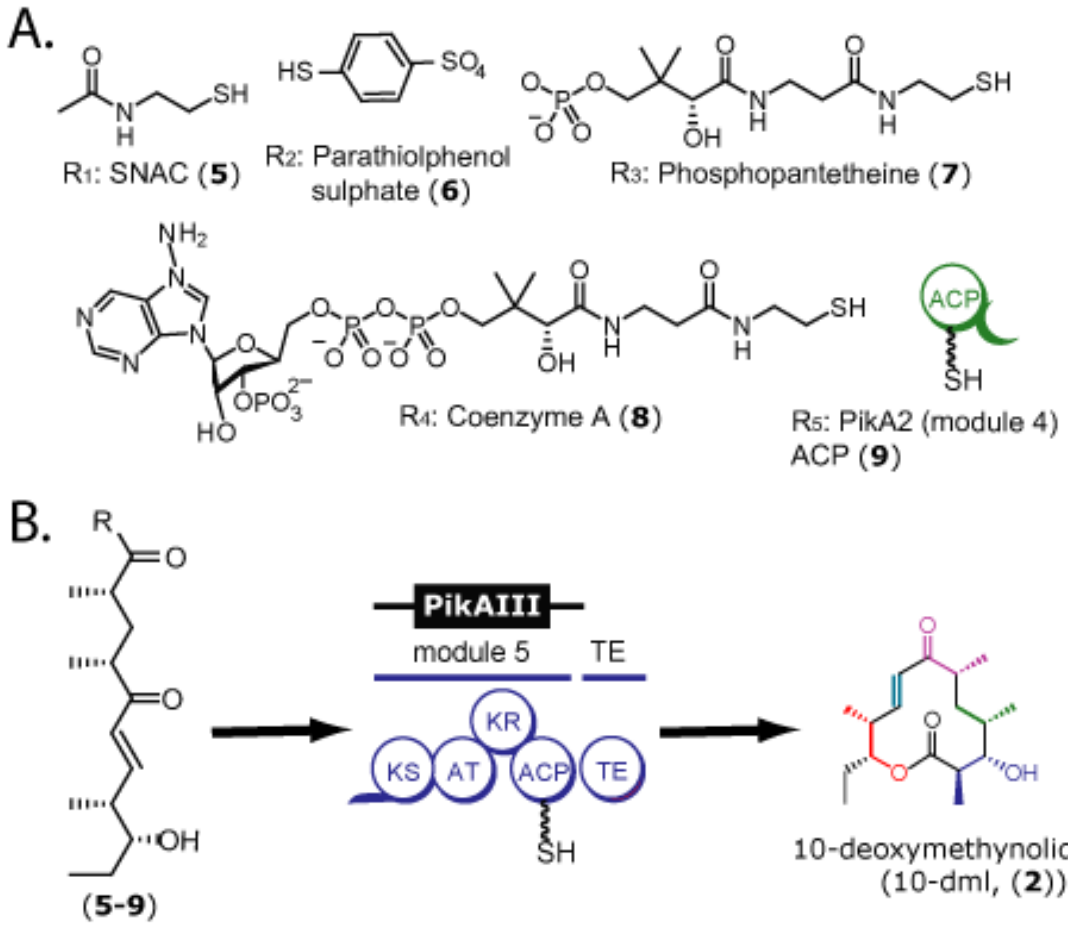


Figure 6-2. Pikromycin pentaketide thioester leaving group effects. (**A**) Five possible pentaketide leaving groups are displayed. (**B**) The loading and extension of the five pentaketide analogs will be catalyzed by the PikAIII TE.

The pikromycin SNAC pentaketide (**5**) has proven to be a valuable tool for probing biosynthesis in the native context such as the chemoenzymatic synthesis of 10-dml (**Figure 6-2B**).[14] However, while the SNAC is certainly a reasonable mimic of the full phosphopantetheine arm is it ideal in terms of enhancing productivity, selectivity, and solubility? Compound (**6**) will contain a highly activated leaving group with little steric hindrance—indeed a similar thiol phenol has been shown to be an optimal leaving group in NRPS-TE reactions.[15] Conversely, Wu and coworkers have shown that loaded (from

the CoA) ACP's and coenzyme A substrates[16] produce orders of magnitude increases in k_{cat}/K_m values as compared to SNAC substrates. The phosphopantetheine itself could also serve as an intermediate between SNAC and CoA—which would be advantageous for large-scale chemoenzymatic synthesis due to the low cost of panthetheine ($40/gram) as compared to CoA ($2,000/gram). Currently three additional pentaketide substrate analogs (**6-8**) are being synthesized to investigate this possibility. These substrates will be reacted with PikAIII-TE and active-site occupancy (**Chapter 2**), fluorescent-based steady-state kinetics (**Chapter 2**), radio-TLC,[17] and LC FTICR-MS analysis of product formation (**Chapter 4**) will be applied to characterize the system.

6.3.2 PikAIII → PikAIV intermodular chain elongation intermediate transfer

The pikromycin biosynthetic pathway is unique in terms of its ability to generate both 12 and 14- membered products. One possible reason for this is the two terminal mono-modular proteins, especially as compared to the dimodular DEBS3, which only produces a 14-membered product. A first in Type I PKS research would be to directly monitor the inter-protein transfer of chain elongation intermediates (**Figure 6-3**).

