2012

Efforts Toward Production of Novel Natural Products From Uncultured Soil Microbes

Jacob Joseph Banik

EFFORTS TOWARDS THE PRODUCTION OF NOVEL NATURAL PRODUCTS FROM UNCULTURED SOIL MICROBES

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Jacob Joseph Banik

June 2012

EFFORTS TOWARDS THE PRODUCTION OF NOVEL NATURAL PRODUCTS FROM UNCULTURED SOIL MICROBES

Jacob Joseph Banik, Ph.D.

The Rockefeller University 2012

The isolation of small molecule natural products from bacteria has led to the identification of many of the antibiotics currently in use today. Within the last 25 years, in fact, natural products (and their derivatives) account for almost three quarters of all antibiotic discoveries (Newman and Cragg). The discovery of novel natural product antibiotics, however, has witnessed a significant drop in numbers over this time period, suggesting that new sources of small molecules may be necessary to keep up with the need for new antibacterial compounds (Newman and Cragg). Due to the emergence of bacteria resistant to commonly used antibiotics, this need seems to be ever increasing, even as the pools of new antimicrobials are dwindling.

One potential source of new small molecule diversity is the large quantity of uncultured bacteria found in any given environment. It is estimated that, in soil, greater than 99% of the bacteria present are recalcitrant to culture. This poses a challenge for the field of natural product discovery, considering the traditional route to small molecule discovery from soil bacteria is to organically extract pure cultures of an individual bacterium. New techniques emerging from the field of metagenomics take a

culture-independent approach to examining the biosynthetic capabilities of soil bacteria by extracting the environmental DNA (eDNA) from the soil and cloning it into a host capable of maintaining and possibly heterologously expressing genes, which encode the production of small molecule secondary metabolites. These techniques have been employed in this thesis to expand not only a family of therapeutically relevant small molecule antibiotics (glycopeptides), but also to develop, in conjunction with new strategies, improvements on existing metagenomics approaches.

My work towards the discovery of novel glycopeptide antibiotics has led to the identification of six novel glycopeptide biosynthetic pathways, via the use of homology-based metagenomics techniques, and the production of 15 novel congeners of this extremely important family of compounds. Additional work in collaboration with members of the Darst Laboratory at the Rockefeller University led to the structural and biochemical characterization of two sulfotransferases, enzymes responsible for the sulfonation of a glycopeptide substrate. This type of chemical modification is rarely seen in glycopeptide biosynthesis, and the collection of three sulfotransferases identified using metagenomics techniques therefore presented a unique opportunity to gain a better understanding of the reactivity and substrate restrictions of these enzymes.

In addition to the work using existing, homology-based metagenomics techniques, an additional strategy was developed, utilizing the

complementation of a well-studied biosynthetic pathway, responsible for the production of an iron-scavenging siderophore in *E. coli*, to select for eDNA clones likely to be rich in secondary metabolite biosynthetic genes. This strategy selects for clones containing 4'-phosphopantetheinyltransferases (PPtases), which are commonly proximally linked to the non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs). PPtases are responsible for the post-translational modification of NRPSs and PKSs, which are genes commonly seen in secondary metabolite biosynthesis. This strategy was used to enrich an eDNA-derived metagenomic library for PKS and NRPS genes by over an order of magnitude compared to an unenriched library.

This thesis is dedicated to

My Family.

None of this would be possible or worth doing without them.

ACKNOWLEDGMENTS

I think it is safe to say that my career as a scientist stems directly from growing up on a farm miles from anything resembling a town. My brother and I rode our bikes up and down the dirt roads near our house, swam in the pond, and of course caused lots of mischief. However, we also spent a lot of time playing in the creeks, fishing, hiking in the woods, building forts, and developing an unending love for being outside. My brother and I have both become scientists, and I can't help but think that it was the combination of our mother's desire to have us outside breathing in "that fresh air" as much as possible, and the instillation in both of us by our father to be ever curious, ever interested, ever mindful of the world around us that led us to this point. I can't thank my mother and father enough for their guidance and inspiration, and my little brother, who was always there to help me get in (and out) of trouble (generally involving picking on our younger sisters, who also need a nod for surviving having us as older brothers). My constantly observing, constantly scrutinizing nature is a direct result of days spent with my brother, Stephen J. Banik II.

My parents birthed my love of nature. My high school science teachers exponentially expanded that love, while honing it into a critically thinking, ever inquisitive young researcher. I credit Mrs. Mary Peropat and Mr. James Bailey, my physics and chemistry teachers, respectively, with pushing me in

the direction of the physical sciences, combining my love of solving math problems with my fascination with atoms, molecules, and the periodic table. I credit Ms. Lisa Higham with providing me with great language fundamentals, which have yet to fail me in my years of writing. However, it was two very special teachers, Mr. and Mrs. Allan and Donna Puskar, my high school biology and ecology teachers, respectively, who showed me what it was like to live science. These people ate, drank, and breathed science. They showed me what it was to truly be a scientist, not just a student in a science class, to immerse oneself in a topic, to a point where the subject and the student were inseparable. To this day, I still regularly correspond with the Puskars. They have become life-long friends, adopted grandparents to my children, and mentors always available to critique an idea. Their impact on my life cannot really be put into words. But, I can definitely say that my life would be infinitely less inspired and intriguing without them in it.

After selecting biochemistry and molecular biology as a major at the University of Massachusetts-Amherst, I enrolled in the Commonwealth College honors program, where I had the fortune to have Prof. Justin Fermann for honors freshman chemistry. Justin was a young, cool guy, who always put a picture of Elvis Presley on his exams and always had a pot of Eight O'clock coffee on, should you stop by his office to chat. He was a great professor, who encouraged and inspired me to add chemistry as a double major. That decision was also helped by a few of my very close fellow

students, including Mike Levine, Gavin Histen, Olaf Aprans, Mike Doherty, and Ethan Sullivan. These solid guys made a kid from Pennsylvania feel welcome in Massachusetts, and gave me a good base of fellow students with which to work on assignments, write up lab reports, and get into college-age mischief.

Around the start of my sophomore year, I was fortunate enough to receive some advice from my uncle, Greg Banik, who had a PhD in chemistry. He told me that I should get into a research lab as soon as possible. So, a few weeks into my sophomore year, I contacted Prof. Mike Maroney to discuss possibly joining his group as an undergraduate researcher. Mike was very welcoming, and within a few weeks, I was in the lab, under the supervision of Sergio Chai, a graduate student, born in Bolivia to Korean parents. Sergio was a night owl, though, and eventually, I was working part time with him and part time with another PhD student, Pete Bryngelson. Pete's attitude towards lab work and his lab mates greatly influenced my graduate career. Even though during my PhD research, I was technically a "student," I took it more like Pete took it, as a job where you work hard, regardless of the habits of those around you, you try to fix equipment yourself first, before asking someone else to do it, and you treat your boss with the respect he deserves. My early days in the Maroney lab were formative, and I have these folks to thank for that.

I studied abroad at the University of East Anglia my junior year, where I had the privilege to meet Prof. Manfred Bochmann, my inorganic chemistry workshop leader, whose insistence that any second or third year chemist should know the first row of the transition elements by heart showed me the type of devotion to one's craft required to be at the top. After finishing up at UMass, where I had the chance to hone my research (and softball) skills alongside a great labmate and friend, Bob Herbst, I returned to UEA, where I started a joint Master's project, under the advisement of both Prof. Bochmann and Prof. Andrew Thomson, who became like a grandfather to me. Always calm, never with a raised voice, never overly critical, but with just enough disappointment expressed at times to really motivate one, for fear of earning the disapproval of such a sage mentor. During my Master's, I worked with some great lab members, including (but not limited to) Andy Mountford, Dale Pennington, Polly Wilson, Yann Sarazin, and Ruth Howard. In the Thomson group, I acquired two extra mothers, Louise Ottignon and Gaye White, whose kindness and generosity I can't even begin to repay, as well as a partner in EPR, Justin Bradley.

Within a year of finishing my M.Sc., I had worked as a long-term substitute chemistry teacher, gotten married, and gotten accepted to grad school. I was accepted into the Tri-I TPCB program, and, after a few months at Cornell taking fall classes, was required to do one last rotation in New York City. I thought this was pointless, as I was convinced that I was going to

join a group in Ithaca. However, I had seen a young faculty member at Rockefeller, Sean Brady, give a rotation lecture. His work seemed fascinating, so I figured I'd give it a shot. I emailed Sean, who was willing to let me rotate, and started from there. When I got to the lab, I was amazed at the caliber of people there. Everyone was hardworking, energetic, and extremely bright. Sean was a chill mentor, who ended every email with "later on" and made you feel like you were in on the ground floor of something huge. The rest, as they say is history. Sean has taken what I thought were excellent research skills and brought them to the next level of understanding and focus. He has been an excellent mentor, and for that I am ever thankful. He has made me the scientist I am today. I must also thank my committee, Hening Lin and my chair Howard Hang for their ability to help me see the forest for the trees. They have given me perspective that is sometimes lost when one is so immersed in a research project. I also must thank my external committee member, Prof. Anthony Hay, for taking the time from his schedule to be an integral part of this process. I need to also thank the people who have made my Rockefeller experience what it is. I have to thank all the Brady lab members, especially, John Bauer, Jeff Kim, Ryan King, Hala Iqbal, and the other Brady lab members who have come and gone. Also, to the members of the Darst lab who were extremely welcoming to the collaborative portions of my project, especially Matt Bick and Seth Darst, many, many thanks. I need to offer a special shout to Jeff Craig, who was my sidekick in

lab over the last two years, and whose influence on my lab etiquette and technique has been extremely beneficial. Lastly, but most definitely not least, my family, which has grown to include an entire new family of in-laws and more recently two loving daughters, Molly and Caroline, and my loving wife, Kerry, who has been by my side through all the ups and downs life has to offer. I thank you from the bottom of my heart. Without my family, not only is none of this possible, but it's also not worth doing. Thank you all so much.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACP	Acyl carrier protein
AHL	Acyl homoserine lactone
ADME	Adsorption, distribution, metabolism and excretion
ATCC	American Type Culture Collection
AB	Anza-Borrego (eDNA-derived cosmid mega-library)
AZ	Arizona (eDNA-derived cosmid mega-library)
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search tool
BHT	Betahydroxytyrosine
CTAB	Cetyl trimethylammonium bromide
dNTPs	Deoxynucleotide triphosphates
DNA	Deoxyribonucleic acid
DPG	3,5-dihydroxyphenylglycine
DMSO	Dimethylsulfoxide
DTT	Dithiothreitol
eDNA	Environmental DNA
ESI	Electrospray ionization
EDTA	Ethylenediaminetetraacetic acid
EC-Δ*entD*	*E. coli* EC100-Δ*entD*
GHL	Glycopeptide helix loop
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPG	Hydroxyphenylglycine
HMW	High molecular weight
HPLC-MS	High pressure liquid chromatography - mass spectrometry
HRMS	High resolution mass spectrometry
kb	Kilobase
LB	Luria Bertani medium
Mb	Megabase
MRSA	Methicillin-resistant *Staphylococcus aureus*
MIC	Minimal inhibitory concentration
MWCO	Molecular weight cut off
Ni-NTA	Nickel-nitrilotriacetic acid
NMR	Nuclear magnetic resonance spectroscopy
NRPS	Non ribosomal peptide synthetase
ORF	Open reading frame
PCP	Peptidyl carried protein
PB	Phosphate binding loop
PAP	3'-phosphoadenosine-5'-phosphate
PAPS	3'-phosphoadenosine-5'-phosphosulfate
PSB	Phosphosulfate binding loop
PCR	Polymerase chain reaction

