

**High-Throughput  
Substrate-Activity Screens  
and Mechanistic Studies  
with Nonribosomal Peptide Synthetases  
Using Large Molecule Mass Spectrometry**

By

Jonathan Robert Blackhall

---

Thesis  
for the  
Degree of Bachelor of Science  
in  
Biochemistry

College of Liberal Arts and Sciences  
University of Illinois  
Urbana-Champaign, Illinois

2006

# **Abstract: High-Throughput Substrate-Activity Screens and Mechanistic Studies with Nonribosomal Peptide Synthetases Using Large Molecule Mass Spectrometry**

The emergence of drug resistant pathogens in recent years has forced pharmaceutical companies to examine alternative means of drug research and development. Natural bioactive compounds, such as penicillin, are produced by various organisms using nonribosomal peptide synthetases (NRPSs), and these compounds act as our defense against pathogenic microbes. Large molecule mass spectrometry is an effective tool to study NRPS systems. This study reports a method of high-throughput substrate screening that drastically improves the efficiency of translating high-resolution mass spectrometry into biochemical insight. Through a set of “proof of concept” experiments, it was shown that natural NRPS substrates can be readily identified, even from a complex mixture of more than 100 compounds. Additionally, non-cognate substrates can be loaded on carrier domains and detected when the natural one is absent. This method for determining “Structure-Activity Relationships” by mass spectrometry was then applied to systems in which the substrate and final products are *unknown*. It was found that the NRPS modules from an orphan gene cluster from *Bacillus subtilis* 168 loaded glycine, alanine, and phenylacetate – giving a first glimpse of what natural product is produced by this novel system. Additionally, this technique was used to examine an amine transfer reaction in the biosynthesis of mycosubtilin, a potent antifungal agent. This combination of high-throughput substrate screening with large molecule mass spectrometry provides an accurate and efficient method to study nonribosomal peptide biosynthesis for accelerated development of novel pharmaceuticals.

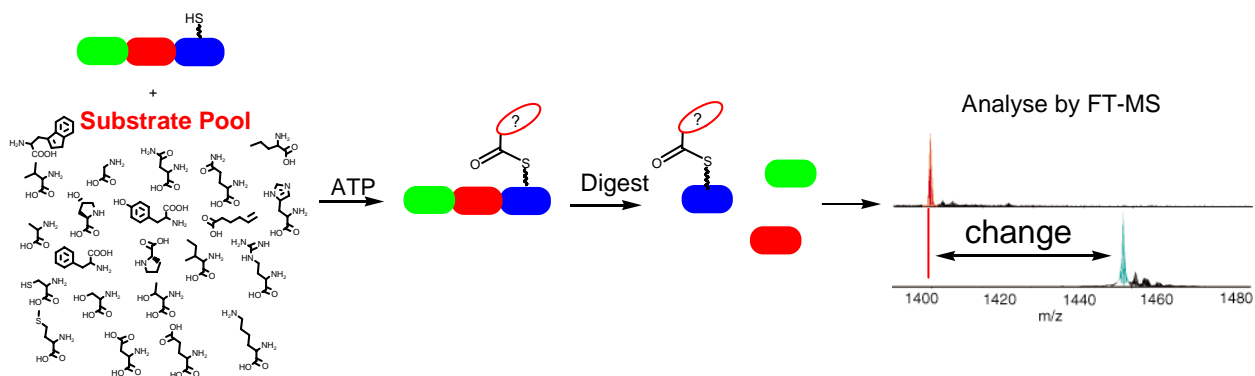
## Acknowledgements

Special thanks to Dr. Pieter Dorrestein, Prof. Neil Kelleher, and the rest of the Kelleher Group. Also, thanks to Prof. Christopher Walsh (Harvard Medical School) and the rest of the Walsh Lab, specifically Dr. Sylvie Garneau-Tsodikova for providing the clorobiocin and coumermycin constructs and Dr. Zachary Aron for providing the MycA constructs. Thanks to Dr. Paul Straight of the Kolter Lab (Harvard Medical School) for providing the PksJ and PksN constructs and Dr. Michael Thomas (University of Wisconsin) for the Atu3673 construct.

Use of the ESI-FTMS, data analysis, and generation of figures represent a joint effort with Dr. Pieter Dorrestein. Portions of this thesis are published in Activity Screening of Carrier Domains within Nonribosomal Peptide Synthetases Using Complex Substrate Mixtures and Large Molecule Mass Spectrometry [Pieter C. Dorrestein, **Jonathan Blackhall**, Paul D. Straight, Michael A. Fischbach, Sylvie Garneau-Tsodikova, Daniel J. Edwards, Shaun McLaughlin, Myat Lin, William H. Gerwick, Roberto Kolter, Christopher T. Walsh, and Neil L. Kelleher. *Biochemistry*; **2006**; 45(6) pp 1537 – 1546] and Characterization of a New Tailoring Domain in Polyketide Biogenesis: The Amine Transferase Domain of MycA in the Mycosubtilin Gene Cluster [Zachary D. Aron, Pieter C. Dorrestein, **Jonathan R. Blackhall**, Neil L. Kelleher, and Christopher T. Walsh. *J. Am. Chem. Soc.*; **2005**; 127(43) pp 14986 – 14987] © American Chemical Society. This work was supported in part by NIH Grants GM 049338 (C.T.W.), GM 020011 (C.T.W.), 067725 (N.L.K.), GM 58213 (R.K.), and CA 83155 (W.H.G.). Additional support came from the NIH Kirschstein NRSA Postdoctoral Fellowships F32-GM073323-01 (P.C.D.) and F32-GM72299-01 (Z.D.A.), National Science Foundation Postdoctoral Fellowship in Microbial Biology DBI-0200307 (P.D.S.), the Hertz Foundation Graduate Fellowship (M.A.F.), and the Jackson Scholars in Biochemistry Award (J.R.B.).

# Table of Contents

	<u>Page</u>
Abstract.....	i
Acknowledgements.....	ii
Table of Contents.....	iii
List of Abbreviations .....	iv
Introduction.....	1
Materials and Methods.....	8
Results and Discussion .....	14
Conclusions.....	23
Bibliography .....	24
Figures and Tables .....	26



*Method for Determining Structure-Activity Relationships by Mass Spectrometry*

## List of Abbreviations

- HPLC: high-pressure liquid chromatography
- ESI: electrospray ionization
- FTMS: Fourier-transform mass spectrometry
- CoA: coenzyme A
- NRPS: nonribosomal peptide synthetase
- PKS: polyketide synthetase
- OCAD: octopole collisional activated dissociation
- IRMPD: infrared multiphoton dissociation
- ECD: electron capture dissociation
- SWIFT: stored waveform inverse Fourier-transform
- TCEP: Tris(2-carboxyethyl)phosphine
- PCP: peptidyl carrier protein
- ACP: acyl carrier protein
- IPTG: isopropyl  $\beta$ -D-thiogalactoside
- OD<sub>600</sub>: optical density at 600 nm
- MIDAS: modular ICR data acquisition system
- THRASH: thorough high-resolution analysis of spectra by Horn
- T: Tesla

## Introduction

### *What are Nonribosomal Peptide Synthetases?*

Nonribosomal peptide synthetases (NRPS) are large, multifunctional enzymes that synthesize small peptides derived from amino acids without the aid of a nucleic acid template. They produce bioactive compounds such as penicillin, vancomycin, epithilone, and yersiniabactin (Figure 1A). They generate peptides with a wide variety of structures including linear, cyclic, and cyclic branched configurations. In addition to the twenty proteinogenic amino acids, NRPS enzymes are able to incorporate additional amino acids including D-amino acids, N-methylated, hydroxy, and halogenated amino acids as well as carboxylates such as salicylate and dihydroxybenzoate. This allows NRPS enzymes to synthesize peptides with an increased structural diversity and with more biofunctional properties than a traditional ribosomal peptide can achieve (1).

### *Essential NRPS Domains*

NRPS enzymes are multimodular proteins that synthesize their products in a similar fashion to fatty acid synthases. The enzymes contain separate domains, each responsible for a particular step in the biosynthesis. Minimally, an NRPS molecule must contain an adenylation domain and a thiolation domain, which represents an enzyme about 650 amino acids long (1). More domains are usually included, with the largest NRPS enzyme known, *tex1*, containing over 21,000 amino acids (2). Many NRPS modules are also found to be freestanding domains *in vivo*, although they are generally located within the same operon or gene cluster. Even when the modules are not connected, however, the domains are able to generate peptide products of defined structure (1).

The primary sequence of the peptide product is determined by the adenylation (A) domains. In NRPS systems each A domain adds one amino acid to the growing peptide chain. The addition of an amino acid first involves the activation of the amino acid to an aminoacyl-adenylate in the presence of  $Mg^{2+}$ -ATP<sup>2-</sup> followed by covalent transfer of the amino acid to the thiolation domain via a 4'-phosphopantetheine (4'-PP) arm (Figure 1C). The activation process is homologous to the activation of tRNA synthetases used in protein translation from mRNA. Analysis of the A domain crystal structure reveals that a large N-terminal domain is responsible for the identification of the substrate and formation of the aminoacyl-adenylate. The smaller C-terminal domain appears to be responsible for the attachment of the activated amino acid to the thiolation domain (1).

The thiolation (T) domain, also known as the peptidyl carrier protein (PCP) in NRPS systems, acts as the point of attachment for the amino acid activated by the A domain. Before amino acid tethering can take place, the T domain must be primed by the addition of 4'-PP by a 4'-PP transferase. In this reaction, the apo form of the T domain is converted to the holo form by the addition of the phosphopantetheinyl portion of CoA to an active site serine, releasing 3',5'-ADP (Figure 1B). Once the PCP has been converted to its holo form, the cysteamine group is ready to attack the activated aminoacyl-adenylate in the A domain. Once the amino acid is tethered to its PCP, further elongation occurs.

The condensation (C) domain catalyzes the elongation reaction. In this reaction, two PCP domains that have been loaded with their respective amino acids can undergo condensation by the nucleophilic attack of the amino group of one attacking the thioester of the other, forming an amide linkage (Figure 1C). Elongation continues by adding one amino acid to the C-terminus of the growing peptide for each PCP domain in the system. Therefore, it follows that the number

of C domains coincides with the number of amide linkages formed to generate the linear forms of intermediates (1). The peptide chain is then released from the final NRPS module either by thioesterase mediated hydrolysis or macrocyclization by a termination (TE) domain.

### *Substrate Specificity of the A Domain*

Since the A domains are responsible for activation of the proper amino acid, their substrate specificity is crucial to the development of accurate peptides. Further analysis of A domains revealed ten amino acid residues that are responsible for its substrate specificity (1). As more information was gathered, it even became possible to alter the specificity of an A domain by mutating these conserved residues. Using these data, the so-called “nonribosomal code” was developed in order to predict substrate specificity of uncharacterized A domains (1). While this “code” is very useful in assigning an amino acid to the modules when the final product is known, it is inadequate in assigning substrates in NRPS systems with unknown final products and, therefore, will have to be experimentally verified.

While the A domain’s specificity is essential for producing accurate peptides, many of them are only moderately specific, especially when compared to their cognates, tRNA synthases (1). In *in vitro* studies, NRPS products have been shown to incorporate incorrect amino acids found in excess through non-specific A domain activation. This gives rise to alternative products that could possibly be useful.

### *Large Molecule Mass Spectrometry in the Investigation of NRPS Biosynthetic Enzymes*

While mass spectrometry has been used since the early twentieth century, it could not be used to study macromolecules such as proteins until more recently. Macromolecules first need

to be desorbed, or transferred to the gas phase, and the techniques for this were not developed until the 1970s. John Fenn and Koichi Tanaka, who shared half of the 2002 Nobel Prize in Chemistry, pioneered the development of two types of macromolecular ionization techniques. Fenn developed electrospray ionization (ESI), which generates positively or negatively charged ions of macromolecules upon desolvation (3). Tanaka developed soft laser desorption (SLD), which uses a laser pulse to blast macromolecules from the solid or viscous phase into positively or negatively charged gaseous ions (4). Once the molecules have been desorbed, mass spectrometry can be performed on the ions. This involves using magnetic and electric fields to measure the ions' mass to charge ratio. Specifically, Fourier-transform mass spectrometry (FTMS) allows for detection of mass to charge ratios with mass accuracies to within 1-2 ppm (5). In these studies, ESI is coupled with FTMS on a custom-built 8.4 T mass spectrometer (6).

Using the high mass accuracy of FTMS, digested NRPS modules exhibiting even minute mass changes can be detected and analyzed (7-9). Many of the steps in nonribosomal peptide biosynthesis feature a growing peptide chain tethered to various sites on the enzyme. Mass spectrometry can be used to determine the exact mass of the enzyme-substrate complex, and when the mass of the enzyme is known, this elucidates the mass of the attached substrate. Thus, mass spectrometry is ideal for analyzing mass changes of the PCP domains upon loading with their amino acid substrates.