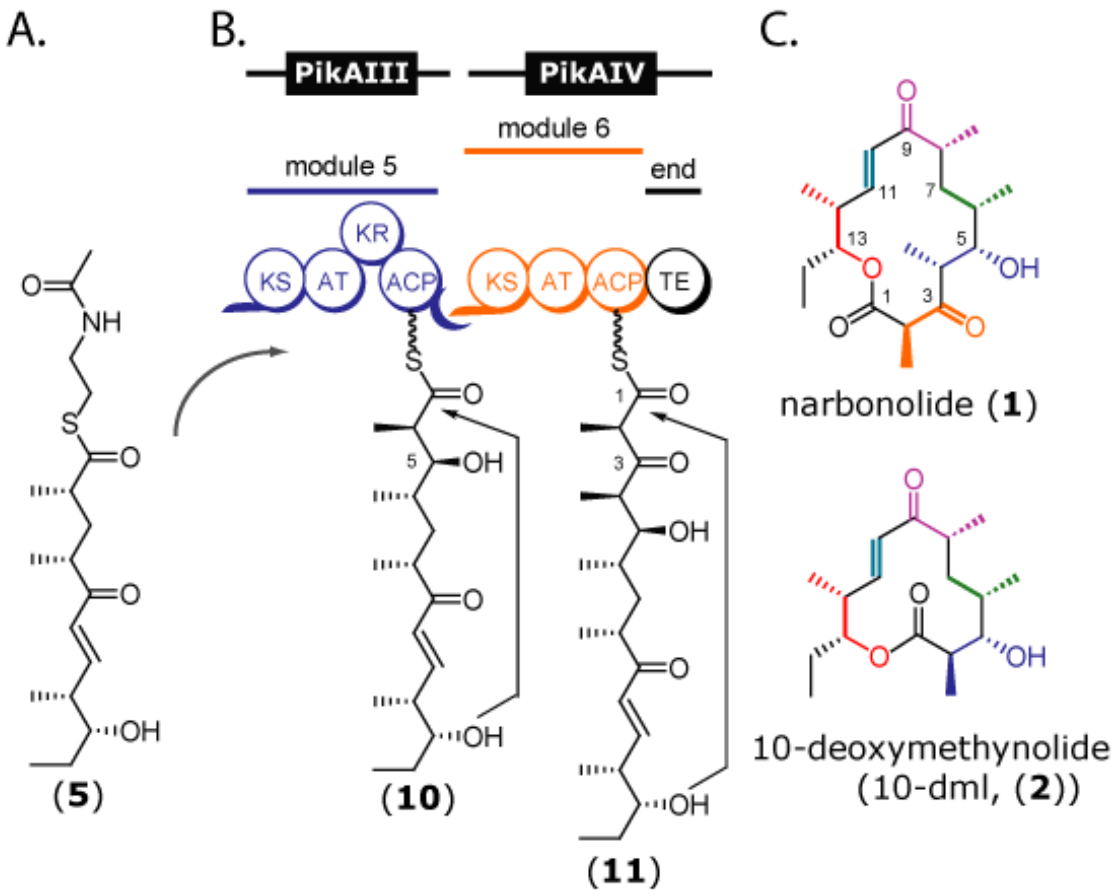


Figure 6-3. PikAIII → PikAIV intermodular chain elongation intermediate transfer. (**A**) The optimal chain elongation intermediate as determined from 6.3.1 will be utilized to directly monitor chain elongation intermediate transfer and active-site occupancy between PikAIII and PikAIV. (**B**) PikAIII and PikIV with enzyme bound hexaketide (**10**) and heptaketide (**11**) intermediates. (**C**) Reaction products narbonolide (**1**) and 10-DML (**2**).

Based on radio-TLC derived steady-state kinetics[17], and crystallography[13] the so-called "reach-back" model has been proposed for both 12- and 14- membered ring formation in the pikromycin biosynthetic pathway. Unlike radio-assays, which cannot easily differentiate multiple signals at once, or crystallography, which is slow and requires large amounts of material, FTICR-MS offers the possibility of a direct read-out of active-site occupancy under a variety of different conditions. One proposed line of research is to directly monitor active-site occupancy (**10-11**) by FTICR-MS. This

methodology may allow for further validation of our model for catalysis in the PikAIII/IV system and will highlight the ability of FTICR-MS to simultaneously monitor multiple biochemical states in a complex system.

6.3.3 DEBS3 and un/natural pentaketides as substrates.

Recently, the Sherman lab has begun to explore *in vitro* biochemistry in the erythromycin system.[11] Indeed, we have determined the rate of catalysis for DEBS3 (modules 5 and 6) with the native SNAC-pentaketide (**Figure 6-4**). However, steady-state kinetics only offers a view of total production. What is occurring on the factory floor as these polyketides are assembled? One potential route to better develop a model for these data is to monitor active-site occupancy with natural and unnatural substrates (**Figure 6-5**).

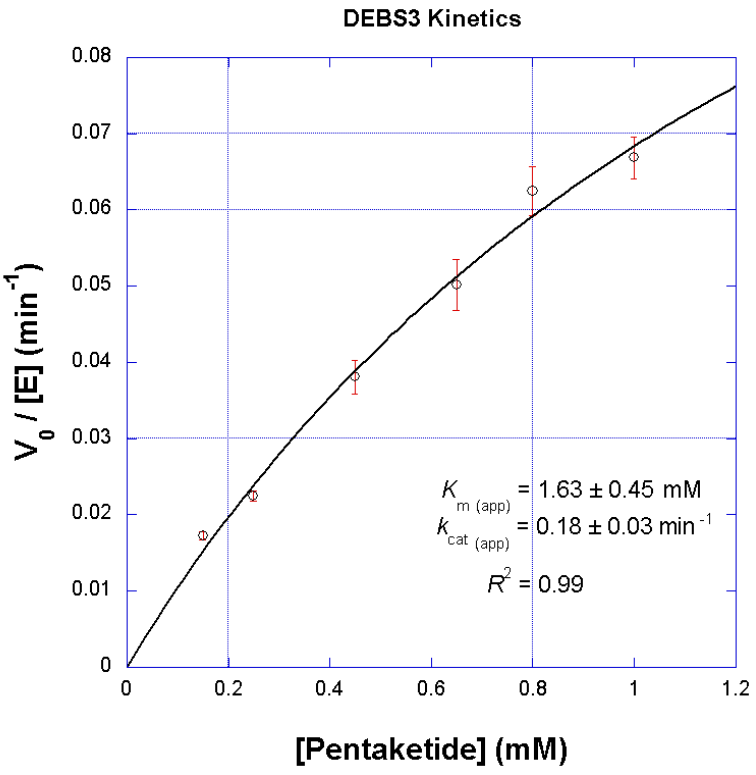
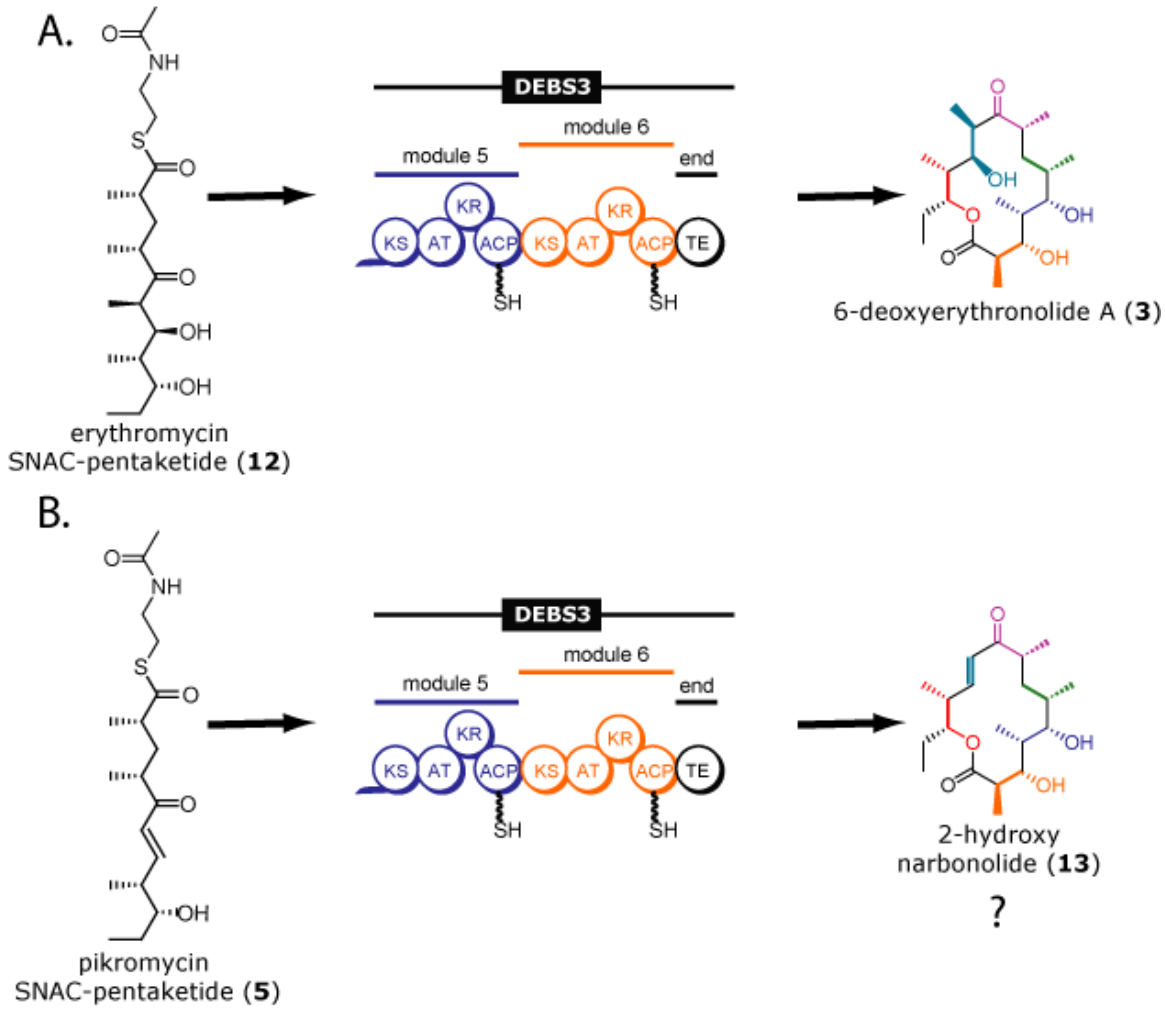


Figure 6-4. Normalized plots for steady-state kinetic parameters of DEBS3 with the native DEBS pentaketide SNAC substrate (12).