PKS	Polyketide synthase
PPTase	4'-phosphopantetheinyltransferase
RAST	Rapid annotation using subsytems technology
TEG	Teicoplanin-like eDNA derived gene cluster
SAM	*Streptomyces* antibiotic producing media
SVM	*Streptomyces* vegetative media
TAR	Transformation associated recombination
TLC	Thin-layer chromatography
UT	Utah (eDNA-derived cosmid mega-library)
VEG	Vancomycin-like eDNA derived gene cluster
VRE	Vancomycin-resistant *Enterococci*
VRSA	Vancomycin-resistant *Staphylococcus aureus*

CHAPTER 1

1 Introduction and Background

1.1 Natural Products

Over two thirds of the antibiotics discovered over the 25 year period spanning 1981-2006 are or are derived from small molecule natural products (Newman and Cragg 2007). Natural products fill many therapeutic niches, with properties ranging from antibacterial and anticancer agents to muscle relaxants and contraceptives. Soil bacteria have been a particularly prolific source of small molecule natural products. In fact, it is estimated that one bacterial genus alone, *Streptomyces,* has been the source of nearly two thirds of all clinically useful antibiotics (Kieser 2000). As there are predicted to be over 10,000 individual bacterial species per gram of soil, the reservoir for new bacterially-derived natural products seems rather promising.

The isolation of small molecules from bacteria recovered from the environment has followed a largely unchanged pattern for decades. The traditional protocol used to isolate and identify natural products from bacteria is to first culture an individual bacterium from the environment, extract bacterial cultures using organic solvents, and isolate pure compounds through activity guided fractionation. Many of the compounds isolated using

this paradigm are not themselves essential to the growth of their bacterial progenitor. Although non-essential to growth, bacteria produce many of these compounds to gain a selective advantage in their natural environment. These molecules are collectively termed secondary metabolites and constitute the majority of bacterial natural products described to date.

Secondary metabolites are used by bacteria for myriad of processes, including communication, self-defense, and iron scavenging (Maplestone, Stone et al. 1992). Bacterial secondary metabolites are often rather complex molecules that are generally produced via the transformation of simple starting materials by various groups of biosynthetic enzymes. These groups of enzymes, which are genomically encoded on either the bacterial chromosome or on stand-alone plasmids, are often clustered with one another into one or more functional operons. It is predicted that this clustering has facilitated the evolution of biosynthetic pathways capable of producing complex secondary metabolites (Maplestone, Stone et al. 1992).

Although natural products have been a rich source of new chemistry, the discovery of new natural product antibiotics has declined over the last several decades (Fox 2006). The factors contributing to this decline are many and varied. The traditional culture-based paradigm itself, however, seems to be one of the major hurdles to new natural product isolation. As one would expect for any population subject to natural variation, there are bound to be both very common and very rare natural products. To identify the rarest

compounds from the milieu of known compounds may require a change in the paradigm, where one can examine all the bacteria in an environmental sample in an expeditious, facile, and unbiased manner. Experimentation over the last several decades has indicated that the majority of the bacteria found in nature are not readily amenable to culture-based strategies (Torsvik, Goksoyr et al. 1990; Torsvik, Salte et al. 1990; Torsvik and Ovreas 2002; Rappe and Giovannoni 2003). Thus, it is now apparent that the traditional culture-dependent paradigm has been omitting the vast majority of bacteria found in the biosphere.

1.2 Uncultured Microbes

Although the traditional culture-based strategy has proven very successful over the course of more than a century of investigation, it excludes a significant portion of the molecules present in most environmental samples. Culture-independent analyses of environmental samples suggest that traditional approaches used to identify microbial metabolites from laboratory grown microorganisms have likely missed the vast majority of bacterial natural products that exist in nature (Hugenholtz, Goebel et al. 1998; Rappe and Giovannoni 2003). In most environments, bacteria that have not yet been cultured are thought to outnumber their cultured counterparts by at least two orders of magnitude (Torsvik, Goksoyr et al. 1990; Hugenholtz, Goebel et al. 1998; Torsvik, Daae et al. 1998; Rappe and Giovannoni 2003). If the diversity of molecules discovered from cultured bacteria is any indication, as

yet uncultured bacteria are likely to be a very rewarding source of previously undiscovered biologically active small molecules.

There are a number of different strategies, using both culture-dependent and culture-independent methods, which are now being developed to access this untapped reservoir of chemical diversity. Typically, culture-independent strategies involve direct cloning of the genetic material from an environmental sample, which contains the genes encoding for small molecule biosynthetic machinery. The genetic diversity of all the bacteria in any given environmental sample has been termed the "metagenome." Culture-independent or "metagenomic" approaches can provide access to previously untapped pools of chemical diversity. It therefore seems likely that the pools opened up through metagenomics represent some of the largest reservoirs of unexamined genetic and chemical diversity remaining in nature. These types of culture-independent techniques are the basis of the research described in this thesis.

1.3 Metagenomics

The foundation of all metagenomic approaches is the isolation and subsequent examination of DNA extracted directly from naturally occurring microbial populations (environmental DNA, eDNA), which avoids the difficulties associated with culturing environmental bacteria (Figure 1) (Handelsman, Rondon et al. 1998). The collections of eDNA used for

metagenomic analyses contain vast quantities of genetic material, which undoubtedly possess the biosynthetic potential to produce a vast array of previously described secondary metabolites. To gain access to this biosynthetic machinery, one must first clone the eDNA into a suitable, cultured host, for archiving and subsequent downstream examination. One method of cloning often used in metagenomics is cosmid cloning, which takes advantage of the size selective packaging of lambda phage to transfect *E. coli* with environmental DNA clones containing between on average 40 kilobases of genetic material (Brady 2007). As previously mentioned, secondary metabolite biosynthetic gene clusters are often clustered in the bacterial genome. Metagenomics is particularly appealing to natural product researchers because of this characteristic clustering, since most secondary metabolite biosynthetic gene clusters are under 100 kb, making it possible to capture biosynthetic gene clusters on individual or, at most, a small number of overlapping eDNA clones (Handelsman, Rondon et al. 1998). Large scale cosmid cloning of environmental DNA provides a platform, from which one can investigate secondary metabolite biosynthetic gene clusters from uncultured bacteria.

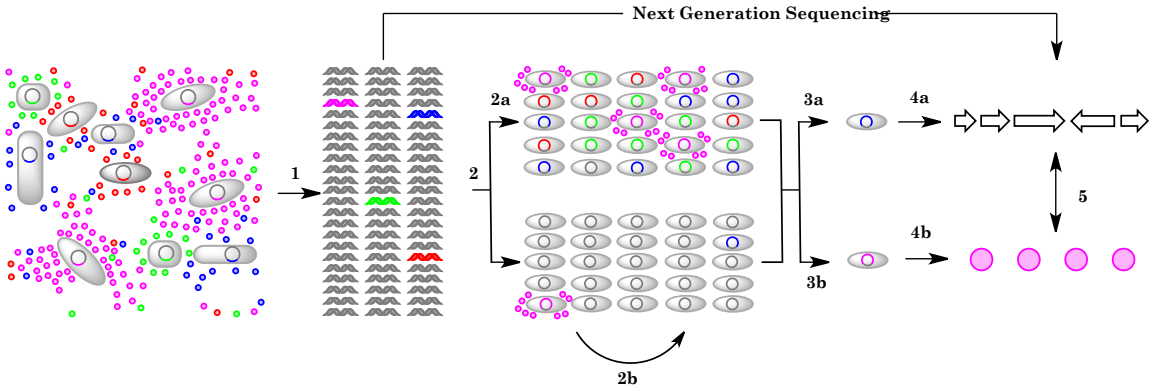


Figure 1: Overview of metagenomic methods.

Environmental DNA isolated directly from an environmental sample (1) is cloned into an easily cultured model bacterial host (**2**). Libraries (or eDNA) can then be enriched for genes of interest (2a), transferred into additional heterologous hosts (2b), or screened directly. The search for bioactive small molecules using a metagenomic approach has generally been conducted using either homology based methods (3a) or functional screening (3b). Novel sequences found in homology-based screens (4a) can be examined for the ability to encode the biosynthesis of novel small molecules in heterologous expression experiments (5). The characterization of hits from functional screens can lead directly to the identification of bioactive small molecules (4b) and their biosynthetic gene clusters (5).

1.3.1 Functional metagenomic screening

Expression-dependent (functional) metagenomic screening strategies have been used to identify eDNA clones that produce bioactive small molecules. In functional metagenomic studies, eDNA libraries are examined in simple high throughput assays designed to identify clones that have phenotypes traditionally associated with the production of small molecules, such as pigmentation, altered colony morphology, or antibiosis. In homology-based studies, libraries are probed to identify clones that contain conserved sequences traditionally associated with secondary metabolite biosynthesis.

Hits identified in these initial high throughput assays are subsequently examined for the ability to confer the production of small molecules to model cultured heterologous hosts.

One of the simplest strategies used to detect eDNA clones that might produce small molecule antibiotics has been to screen libraries hosted in *E. coli* for clones that generate zones of growth inhibition against test microbes in top agar overlay assays. The isolation of clone specific metabolites produced by antibacterially active eDNA clones identified from bacterial top agar overlay assays has led to the characterization of a variety of new long-chain *N*-acylated amines (**1**), as well as a new isonitrile functionalized indole antibiotic (**2**) (Figure 2) (Brady and Clardy 2000; Brady and Clardy 2005). Small molecule antibiotics have also been found by examining pigmented eDNA clones, as well as through the direct examination of culture broth extracts from randomly selected clones (Wang, Graziani et al. 2000; Brady, Chao et al. 2001; MacNeil, Tiong et al. 2001; Gillespie, Brady et al. 2002; Lim, Chung et al. 2005; Long, Dunlap et al. 2005). Compounds with bioactivity identified from these types of studies include the antibiotic pigments violacein, indigo (**3**) and the turbomycins (**4**), all recovered from soil libraries, as well as the known cyclic peptides patellamide D (**5**) and nocardamine (**6**), isolated from marine sponge and soil libraries, respectively (Figure 2).