Furthermore, the use of tandem mass spectrometry (MS/MS) can localize the additional mass of the substrate to the exact active site serine of the PCP. This is done by using methods such as octopole collisional activated dissociation (OCAD), infrared multiphoton dissociation (IRMPD), and electron capture dissociation (ECD) to fragment the NRPS module. OCAD involves the introduction of neutral gas molecules into the chamber to induce collisional

dissociation, while IRMPD uses laser photons. Both of these methods are considered harsh fragmentation methods (5). ECD, which uses electrons to bombard the gaseous ions, is a much more gentle fragmentation technique (5). The fragments are analyzed by FTMS using ProSight PTM software in order to determine which portion of the module each fragment represents and also whether that fragment contains the 4'-PP tethered substrate (10). Comparison of many fragment ions allows for increased localization of the active site serine of the PCP (10). This technique provides efficient localization of the precise covalently modified serine with nearly flawless accuracy.

#### *Fluorescent Derivatives in Active Site Mapping*

Due to the large size of many NRPS enzymes, it is often necessary to digest them with a protease prior to analysis by mass spectrometry. After HPLC purification of the various fragments, visualization of the enzyme's active site by mass spectrometry becomes a more arduous process. Each fragment must be separately analyzed by mass spectrometry to determine its mass. This is compared to the masses of all the possible fragments in order to determine which portion of the NRPS module it represents. Sifting through all of the possible fragments in search of one or two possible species containing the active site serine can be quite time consuming.

The use of a BODIPY (Molecular Probes) fluorescent derivative of CoA has made this active site mapping process much more efficient (Figure 2A) (11). The HPLC fractions can be analyzed directly by irradiating them with UV light. Samples that exhibit high fluorescence contain the BODIPY-CoA derivative and can be further analyzed by mass spectrometry (11). This is an accurate and efficient assay to determine which HPLC fractions contain NRPS carrier

domain active sites that have been loaded with BODIPY-CoA. While this technique is quick and relatively inexpensive, without further analysis by mass spectrometry, use of the fluorophore is limited. The active site of the enzyme is only localized to an HPLC fraction, and amino acid loading cannot be observed. However, this does provide an efficient method for mapping NRPS active sites that can be further analyzed by mass spectrometry in order to localize the active site to a single residue by tandem mass spectrometry (7).

### *Why study NRPS?*

The diversity of nonribosomal peptides leads to a subsequent diversity in their biological utility including medicinal, agricultural, and biological applications. While these peptides display a wide variety of functions, their use in the field of medicine has shown promise, for example the use of penicillin to treat microbial infections. In fact, 75% of all antimicrobial agents and 50% of all commercial drugs are derived from secondary metabolites (12). Many nonribosomal peptides display antimicrobial, antiviral, antitumor, and immunosuppressive properties. Due to the emergence of drug-resistant microbes, viruses, and tumors, a large fraction of modern drugs are no longer effective treatment methods (13). The underlying modular biosynthesis of the peptides is incredibly useful for controlled manipulation of the final product, which can aid in the development of novel bioactive compounds, and several companies are currently taking this approach. For example, Kosan Biosciences, Ecopia Biosciences, and Biotica are all attempting to generate novel or modified bioactive compounds as potential drug treatments.

By understanding the biosynthetic mechanisms involved in synthesizing nonribosomal peptides, attempts can be made to mimic them in research or to modify them in order to alter and

improve their activities. The antibiotic clorobiocin, for example, is a nonribosomally produced antibiotic that could be medicinally valuable due to its ability to combat methicillin resistant *Staphylococcus aureus* (MRSA). However, its insolubility causes it to be toxic in humans (13). By understanding the biosynthetic mechanism involved, an opportunity arises to adjust the amino acids added and improve solubility without eliminating its antimicrobial properties. A 3-4 fold improvement in solubility would likely make this an effective drug in MRSA treatment (13).

#### *Possible Medical Science Advances from NRPS*

Due to the incredible medicinal importance of NRPS derived compounds, one of the ultimate goals of studying these systems is to understand how they work so that they may be modified to produce novel peptides. One method for achieving this goal is to rearrange the domains in an NRPS cluster in order to generate an altered peptide or to design one completely from scratch. There are four main ways to accomplish this. The first three involve fusion between domains of separate modules. Hybridization can occur between the C and A domains, the A and T domains, or T and C domains. Unfortunately, there is a loss of productivity when using domain hybridization, especially in C-A hybrids (14). Another method for generating alternative peptide products is to alter the specificity of A domains. This can be accomplished by mutating one or more of the ten anchor amino acids in the A domain, which will alter its substrate specificity. Using the “nonribosomal code”, it would even be possible to select a particular amino acid to substitute into the novel peptide product (14).

In addition, the moderate substrate specificity of A domains coupled with their ability to incorporate over 100 nonproteinogenic amino acids allows for alternative product formation without altering the NRPS modules themselves. This technique could be used to form products

with improved effects, such as improved antimicrobial properties, or to improve ancillary issues such as solubility (13). Without altering the modules themselves, it is feasible that this technique could be used without a significant loss in the productivity of the system, which is significant in pharmaceutical development.

One of the foremost difficulties in the development of novel bioactive compounds is that these systems are not understood well enough to utilize their full medicinal capabilities. New tools and methods need to be developed to study these systems. Mass spectrometry is one such method that has shown promise as an unbiased assay in these studies (13). Considering A domain specificity will likely be the most fruitful aspect of NRPS systems in the development of novel bioactive compounds, it is important to hone resources into this area of research. The following studies demonstrate the utility of mass spectrometry in studying NRPS systems, specifically the substrate specificity of the A domain.

## **Materials and Methods**

### **NRPS Basics**

#### *Preparation of NRPS Enzymes*

*E. coli* strain BL21(DE3) star containing plasmids encoding genes for the proteins Sfp ( $Am^R$ ), CloN4 ( $Am^R$ ), CloN5 ( $Am^R$ ), CouN4 ( $Am^R$ ), CouN5 ( $Km^R$ ), or EntB(ArCP) ( $Am^R$ ) was obtained from the Walsh lab at Harvard Medical School. The *E. coli* was grown in Luria-Bertani (LB) media at 37°C until the OD<sub>600</sub> reached 0.6. The cells were induced with 100 mg/L IPTG (Promega) and were overexpressed for 4-6 hours at 25°C. After harvesting the cells by centrifugation at 6000 rpm (Sorvall RC-5C Plus centrifuge with the SLA-3000 rotor), the cells were resuspended in a 50 mM phosphate buffer (pH 7.5) containing 100 mM NaCl and 10 mM

imidazole. The cells were lysed by sonication in the presence of lysozyme (Sigma-Aldrich), and the lysate was clarified by centrifugation at 16,000 rpm (SS-34 rotor) for 25 minutes. The clarified lysate was applied to a column containing Ni-NTA Superflow nickel affinity resin (Qiagen) and washed with a 50 mM phosphate buffer (pH 7.5) containing 100 mM NaCl and 20 mM imidazole. The column was eluted with a 50 mM phosphate buffer (pH 7.5) containing 100 mM NaCl and 250 mM imidazole. Fractions containing protein as determined by the BioRad assay were buffer exchanged into 50 mM Tris with 1 mM TCEP using a PD-10 gel filtration column (Amersham Biosciences). The purified proteins were stored in 10% glycerol at -80°C until used.

#### *Preparation of the holo PCP*

All of the acylation reactions of the PCP domains (e.g. CloN5) were carried out in a similar fashion. Approximately 25  $\mu$ M PCP domain was reacted in the presence of 3.6  $\mu$ M Sfp, 8 mM MgCl<sub>2</sub>, and 250  $\mu$ M coenzyme A (trilithium salt, Sigma-Alrich) in a total reaction volume of 100-500  $\mu$ L for one hour. Depending on the size of the PCP, a trypsin digestion was used in order to facilitate analysis by mass spectrometry (Tables 1 & 2). The proteins were digested with 85 units of sequencing trypsin (pH 8) for 5-10 minutes, followed by quenching with 50  $\mu$ L of 10% formic acid. One unit of trypsin is defined as the amount of sequencing grade modified trypsin (Promega) required to produce a  $\Delta A_{253}$  of 0.001/min at 30°C with the substrate  $\alpha$ -benzoyl-L-arginine ethyl ester. The PCP domains were purified via HPLC (HP1100, Jupiter 5  $\mu$ m C4 300 Å column from Phenomenex) using either a 30-minute or a 60-minute gradient (Table 3), collecting fractions at 1-minute intervals. The fractions were frozen at -80°C and lyophilized.

### *Fluorescent Active Site Mapping*

Holo PCPs were prepared similarly to the previous description, except with the use of BODIPY-FL-*N*-(2-aminoethyl)maleimidyl-S-CoA (BODIPY-CoA) in place of 250  $\mu$ M CoA. The enzymes were digested with trypsin and purified by HPLC, as done previously, collecting one fraction per minute. The resulting fractions were examined under UV light in order to determine which fractions fluoresced. These fractions were further analyzed by mass spectrometry in order to further characterize the active site. As an alternative to HPLC separation, an SDS-PAGE (100 V, 15% polyacrylimide gel, BioRad) separation was also used to verify loading of BODIPY-CoA onto a protein fragment approximately the same size as observed by mass spectrometry.

### *ESI-FTMS Analysis*

The lyophilized, HPLC-purified fractions were resuspended in 100  $\mu$ L of 78% ACN, 0.1% acetic acid or 49% methanol, and 1% formic acid and analyzed by a custom-built 8.5 T ESI-FTMS equipped with a front-end quadrupole (6). The samples were introduced to the FTMS via a NanoMate 100 automated nanospray (Advion Biosciences). The ions were allowed 500 ms per scan to accumulate, and 50-200 scans were acquired per spectrum. The FTMS was externally calibrated with commercial ubiquitin (Sigma-Aldrich) to a monoisotopic  $M_r$  of 8560.65 Da. The MIDAS data station was used to collect the data and calculate the protein masses (15). THRASH was used in conjunction with manual deconvolution to produce a mass peak list of all the observed ions (16).

## Substrate Identification

### *Preparation of E. coli Metabolomes*

*E. coli* was grown in 150ml LB at 37°C until the OD<sub>600</sub> reached 0.8. The cells were harvested by centrifugation at 6000 rpm (SLA-3000 rotor), and the pellet was resuspended in 2.5 ml of 25 mM Tris (pH 7.6). The cells were lysed by sonication in the presence of lysozyme and the lysate was clarified by centrifugation at 16,000 rpm (SS-34 rotor) for 25 minutes. This crude extract was gel filtered in a PD-10 column by adding 1 ml of clarified lysate to a column equilibrated with 25 ml of 25 mM Tris (pH 7.6). The proteins from the lysate were washed away with 5 ml of 25 mM Tris (pH 7.6), and the small molecules were eluted by adding another 3 ml of 25 mM Tris. These filtered crude extracts were frozen at -80°C, lyophilized, and resuspended in 50-100 µL of 50 mM Tris (pH 7.6) per mL of extract.

### *High-Throughput Substrate Screens*

The holo PCP domains were reacted for 30 minutes in the presence of their corresponding A domain (e.g. CloN4) at a concentration of 2 µM, 4 mM ATP, and 10-50 µL of an *E. coli* metabolome. The reaction was terminated by quenching 1:1 (v/v) with 10% formic acid. Each system was examined in the presence of each of five *E. coli* metabolomes and, in the case of EntB, an extra metabolome grown in the iron-limited conditions. Alternatively to the use of *E. coli* metabolomes, a mixture of 21 proteinogenic amino acids in concentrations of 0.5-1.0 mM was also screened, and an algal hydrolysate mixture was used at 0.2-0.5 mg/ml. See Table 1 for a complete list of substrate screens performed.

### *Active Site Mapping of Orphan Gene Clusters*

The holo forms of the PksJ AT1, PksJ AT2, PksN, and Atu3673 proteins were generated, digested, HPLC-purified, and analyzed by ESI-FTMS as previously described. Additionally, BODIPY-CoA was used to map the active sites as previously described. The HPLC spectra at 220 nm and 400 nm were compared to predict which fractions contained the active site. In order to visualize the active sites by FTMS, the peak list generated by THRASH for each fluorescent HPLC fraction was imported into the PAWS software. This software allowed for the determination of the tryptic fragment containing the active site with BODIPY-CoA bound. Once a fragment was determined to cover the predicted active site, they were further analyzed by MS/MS using the OCAD and IRMPD fragmentation techniques. This allowed for the further localization of the active site serine residue.