6-5. Chemoenzymatic synthesis of two macrolide antibiotics by DEBS3 with pentaketide substrates. (**A**) The production of 6-deoxyerythronolide A by DEBS3 from the erythromycin SNAC-pentaketide (**12**). (**B**) 2-hydroxy narbonolide is a potential product from the reaction of DEBS3 and the pikromycin SNAC-pentaketide.

As depicted in **Figure 6-5** we hope to monitor active-site occupancy in the DEBS3 module with both native (**Figure 6-5A**) and unnatural substrates (**Figure 6-5B**). Early experiments have proven that all DEBS3 active-site peptides can be monitored from a single experimental sample. While this preliminary result is our first glimpse into chain-elongation intermediate processing in this system it is intriquing. Further validation

by MS/MS of all loaded species is ongoing as is LC FTICR-MS/MS analysis of cyclic and linear on- (**13**) and off- pathway products from the unnatural substrate (**5**).

6.3.4 Component exchange: pikromycin, erythromycin, and tylosin

The final proposed experiment in this investigation of I PKS systems is exchange components between all three modular systems. This is illustrated below in **Figure 6-6**.

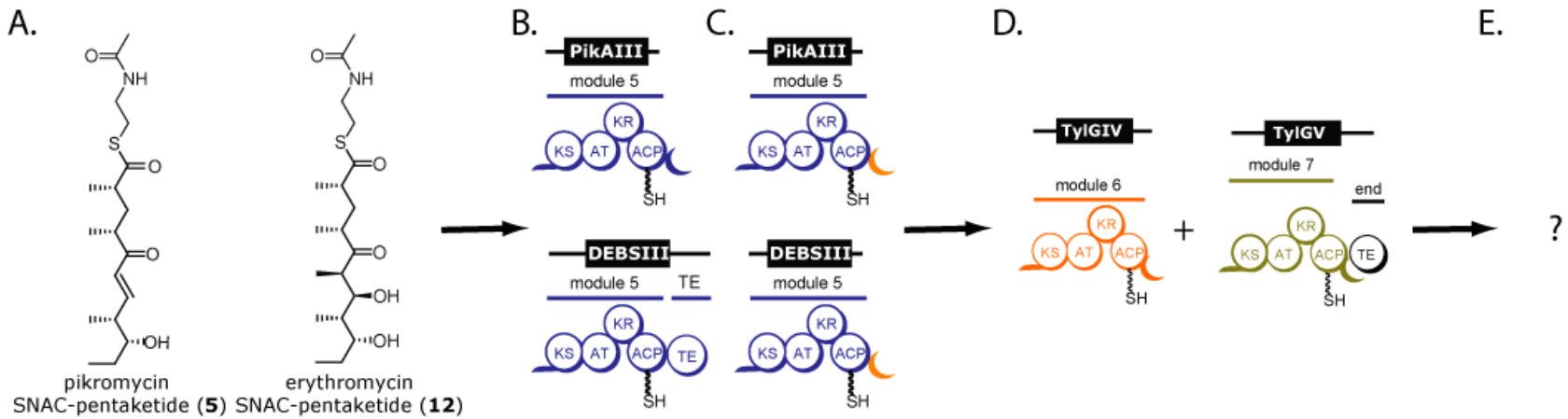


Figure 6-6. Exploring non-native module pairing with Pik, DEBS and Tyl. (**A**) Pikromycin (**5**) and erythromycin (**12**) SNAC-pentaketide substrates. (**B**) PikAIII and DEBSIII module 5 enzyme variants (**C**) PikAIII and DEBSIII module 5 variants with engineered tylosin docking domain 5 C-terminal docking domains. (**D**) Tylosin modules 6 and 7. (**E**) Potential linear and cyclic polyketide products.

We hope to exchange components between all three in-lab model type I PKS systems with the goal of better understanding contributing factors to productive versus non-productive parings (**Figure 6-6**). As illustrated both SNAC pentaketides (**5**, **12**) will be used to probe pikromycin and DEBS module 5 (**Figure 6-6B, Figure 6-6C**). The key variable in this experiment is whether or not the module 5's contain the engineered docking domain (**Figure 6-6B**) to be promote interactions with the downstream tylosin modules 6 and 7 (**Figure 6-6D**). Docking domains serve as specificity determining factors in ordering Type I PKS biochemical reactions.[18] This experiment will seek to determine if the correct docking domains can facilitate transfer of unnatural substrates to mispaired modules. Both enzyme-active sites and production of on- or off- pathway linear and cyclic products will be monitored by LC FTICR-MS/MS. Currently all proteins and substrates have been prepared for these studies.

6.4 Chemoenzymatic synthetic methods with FTICR-MS product analysis

Chemoenzymatic synthesis offers potential advantages over traditional chemical techniques in terms of selectivity, catalytic-efficiency, and "green" footprint.[19,20] Yet these powerful tools are not widely employed due to difficulty such as characterizing target substrates. Here, two systems are presented for screening the productivity of chemoenzymatic systems with small amounts of material in an automated fashion.

6.4.1 Cryptophycin combinatorial biosynthesis in a microfluidic device

The investigations into the full flexibility of the cryptophycin biosynthetic system were proposed as an RO1 through joint efforts in the Sherman and Dordick laboratory. This research plan was primarily planned and written by CMR. At this time, further preliminary efforts may be required prior to resubmission of this proposal (**Figure 6-7**).