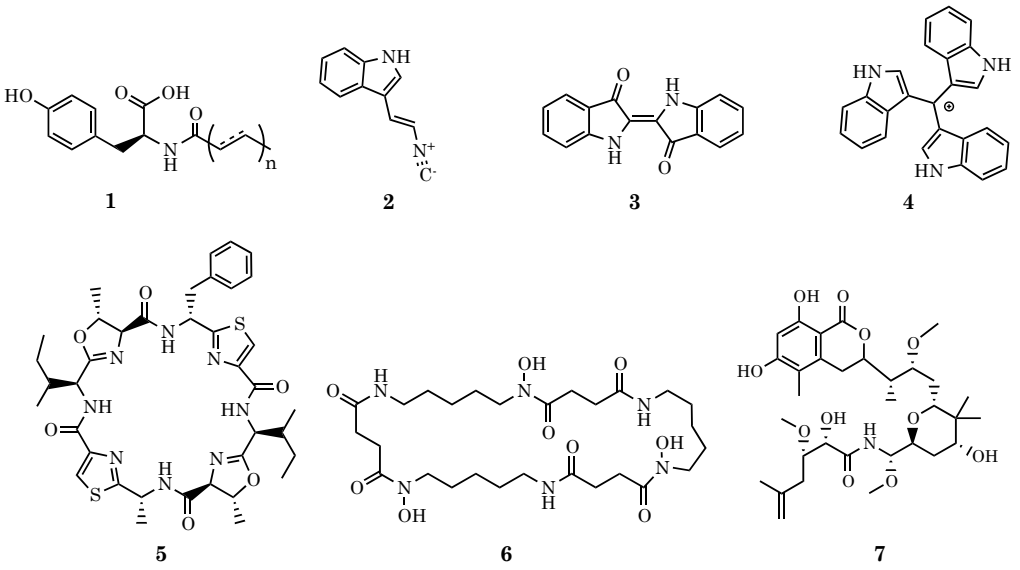


Figure 2: Natural products heterologously produced in model cultured bacteria from metagenomic derived genes and gene clusters.

N-acyltyrosine (**1**), isocyanide functionalized indole (**2**), indigo (**3**), turbomycin A (**4**), cyanobactin patellamide D (**5**), nocardamine (**6**) and psymberin (**7**).

Functional metagenomics has also been used to identify clones that produce proteins with potential anti-infective properties. Using an acylhomoserine lactone synthase promoter fused to a *lacZ* reporter, Schipper et al. identified three eDNA derived AHL lactonases that are capable of inhibiting biofilm formation by *Pseudomonas aureginosa* (Schipper, Hornung et al. 2009). And, an examination of bacteriophage DNA isolated from bat guano and earthworm guts by Schmitz, et. al. led to the discovery of three new lysins capable of halting *Bacillus anthracis* proliferation (Schmitz, Daniel et al. 2008). In this work, the authors were able to functionally access phage lysins by inducing the expression of genes cloned from environmental samples using a vector associated *ara*BAD promoter.

Although nearly all small molecule focused functional metagenomic studies have been carried out in *E. coli*, it is likely that the majority of the biosynthetic diversity present in an environmental sample is not functionally accessible using a single heterologous host. A computational analysis of promoters and ribosomal binding sites used by a taxonomically diverse group of sequenced bacteria found that at most, 40% of the enzymatic activities present within a typical metagenomic sample could be accessed using *E. coli* as a heterologous host (Gabor, Alkema et al. 2004). The successful expression of entire biosynthetic gene clusters, which requires the coordinated production of multiple proteins, is likely to occur at an even lower frequency. Vector-host pairs that allow for the introduction and screening of metagenomic libraries in phylogenetically diverse bacteria have the potential to expand the number and type of compounds found from metagenomic studies. While cosmid and BAC vectors capable of replicating in a variety of Gram-positive and Gram-negative hosts have been described in the literature, until recently, none of these had been used in an extensive broad-host- range small molecule focused screen of metagenomic libraries (Courtois, Cappellano et al. 2003; Martinez, Kolvek et al. 2004). Two RK2-derived broad-host-range vectors (pJWC1 and pRS44) were recently constructed with this specific purpose in mind (Aakvik, Degnes et al. 2009; Craig, Chang et al. 2009). Craig, et al. demonstrated the utility of pJWC1 by screening metagenomic libraries for eDNA clones that confer antibacterial activities to

any of six different host Proteobacteria, including *Agrobacterium tumefaciens, Burkholderia graminis, Caulobacter vibrioides, Escherichia coli, Pseudomonas putida,* and *Ralstonia metallidurans* (Craig, Chang et al. 2009; Craig, Chang et al. 2010). This study found that distinct collections of eDNA clones within the same metagenomic library are likely to confer detectable phenotypes to different hosts and that eDNA clones infrequently confer the same phenotype to two different hosts.

1.3.2 Metagenomic analysis of microbial endosymbionts

Many bioactive natural products that were originally isolated from extracts derived from multicellular organisms are now thought to be products of as yet uncultured microbial symbionts. Metagenomics provides a strategy for cloning the biosynthetic gene clusters of these metabolites, which may in turn provide a renewable source of compounds that have often been difficult to isolate in sufficient quantities to permit extensive biological testing. The biosynthetic gene cluster for pederin, an anticancer agent originally isolated from the beetle *Paederus fuscipes*, was recovered from a cosmid library constructed using beetle-derived metagenomic DNA and this pathway was shown to originate from an uncultured symbiotic *Pseudomonas spp.* (Soldati 1966; Piel 2002). As additional pederin-like structures had also been isolated from marine sponge extracts, it was hypothesized that these other molecules might originate from bacterial symbionts as well. In two separate studies, the Piel group reported the cloning of gene clusters encoding the biosynthesis

of the pederin relatives onnamide and psymberin (**7)** from symbionts associated with field collected *Demospongiae* sponges (Figure 2) (Piel, Hui et al. 2004; Fisch, Gurgui et al. 2009). While it has not yet been possible to heterologously express these gene clusters in the laboratory, Zimmermann et al., reported the use of a recombinant *O*-methyltransferase, PedO, from the pederin biosynthetic gene cluster to site-specifically methylate mycalamide A resulting in the production of a derivative (**8**) that exhibits enhanced antitumor activity (Figure 3) (Zimmermann, Engeser et al. 2009). In related work using libraries constructed from DNA extracted from uncultured cyanobacterial symbionts associated with marine *Didemnidae* sponges, two separate groups have reported the cloning and heterologous expression of biosynthetic gene clusters for a number of patellamides, cytotoxic cyclic peptides originally isolated from sponge extracts (Figures 2 and 3) (**5,9**) (Long, Dunlap et al. 2005; Schmidt, Nelson et al. 2005).

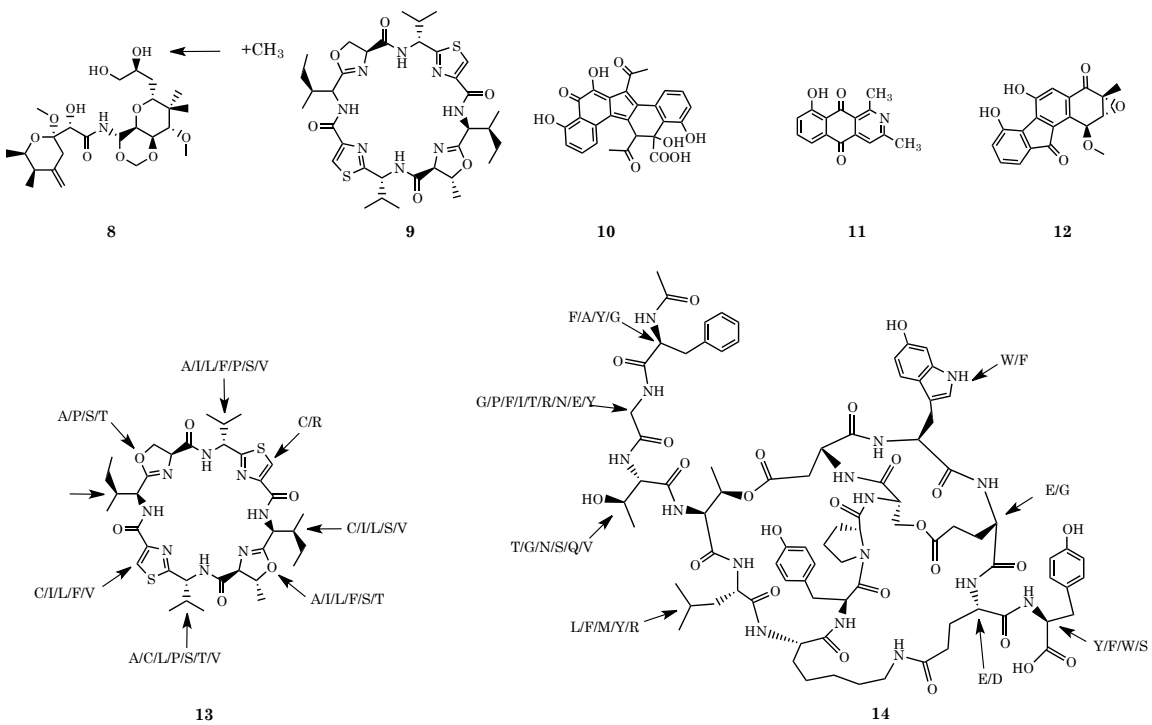


Figure 3: Secondary metabolites identified as a result of recent metagenomic efforts.

O-methylated mycalamide A (**8**), patellamide A (**9**), erdacin (**10**), utahmycin A (**11**), fluostatin F (**12**), patellamides (**13**), and microviridins (**14**). The cyanobactin and microviridin precursor peptide diversity found in metagenomic studies is displayed on patellamide A and microviridin B, respectively.

Most culture independent symbiont studies have focused on anticancer agents. In future studies, the same general approach will undoubtedly prove useful for investigating symbiont-derived antimicrobials. It was recently shown that a symbiotic *Streptomyces* species associated with leaf-cutting ants (*Acromyrmex*) produces the potent antifungal candicidin, which is active against the pathogenic fungus (*Escovopsis*), but non-active against the symbiotic fungus (*Leucoagaricus*) that the ants maintain as their main food

source (Haeder, Wirth et al. 2009). While the producing organism in this study could be cultured, collections of uncultured microbial symbionts that promise to be rich sources of future small molecule metagenomic studies have been found in environments ranging from the human gut to marine snails (Peraud, Biggs et al. 2009; Qin, Li et al. 2010).

1.3.3 Homology-based metagenomic screening

Expression-independent or homology-based screens rely, as the name suggests, not on the expression of genes for detection of a small molecule as in functional metagenomics, but in the homology of previously unknown, eDNA-derived secondary metabolite biosynthetic pathways to sequenced pathways of known metabolites. One of the first groups of molecules targeted using this type of strategy was the iterative (Type II) polyketides (Seow, Meurer et al. 1997). Type II polyketides utilize a small collection of biosynthetic enzymes to produce a vast array of malonyl-CoA-derived scaffolds, which are modified and decorated by various downstream enzymes.