### **Amine Transfer**

#### *A4N7 Purification*

*E. coli* strain BL21(DE3) star containing a plasmid encoding the gene for the protein A4N7 (Km<sup>R</sup>) was obtained from the Walsh lab at Harvard Medical School. The *E. coli* was grown in LB media at 37°C until the OD<sub>600</sub> reached 0.4. They cultures were then temperature-shifted to 15°C for one hour. The cells were induced with 100 mg/L IPTG and were overexpressed for 18-24 hours at 15°C. After harvesting the cells by centrifugation at 6000 rpm (SLA-3000 rotor), the cells were resuspended in a 25 mM Tris lysis buffer (pH 8) containing 400 mM NaCl, 10% glycerol, and 10 mM imidazole. The cells were lysed by sonication in the presence of lysozyme, and the lysate was clarified by centrifugation at 16,000 rpm (SS-34 rotor) for 60 minutes. The clarified lysate was applied to a column containing NTA Superflow nickel

affinity resin and washed with a 25 mM Tris buffer (pH 8) containing 400 mM NaCl, 10% glycerol, and 20 mM imidazole. The column was eluted with a 25 mM Tris buffer (pH 8) containing 400 mM NaCl, 10% glycerol, and 250 mM imidazole. Fractions containing protein were desalted into lysis buffer using a PD-10 gel filtration column. The purified proteins were stored at -80°C until used.

#### *Visualization of Amine Transfer*

MycA4N7 (185 µg) was acylated in a similar fashion to that previously described, using 466 µM Acetoacetyl-CoA (Sigma-Aldrich) in place of CoA. The active site was mapped in a similar fashion to PksN and PksJ, including digestion, HPLC purification, and analysis by ESI-FTMS except the BODIPY-CoA fluorescent technique was not used. The ACP<sub>2</sub> active site was observed in the 37<sup>th</sup> minute, and the PCP<sub>1</sub> active site eluted in the 32<sup>nd</sup> minute. To visualize the amine transfer, the holo PCP was reacted with a mixture of all proteinogenic amino acids (1.33 mM final concentration of each) for 30 minutes. Some precipitation was observed during this reaction. The reaction was then analyzed by mass spectrometry as described previously. The ACP<sub>2</sub> active site was observed in the 36<sup>th</sup> minute, and the PCP<sub>1</sub> active site eluted in the 31<sup>st</sup> minute.

#### *Determination of the Amine Donor*

Once the 1 Da mass shift due to amine transfer was visualized by mass spectrometry, a high-throughput screen was assembled to determine the precise amine donor. The HPLC spectra were used to monitor the amine transfer by observing a peak shift from the 37<sup>th</sup> to the 36<sup>th</sup> minute and the 32<sup>nd</sup> to the 31<sup>st</sup> minute. The donor screen was done in two steps. First, five groups of 3-

4 amino acids (2.0 mM concentration) were incubated with acetoacetyl-S-MycA4N7 for 30 minutes. The samples were analyzed for an HPLC peak shift. All of the groups displayed amine transfer, but the two best groups contained Ala, Glu, Arg, Gln, Met, Tyr, Ser, and Trp. These eight amino acids were then screened separately, which revealed Gln as the most efficient amine donor. Because glutamine contains two amine groups, it was necessary to determine whether the transferred amine came from the amide side chain or the alpha-amino position. Amino acid solutions with the two corresponding  $^{15}\text{N}$ -labeled amine groups were prepared and screened separately (2.0 mM concentration).

#### *Time Course of Amine Transfer*

Since a 2 Da mass increase was observed in both the ACP<sub>2</sub> and PCP<sub>1</sub> active sites, the relative time course of amine transfer to each site was examined. This was done by preparing a large scale reaction (1.6 ml total volume) of acetoacetyl-S-MycA4N7 in the same manner as done previously. Upon adding 2 mM  $^{15}\text{N}$ -Gln, portions at various time points (0.33, 0.66, 1, 2, 4, 8, 16, and 32 min) were removed and immediately digested with trypsin and quenched with 10% formic acid as done previously. The time points were then HPLC-purified and analyzed by ESI-FTMS.

## **Results and Discussion**

#### *Substrate Identification*

Since the A domain of NikP1 was found to be selective for its natural substrate L-histidine in the presence of a complex pool of amino acids (13), additional NRPS systems were screened for substrate loading selectivity in order to determine if this could be developed as a

general method of high-throughput screening (Figure 2). After generating the phosphopantetheinylated form of the PCPs CloN5 and CouN5, the domains were incubated with their respective A domain (CloN4/CouN4), ATP, and a substrate pool consisting of 19 proteinogenic L-amino acids, glycine, L-selenocysteine, and 4-*trans*-hydroxy-L-proline. Upon quenching the reaction with acid, HPLC purification, and analysis by ESI-FTMS, a mass shift of 96.8 Da was observed from that of the holo form with CloN5 and a 97.0 Da increase was seen with CouN5 (Figure 3A,B,F). This agreed with a mass addition of that of L-proline, which has a calculated mass addition of 97 Da upon loading (Table 1).

The analysis was repeated with a substrate pool of the same composition except lacking L-proline, and a mass shift of 112.9 Da was observed. This agreed with a mass addition of 4-*trans*-hydroxy-L-proline (113.0 Da), which demonstrated that the CloN4 and CouN4 A domains are selective for L-proline, but they can also activate alternative substrates (Figure 4). The masses of these substrates can then be determined by FTMS. The alternative substrate 4-*trans*-hydroxy-L-proline was not known to load previously, but its loading was verified as a viable substrate using the classic radioactive pyrophosphate exchange assay commonly used to detect activation in NRPS systems (13). This type of alternative substrate loading is desirable in some instances, such as to generate a more soluble aminocoumarin antibiotics, which are the natural products of the clorobiocin and coumermycin systems. If the alternative substrate can be loaded *in vivo*, its additional hydroxyl group may be incorporated into a final aminocoumerin product with greatly improved solubility. This suggests a possible feeding study to determine if this effect can be observed *in vivo*. If this effect is still not great enough to preclude toxicity in humans, other modules in the system could be probed in a similar fashion to determine if they can load alternative substrates.

In an additional study, the CloN4 and CouN4 adenylation domains were screened using two undefined substrate pools, an *E. coli* metabolome and a commercially available algal hydrolysate. Both of these screens yielded the mass increase equivalent to the addition L-proline onto the respective holo PCP (Figure 3C,G). These results demonstrate that when the substrate L-proline is present, it is the primary species activated by the A domain. In the absence of L-proline, however, an alternative substrate 4-*trans*-hydroxy-L-proline can be activated by the domain as well.

The enterobactin system, consisting of the carrier protein EntB (ArCP) and the A domain EntE, was studied with a similar substrate screen. When the natural substrate 2,3-dihydroxybenzoic acid was present in the screen, it was the only observed product (+136.2 Da) (Figure 3D,E). However, in the absence of the natural substrate, an alternative substrate (+119.0 Da) loaded. This was seen in a substrate screen consisting of 19 proteinogenic L-amino acids, glycine, L-selenocysteine, and 4-*trans*-hydroxy-L-proline, but the mass increase could not be attributed to one of the 98-99% pure amino acids in the mixture. The most likely explanation for this is that an alternative substrate was present in the 1-2% impurity of the amino acid mixture. This also agrees with the fact that there was incomplete loading of this 119 Da species onto the PCP, and thus it remained partially in the holo form. The EntB/EntE loading system was also analyzed with two different *E. coli* metabolomes. One was generated from *E. coli* grown in standard conditions while another was generated in iron-limiting conditions. No loading was observed on the PCP domain in the presence of either metabolome. This suggests that either the substrate was not present in the metabolomes or that it was not present in a high enough concentration to observe loading by FTMS.

The method was developed by first examining the proof of concept experiments. In the CloN5, CouN5, and EntB systems, the natural substrates were already known. By combining the work done on these systems with work done on the systems of NikP1, HMWP2, JamC/JamA, and PchE/PchD, it is clear that as long as the natural substrate of the system was present, it would be loaded as the only major product and noncognate substrate loading was negligible (13). This was even shown to be true in the presence of undefined substrate pools such as the algal hydrolysate and the *E. coli* metabolome.

After verifying that these systems would load their known natural substrate even from a crude reaction mixture, the concept was applied to A domains in which the substrate is not known. Since new NRPS modules are constantly being discovered, this method provides an efficient and effective assay to determine their native substrate specificity. To demonstrate the benefits of this assay, three uncharacterized NRPS modules from a *Bacillus subtilis* orphan gene cluster were analyzed to determine their natural substrate. Since the final product of this NRPS/PKS hybrid cluster was unknown, bioinformatic techniques were used to hypothesize what the domain might load. According to this prediction, PksJ AT2 should have loaded glycine and PksN should have loaded cysteine or possibly serine. PksJ AT1 could not be reliably predicted, but it showed homology to a microcystin biosynthetic enzyme that loads a phenyl propionate, such as phenylalanine or phenylacetate (17). Additionally, there had been some speculation as to this gene cluster's involvement in the biosynthesis of difficidin, but the lack of amino acids in its final structure and the lack of experimental evidence made this hypothesis unlikely (Figure 1A). Furthermore, it was possible that the NRPS modules in this cluster were not active, so an activity screen was necessary to establish whether the modules could load a substrate.

The PksN A/T didomain was first analyzed by fluorescent active site mapping and ESI-FTMS in order to map its active site (Figure 5). A fragment eluting at the 38th minute by HPLC was seen to fluoresce. The fraction was analyzed by FTMS in order to verify that it contained the active site of the PCP with the mass addition of BODIPY-CoA (Figure 6). The reaction was completed again using unlabeled CoA in order to verify the elution time of the fragment, and a mass increase of 340 Da was localized to the active site serine by FTMS, consistent with the addition of 4'-PP. SWIFT was used to isolate peaks containing the active site fragments among contaminants or other fragments not containing the active site. The PksN didomain was then analyzed in a substrate screen using an algal hydrolysate, which acted as an unbiased substrate mixture. After phosphopantetheinylating the apo PksN and incubating it in the presence of the hydrolysate and ATP, analysis by FTMS revealed a mass increase of 71.1 Da (Figure 6C). This is consistent with the addition of L-alanine onto the holo PCP. The mass addition was further verified by MS/MS, and it was localized to the active site serine. Since serine or cysteine was predicted to be the natural substrate of the domain, PksN was also analyzed in the presence of a mixture of serine, cysteine, alanine, and threonine. A mass increase of 71 Da was still observed (Figure 6D). In the presence of only serine and threonine, a mass shift of 87 Da was observed, consistent with serine loading (Figure 6F). In the presence of cysteine alone, no mass increase from that of the holo PCP was observed. When alanine is present as a substrate, it out-competes the alternative substrate serine. The predicted substrate cysteine was never seen to load, even in the absence of other competitors. This not only provides evidence for the imprecision of bioinformatics predictions, but it also further demonstrates the ability of NRPS modules to load alternative substrates.

Two additional didomains, PksJ AT1 and AT2, were analyzed in a similar fashion (Figures 5 & 7). The active sites were mapped using BODIPY-CoA and were found to elute in the 36th (holo AT1) and 26<sup>th</sup>-27<sup>th</sup> (holo AT2) minute by HPLC, and this was verified mass spectrometry by a mass increase of 754.3 Da from that of the apo form (Figure 7A,B). The holo PCPs were generated and screened in the presence of an algal hydrolysate. A mass increase of 118.04 Da was observed in the PksJ AT1 didomain, which indicated the addition of phenylacetate, and MS/MS was used to localize the addition to the module's active site (Figure 7C). This was confirmed by analysis in the presence of phenylacetate alone. This domain did not load phenylpropionate, phenyllactate or any amino acids, including phenylalanine. In the case of the PksJ AT2 didomain, a mass increase of 57.3 Da was observed in the presence of the hydrolysate. This indicated the addition of glycine to the module, which was further localized to the active site using tandem MS. Furthermore, screening with glycine alone confirmed the ability of PksJ AT2 to load this substrate (Figure 7F,G).

These findings provide two important observations concerning the NRPS modules of the *B. subtilis* orphan gene cluster. Since loading was observed in the systems, it provides evidence that the modules are, in fact, active in the gene cluster. It also suggests that these modules contribute to an unknown natural product and not difficidin, as speculated, because difficidin lacks amino acids in its final structure. It cannot be confirmed that these modules are not involved in the formation of a modified difficidin product or another natural product such as hydroxyl-mycotrienin A (BMJ958-62F4), which contains amino acids. Additionally, these results demonstrate an important application of the substrate screening method being developed. It is possible to determine the mass of an unknown substrate upon loading.

Another uncharacterized NRPS module, Atu3673, from the organism *Agrobacterium tumefaciens* is involved in the biosynthesis of an unknown siderophore. In order to learn more about the structure of the final siderophore product, this module was also examined using our method. Initially, SDS-PAGE was used instead of the traditional HPLC because no fluorescence or noticeable differences in the spectra were observed. The SDS-PAGE demonstrated that a tryptic fragment was being loaded with the fluorophore (Figure 8). Further analysis of an HPLC fraction by FTMS did verify incorporation of BODIPY-CoA by a 754.3 Da mass increase compared to the apo form. Analysis of this module in the presence of complex mixtures, such as the algal hydrolysate and mixtures of amino acids, did not yield clear results as to the identity of the substrate. Ongoing studies may reveal the substrate of this domain.