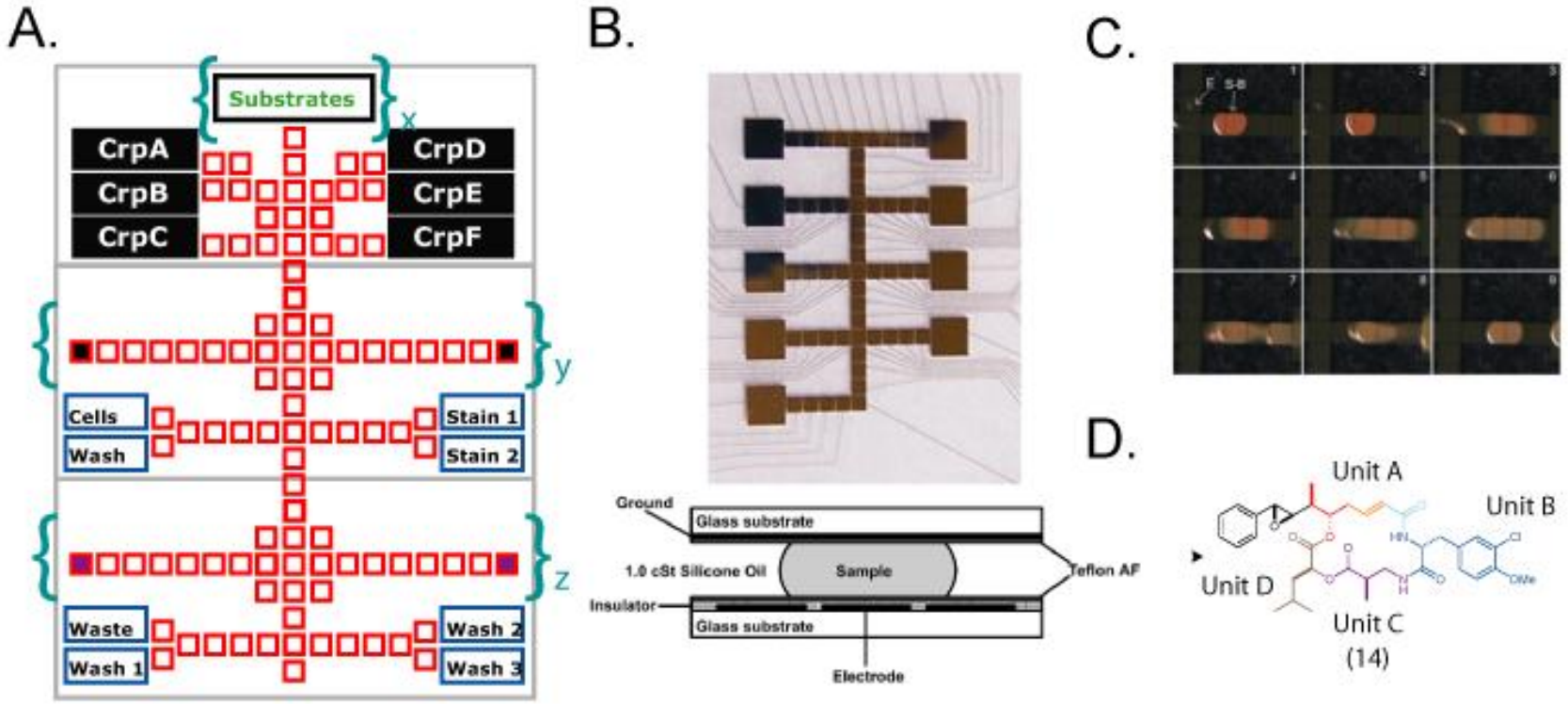


Figure 6-7. A digital microfluidics platform for chemoenzymatic synthesis of cryptophycin analogs with integrated biological and structural analytics. (**A**) Schematic diagram of the device. (**B**) A functional device prototype. (**C**) Video frames illustrating droplet mobilization and splitting within the device. (**D**) A model cryptophycin illustrating units A-D.

The goal of the proposed research is to develop a digital microfluidic device for the assembly and tailoring of novel compounds based on the broad catalytic properties of natural product biosynthetic enzymes. We will employ cross-disciplinary approaches toward development of a new chip-based method that harnesses the versatile catalytic activities of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) modules and allied tailoring enzymes for creation of a diverse range of biologically active, structurally complex, small molecules.

A growing body of work has demonstrated the broad ability of engineering modular PKS enzymes to generate novel compounds with unique biological activities.[21] Rationally designed biosynthesis of novel analogs, at high yields, is possible, as with the recently approved antibiotic daptomycin.[22] In another case, 14 modules from eight PKS clusters were combined into 154 bimodular combinations resulting in 72 products with engineered secondary metabolites produced in 46% percent of the cases tested. Despite this advance, production levels for unnatural megasynthases were often reduced 10-1,000 times compared to native modular pairings.[23] These findings highlight the fundamental difficulty in achieving efficient, non-native biomolecular recognition and have motivated us to develop a biochip-based approach to overcome these limitations. An essential concept behind the microfluidic approach is its ability to off-load/on-load product and substrate between separated modules, thus obviating the need to mediate productive protein-protein interactions in these multi-component systems. Our approach is based on a new strategy to overcome the limitations of molecular recognition that control module-to-module substrate transfer in PK/NRP enyzmes. This will enable facile modular

exchange from a diverse selection of metabolic pathways to enhance chemical diversity of the reaction products.

The core hypothesis of the proposed work is that engineering and physically separating individual PKS/NRPS modules into discrete reservoirs on a digital microfluidic platform will expand the capability of these enzyme systems and enable the construction of multienzyme pathways on individual microchips. This in turn can provide a unique environment to exploit the functional diversity and activity encoded by PKS/NRPS biosynthetic enzymes to generate synthetic natural product pathways for creation of novel structures and biological activities. As a result, a roadmap for large-scale biosynthesis of novel natural products will be provided. The key focus involves development of methods for effective module→module transfer of acyl-enzyme intermediates mediated by a digital microfluidic platform rather than module-module interactions.

Specific Aim 1. Engineer PKS/NRPS biosynthetic enzymes to function efficiently as distinct modules *in vitro*. To develop a microfluidic PKS/NRPS system, growing chain elongation intermediates must be transferred between physically separated modules. Strategies **(1a-c)** will be evaluated for transfer of activated thioester intermediates in the cryptophycin biosynthetic pathway as a model system.

a) Transthioesterification reactions will be utilized to off-load intermediates from biosynthetic PKS/NRPS modules to free CoA-SH.

b) Discrete acyl- or peptidyl carrier proteins will be used *in trans* to diffusively transfer biosynthetic intermediates between modules.

c) Acyl carboxy acid intermediates will be released by terminal thioesterase domains and reactivated by CoA-ligases to acyl-CoA intermediates.

d) The cryptophycin biosynthetic pathway will be reconstituted *in vitro* based on the most effective strategy determined from **1a-c**.

Specific Aim 2. Develop a digital microfluidic device for cryptophycin pathway reconstitution. The outcome of this phase of the work is development of a microfluidic platform that includes magnetic immobilization of PKS/NRPS modules. This design will enable effective electrowetting and fluidic transport through the platform for facile handling of biosynthetic substrates and intermediates.

a) A digital microfluidic device for PKS/NRPS biosynthesis *in vitro* will be fabricated and tested.

b) Distinct polypeptides of the cryptophycin PKS/NRPS will be immobilized on various supports to optimize enzyme activity and stability.

c) Immobilized PKS/NRPS proteins will be loaded into a reservoir of the digital microfluidic device and activity and stability determined.

d) Substrates will be added to the active, immobilized, and ordered PKS modules using the method determined in **1a-d** to provide an active, reconstituted PKS pathway on a microfluidic platform.

Specific Aim 3. Combinatorially manipulate the modular pathway to synthesize novel hybrid PKS/NRPS products. By altering the order and composition of the PKS/NRPS biosynthetic modules in the microfluidic device, changing input substrates

(e.g. starter/extender units), and utilizing tailoring enzymes, a suite of novel natural product analogs will be generated.

a) Evaluate the use of alternative starter units and elongation units on efficiency of product formation.

b) Perform combinatorial pathway assembly on the microscale with additional PKS/NRPS protein modules.

c) Perform post-PKS/NRPS tailoring reactions on the microscale with immobilized tailoring enzymes.

d) 3a-c will be combined for biosynthesis of a library of PKS/NRPS analogs in a high-throughput approach.