This "minimal PKS," two ketosynthase genes (KS_α and KS_β) and an acyl carrier protein, possesses a conserved organization, and were therefore an ideal target to identify using homology-based metagenomic screening (Macpherson, Manning et al. 1994). Homology-based screening utilizes degenerate oligonucleotides to amplify sequence homologs via the polymerase chain reaction (PCR) (Seow, Meurer et al. 1997). Two reports had described the identification of type II PKS genes from environmental DNA samples, but

it was not until 2009 that a pathway encoding the biosynthesis of a type II PKS was expressed and a molecule characterized. This new molecule, erdacin (**10**) possesses a scaffold previously unseen in type II PKS biosynthesis, lending credence to the idea that uncultured bacteria may contain chemical diversity unlike that seen in cultured bacteria (Figure 3) (King, Bauer et al. 2009). Since this study, several other studies have been published, detailing the discovery of additional novel type II PKS-derived compounds, such as utahmycin A (**11)** and fluostatin F (**12**) (Figure 3) (Bauer, King et al. 2010; Feng, Kim et al. 2010). Taken together, these reports offer a glimpse of the previously unseen chemistry homology-based metagenomic screening has the potential to uncover.

In an expansion of its earlier cyanobactin (patellamide) research (Schmidt, Nelson et al. 2005), the Schmidt group recently reported the PCR amplification of 30 genes encoding novel patellamide-like precursor peptides from uncultured *Prochloron spp.* symbionts living in consortia with marine sponges (**13**) (Figure 3) (Donia, Hathaway et al. 2006; Schmidt and Donia 2009). In another PCR based study, in this case using DNA isolated from uncultured freshwater cyanobacteria of the genera *Microcystis*, Ziemert, et al. identified 15 new variants of the gene that encode for the precursor to the microviridin peptide (**14**) (Figure 3) (Ziemert, Ishida et al. 2010). Microviridins are ribosomally synthesized tricyclic depsipeptide proteases inhibitors produced by a number of cyanobacteria. The discovery of both the

new microviridin and patellamide-like precursor peptides should aid in attempts to generate additional members of these two important cyclic peptide families.

Most homology-based screens have been carried out using degenerate PCR primers designed to recognize conserved sequences within secondary metabolite biosynthetic genes. As more metagenomic sequencing data appears in publicly available databases, it should also be possible to use purely bioinformatics based search strategies to identify new natural product biosynthetic enzymes and gene clusters. An *in silico* examination of data from the Global Ocean Metagenomic Survey for genes involved in the biosynthesis of lantibiotic type antibiotics uncovered more than 20 novel lantibiotic cyclases (Li, Sher et al. 2010). This class of peptide cyclase is used in the biosynthesis of potent cyclic peptide antibiotics (lantibiotics) from short linear ribosomally synthesized peptides. Environmental DNA derived lantibiotic cyclases could one day aid in the enzymatic synthesis of new lantibiotic variants. The examination of eDNA libraries and metagenomic sequencing data for relatives of known biosynthetic systems is likely to be a generally applicable strategy for identifying new structural variants of many bacterially derived antibiotics, potentially providing ready access to compounds with improved pharmacological properties and improved spectra of activity.

1.4 Library Enrichment Strategies

The development of generic gene enrichment strategies would undoubtedly simplify the screening of large eDNA libraries and likely increase the utility of metagenomics as a tool for the discovery of novel bioactive small molecules. A number of DNA hybridization strategies, including subtractive hybridization, PCR denaturing gradient gel electrophoresis (PCR-DGGE), "biopanning" and fluorescence in-situ hybridization coupled with cell sorting have been explored for enriching eDNA samples for genes of interests with varying degrees of success (Gray, Richardson et al. 2003; Kalyuzhnaya, Zabinsky et al. 2006; Quinton, Stephanie et al. 2007; Chew and Holmes 2009; Morimoto and Fujii 2009). Zhang, et al. recently reported a phage display-based strategy for specifically enriching eDNA libraries for clones containing two important classes of natural product biosynthetic genes, polyketide synthases and non-ribosomal peptide synthetases (NRPS) (Zhang, He et al. 2009). Their selection strategy takes advantage of the fact that PKS and NRPS proteins are posttranslationally modified with the addition of a phosphopantetheine prosthetic group. Phage that display either PKS or NRPS proteins on their surface could therefore be collected by first incubating the phage library with a recombinant phosphopantetheinyltransferase and a biotinylated phosphopantetheine analog, and then panning with streptavidin. At the moment, complete biosynthetic gene clusters are not accessible using this

strategy. It does, however, provide a promising method for recovering individual NPRS and PKS megasynthases from complex eDNA samples.

CHAPTER 2

2 Homology-based metagenomic efforts towards the production of novel glycopeptide antibiotics

2.1 Chapter Summary

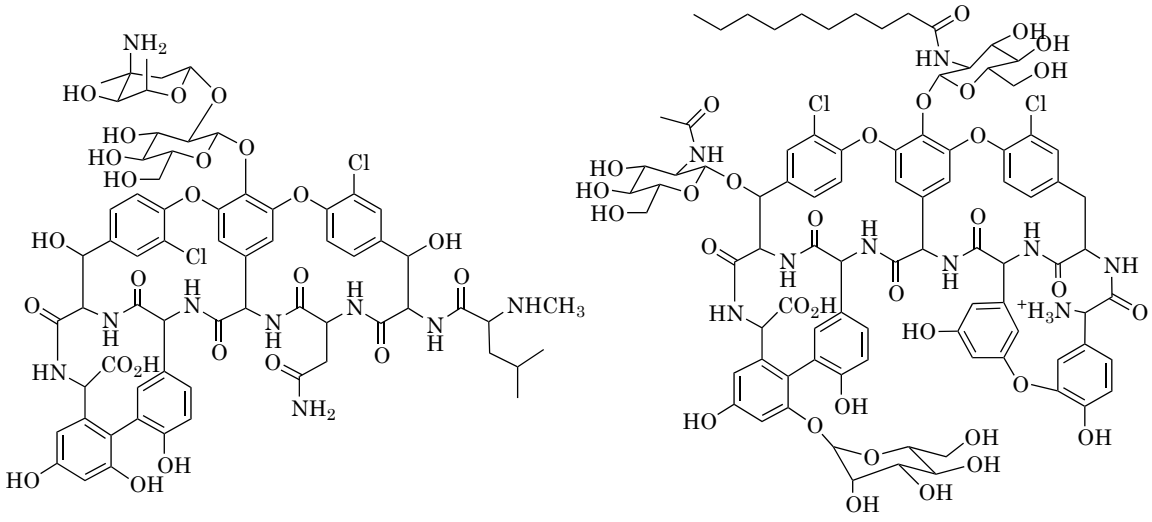
Homology-based metagenomics efforts have the potential to identify novel biosynthetic pathways related to any previously sequenced biosynthetic gene cluster. This chapter details the application of homology-based metagenomics to the identification of novel glycopeptide antibiotic biosynthetic gene clusters from three eDNA-derived cosmid mega-libraries. Using degenerate oligonucleotides designed to specifically target a biosynthetic enzyme unique to glycopeptide biosynthesis, *oxyC*, a preliminary screen was carried out to assess the prevalence of *oxyC* in a range of eDNA samples from varied geographic locations. This screen, which used 11 eDNA samples from three different continents as PCR template, discovered *oxyC* homologs in every sample examined.

After the preliminary screen, three eDNA cosmid mega-libraries, each containing in excess of ten million unique members, were screened using the same degenerate PCR primers from the preliminary screen. This screen led to the identification, recovery, and sequencing of six novel glycopeptide

biosynthetic pathways. These pathways were subsequently used to produce fifteen novel glycopeptide congeners, through a combination of *in vitro* and *in vivo* techniques. These compounds maintained their antibiotic activity against *Staphylococcus aureus* and *Enterococcus faecalis* in antibacterial test assays.

2.2 Introduction

Glycopeptide antibiotics are soil-bacterially derived secondary metabolites, which display potent bacteriostatic activity against a range of Gram-positive bacteria. Two of the most commonly employed glycopeptides, vancomycin (**15**) and teicoplanin (**16**), are widely used to treat nosocomial Gram-positive bacterial infections in humans (Figure 4). More recently, vancomycin has been quite heavily administered to treat the emergence of *Staphylococcus aureus*, which displays resistance to the more commonly used beta-lactam antibiotics (methicillin-resistance *S.aureus,* MRSA). However, the emergence of strains of *S. aureus* and *Enterococci* resitant to vancomycin (vancomycin-resistant *S. aureus*, VRSA and vancomycin-resistant *Enterococci,* VRE, respectively) threaten to render glycopeptides obsolete. Recent efforts in glycopeptide discovery have focused primarily on semi-synthetic efforts, as the identification of novel glycopeptide congeners from soil bacteria has waned over the last few decades.

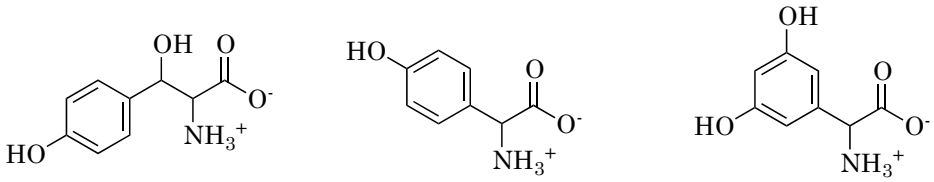


15 16

Figure 4: Vancomycin (15) and teicoplanin (16)

There are over 150 known glycopeptides isolated from bacterial cultures, yet there are only four different heptapeptide scaffolds seen in glycopeptide biosynthesis. These scaffolds are initially biosynthesized by non-ribosomal peptide synthetases (NRPS), large, multimodular enzymes, which produce linear peptides. The sequence of the peptide produced by the NRPS is determined by the amino acid sequence of the NRPS itself (Stachelhaus, Mootz et al. 1999). Glycopeptides often contain pathway encoded, non-canonical amino acids, such as β-hydroxytyrosine (**17**), 4-hydroxyphenylglycine (**18**), or 3,5-dihydroxyphenylglycine (**19**) (Figure 5). These linear heptapeptides are subsequently oxidatively cross-linked, to form either three or four macrocycles. The majority of glycopeptide diversity is provided by an extensive repertoire of tailoring enzymes, which site-

specifically perform an array of biosynthetic modifications, including, but not limited to, glycosylation, acylation, methylation, and sulfonation. This chapter describes the homology-based metagenomics search for novel glycopeptide biosynthetic gene clusters, and the subsequent use of enzymes found in these clusters to produce novel glycopeptide congeners both *in vitro* and *in vivo*.



17 **18** **19**

Figure 5: Non-canonical amino acids common in glycopeptide biosynthesis. β-hydroxytyrosine (BHT) (**17**), hydroxyphenylglycine (HPG) (**18**), and di-hydroxyphenylglycine (DPG) (**19**).