#### *Amine Transfer*

Since MycA4N7 represents a hybrid between a nonribosomal peptide synthetase (NRPS) and a polyketide synthetase (PKS), it contains two carrier protein active sites, the usual PCP and the polyketide acyl carrier protein (ACP). In order to map the two active sites, a slightly different technique needed to be used. Since a truncated version of the protein was used, the true  $\beta$ -keto acid could not be generated *in vitro*, an alternative substrate for the 4'-PP transferase was used to activate the carrier proteins: acetoacetyl-CoA. This was used because it mimics the natural substrate of the ACP active site, a ketide on a long fatty acid chain (Figure 9A). Upon acylating the ACP and PCP, the active sites were mapped by subsequently digesting with trypsin and separating the fragments by HPLC as was done previously. Since the BODIPY-CoA fluorophore was not used, changes in the HPLC spectra between the apo and holo forms of the enzyme were analyzed to predict which fragments contained the acylation. Analysis by mass

spectrometry revealed that the ACP active site eluted in the 37<sup>th</sup> minute and the PCP active site eluted in the 32<sup>nd</sup> minute. Tandem MS was used to localize the 424 Da mass increase to each active site serine (Figure 10).

An amine transfer (AMT) domain is found between the ACP and PCP in MycA4N7, suggesting that an amine transfer takes place to replace the  $\beta$ -keto group of acetoacetate through a pyridoxal 5'-phosphate (PLP) dependant reaction (Figure 9B). It was unknown whether the four-domain construct was active, however. As an application of our method of high-throughput substrate screens, a mixture of 21 amino acids was added to acetoacetyl-S-MycA4N7, and the products were digested and purified by HPLC. A shift in the peak from the ACP active site occurred from the 37<sup>th</sup> minute to the 36<sup>th</sup> minute (Figure 12A), and an analogous shift occurred in the PCP peak from the 32<sup>nd</sup> minute to the 31<sup>st</sup>. Analysis by FTMS confirmed that the amine transfer and thus the mass increase of 1 Da was observed on both the ACP and PCP products. ECD fragmentation was used to further demonstrate that the mass addition was localized to the active site serine, and analysis of the c56 and c57 fragment ions display the 1 Da shift unambiguously (Figure 11).

In order to determine the precise amine donor, subsets of amino acids were screened until the most efficient amine donor was found to be glutamine. It gave approximately twice the amount of amine transfer compared to its nearest competitors, alanine and methionine, as determined by HPLC (Figure 12B,C). HPLC proved to be the more reliable assay in these screens due to the overt peak shift that occurred upon amine transfer and the fact that partial transfer could be observed. The determination of glutamine as the amine donor was further solidified by a UV-Vis assay (18). Since glutamine contains both an alpha and amide amine group, further analysis was necessary to determine the precise amine being transferred. The <sup>15</sup>N-

labeled amine for each position was examined so that a 2 Da mass increase could be observed compared to the usual 1 Da increase. FTMS confirmed only the  $\alpha$ - $^{15}\text{N}$ -amine group gave rise to a 2 Da mass increase from the acylated forms of the ACP and PCP (Figure 13).

As a further mechanistic study, the kinetics of the amine transfer were examined using  $\alpha$ - $^{15}\text{N}$ -amino glutamine in a time course reaction. Examination of the timing of the amine transfer would help to determine whether transfer was taking place on the ACP or the PCP. Analysis of the HPLC peaks and FTMS results revealed that by 32 minutes the ACP site contained nearly 50% loading with the  $^{15}\text{N}$ -amine while the PCP had an estimated amine incorporation of less than 10% (Figure 14). Longer incubation times did not increase the amount of conversion, suggesting that the reaction is reversible. The reversibility was confirmed by observing the conversion of  $\beta$ -aminobutyryl-S-ACP to acetoacetyl-S-ACP (18). The small amount of transfer seen on the PCP is likely due to non-enzymatic hydrolysis of acetoacetyl-S-PCP and the subsequent transfer of the  $\beta$ -aminobutyrate to the PCP.

All of these studies, but most significantly those of MycA, demonstrate the utility of FTMS in combination with high-throughput substrate screens. FTMS is currently the most accurate method to observe changes in masses of protein domains. The instrument used in these experiments was custom-built and has a mass accuracy of 5-25 ppm while commercial instruments can have mass accuracies up to 2 ppm. This type of high accuracy is generally necessary to observe substrates tethered to carrier domains such as those presented here. The increasing popularity of FTMS is creating more opportunities for this method to be used. Additionally, the high power of FTMS allowed for the visualization of a 1 Da mass difference in the MycA studies. This incredible resolution demonstrates the tremendous power that FTMS can

contribute to studies of this nature. The precise amine donor would probably not have been determined as efficiently using alternative techniques, and it may not have been possible at all.

## Conclusions

By combining high-throughput substrate screens with the high resolution of FTMS, our technique provides an important tool in the characterization and development of novel bioactive compounds. The substrate specificity of the systems allows for the detection of the natural substrate loading from a complex mixture of components. As long as the natural substrate was present in the mixture, it was seen to load as the only major product. Additionally, in the absence of the natural substrate the technique demonstrates the ability of some systems to load additional substrates. This could prove to be extremely beneficial in generating alternative bioactive compounds with adjusted properties, such as increased solubility. These properties are of vital importance to the pharmaceutical industry.

Thus, the use of high-throughput substrate screens in conjunction with FTMS and fluorescent active site mapping represents an extremely effective combination of techniques that now provides a method of highly accurate and efficient data acquisition. While the main evidence of this technique is demonstrated in NRPS systems here, it is feasible to use them in the analogous PKS and fatty acid synthase systems as well as other systems that involve covalent tethering of substrates to proteins. The technique provides a prospect for future pharmaceutical development of “unnatural” products from these systems that exhibit additional benefits not observed in their natural products. In addition, simply using the technique to study the biosynthetic mechanisms of these systems provides deep insight into how natural products are formed. Studies in adenylation domain specificity and biosynthetic mechanisms in combination

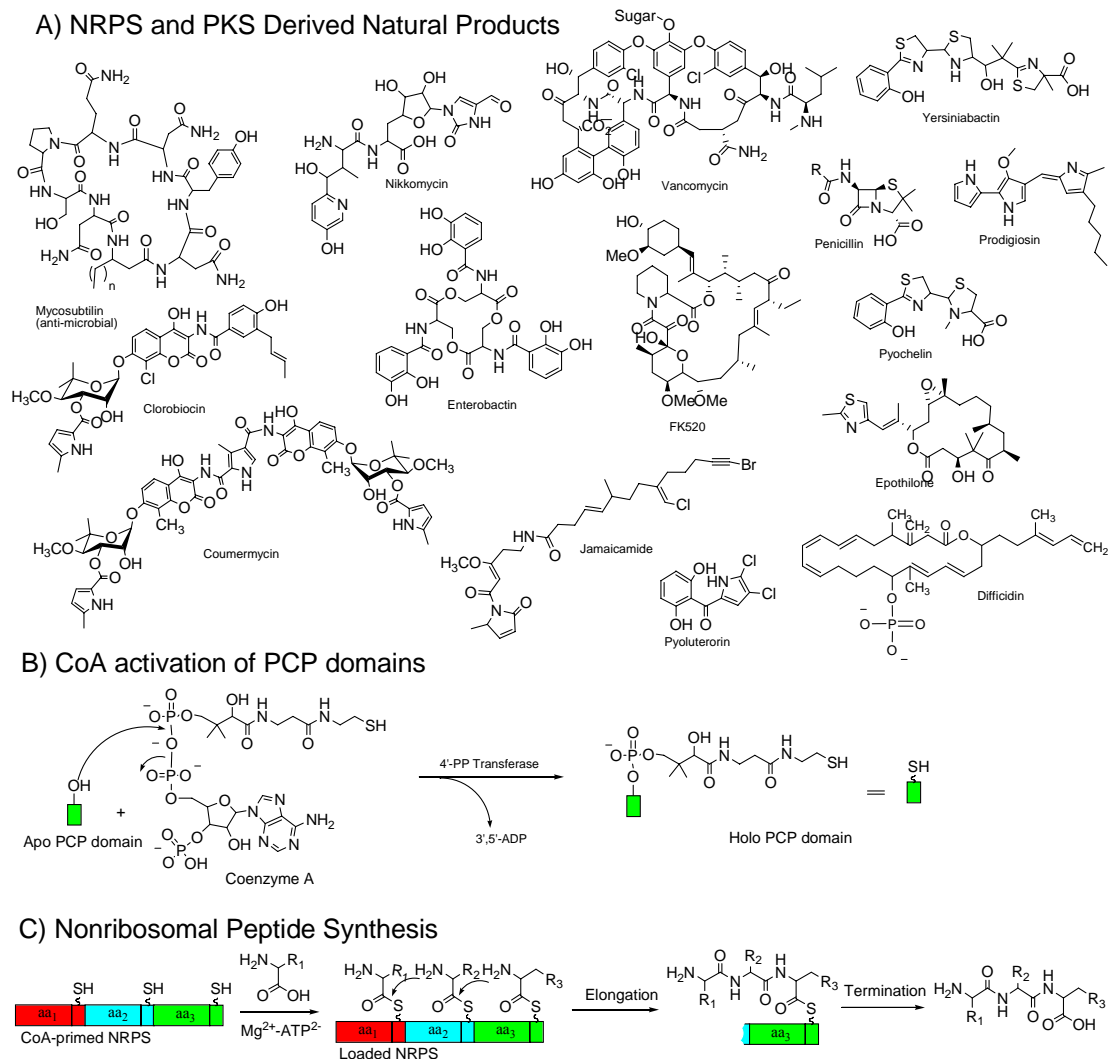
with genetic engineering could present a final outcome of these systems: a “homemade” NRPS gene cluster with its own “unnatural” bioactive product. A deeper understanding of these systems is necessary before this type of product manufacturing can occur in a routine fashion, but our method makes this process easier, faster, and more accurate than ever before!

## Bibliography

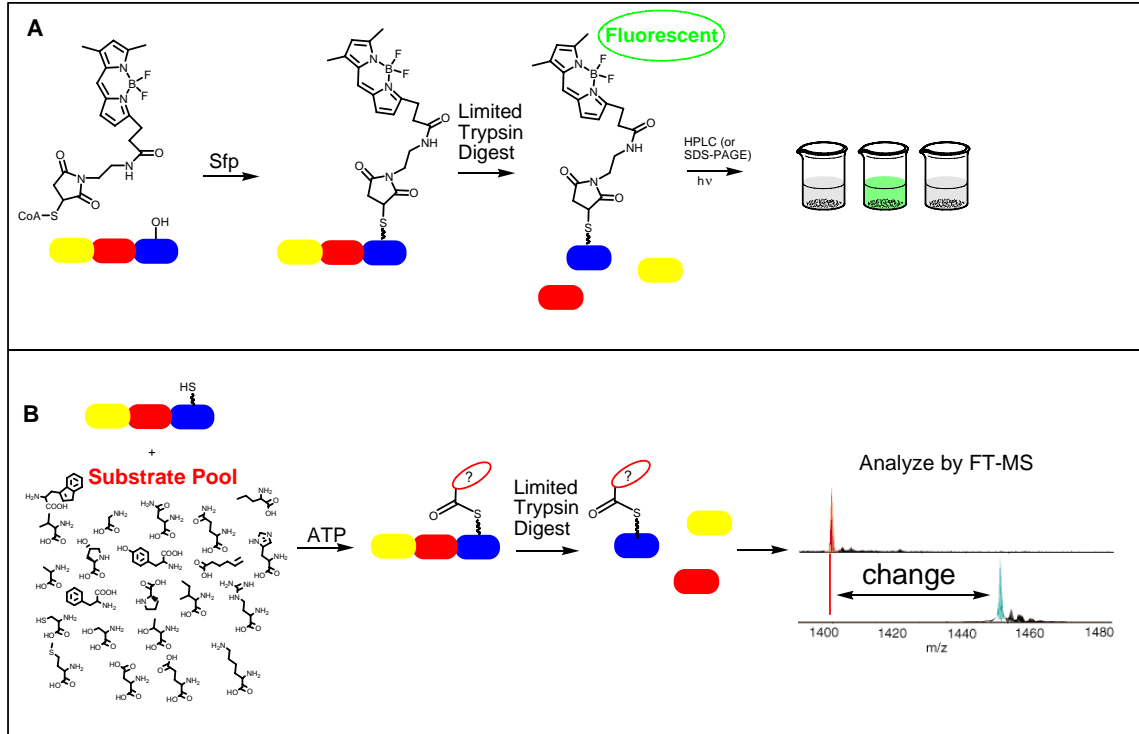
1. Marahiel, M. A., Stachelhaus, T., and Mootz, H. D. (1997) *Chem. Rev.*, **97**, 2651-2673
2. Wiest, A., Grzegorski, D., Xu, B.-W., Goulard, C., Rebuffat, S., Ebbole, D. J., Bodo, B., and Kenerley, C. (2002) *J. Biol. Chem.* **277**, 20862–20868
3. Fenn, J. B. (2003) *Angew. Chem. Int. Ed. Engl.* **42(33)**, 3871-3894
4. Tanaka, K. (2003) *Angew. Chem. Int. Ed. Engl.* **42(33)**, 3860-3870
5. Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998) *Mass Spectrom. Rev.* **17(1)**, 1-35
6. Patrie, S. M., Charlebois, J. P., Whipple, D., Kelleher, N. L., Hendrickson, C. L., Quinn, J. P., Marshall, A. G., and Mukhopadhyay, B. (2004) *J. Am. Soc. Mass Spectrom.* **15**, 1099-1108
7. Mazur, M. T., Walsh, C. T., and Kelleher, N. L. (2003) *Biochemistry.* **42(46)**, 13393-13400
8. Hicks, L. M., O'Connor, S. E., Mazur, M. T., Walsh, C. T., and Kelleher, N. L. (2004) *Chem. Biol.* **11(3)**, 327-335
9. Garneu, S., Dorrestein, P. C., Kelleher, N. L., and Walsh, C. T. (2005) *Biochemistry.* **44(8)**, 2770-2780