Specific Aim 4. Integrated product analysis and biological screening of engineered natural products. Direct screening of product structures will be accomplished through digital microfluidic-coupled DIOS FTICR mass spectrometry with online sample preparation. Product activity screening will be accomplished with a digital microfluidic coupled cell-based cytotoxicity assay.

a) Fabrication and optimization of a microfluidic coupled DIOS FTICR MS assay with sample preparation.

b) Development of a microfluidic whole cell viability assay using fluorescence.

c) Coupling of **4a-b** with the digital microfluidic device for PKS biosynthesis as developed in **1-3**.

6.4.2 RhFRED-PikC substrate screening by LC FTICR-MS/MS

The pikromycin gene product PikC P450 has been proven to be a powerful tool for direct activiation of C-H bonds.[24] Indeed, when coupled with an activating domain (RhFRED), it can function as a self-sufficient catalytic unit.[3] A novel system for screening the catalytic flexibility of this enzyme on the large scale is proposed (Figure 6-8).

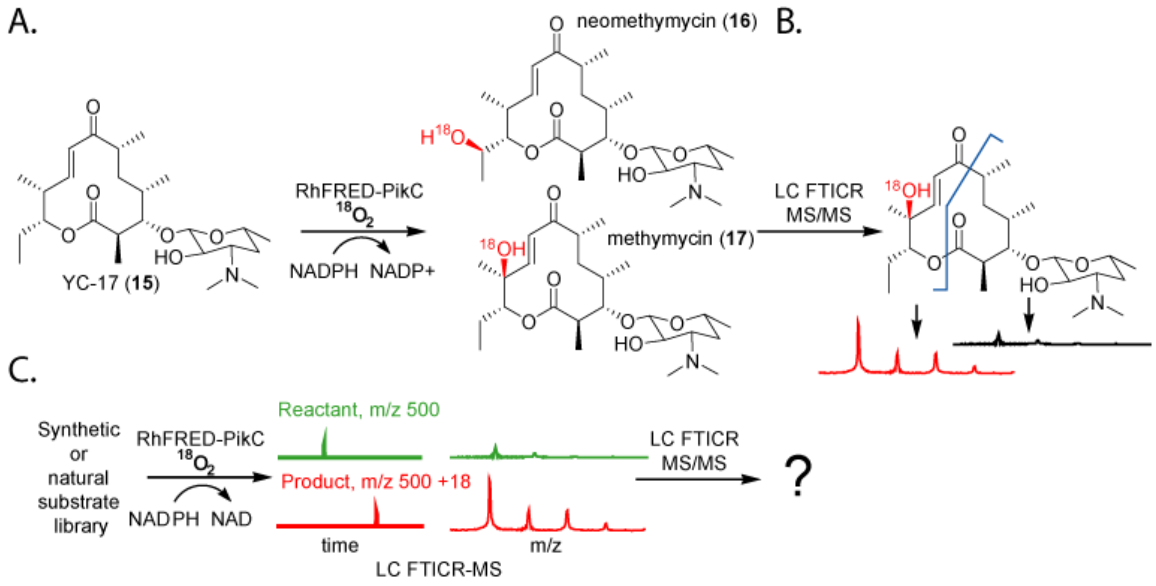


Figure 6-8. An FTICR-MS/MS platform for high-throughput screening of substrates for C-H bond activation by Rh-FRED PikC with $^{18}O_2$. (**A**) The oxidation of YC-17 (**15**) by the P450 RhFRED-PikC to neomethymycin (**16**) and methymycin (**17**). (**B**) The ^{18}O label results in a shift of in the isotopic distribution at the monoisotopic +2 peak, this allows for localization of the modification in MS/MS experiment. (**C**) A large synthetic or natural product library could be screened for product generation in MS1 through a shift of +18 Da and a change in isotopic distribution. MS/MS then allows for localization of modification.

RhFRED serves as a self-sufficient catalyst for hydroxylation of un-activated C-H bonds, and in the presence of $^{18}O_2$ and equimolar $^{16}O_2$should introduce a +18 Da mass shift to the target molecule (**Figure 6-8A**). This will provide a unique signature in the MS spectra in terms of mass-shift and isotopic pattern that will also allow for localization

within MS/MS (**Figure 6-8B**) fragmentation techniques such as CID, IRMPD, ECD, and EDD.[25-28] Since modern high-mass accuracy/high-resolution mass spectrometers operate routinely with sub-nanogram amounts of material, it should be possible to rapidly screen hundreds of potential natural or synthetic compounds as potential substrates for this enzyme (**Figure 6-8C**). This screen will both help to better characterize the substrate profile for this enzyme and to identify potential commercial applications for this catalytic C-H bond activation.

6.5 ET-743 and the Etu biosynthetic pathway

Early research efforts in the ET-743 biosynthetic pathway have been fruitful—allowing us to identify key metabolites, genes, and proteins within the tunicate associated bacteria (**Figure 6-9A, Chapter 5**). Future efforts will focus on verifying individual enzyme activities through *in vitro* biochemistry, and developing a deeper understanding of enzyme mechanisms through crystallography. As well, chemical probes will be applied to test biochemical activity in the native in-vivo context (**Figure 6-9B**). Eventually, a full understanding of this system will be applied to develop a minimal set of genes for ET-743 biosynthesis and either a chemoenzymatic or fermentation based route for economical production of ET-743 and analogues to meet the clinical demand for this compound (**Figure 6-9C**).

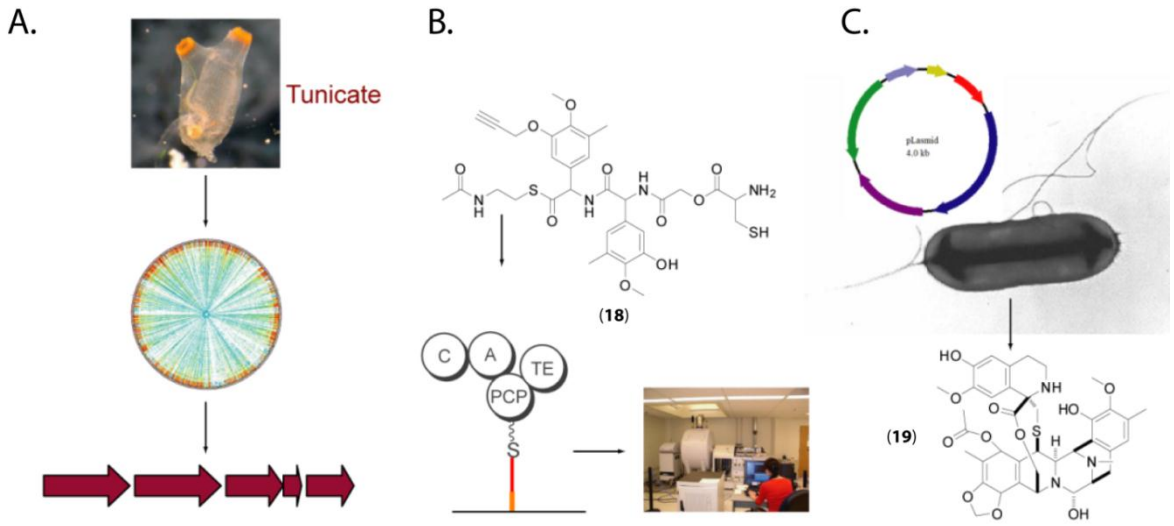


Figure 6-9. An outline of the long-term goals for the ET-743 project. (**A**) Tunicate collection, metabolomics, metagenomics, metaproteomics. (**B**) Biochemical validation *in vitro* (biochemistry) and *in vivo* (activity based protein profiling with hypothetical probe (**18**)). (**C**) Heterologous expression of a minimal set of ET-743 biosynthetic genes in a well-behaved host could allow cheap fermentation of ET-743 (**19**) and analogues.