2.3 Results

2.3.1 Crude eDNA Screening and eDNA-derived Cosmid Mega-Library Construction

To identify potentially novel glycopeptide biosynthetic gene clusters from environmental samples using a homology-based metagenomics approach first required identification of a gene unique to glycopeptide biosynthesis, that would limit background while not excluding possible pathway hits. The selected target gene encodes the production of an oxidative enzyme, whose

activity is responsible for the direct C-C coupling of the HPG at position 5 and the DPG at position 7 during macrocycle formation in heptapeptide maturation (Figure 6). This gene, *oxyC*, is common to all sequenced glycopeptide biosynthetic gene clusters. Importantly, all known *oxyC* genes display greater than 75% DNA sequence homology to each other, yet are easily distinguishable from genes encoding other oxidative enzymes. This high level of homology allowed for the design of degenerate oligonucleotides to specificaly target oxyC using a PCR based screen.

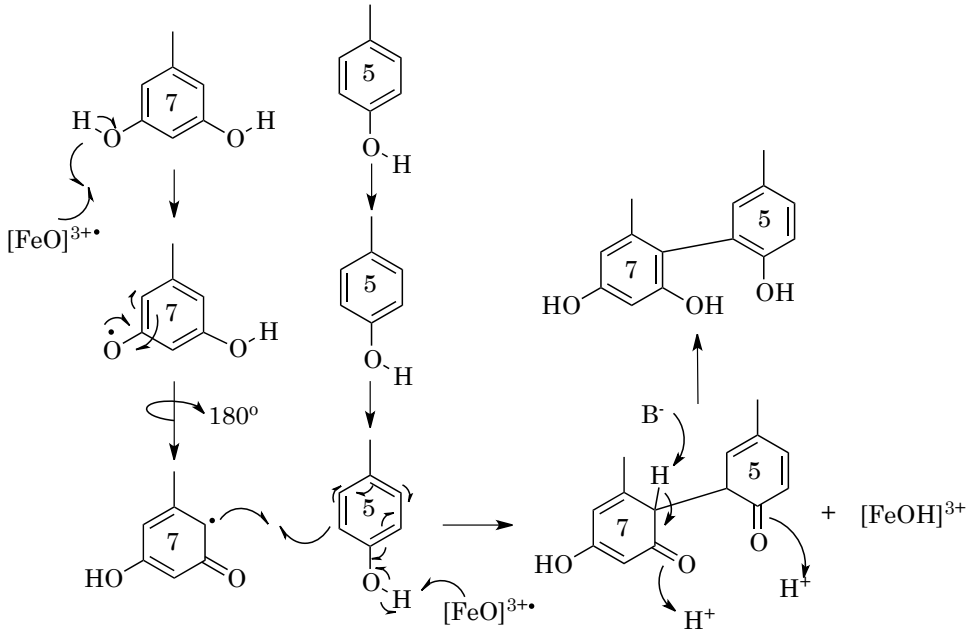


Figure 6: A possible mechanism for the C-C coupling reaction performed by the P450 mono-oxygenase *oxyC*

The first investigation using these PCR probes was carried out on a group of geographically diverse, crude eDNA samples. These samples were

collected from several locations in Africa, Central America, and North America. Nested primers were designed to specifically target these genes in these extremely complex mixtures. These primers identified novel *oxyC*-like genes in every sample examined (Figure 7). However, it would take a significant increase in the efficiency of eDNA cosmid cloning to transition from the amplification of gene segments to the recovery of intact glycopeptide gene clusters.

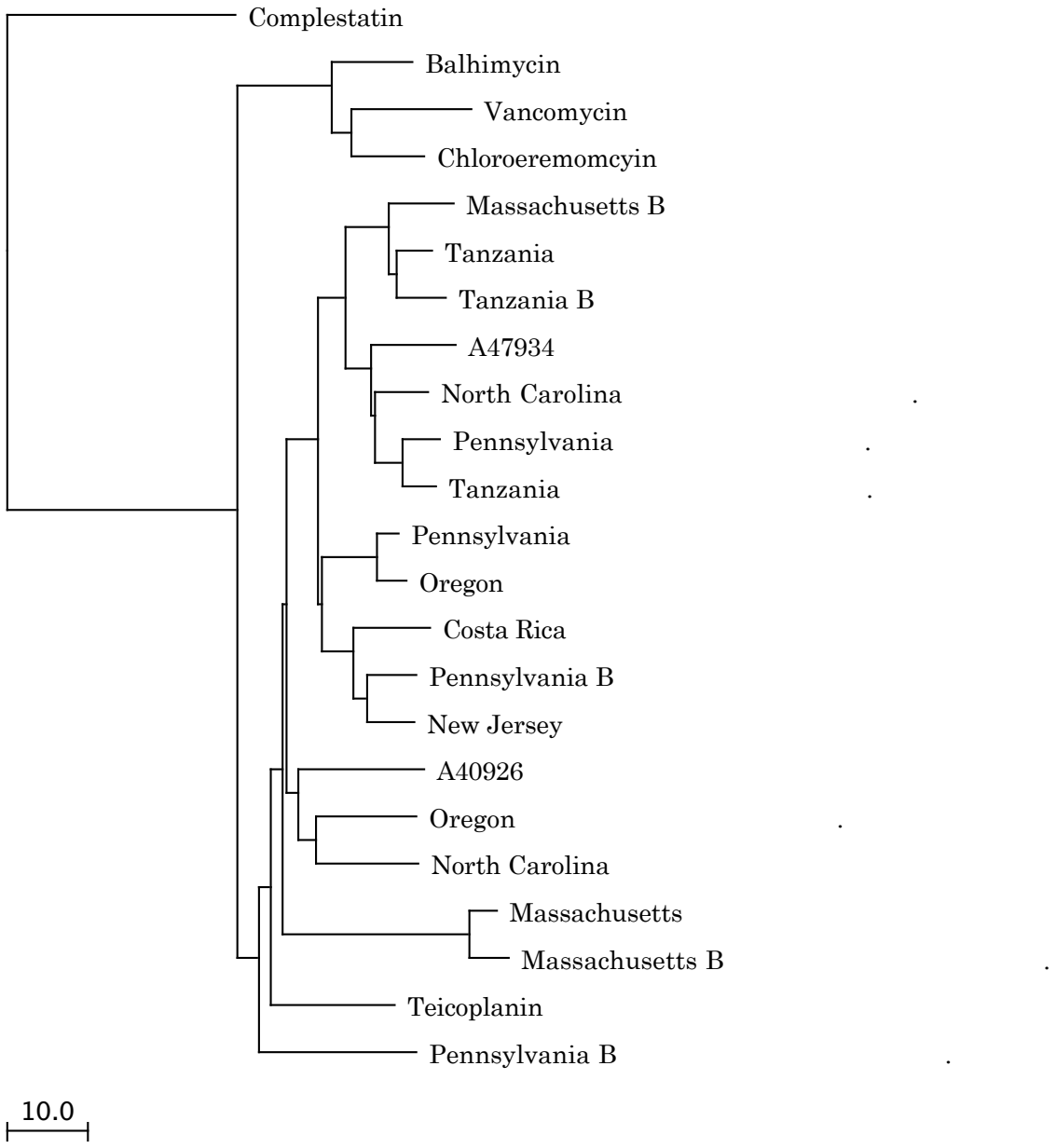


Figure 7: Phylogenetic tree of *oxyC* amplicons identified from crude eDNA samples
Phylogenetic tree of *oxyC* amplicons amplified from crude eDNA samples. Trimmed sequences for several known *oxyC* genes are also included for comparison, as well as an oxidative enzyme from the biosynthesis of a compound closely related to glycopeptides, complestatin for an outgroup.

It was hypothesized that if a large enough collection of eDNA clones could be constructed from a single environmental sample, complete glycopeptide biosynthetic gene clusters could be identified on overlapping clones, and recovered for sequencing. However, at that time, the largest eDNA-derived cosmid libraries were comprised of at most several thousand clones (Courtois, Cappellano et al. 2003). It was elsewhere determined that seqeunce redundancy was not observed until a cosmid library contained in excess of one million clones (Kim, Feng et al. 2010). Glycopeptide biosynthetic gene clusters are generally in excess of 70,000 bp, which would require multiple overlapping cosmid clones to recover in totality. The recovery of intact glycopeptide biosynthetic gene clusters would require a library several orders of magnitude larger than those previously constructed. Through trial and error, it was determined that the eDNA purified from samples, which are much less rich in organic matter, e.g. desert soils displays a much higher cloning efficiency than soils very rich in organic matter, e.g. soil from a deciduous forest. This discovery led to the construction of multiple eDNA-derived cosmid "mega-libraries," which each contained in excess of ten million individual clones, the equivalent of 100,000 4-mb bacterial genomes.

To aid in screening, each library was constructed as a collection of 5,000 clone sublibrary pools. Each library contained in excess of 2,000 of these sublibraries, which were arrayed into 8x8 grids, again to aid in library screening. Each 8x8 grid was pooled in two ways: all sixty-four sublibraries in

one aliquot, termed the "grand pool" and all eight sublibraries in one row pooled in one aliquot. PCR screening proceeded through an initial screen of all grand pools. A hit in a grand pool, identified by PCR screening and subsequent amplicon seqeuencing, led to subsequent screening of the pertinent rows of pooled sublibraries, followed by screening of the eight individual sublibraries. Once a single sublibrary was identified as containing a clone of interest, that sublibrary was serially diluted over several rounds to isolate one individual clone from the entire sublibrary aliquot of over 5,000 clones.

2.3.2 eDNA-derived Cosmid Mega-Library Screening and Glycopeptide Biosynthetic Gene Cluster Recovery

Using the same *oxyC*-based degenerate primers, PCR screening was carried out on all of the grand pools from each of three cosmid mega-libraries, one constructed from soil collected from a desert in Utah (UT), one constructed from soil collected from the Anza-Borrego desert in California (AB), and a third library constructed from soil collected from a desert in Arizona (AZ). The screening of the UT cosmid mega-library led to the identification of two *oxyC*-like hits (UTA15 and UTD30). These eDNA-derived hits were then added to the previously known sequences to design a second set of degenerate PCR primers. The screening of the AB and AZ libraries with both sets of degenerate primers led to the identification of three *oxyC*-like hits in the AB library (AB37, AB878, and AB915) and one *oxyC*-like hit in the

AZ library (AZ205). The cosmid clones containing these *oxyC*-like sequences were recovered and sequenced using 454 pyrosequencing (Roche). A phylogenetic analysis of the full-length *oxyC* sequences from these recovered clones revealed that the *oxyC* sequences recovered using culture-independent methods appeared as diverse as those recovered from cultured bacteria (Figure 8).