10. LeDuc, R. D., Taylor, G. K., Kim, Y.-B., Januszyk, T. E., Bynum, L. H., Sola, J. V., Garavelli, J. S., and Kelleher, N. L. (2004) *Nucleic Acids Res.* **32**, W340-W345
11. McLoughlin, S. M., Mazur, M. T., Miller, L. M., Yin, J., Liu, F., Walsh, C. T. and, Kelleher, N. L. (2005) *Biochemistry.* **44(43)**, 14159-14169
12. Paterson, I. and Anderson, E. A. (2005) *Science.* **310(5747)**, 451-453
13. Dorrestein, P. C., Blackhall, J. R., Straight, P. D., Fischbach, M. A., Garneau-Tsodikova, S., Edwards, D. J., McLoughlin, S., Lin, M., Gerwick, W. H., Kolter, R., Walsh, C. T., and Kelleher, N. L. (2006) *Biochemistry.* **45(6)**, 1537-1546
14. Schwarzer, D., Finking, R., and Marahiel, M. A. (2003) *Nat. Prod. Rep.* **(20)3**, 275-287
15. Senko, M. W., Canterbury, J. D., Guan, S., and Marshall, A. G. (1996) *Rapid Commun. Mass Spectrom.* **10**, 1839-1844
16. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) *J. Am. Soc. Mass Spectrom.* **11(4)**, 320-332
17. Hicks, L. M., Moffitt, M. C., Beer, L. L., Moore, B. S., and Kelleher, N. L. (2006) *Chem. Biol.* **1(2)**, 93-102
18. Aron, Z. D., Dorrestein, P. C., Blackhall, J. R., Kelleher, N. L., and Walsh, C. T. (2005) *J. Am. Chem. Soc.* **127(43)**, 14986-14987

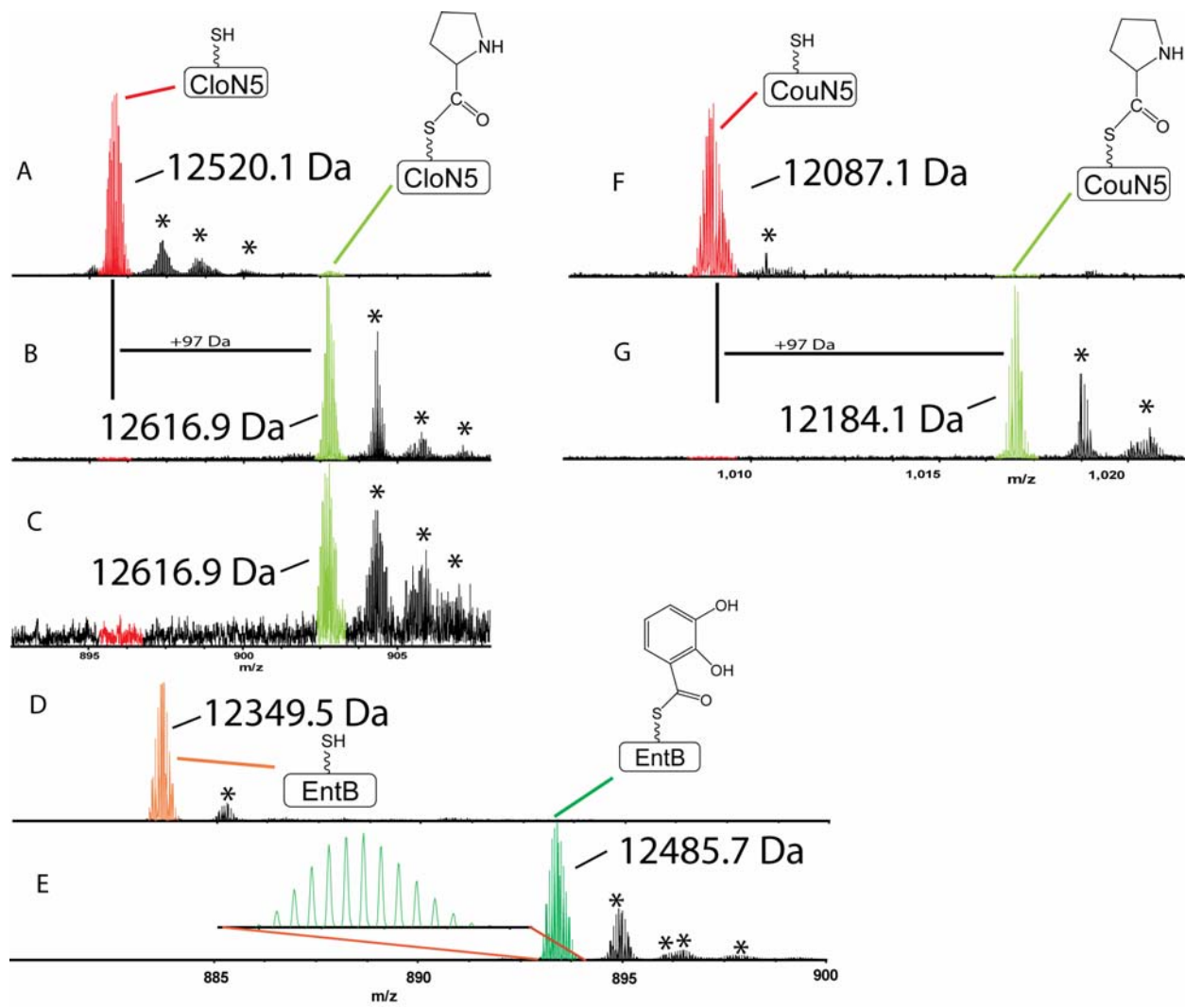
## Figures and Tables



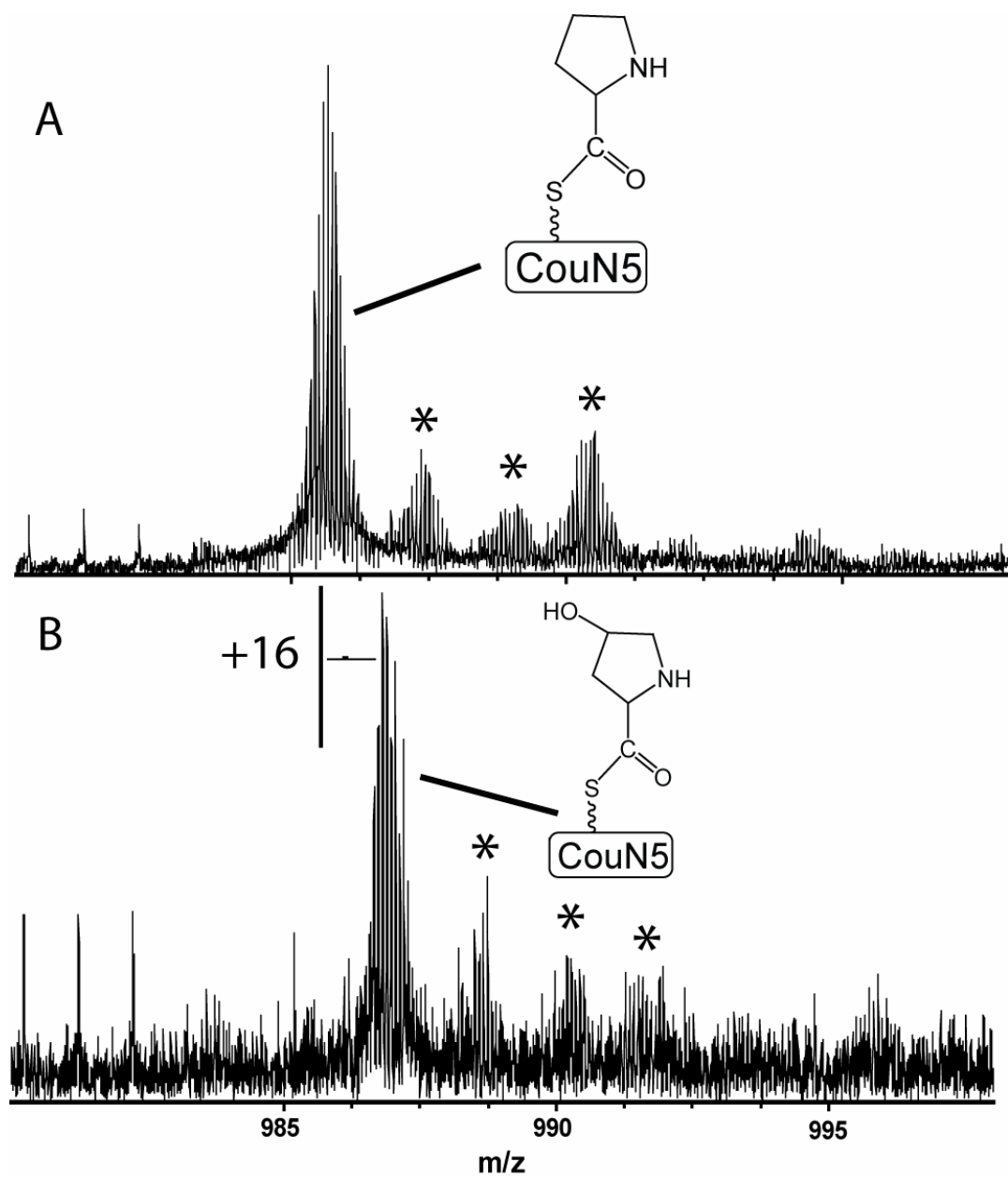
**Figure 1: Basic Features of Nonribosomal Peptide Synthases.** (A) Some natural products of NRPS systems, including many of the systems in this study. (B) The mechanism of converting the apo PCP domain to its holo form by a 4'-phosphopantetheinyl transferase. (C) The mechanism of elongation by NRPS systems, including amino acid identification by the A domain and tethering to the primed PCP, the elongation reaction, and termination by a thioesterase.



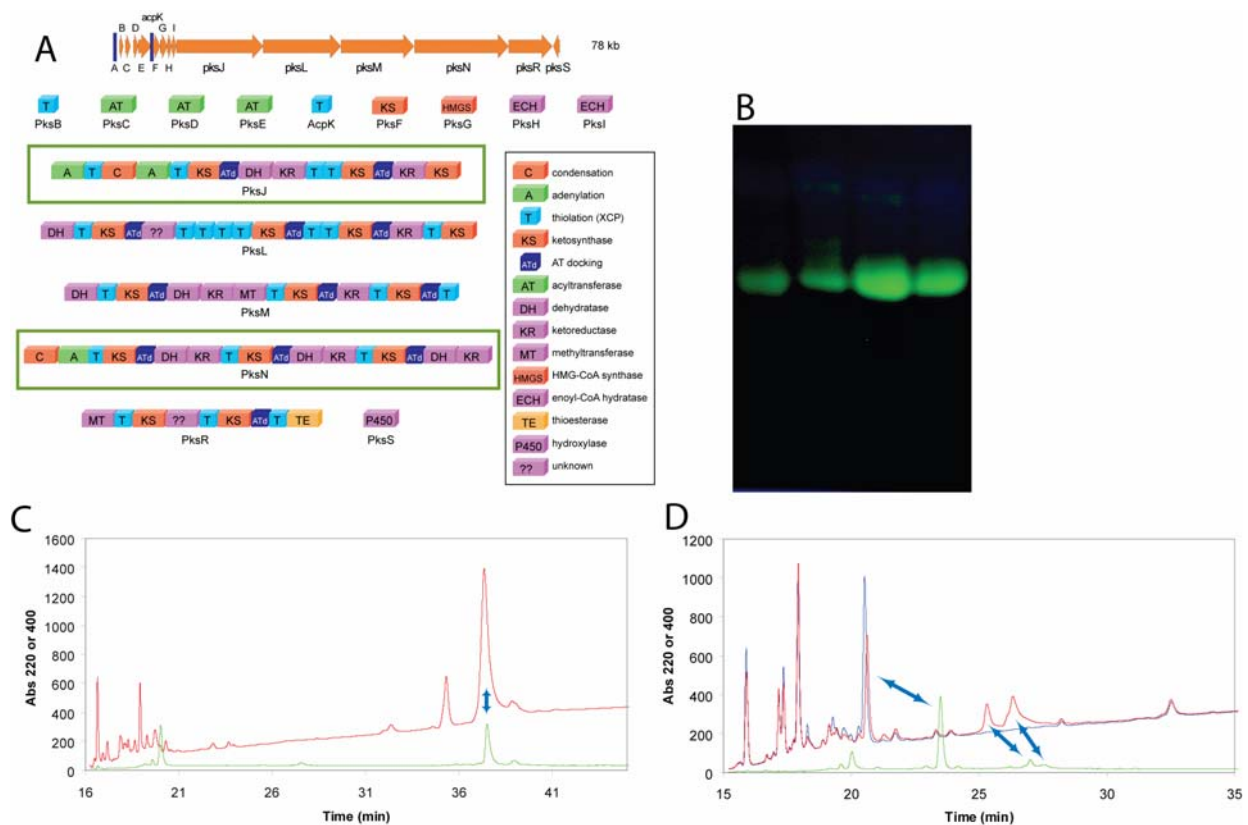
**Figure 2: Outline of the General Approach for Determining Structure-Activity Relationships by Mass Spectrometry.** (A) Fluorescent labeling of the PCP using BODIPY-FL-*N*-(2-aminoethyl)maleimidyI-S-CoA provides an efficient technique for determining which fragment contains the PCP active site. (B) Once the active site has been mapped, the A domain loads an amino acid from a complex substrate mixture to the PCP. The mass addition can be determined by FTMS.