6.5.1. *In vitro* biochemistry and crystallography.

Currently a series of Etu genes have been successfully cloned with more efforts ongoing (**Table 6-1**). Exhaustive attempts at expression in *E. coli* with various tools (fusion proteins, increased rare codons tRNA's, chaperones, modified expression conditions, alternative constructs) have only yield one poorly soluble, low-yielding product (<1 mg/L). This is likely due to an extremely poor match between the *E. frumentensis* and *E. coli* codon usage ("The worst genes I have ever seen"-personal communication, Clay Brown). Currently, two Etu genes have been synthesized with optimized codons for *E. coli* expression. In parallel, efforts are underway to express Etu proteins in yeast, as the codon usage profiles are coincidently similar.

	E. coli (WT)		*E. coli* (synthetic genes)		Yeast	
	Cloned?	**Expressed?**	**Cloned?**	**Expressed?**	**Cloned?**	**Expressed?**
EtuA1 (CAT)	Y	N			Y	?
EtuA2 (CATRE)	Y	N			Y	?
EtuA2 (A)1	Y	N			Y	?
EtuA2 (A)2	Y	N			Y	?
EtuA2 (A)3	Y	N			Y	?
EtuA2 (RE)1	Y	Y			Y	?
EtuA2 (RE)2	Y	N				
EtuA2 (RE)3	Y	N				
EtuA2 (RE)4	Y	N				
EtuA2 (T)	Y	N			Y	?
EtuA2 (TRE)	Y	N			Y	?
EtuF3			Y	Y		
EtuH	Y	N			Y	?
EtuM1	Y	N			Y	?
EtuM2	Y	N			Y	?
EtuO			Y	Y		

Table 6-1. Cloning and expression efforts for Etu genes. Results are provided for cloning and expression of the wild-type and synthetic genes in *E. coli* and yeast. Y = success, N = failure, ? = unknown results.

6.5.2. Activity based protein profiling for natural product systems.

With key genes in the ET-743 biosynthetic pathway identified a further goal is to link proposed biochemical activity to function in the *in vivo* tunicate symbiont system. We propose to employ Et-743 early pathway biosynthetic precursors as chemical probes (**Figure 6-10**) to isolate and sequence key biosynthetic enzymes—thus verifying their assigned functions (**Figure 6-11**). This methodology, developed by the Cravat laboratory as activity based protein profileing (ABPP), has proven to useful for targeting enzymes in a variety of different contexts.[29-31]

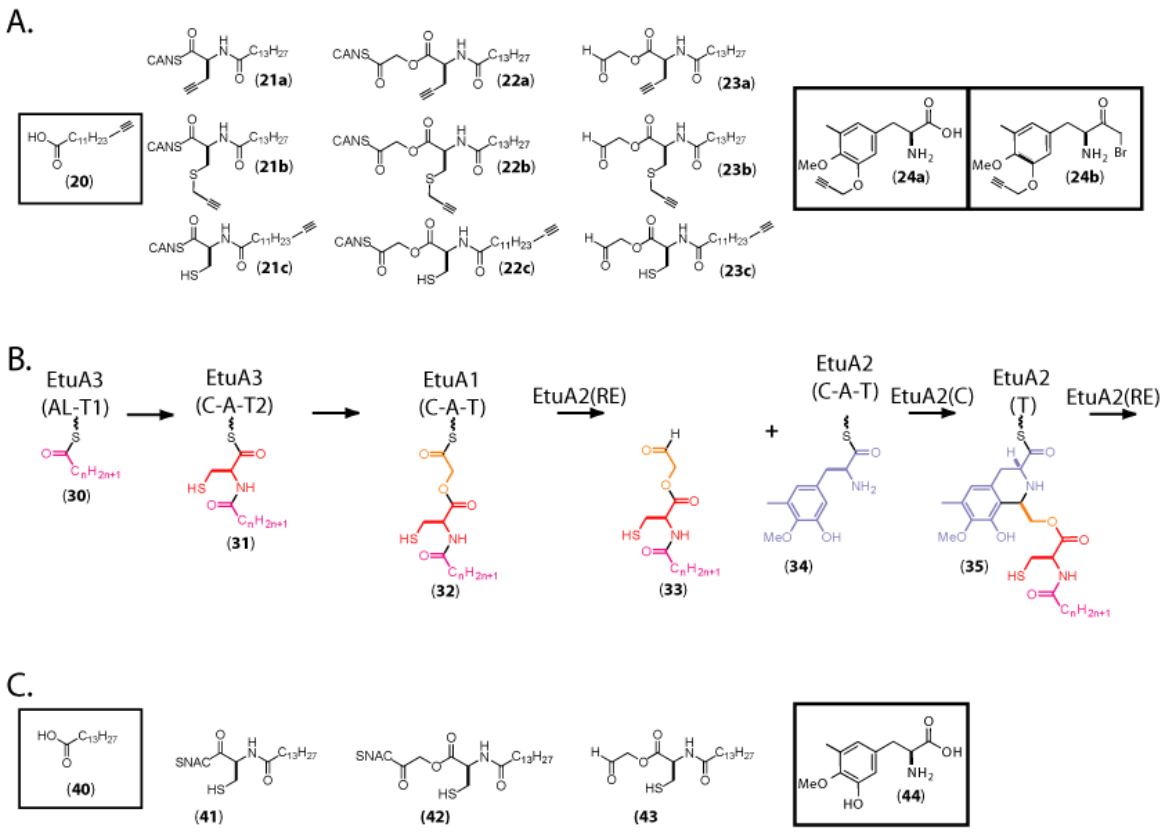


Figure 6-10. Probing the ET-743 biosynthetic pathway with ABPP. (**A**) Synthetic targets (**20-24**) for activity based protein probes in the ET-743 biosynthetic pathway. (**B**) A portion of the ET-743 biosynthetic pathway (EtuA3→EtuA1→EtuA2) with key intermediates to be targeted (**30-35**). (**C**) Biochemical probes **40-44**, SNAC and aldehyde analogs. Probes in boxes are commercially available (**20**, **40**) or have been synthesized (**24a**, **24b**, **44**).

Chemical probes for activity based labeling (**20-24**) based upon key biosynthetic intermediates (**30-35**) are illustrated (**Figure 6-10**). The series of probes above are targeted towards each of the modules in the NRPS portion of the ET-743 biosynthetic pathway, and are currently being synthesized. Free-acid (**20, 24a**), thioester (**21-22**), aldehyde (**23**) and alpha-keto halide (**24b**) probes are being explored as click reagents with three possibilities for alkyne incorporation (**a-c**). The key common feature among these probes, the alkyne handle, will allow ABPP methodology to be applied to pull-

down and characterize the proteins EtuA1-3 (**Figure 6-11**). Native biochemical probes are also being prepared for *in vitro* biochemical experiments (**40-44**).