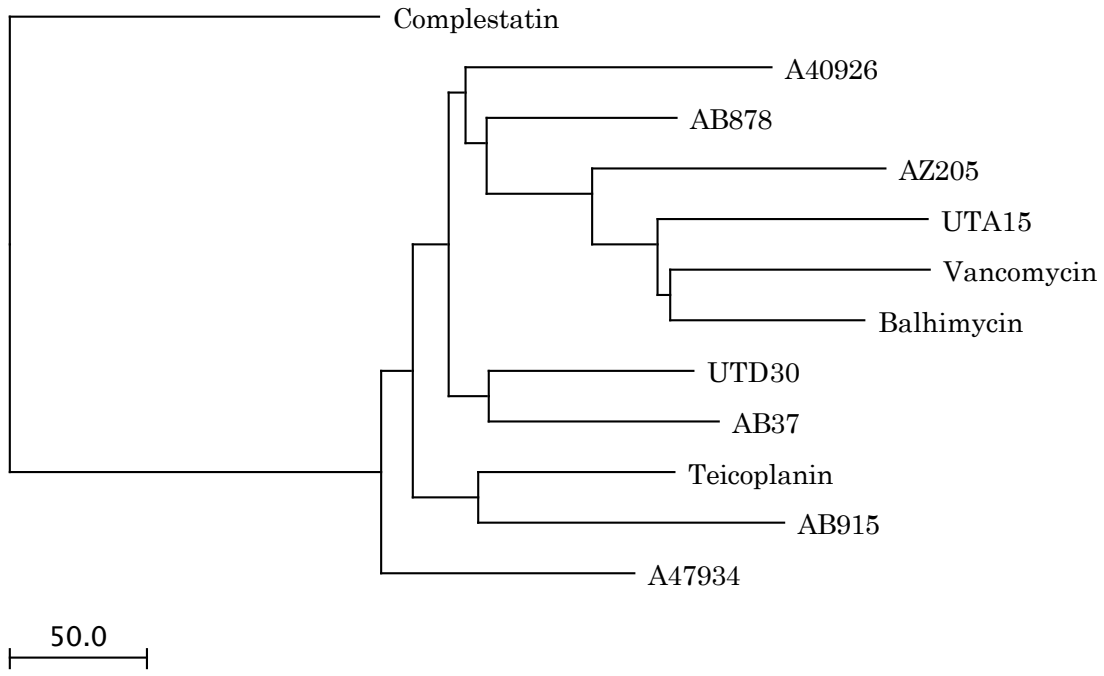


Figure 8: Phylogenetic analysis of sequenced *oxyC* genes

oxyC sequences identified from cultured bacteria (from the A40926, A47934, Balhimycin, Teicoplanin, and Vancomycin biosynthetic gene clusters) and uncultured bacteria (from the UT, AB, and AZ eDNA-derived cosmid mega-libraries). Again, an oxidative enzyme from the pathway encoding a compound related to glycopeptides, complestatin, has been included as an outgroup.

Upon successful recovery and sequencing of the eDNA clones containing *oxyC* genes, new primers, specific to the ends of the eDNA insert for each cosmid clone containing glycopeptide biosynthetic machinery, were designed to identify and recover overlapping clones to recover and sequence entire gene clusters. This process of recovery and sequencing was repeated until the pathway was predicted to be complete, or no overlapping clone could be found.

2.3.3 Glycopeptide Pathway Annotation and Sequence Analysis

It was predicted that five of the six glycopeptide pathways identified in these megalibrary screens are complete, with one, the UTD30 pathway, being truncated due to the absence of overlapping clones in the UT mega-library. The UT library was the first library screened using the degenerate *oxyC* specific primers. Of the two *oxyC* sequences found in the UT library, one (UTA15) is more closely related to those found in sequenced vancomycin-like gene clusters (vancomycin-like eDNA derived gene cluster, VEG cluster), whereas the other (UTD30) resembles those found in sequenced teicoplanin-like gene clusters (teicoplanin-like eDNA derived gene cluster, TEG cluster). Cosmids containing the more common VEG pathway appeared in approximately 1 of every two million clones, whereas clones containing the less common TEG pathway were found in one of every 5–10 million clones. For the common pathway, ~100 kb of continuous sequence were

reconstructed from 6 overlapping cosmid clones. Almost 75 kb of this sequence are predicted to be part of a glycopeptide antibiotic biosynthetic gene cluster. For the less common TEG pathway, ~50 kb of continuous sequence were reconstructed from three overlapping cosmid clones. Although the TEG gene cluster is truncated at one end, a comparison of the two gene clusters, which are closely related at both ends but differ significantly in the central finishing enzyme regions, suggests that the missing region is likely to contain conserved biosynthetic machinery that is common to most glycopeptide gene clusters.

The clusters found in the AB and AZ clusters were well represented in these mega-libraries, which aided significantly in the complete recovery of all pathways. After full sequencing and annotation, each set of overlapping cosmid clones could be assembled into a single 100-120 kb contig that appears to contain a full glycopeptide biosynthetic gene cluster (Figure 9) All three gene clusters recovered from the AB library and the one gene cluster recovered from the AZ library are predicted to contain a complete complement of glycopeptide biosynthetic genes and show the same conserved gene cluster architecture that is seen in most previously sequenced glycopepetide gene clusters (Figure 9). Detailed description of all ORFs in the six eDNA-derived glycopeptide gene clusters are provided in the Appendix (Tables 7-12).

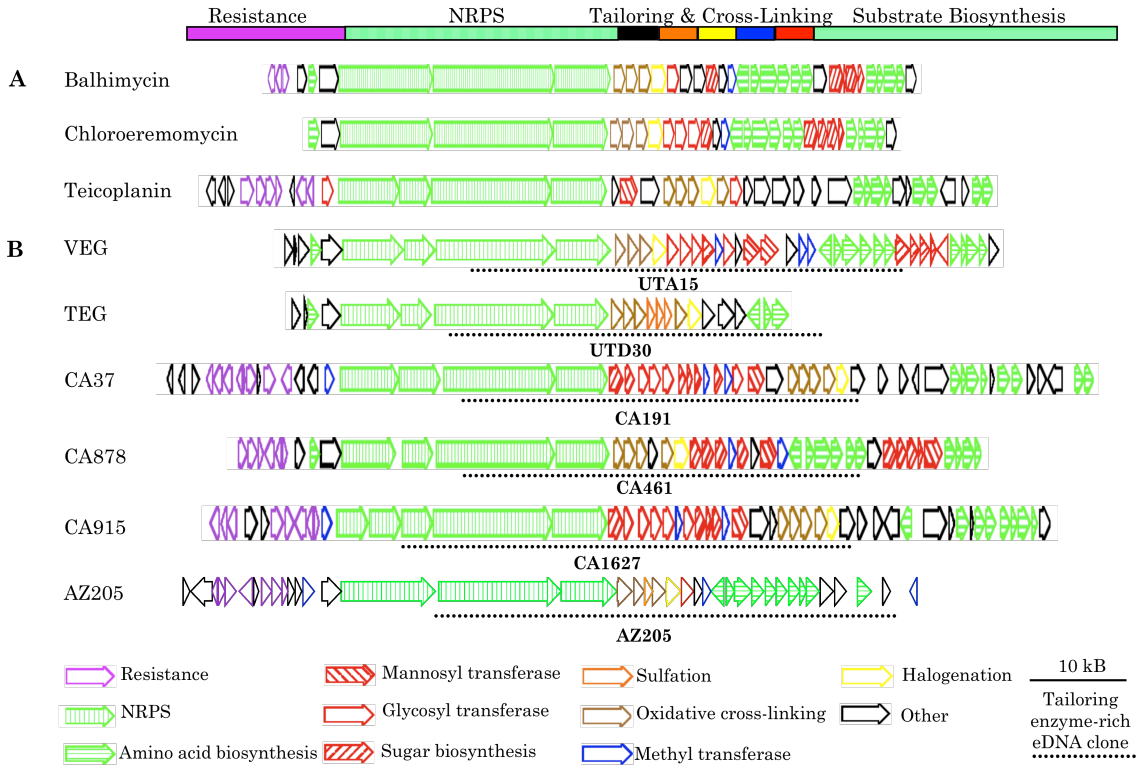


Figure 9: Annotated glycopeptide biosynthetic gene clusters

A) Annotated pathways for select glycopeptides originally isolated from cultured soil bacteria. B) Six glycopeptide clusters that were recovered as a part of the work detailed in this thesis using culture-independent methods

NRPS modules are predicted to carry out a single condensation reaction that extends the growing peptide by one amino acid (Marahiel, Stachelhaus et al. 1997) At a minimum, a module contains a condensation domain for condensing an incoming amino acid with the growing peptide, an adenylation domain for selecting and activating the incoming amino acid and a thiolation domain that carries the growing peptide. The amino acid incorporated by a module is determined by a small number of variable

residues found in the adenylation domain (Table 1) (Stachelhaus, Mootz et al. 1999). Table 1 shows these key amino acid specificity residues from each adenylation domain found in the six eDNA-derived glycopeptide biosynthetic pathways.

A detailed analysis of the six eDNA derived gene clusters indicates that each is predicted to encode the biosynthesis of a previously unknown glycopeptide congener. As expected, all clusters contain large NRPS systems that are predicted to encode heptapeptides (Figure 9). The NRPS system in the VEG pathway is predicted to encode the biosynthesis of a Type II heptapeptide, HPG-BHT/TYR-HPG-HPG-HPG-BHT/TYR-DPG, with only three predicted macrocyclizations and the TEG pathway is predicted to encode the biosynthesis of a Type III heptapeptide HPG-BHT/TYR-HPG-HPG-HPG-BHT/TYR-DPG with four predicted oxidative macrocyclizations (Table 1). Similary, based on a comparison of these residues to those seen in adenylation domains from sequenced glycopeptide NRPS modules, all three AB gene clusters are expected to encode the production of a Type III HPG-BHT/TYR-HPG-HPG-HPG-BHT/TYR-DPG-like heptapeptide (Table 1). The pathway recovered from the AZ library is simlar to the VEG pathway, with a predicted Type II, HPG-BHT/TYR-HPG-HPG-HPG-BHT/TYR-DPG backbone with only three predicted oxidative macrocyclizations. Potential ambiguities in the predicted peptide arise from the mode of BHT incorporation, which has been reported to occur by two distinct mechanisms during glycopeptide

biosynthesis. BHT can be biosynthesized and incorporated directly into the glycopeptide backbone through BHT-specific adenylation domains, or it can be produced by oxidizing tyrosine in a reaction that takes place after the selective incorporation of tyrosine into the growing peptide (Puk, Bischoff et al. 2004; Stinchi, Carrano et al. 2006). Both AB37 and AB915 are predicted to generate BHT through the latter, tyrosine-oxidation strategy. NRPS modules 2 and 6 from both gene clusters are predicted to incorporate TYR into the growing peptide. However, as can be seen in a similar analysis of both the teicoplanin and A47934 NRPS modules, the final residues at these positions will be BHT if the tyrosines are subsequently oxidized.