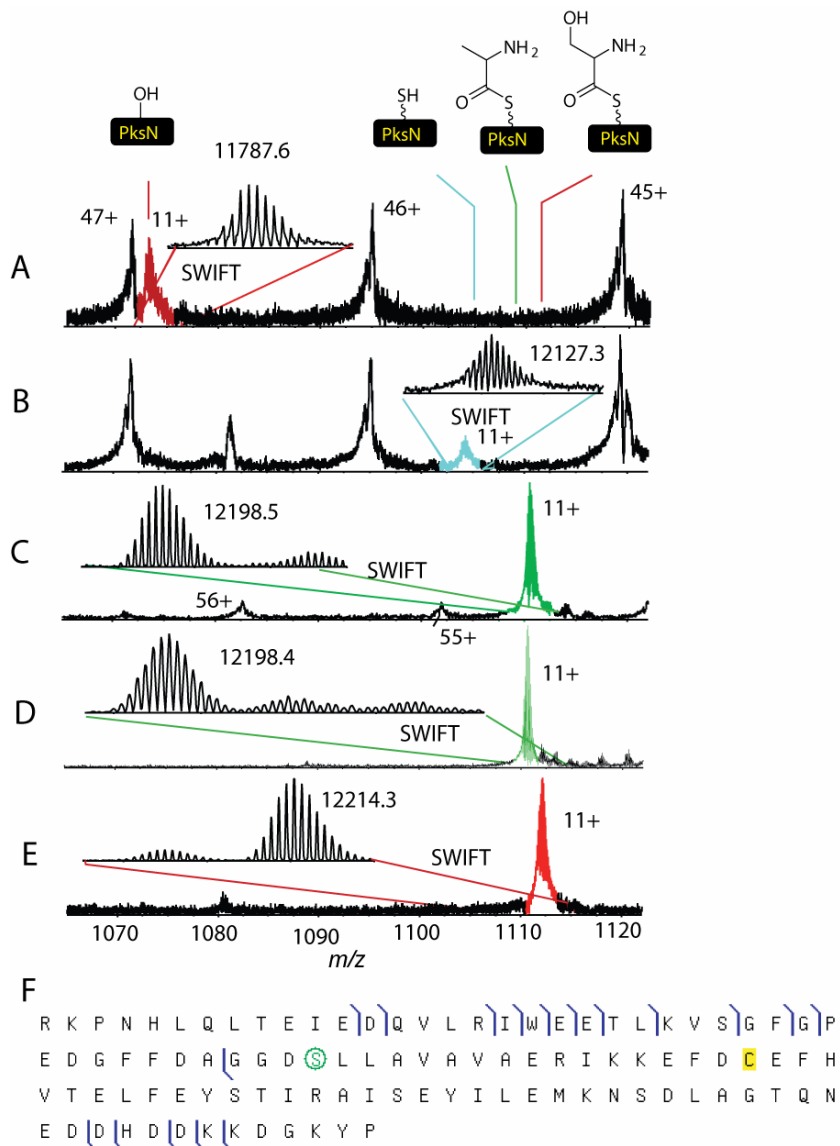
**Figure 3: Loading of CloN5, CouN5, and EntB with Their Natural Substrates.** (A) Holo CloN5. (B) CloN5 after incubation with 21 amino acids. (C) CloN5 after incubation with an *E. coli* metabolome. (D) Holo EntB. (E) EntB in the presence of 21 amino acids and 2,3-dihydroxybenzoic acid. (F) Holo CouN5. (G) CouN5 in the presence of an *E. coli* metabolome. The peaks with asterisks (\*) over them represent noncovalent adducts of metal ion such as Na<sup>+</sup> and K<sup>+</sup>.



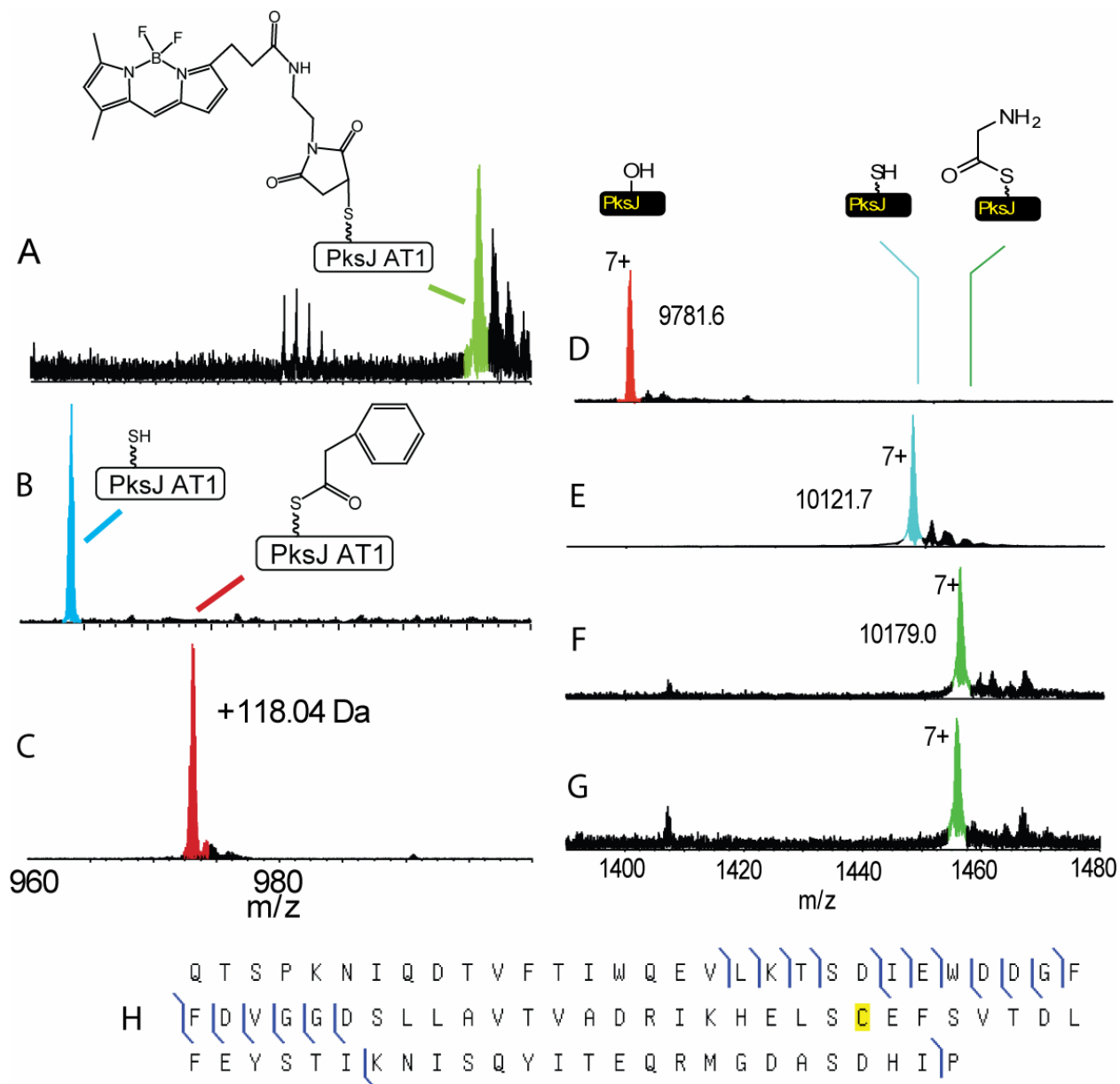
**Figure 4: Alternative Substrate Loading onto CouN5.** Loading of the natural substrate L-proline (A) compared to the alternative substrate 4-*trans*-hydroxy-L-proline in the absence of L-proline (B) can be visualized by a mass difference of 16 Da using FTMS. The asterisks (\*) represent noncovalent adducts of metal ions.



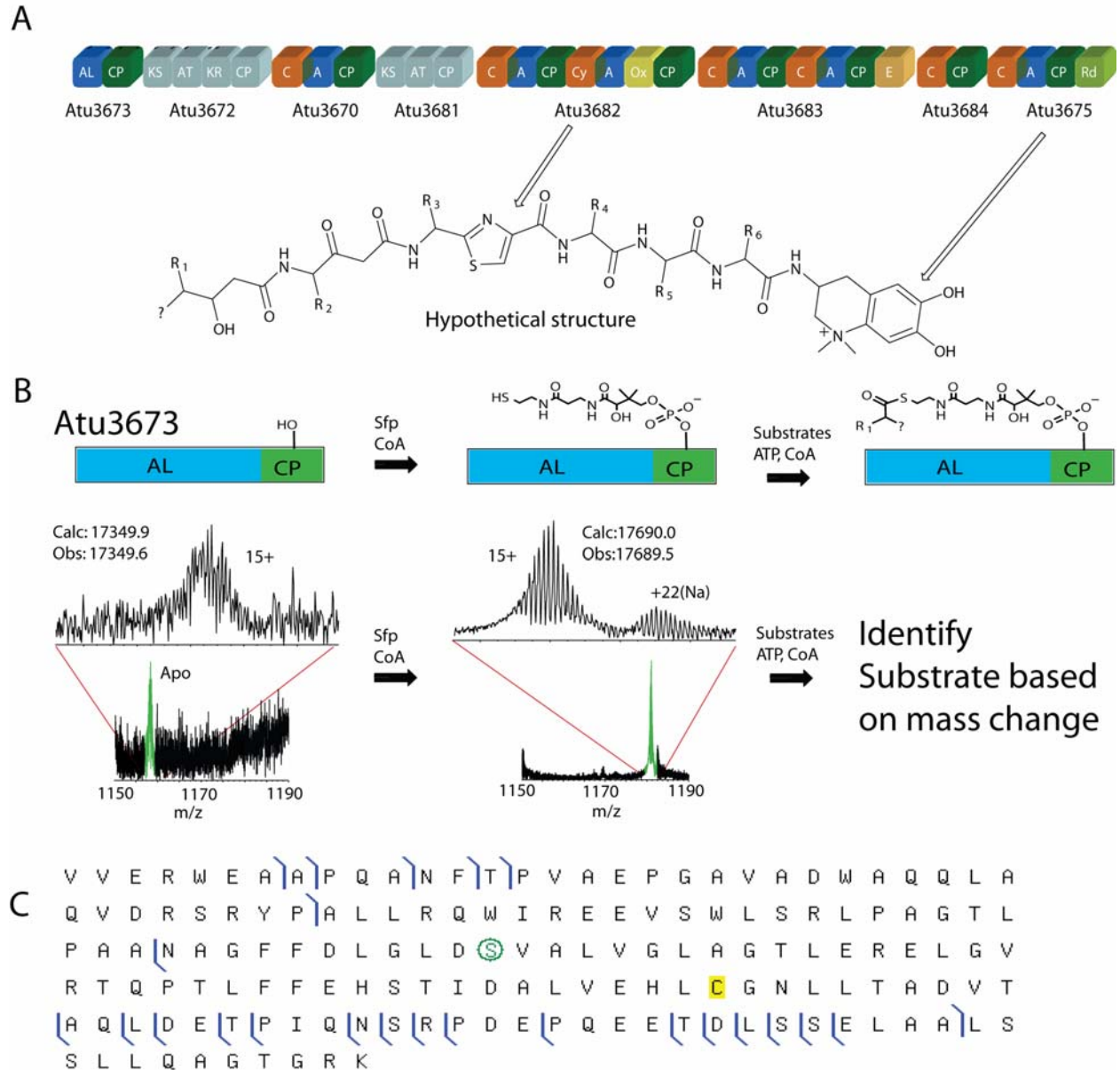
**Figure 5: Fluorescent Active Site Mapping of PksN and PksJ AT2.** (A) A diagram of the orphan gene cluster from *B. subtilis*. (B) An SDS-PAGE gel containing PksN (left 2 lanes) and PksJ AT1 (right 2 lanes) after acylation with BODIPY-CoA. Irradiating the gel with UV light confirmed the existence of fragments containing the active sites. (C) A comparison of the HPLC spectra of holo PksN at 220 nm (red trace) and BODIPY-PksN at 400 nm (green trace) revealed that the active site elutes at approximately the 38<sup>th</sup> minute. (D) A comparison of the HPLC spectra of apo PksJ AT2 at 220 nm (blue trace), holo PksJ AT2 at 220 nm (red trace), and BODIPY-PksJ AT2 at 400 nm (green trace) revealed two active site fragments, eluting during the 26<sup>th</sup> and 27<sup>th</sup> minutes. The additional large peak in the green trace did not correspond to an active site fragment.