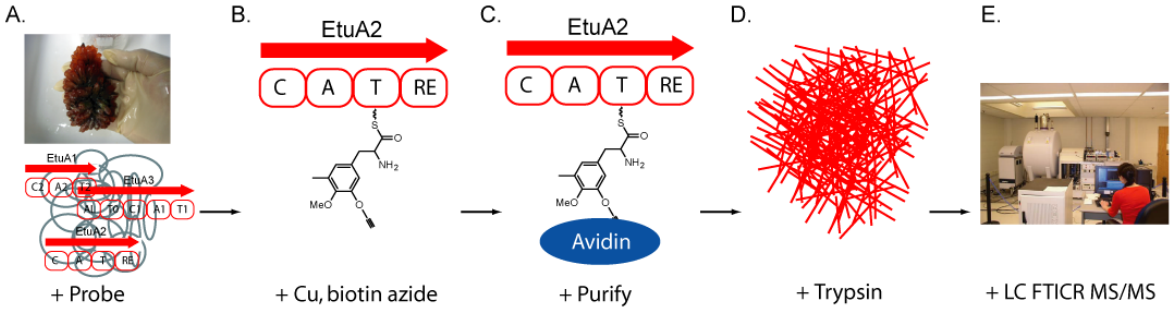


Figure 6-11. Activity based protein profiling in the ET-743 biosynthetic system. (**A**) Tunicate collection and probe addition pre- and/or post- lysis. (**B**) Alkyne-labeled probes (**Figure 6-11**) will covalently react with targeted natural product synthases. Cu (I) catalyzed 2+3 cyclo-addition reaction allows for the bioorthogonal ligation of a biotin handle. (**C**) Biotin-avidin affinity chromatography can be used to target reacted enzymes-substrate complexes. (**D**) A trypsin digest results in a mixture of peptides. (**E**) The peptides can be identified and linked to expressed proteins with high-performance nLC FTICR-MS/MS.

The ET-743 biosynthetic probes (**20-24**) will be added to the live animal or an active whole-cell lysate (**Figure 6-11A**). This strategy will take advantage of the thioester linkage formed between the phosphopantetheine arm of the peptidyl carrier protein in each of three NRPS modules and loaded ET-743 precursor, where a biotin can then be appended through a 3+2 Cu(I) catalyzed cyclo-addition "Click-Chemistry" (**Figure 6-11B**). This protein-probe-biotin covalent complex can then be purified through biotin-avidin affinity chromatography (**Figure 6-11C**). The purified probe labeled sample can then be digested with trypsin (**Figure 6-11-D**) and then subjected to a proteomics nLC FTICR-MS/MS workflow (**Chapter 5**) for identification of the labeled proteins. Identification of any of the putative ET-743 biosynthetic proteins will serve as proof of their assigned function in the *in vivo* context. This strategy draws upon the ability of the Williams laboratory to synthesize intermediate-based affinity probes, the experience of

the Sherman laboratory in biosynthetic pathway identification and characterization, and the Håkansson laboratories expertise in probing complex biological systems with mass spectrometry.

These same activity based probes can also be used in imaging experiments by substituting an azide fluorophore for the azide biotin.[6] This technique then allows for biosynthetic proteins to be localized within the context of the tunicate tissue, where the bacteria can be localized with techniques such as FISH (**Figure 6-12**).[32]

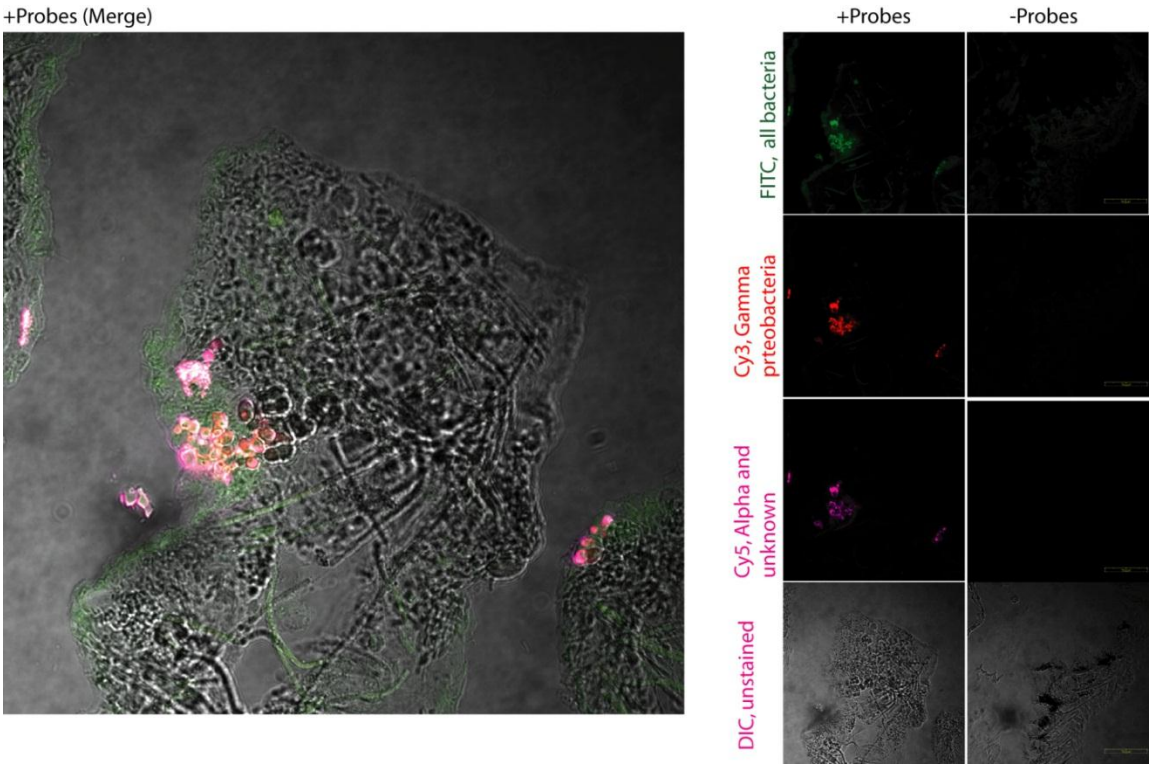


Figure 6-12. Fluorescent *in situ* hybridization (FISH) analysis of an *E. turbinata*. 16S DNA probes are used with a general bacterial probe (FITC), a probe for the γ-proteobacteria *E. frumentensis* (Cy3), and the alpha-proteobacteria and other bacteria strain identified (Cy5).

Localization of the biosynthetic proteins with ABPP methodology, and key species with FISH (**Figure 6-12**), could be further complimented with the use of imaging

mass spectrometry to localize small molecules within the same system,[7,33,34] allowing for delineation of the biological system at the species, protein, and small molecule level—a thus far unexplored level of integration. These same techniques could also be applied to identify ET-743 producing bacteria in high-throughput attempts to culture the producing bacteria in a microfluidic device (**Figure 6-13**).