Table 1: NRPS adenylation domain predictions

(* indicates possible side of ambiguity, based on biosynthesis of BHT and TYR)

	Binding Pocket	**Substrate Prediction**
TEG Pathway		
Module 1	DAFSLGLL	DAFLLGLL (A47934-M1-HPG)
Module 2	DTSKVAAI	DTSKVAAI (Cep-M2-BHT)*
Module 3	DPFFQGTF	DPYLGGT (Cep-M7-DPG)
Module 4	DIFLLGLL	DIFHLGLL (Cep-M4-HPG)
Module 5	DALLLGLL	DAVHLGLL (Cep-M5-HPG)
Module 6	DASTLGAI	DASTLGAI (Cep-M6-BHT)*
Module 7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)
VEG Pathway		
Module 1	DAFTLGLL	DAFLLGLL (Tcp-M1-HPG)
Module 2	DTSKVAAI	DTSKVAAI (Cep-M2-BHT)*
Module 3	DAFSLGLL	DIFHLGLL (Cep-M4-HPG)
Module 4	DIFLLGLL	DIFHLGLL (Cep-M4-HPG)
Module 5	DAVLLGIG	DIFHLGLL (Cep-M4-HPG)
Module 6	DASTLGAI	DASTLGAI (Cep-M6-BHT)*
Module 7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)
AB 37 Pathway		
Module1	DAFLLGLL	DAFLLGLL (Tcp-M1-HPG)
Module2	DASTVAAV	DASTVAAV (A47934-M2-TYR)*
Module3	DAYFSGSL	DAYFLGTL (Tcp-M3-DPG)

Module4	DIFTLGLL	DIFHLGLL (Cep-M4-HPG)
Module5	DALLLGLL	DALLLGVG (Tcp-M5-HPG)
Module6	DASTVAAV	DASTVAAV (A47394-M2-TYR)*
Module7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)

AB 878 Pathway

Module1	DAFNLGLL	DAFLLGLL (Tcp-M1-HPG)
Module2	DTSKVAAI	DTSKVAAI (Cep-M2-BHT)*
Module3	DAYFQGTF	DAYFLGTL (Tcp-M3-DPG)
Module4	DIFTLGLL	DIFHLGLL (Cep-M4-HPG)
Module5	DAVLLGLL	DAVHLGLL (Cep-M5-HPG)
Module6	DASTLGAI	DASTLGAI (Balhi-M6-BHT)*
Module7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)

AB 915 Pathway

Module1	DAFSLGLL	DAFLLGLL (Tcp-M1-HPG)
Module2	DASTVAAV	DASTVAAV (A47934-M2-TYR)*
Module3	DAFFSGSL	DAFLLGLL (A47934-M1-HPG)
Module4	DIFLLGLL	DIFHLGLL (Cep-M4-HPG)
Module5	DALLLGLL	DALLLGVG (Tcp-M5-HPG)
Module6	DASTIAAV	DASTIAGV (Tcp-M6-BHT)*
Module7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)

AZ 205 Pathway

Module1	DAFGQGLV	DAFLLGLL (Tcp-M1-HPG)
Module2	DASTVAAV	DASTVAAV (A47934-M2-TYR)*
Module3	DVLLVGTI	DALLLGVG (Tcp-M5-HPG)
Module4	DIFTLGLL	DIFHLGLL (Cep-M4-HPG)
Module5	DALLLGLL	DAFLLGLL (Tcp-M1-HPG)
Module6	DASTLGAI	DASTLGAI (Balhi-M6-BHT)*
Module7	DPYLGGTL	DPYLGGTL (Cep-M7-DPG)

The VEG cluster contains a complete complement of the genes predicted to be necessary for the biosynthesis of an oxidatively cross-linked heptapeptide. This includes all of the biosynthetic machinery required for the production of the three non-canonical amino acids (HPG, BHT and DPG) found in the peptide core, a halogenase for the production of chloro-BHT, and the three oxidative enzymes that oxidatively cross-link the heptapeptide core (Figure 9) (Choroba 2000; Hubbard, Thomas et al. 2000; Chen, Tseng et al.

2001; Puk, Bischoff et al. 2004; Sosio, Kloosterman et al. 2004; Stinchi, Carrano et al. 2006; Sri, Ellen et al. 2009). In addition to the core biosynthetic machinery, the VEG cluster contains seven glycosyltransferases and three methyltransferases. Two of the three methyltransferases are predicted N-methyltransferases suggesting that the product of the VEG pathway is doubly N-methylated at the N-terminus.

Four of the glycosyltransferases in the VEG gene cluster are related to glycosyltransferases from chloroeremomycin biosynthesis, two are related to putative mannosyltransferases from teicoplanin-like biosynthetic gene clusters, and the seventh is related to the glycosyltransferase CalG1 from calicheamycin biosynthesis. Two of the related glycosyltransferases from chloroeremomycin biosynthesis use derivatives of vancosamine as substrates. Homologs of the five enzymes required for the biosynthesis of a vancosamine derivative are also found in the VEG gene cluster (Figure 9) (Chen, Thomas et al. 2000). To the best of our knowledge, no glycopeptide congener containing the unique assortment of functionality encoded by the VEG gene cluster (doubly N-methylated, vancosamine functionalized, Type II glycopeptide) has been characterized from cultured bacteria.

While the VEG gene cluster is rich in methyl- and glycosyltransferases, the TEG gene cluster is rich in sulfotransferases (Figure 9). In addition to containing biosynthetic machinery that is required for the production of an oxidatively crosslinked heptapeptide core, the TEG gene

cluster contains three predicted sulfotransferases. Although a large number of glycosylated, halogenated, and alkylated glycopeptide congeners have been identified from studying cultured bacteria, only two sulfated, and no polysulfated, glycopeptides have been reported (Boeck and Mertz 1986; Holdom 1988). The presence of three sulfotransferases in the TEG cluster suggested that it was likely to encode the biosynthesis of the first polysulfated glycopeptides.

The AB37 and AB915 gene clusters each contain genes encoding four glycosyltransferases, a mannosyltransferase, a halogenase, and three methyltransferases (Figure 9). Both gene clusters also contain genes encoding enzymes used in the biosynthesis of a number of different rare glycopeptide-associated sugars, including vancosamine, acosamine or ristosamine (Debono, Merkel et al. 1984; Heald, Mueller et al. 1987; Chen, Thomas et al. 2000). It is not possible from the DNA sequence alone to tell which of these sugars the two gene clusters will actually produce. Of the known glycopeptide congeners that contain five sugars, only actaplanin contains the same cross-linked heptapeptide that is predicted to arise from the AB37 and AB915 gene clusters (Debono, Merkel et al. 1984). These gene clusters may therefore encode the biosynthesis of actaplanin itself or more highly glycosylated and methylated actaplanin-like glycopeptides.

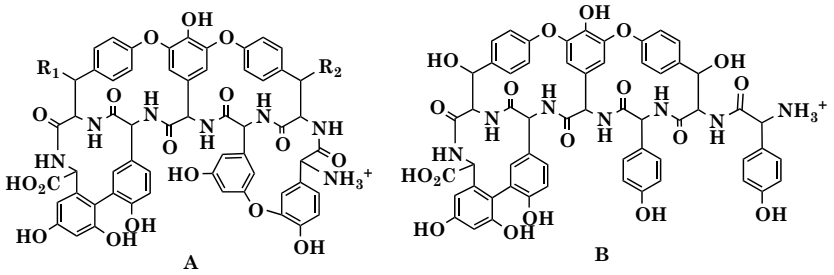
The AB878 gene cluster contains genes encoding four glycosyltransferases, two mannosyltransferases, a halogenase, and two

methyltransferases. This gene cluster also contains homologs of the same sugar biosynthetic genes that are seen in AB37 and AB915. Based on the collection of tailoring enzyme genes present in AB878, this gene cluster would appear to encode the biosynthesis of a ristocetin-like congener. Ristocetin contains the same heptapeptide as is predicted to arise from the AB878 gene cluster as well as six sugars, including a ristosamine, and two methyl substituents (Fehlner, Hutchinson et al. 1972). The presence of a halogenase in AB878 suggests that this cluster probably does not produce ristocetin itself, but instead produces a halogenated ristocetin-like congener.

The AZ205 cluster is a mixture of the methyl/glycosyl-transferase rich (VEG/AB37/AB878/AB915) pathways and the sulfotransferase-rich (TEG) pathway. It possesses all of the required resistance elements, NRPS genes, and substrate biosynthetic machinery, as well as a predicted N-methyltransferase, glycosyltransferase, and sulfotransferase. The predicted product of this pathway, an N-methylated, glycosylated, sulfonated type III glycopeptide, is different from any previously reported glycopeptide congener.

For sequenced gene clusters, the number and type of tailoring enzymes found in these clusters is a reliable predictor of the features that are seen in the most highly functionalized congener it produces (Figure 10) (Goldstein, Selva et al. 1987; van Wageningen, Kirkpatrick et al. 1998; Pelzer, Sussmuth et al. 1999; Pootoolal, Thomas et al. 2002; Sosio, Kloosterman et al. 2004). The inventory of tailoring enzymes found in cryptic eDNA-derived

glycopeptide gene clusters should in turn be a reasonable predictor of the features found in the congeners produced by these gene clusters. This type of analysis does not allow for precise structural predictions to be made because it does not account for differences in regioselectivity, subtle differences in substrate specificities or the functional order of the tailoring enzymes. It can, however, provide an indication as to whether or not a cryptic gene cluster has the potential to generate a glycopeptide congener with a unique number or type of prosthetic groups.



A B

Source	Gene Cluster	Peptide	Skeleton		Transferases						Halogenase
			R_1	R_2	Glycosyl	Mannosyl	Methyl	N-Methyl	Sulfo	Acyl	
Cultured	Teicoplanin	A	OH	H	2	1	0	0	0	1	2
	Cluster		H or OH	H or OH	2	1	0	0	0	1	1
	A47934	A	OH	H	0	0	0	0	1	0	3
	Cluster		H or OH	H or OH	0	0	0	0	1	0	2
	A40926	A	OH	H	1	1	0	1	0	1	2
	Cluster		H or OH	H or OH	1	1	0	1	0	1	1
eDNA	AB37	A	H or OH	H or OH	2	1	2	0	0	0	1
	AB915	A	H or OH	H or OH	4	1	2	0	0	0	1
	AB878	A	H or OH	H or OH	4	2	2	0	0	0	1
	TEG	A	H or OH	H or OH	0	0	0	0	3	0	1
	VEG	B	H or OH	H or OH	4	2	1	2	0	0	1
	AZ205	B	H or OH	H or OH	1	0	2	1	1	0	1

Figure 10: NRPS and Tailoring Enzyme Inventory

An inventory of the number and type of tailoring enzymes (or functional group transferases) found in sequenced gene clusters from cultured soil bacteria as well as gene clusters cloned from eDNA, is shown. For gene clusters associated with known metabolites, the number of tailoring-enzyme-derived functional groups found on the encoded metabolite is also shown

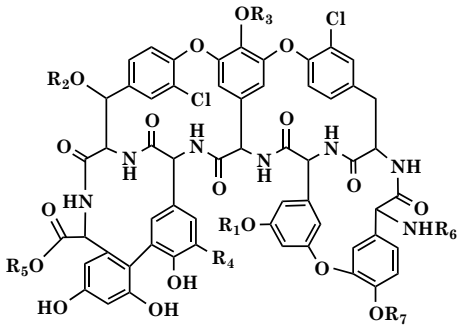
A similar detailed accounting of the enzymes in each of the eDNA-derived glycopeptide biosynthetic gene clusters allowed for a comparison of the predicted structures to known glycopeptide congeners (Figure 10). In addition to the seven NRPS modules that are predicted to produce the heptapeptide, each gene cluster encodes three (or four) predicted monooxygenases that are homologs of OxyA-C (or A-D) (Figure 9). OxyA-D

catalyze the four oxidative couplings that are needed to generate the four macrocycles seen in many glycopeptides (Pootoolal, Thomas et al. 2002).