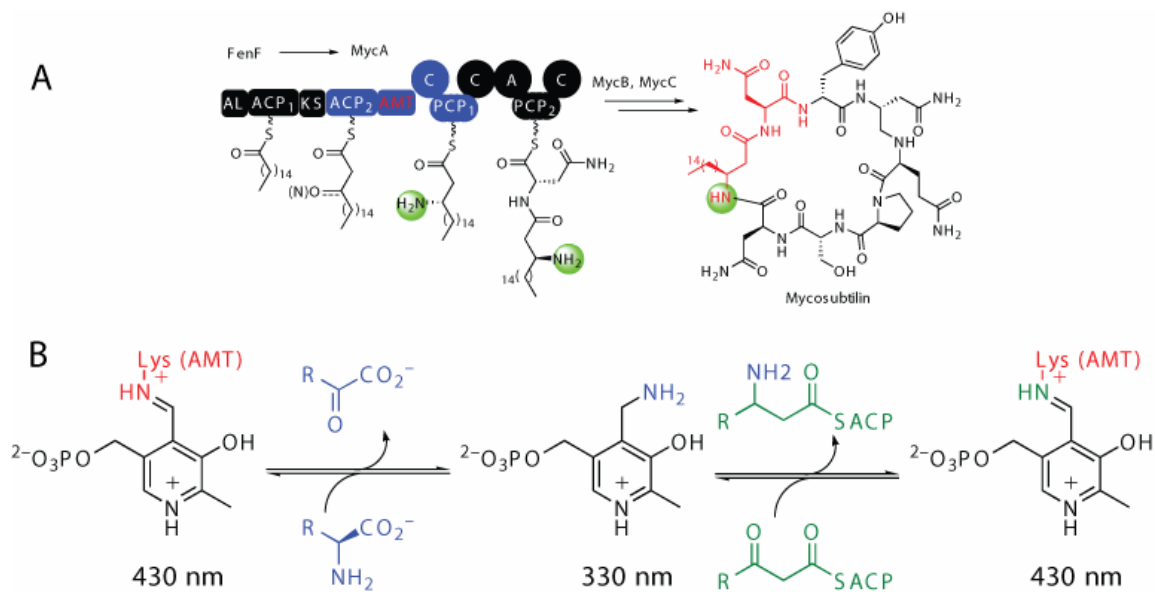
**Figure 6: FTMS Identification of Substrates for PksN.** Apo PksN (A) is converted to its holo form (B) before loading with its substrate alanine from the algal hydrolysate (C) and a mixture of Ala, Cys, Ser, Thr (D). An alternative substrate, serine, was loaded upon incubation with Ser and Thr only (E). When contaminant fragments were present, SWIFT isolation allowed for better resolution of the desired fragments (A-E). (F) IRMPD verification of the active site map for holo-PksN. The green circle indicates the location of the 4'-PP moiety.



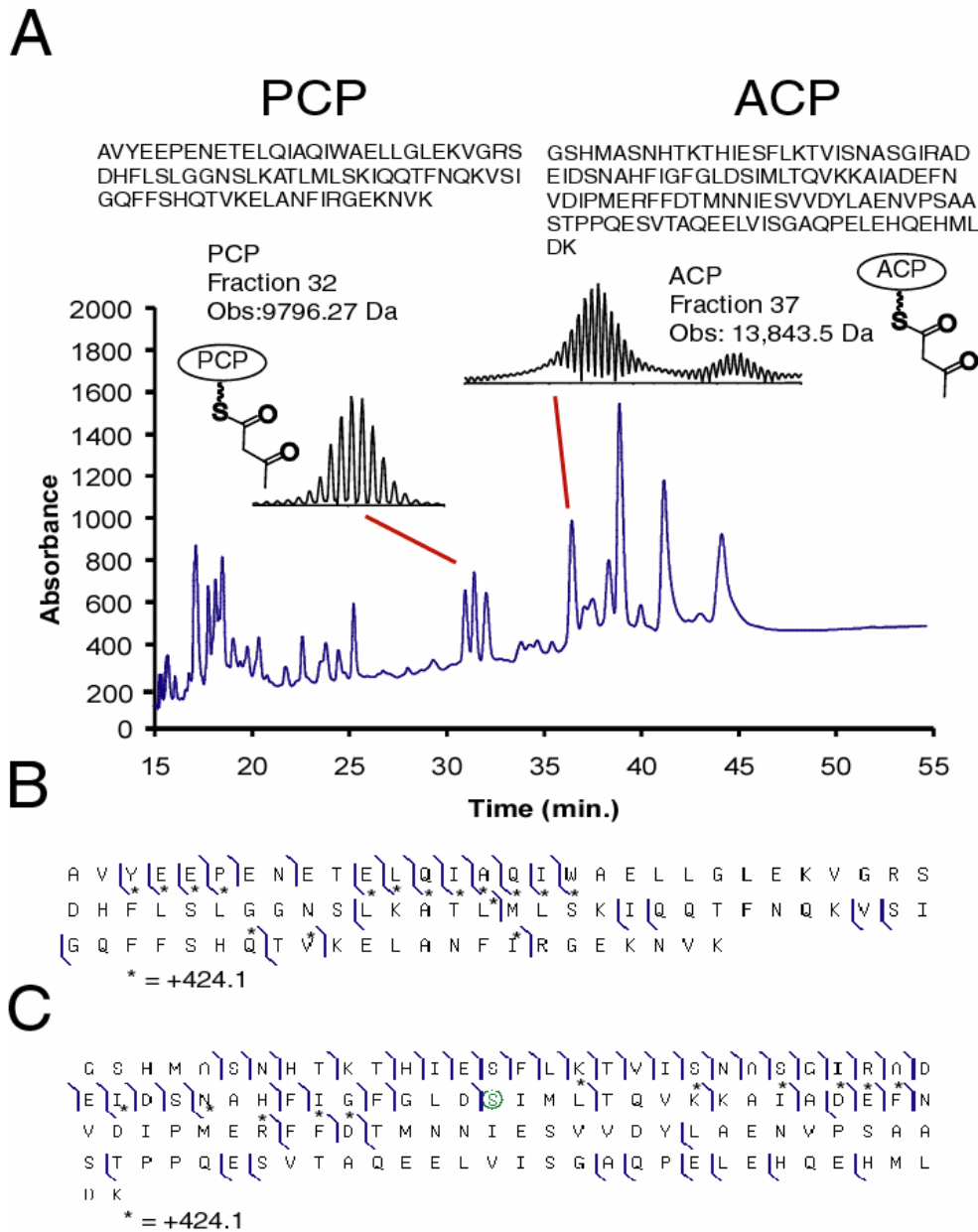
**Figure 7: FTMS Identification of Substrates for PksJ AT1 and AT2.** (A) BODIPY-PksJ AT1 purified by HPLC can be observed by FTMS. (B) Holo PksJ AT1. (C) Phenylacetate-loaded PksJ AT1 from the algal hydrolysate substrate pool. PksJ AT2 is activated from its apo form (D) to its holo form (E) before loading with its substrate glycine via the algal hydrolysate (F) or glycine alone (G). (H) OCAD verification of the active site map for holo PksJ AT2.



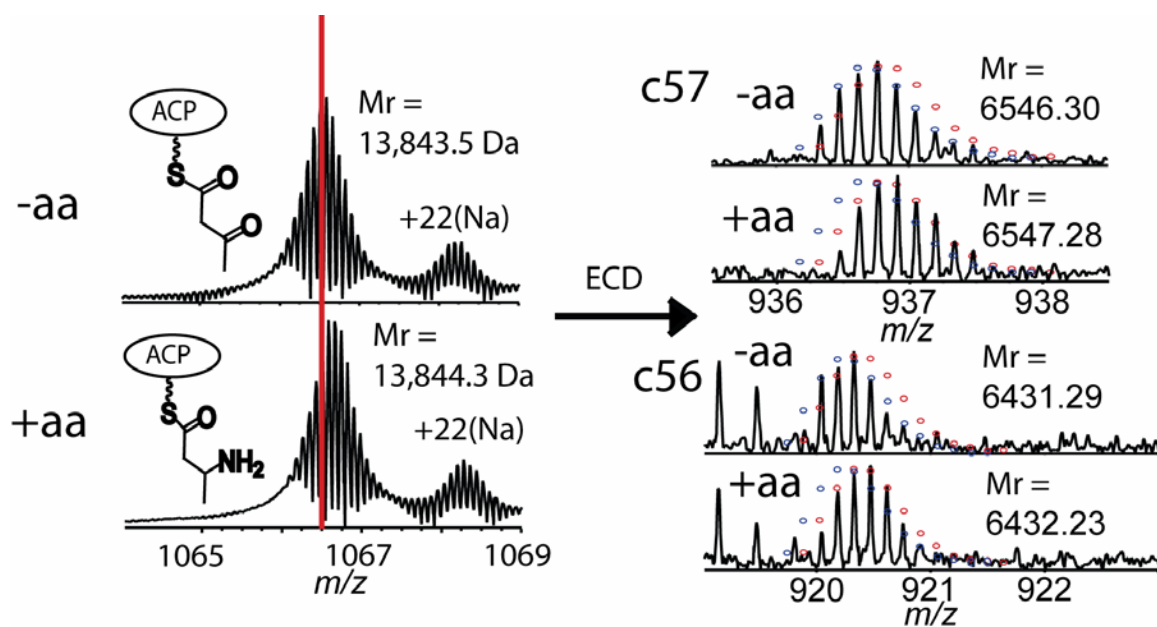
**Figure 8: FTMS Active Site Mapping of Atu3673.** (A) A diagram of the Atu3673 gene cluster from *A. tumefaciens*, including a hypothetical product. (B) The active site fragment was mapped by visualizing the apo and holo forms using FTMS. A substrate-loaded form was not observed, however. (C) OCAD verification of the active site map for holo Atu3673. The green circle represents the active site serine.



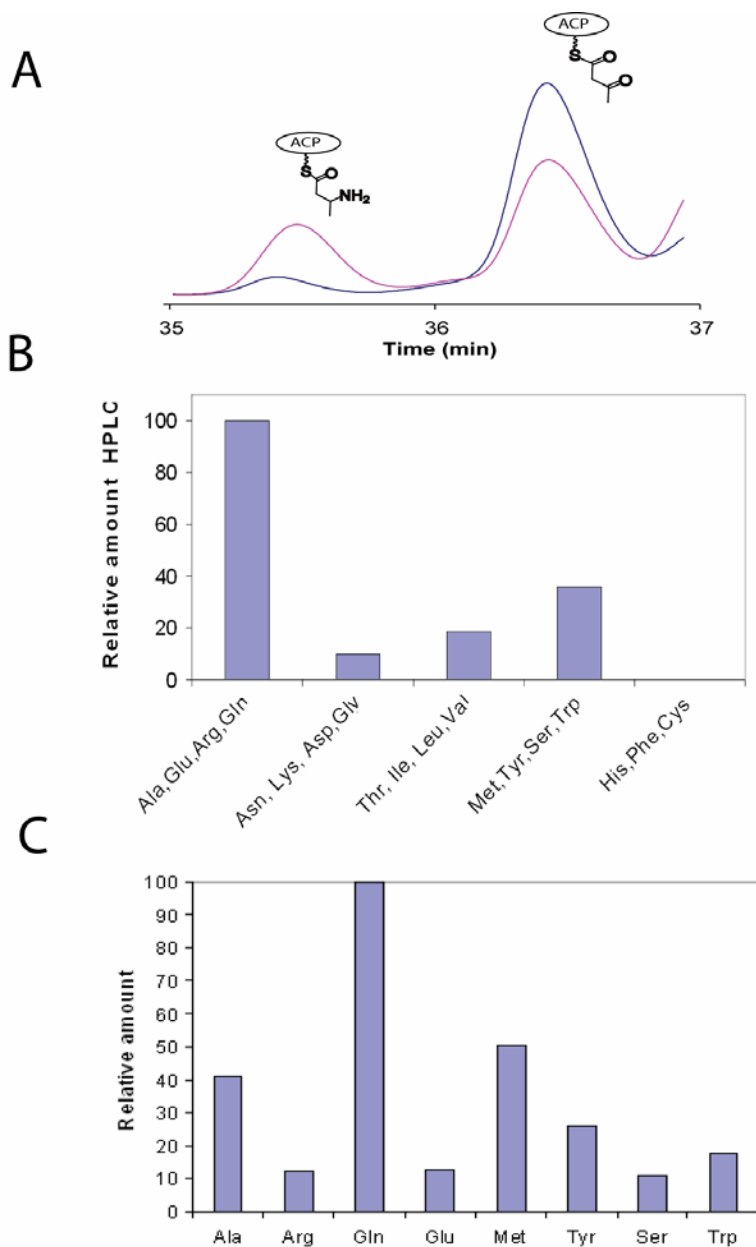
**Figure 9: Mycosubtilin Reaction Scheme.** (A) The MycA hybrid NRPS/PKS with the truncated MycA4N7 construct highlighted in blue. The amine that is transferred by the AMT is circled. (B) The amine transfer is a pyridoxal 5'-phosphate dependent reaction.