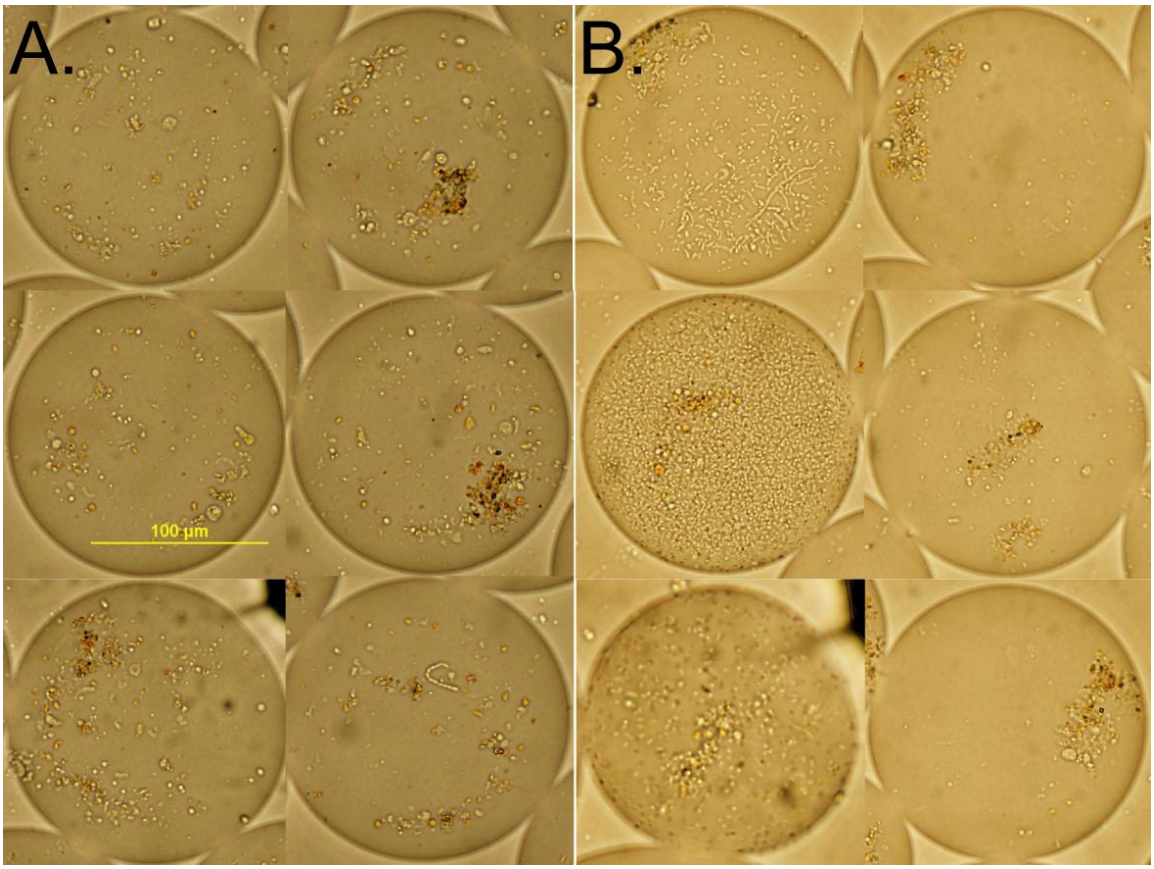


Figure 6-13. Culturing *E. turbinata*-derived bacteria in a microfluidic device. (**A**) 6 droplets at 0 hours. (**B**) 6 droplets at 26 hours. Data: (Jihyang Park and Prof. Lin)

6.6 Conclusion

Four chapters of this manuscript have described the detailed investigation of biosynthetic systems with mass spectrometry, and the final chapter has detailed ongoing

and future investigations inspired by these efforts. This work has proven that the application of powerful analytical techniques, primarily FTICR-MS, allow for natural product biosynthetic systems to be dissected and characterized at an unprecedented level of detail. The techniques have supported the development of mechanistic models, supported chemoenzymatic generation of natural products, and allowed for novel symbiont derived pathways to be characterized. Clearly, the future is wide open for many of these avenues of investigation.

6.7 References

1. Beck, Z. Q.; Aldrich, C. C.; Magarvey, N. A.; Georg, G. I.; Sherman, D. H. *Biochem,* **2005**, *44*, 13457.
2. Magarvey, N. A.; *et al. ACS Chem Biol,* **2006**, *1*, 766.
3. Li, S; Podust, L. M.; Sherman, D. H. *J Am Chem Soc,* **2007**, *129*, 12940.
4. Koketsu, K.; Watanabe, K.; Suda, H.; Oguri, H.; Oikawa, H. *Nat Chem Biol*, *6*, 408.
5. Cravatt, B. F.; Wright, A. T.; Kozarich, J. W. *An Rev of Biochem*, **2008**, *77*, 383.
6. Jessani, N.; Cravatt, B. F. *Cur Op Chem Biol*, **2004**, *8*, 54.
7. Esquenazi, E.; Yang, Y.-L.; Watrous, J.; Gerwick, W. H.; Dorrestein, P. C. *Nat Prod Rep,* **2009**, *26*, 1521.
8. Eid, J.; *et al. Science,* **2009**, *323*, 133.
9. Xue, Y.; Zhao, L.; Liu, H. W.; Sherman, D. H. *Proc Natl Acad Sci USA*, **1998**, *95*, 12111.

10. Khosla, C.; Tang, Y.; Chen, A. Y.; Schnarr, N. A.; Cane, D. E. *Annu Rev Biochem,* **2007**, *76*, 195.

11. Mortison, J. D.; Kittendorf, J. D.; Sherman, D. H. *J Am Chem Soc*, **2009** 15784.

12. Cundliffe, E.; *et al*. *Ant van Leeuwenhoek,* **2001**, *79*, 229.

13. Akey, D. L.; *et al. Nat Chem Biol,* **2006**, *2*, 537.

14. Aldrich, C. C.; Beck, B. J.; Fecik, R. A.; Sherman, D. H. *J Am Chem Soc,* **2005**, *127*, 8441.

15. Sieber, S. A.; Tao, J.; Walsh, C. T.; Marahiel, M. A. *Angewandte Chemie,* **2004**, *116*, 499.

16. Wu, N.; Cane, D. E.; Khosla, C. *Biochem,* **2002**, *41*, 5056.

17. Kittendorf, J. D.; Beck, B. J.; Buchholz, T. J.; Seufert, W.; Sherman, D. H. *Chem Biol,* **2007**, *14*, 944.

18. Buchholz, T. J.; *et al. ACS Chem Biol,* **2009**, *4*, 41.

19. Davis, B. G.; *et al*. *Nat Product Rep,* **2001**, *18*, 618.

20. Schmid, A.; *et al. Nature,* **2001**, *409*, 258.

21. Menzella, H. G.; Reeves, C. D. *Curr Opin Microbiol*, **2007**, *10*, 238.

22. Doekel, S.; *et al. Microbiol,* **2008**, *154*, 2872.

23. Menzella, H. G.; *et al. Nat Biotech* **2005**, *23*, 1171.

24. Sherman, D. H; *et al. J Biol Chem*, **2006**, *281*, 26289.

25. Laskin, J.; Futreil, J.H. *Mass Spec Rev,* **2003**, *22*, 158.

26. Little, D. P. *Anal Chem,* **1994**, *66*, 2809.

27. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F.W. *J Am Chem Soc,* **1998**, *120*, 3265.

28. Budnik, B. A.; Haselmann, K.F.; Zubarev, R.A. *Chem Phys Let*, **2001**, *342*, 299.

29. Speers, A. E.; *et al*. *J Am Chem Soc*, **2003**, *125*, 4686.

30. Weerapana, E. *Nature Protocols*, **2007**, *2*, 1414.

31. Everley, P. A.; *et al. Mol Cell Prot*, **2007**, *6*, 1771.

32. Parez-Matos, A. E.; Rosado, W.; Govind, N. S. *Antonie van Leeuwenhoek*, **2007**, *92*, 155.

33. Esquenazi, E.; *et al. Mol Biosystems*, **2008**, *4*, 562.

34. Esquenazi, E.; Dorrestein, P. C.; Gerwick, W. H. *Proc Nat Acad Sci USA*, **2009**, *106*, 7269.