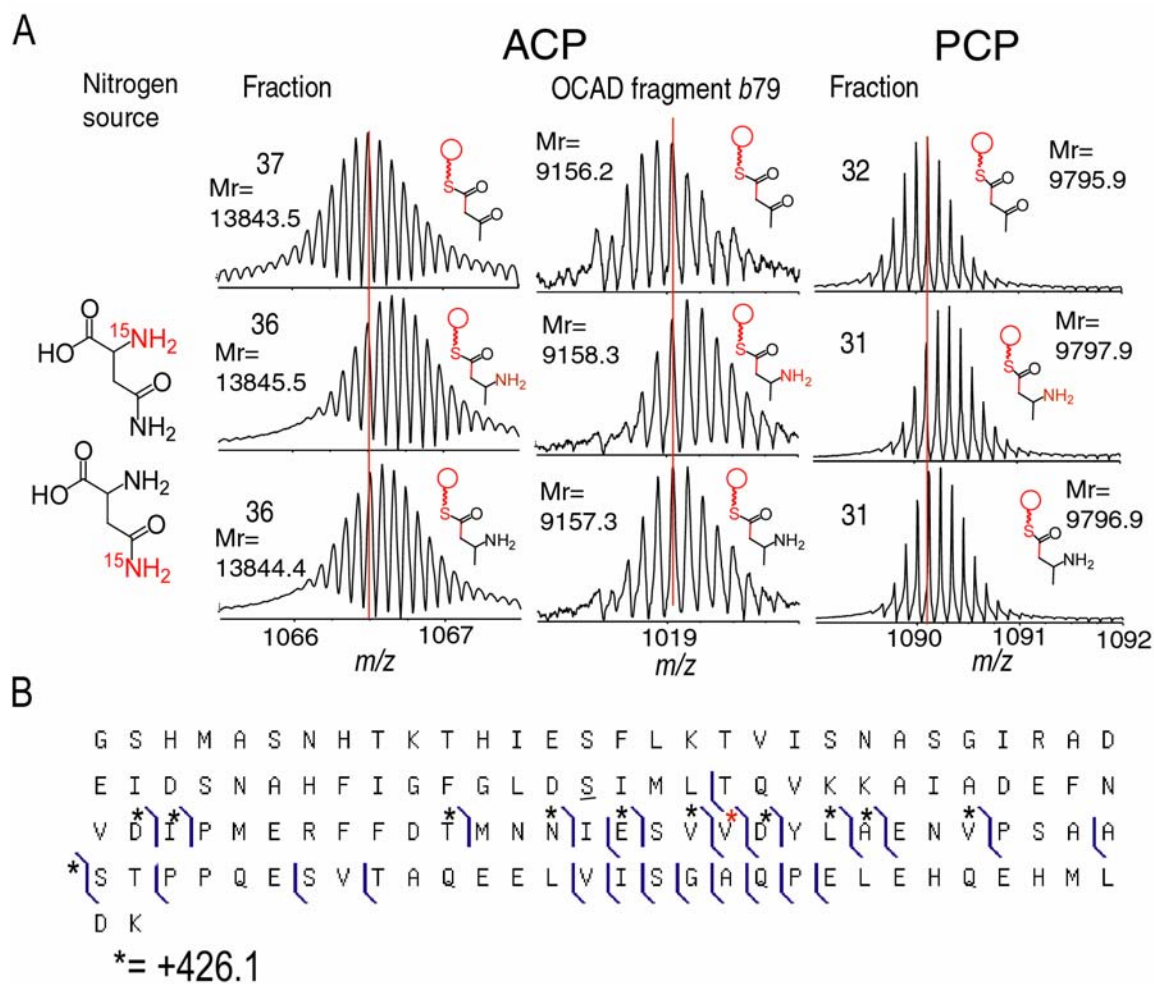
**Figure 10: Active Site Mapping of PCP<sub>1</sub> and ACP<sub>2</sub>.** (A) HPLC fractions from the 32<sup>nd</sup> minute and the 37<sup>th</sup> minute correspond to the PCP<sub>1</sub> and ACP<sub>2</sub> tryptic fragments as verified by FTMS, including the fragments' sequences. Analysis of tandem MS (MS/MS) fragments in ProSight PTM localized the active sites of the PCP (B) and the ACP (C) to fragments with the 424.1 Da mass increase due to acetoacetyl-CoA loading. The green circle represents the unambiguous localization of the active site serine.



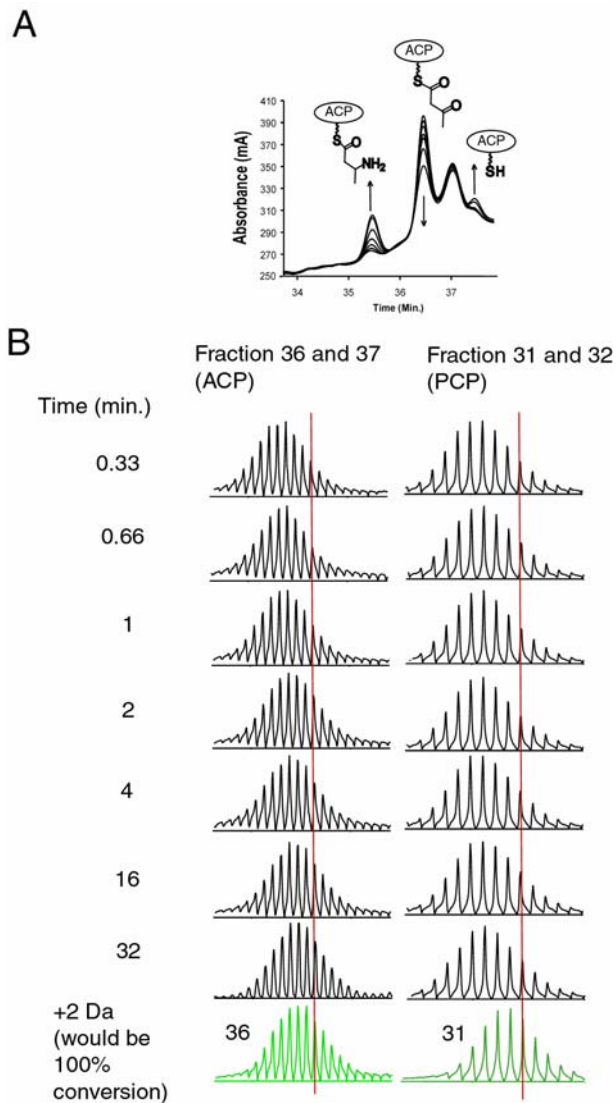
**Figure 11: FTMS Detection of Amine Transfer.** FTMS of the ACP tryptic fragment revealed a 1 Da mass increase in the presence of amino acids, suggesting that amine transfer had taken place (left). The red line is simply to aid in visualizing the mass increase. The result was further supported by examining two ECD fragment ions, c57 and c56 (right). The red dots represent a theoretical fit for aminoacetyl-S-ACP<sub>2</sub> while the blue dots represent acetoacetyl-S-ACP<sub>2</sub>.



**Figure 12: Determination of the Most Efficient Amine Donor.** (A) The HPLC spectra were used to assay the efficiency of amine transfer by an overt peak shift. (B) Small groups of amino acids were screened separately. (C) Amino acids from the most efficient groups were screened separately to determine the most efficient single amino acid donor.



**Figure 13: Determination of Alpha or Amide Amine Transfer from Glutamine.** (A) A comparison of FTMS data for acetoacetyl-S-ACP<sub>2</sub> with one of its OCAD fragments and PCP<sub>1</sub> in the absence of amino acids (top). The same analysis in the presence of <sup>15</sup>N- $\alpha$ -Gln (middle) and <sup>15</sup>N-amide-Gln (bottom). (B) Prosight PTM map of OCAD fragments for <sup>15</sup>N-aminoacetyl-S-ACP<sub>2</sub>. The red asterisk (\*) represents the *b*79 OCAD fragment ion.



**Figure 14: Time Course Examination of the Amine Transfer Reaction with  $^{15}\text{N}$ - $\alpha$ -Gln. (A)**

The HPLC spectra at various time points exhibit a decrease in the level of acetoacetyl-S-ACP<sub>2</sub> and an increase in the levels of aminoacetyl-S-ACP<sub>2</sub> and ACP<sub>2</sub> where acetoacetate has been hydrolyzed.

(B) The mass spectra of the ACP<sub>2</sub> and PCP<sub>1</sub> at various time points compared to the theoretical 2 Da mass increase, signifying 100% amine transfer. The red lines are simply to aid in visualizing the mass increase over time.

**Table 1: Identification of Loading from Substrate Pools in Known Systems**

System	Holo Mass (calc) (Da)	Acylation substrate source	Observed mass after acylation (major species only) (Da)	Mass change (observed – holo <sub>obs</sub> ) (Da)	Substrate name based on mass change (asterisk denotes a natural substrate)
Clorobiocin	12520.1	<i>a</i>	12616.9	96.8	L-proline*
CloN5/CloN4	(12520.1)	<i>b</i>	12633.0	112.9	4-trans-hydroxy-proline
		<i>d</i>	12617.2	97.1	L-proline*
		<i>e</i>	12617.1	97.0	L-proline*
Coumermycin	12087.1	<i>a</i>	12184.1	97.0	L-proline*
CouN5/CouN4	(12087.1)	<i>b</i>	12200.2	113.1	4-trans-hydroxy-proline
		<i>d</i>	12184.1	97.0	L-proline*
Enterobactin	12349.5	<i>a</i>	12468.5	119.0	Unknown
EntB(ArcP)/EntE	(12349.3)	<i>c</i>	12485.7	136.2	2,3-dihydroxybenzoic acid*
		<i>d</i>	No change	No change	Not loaded
		<i>f</i>	No change	No change	Not loaded

*a.* 19 proteinogenic L-amino acids, glycine, L-selenocysteine, and 4-*trans*-hydroxy-L-proline. *b.* Same as (a) but without 4-*trans*-hydroxy-L-proline. *c.* Same as (a) but with 2,3-dihydroxybenzoic acid. *d.* *E. coli* metabolome. *e.* Commercially available algal hydrolysate. *f.* *E. coli* metabolome grown in iron-limiting conditions.

**Table 2: Identification of Substrate Loading in Orphan Gene Clusters**

System	Digestion method	Mass holo (calc) (Da)	Acylation substrate source	Observed mass after acylation (major species only) (Da)	Mass change (observed – holo <sub>obs</sub> ) (Da)	Substrate name based on mass change (asterisk denotes a natural substrate)
PksN	Trypsin	12127.3	<i>a</i>	12198.4	71.1	L-alanine*
		(12127.9)	<i>b</i>	12198.5	71.2	L-alanine*
			<i>c</i>	12214.3	87.0	L-serine
PksJ AT1	Trypsin	12511.0	<i>a</i>	No change	No change	Not loaded
			<i>d</i>	12629.04	118.04	phenylacetate*
PksJ AT2	Trypsin	10121.7 (10121.8)	<i>a</i>	10179.0	57.3	glycine*
Atu3673	Trypsin	17690.0 (17689.5)	<i>a</i>	No change	No change	Not loaded

*a.* Commercially available algal hydrolysate. *b.* A mixture of Ala, Cys, Ser, Thr. *c.* A mixture of Ser and Thr. *d.* A mixture of phenylacetate, phenylpropionate, and phenyllactate.

---

**Table 3: The 60-minute and 30-minute HPLC Purification Gradients**

---

	0.00	10.0	15.0	55.0	60.0	60.1	60.2	62.6	63.0	65.0	66.0
	min	min	min	min	min	min	min	min	min	min	min
%A	90	90	70	30	10	10	95	95	5	5	90
%B	10	10	30	70	90	90	5	5	95	95	10

---

	0.00	5.0	30.0
	min	min	min
%A	90	90	10
%B	10	10	90

---

*Note that the unusual gradient above 60 minutes was used to wash the column prior to injection of the next injection to avoid contaminations from the previous run.*

---