2.3.4 Novel Glycopeptide Congener Production

The collection of novel sulfotransferases identified in the TEG pathway provided an opportunity to rapidly produce a considerable collection of glycopeptide congeners. The substrate upon which these sulfotransferases act was predicted based on the complete sequence of the NRPS genes recovered as part of the TEG pathway to be either the teicoplanin aglycone or the dehydroxylated teicoplanin aglycone. The teicoplanin aglycone is readily available and therefore we elected to use this substrate in *in vitro* studies with the TEG sulfotransferases. The three sulfotransferases (Teg12, 13, and 14) in the TEG cluster are homologs of StaL, a 3'-phosphoadenosine 5'-phosphosulfate (PAPS) dependent sulfotransferase found in the A47934 glycopeptide gene cluster (Pootoolal, Thomas et al. 2002). StaL has been used to generate monosulfonated glycopeptides *in vitro* (Lamb, Patel et al. 2006). For our *in vitro* glycopeptide sulfonation studies TEG12, 13, and 14) were PCR amplified and cloned into the pET28a expression vector. Each recombinant 6-HIS-tagged sulfotransferase was then Ni-affinity purified from cultures of *E. coli* BL21(DE3). In the presence of PAPS and the teicoplanin aglycone each predicted sulfotransferase produces a unique monosulfonated glycopeptide derivative, sulfo-teicoplanin aglycones A-C (**20-22**)(Figure 11). In reactions with two sulfotransferases, the three possible

disulfonated derivatives are formed (sulfo-teicoplanin aglycones D-F (**23-25**)), and in a reaction with all three sulfotransferases a tri-substituted derivative is produced, sulfo-teicoplanin aglycone G (**26)** (Figure 11).



-	R1	R2	R3	R4	R5	R6	R7	
20	SO3H	H	H	H	H	H	H	Teicoplanin Aglycone + TEG12
21	H	SO3H	H	H	H	H	H	Teicoplanin Aglycone + TEG13
22	H	H	SO3H	H	H	H	H	Teicoplanin Aglycone + TEG14
23	SO3H	SO3H	H	H	H	H	H	Teicoplanin Aglycone + TEG12,13
24	SO3H	H	SO3H	H	H	H	H	Teicoplanin Aglycone + TEG12,14
25	H	SO3H	SO3H	H	H	H	H	Teicoplanin Aglycone + TEG13,14
26	SO3H	SO3H	SO3H	H	H	H	H	Teicoplanin Aglycone + TEG12,13,14
29	H	H	H	Cl	CH3	H	SO3H	A47934 + A15 Cosmid
30	H	H	SO3H	Cl	H	H	SO3H	A47934 + D30 Cosmid
31	H	H	Glucose	Cl	H	H	SO3H	A47934 + AB191 Cosmid
32	H	H	H	Cl	H	CH3	SO3H	A47934 + AZ205 Cosmid
33	H	SO3H	H	Cl	CH3	H	SO3H	A47934 + AZ205 Sulf
34	H	SO3H	SO3H	Cl	H	H	SO3H	A47934 + D30 Cosmid + AZ205 Sulf
35	H	SO3H	H	Cl	H	H	SO3H	A47934 + AB191 Cosmid + AZ205 Sulf
36	H	SO3H	H	Cl	H	CH3	SO3H	A47934 + AZ205 Cosmid + AZ205 Sulf

Figure 11: New sulfonated glycopeptide derivatives produced using combinations of *in vivo* and *in vitro* methods using eDNA-derived tailoring enzymes

Mass spectrometry and 1 and 2D NMR were used to identify the sulfonation site in each monosulfonated product (Figures 20-31 in the Appendix) The sulfonation patterns seen in the di- and tri- sulfonated teicoplanin aglycone analogs were then inferred from the sulfonation specificities of the sulfotransferases used to synthesize these derivatives. Upon fragmentation by negative ion ESI-MS/MS, each of the monosulfonated

aglycone derivatives produces a daughter ion with an m/z=906. This fragment, which is not produced by the teicoplainin aglycone (Figure 29 in the Appendix), corresponds to a sulfonated product that has lost the macrocycle formed by amino acids five and seven. The presence of this sulfonated fragment eliminates three of the seven hydroxyls on the teicoplainin aglycone as candidates for sulfonation. To determine which of the four remaining hydroxyls are sulfonated, HPLC purified sulfo-teicoplanin aglycones, derived from milligram scale sulfonation reactions, were analyzed by 1 and 2D NMR, and these spectra were compared to those obtained with the teicoplanin aglycone (Tables 13-16 in the Appendix) (Malabarba, Ferrari et al. 1986; Boger, Weng et al. 2000). Compared to their hydroxylated counterparts, sulfonated compounds show significant deshielding (~0.5 ppm) of aromatic protons ortho to the site of sulfonation as well as deshielding of aliphatic protons that are attached to a sulfonated carbon. (Boeck and Mertz 1986; Pretsch, Bühlmann et al. 2000; Yates, Santini et al. 2000) Sulfonation of the hydroxyls on the amino acids at positions three, six and four by TEG12, 13, and 14 respectively, could be inferred from the sulfate induced deshielding of protons present in the ^{1}H NMR spectra of sulfo-teicoplanin aglycones A, B and C (Figures 19, 24, and 27, respectively in the Appendix). In sulfo-teicoplanin aglycone A (the TEG12 product), the two protons ortho to the hydroxyl on the resorcinol side chain of amino acid three are deshielded by 0.51 and 0.57 ppm relative to their chemical shifts in the aglycone. In

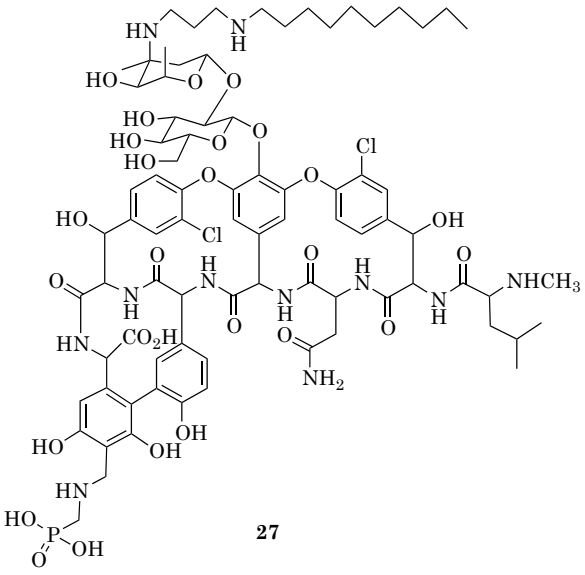
sulfo-teicoplanin aglycone B (the TEG13 product), the beta carbon proton of amino acid six is deshielded by 0.59 ppm relative to the aglycone. In sulfo-teicoplanin aglycone C (the TEG14 product), no protons ortho to a potential sulfonation site are significantly deshielded compared to the aglycone. The hydroxyphenylglycine at position four is the only potential sulfation site that is not predicted to undergo significant proton chemical shifts changes upon sulfonation. The cup shaped conformation of glycopeptides places a sulfate at position four in close proximity to the protons on amino acid two, which could explain the 0.43 ppm deshielding of the proton meta to the chlorine on amino acid two in sulfo-teicoplanin aglycone C. After literature examination, it is believed that all three sites sulfonated by the TEG sulfotransferases differ from those sulfonated in known glycopeptide congeners.

In antibacterial assays run against a wild-type (ATCC 6538P) and methicillin resistant (USA300) *Staphylococcus aureus*, the sulfated teicoplanin aglycone analogs show similar activity to vancomycin and teicoplanin (Sieradzki, Leski et al. 2003; Diep, Gill et al. 2006). The general trend observed within this family of metabolites is that each successive sulfate addition increases the MIC slightly (Table 2). While sulfates have rarely been seen in naturally occurring glycopeptide antibiotics, a related negatively charged substituent, a phosphono group appears on the DPG at position seven in telavancin (**27**), a second-generation semisynthetic glycopeptide antibiotic (Figure 12) (Higgins, Chang et al. 2005). The

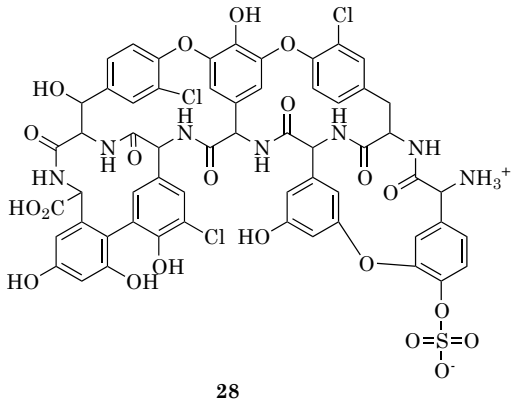
phosphono group, which likely mimics the naturally occurring sulfate functionality that has been seen in a small number of known glycopeptide congeners, is reported to markedly improve the ADME (absorption, distribution, metabolism, and excretion) characteristics of telavancin.

Table 2: MIC (μg/mL) for Compounds 20-26, 28-36, and the Teicoplanin Aglycone

Compound	*S. aureus*		*E. faecalis*	
	ATCC 6538P	USA300	ATCC 47077	EF18
20	2	2	16	>16
21	1	1	16	>16
22	3	4	>16	>16
23	3	2	>16	>16
24	8	8	>16	>16
25	8	8	>16	>16
26	16	8	>16	>16
28	2	2	16	>16
29	1	1	8	>16
30	3	8	>16	>16
31	3	8	>16	>16
32	2	2	8	>16
33	2	4	16	>16
34	16	16	>16	>16
35	8	16	16	>16
36	2	4	16	>16
teicoplanin aglycone	1	1	8	>16



27



28

Figure 12: Telavancin (27) and A47934 (28)

An additional discovery, which came out of the glycopeptide pathway sequencing and annotation was the observation of the high level of conservation in the ordering of genes in the eDNA-derived pathways (Figure 9). There even appeared to be conserved regions, where, for example, genes conferring self-resistance are clustered together, in the same relative location from pathway to pathway. This held true for not only resistance elements, but also for the large NRPS systems, the genes responsible for the biosynthesis of non-canonical amino acids, as well as the genes encoding the "tailoring enzymes," the proteins responsible for the majority of the glycopeptide chemical diversity. These tailoring enzymes convert the linear heptapeptide precursor peptide to the fully cross-linked, modified active product. Therefore, the sequencing data not only provided information about

the types and numbers of transformations carried out in each pathway, but also provided insight into the conservation of these pathways, which proved ultimately very useful for the production of novel glycopeptide congeners. With such an arrangement, it is possible to capture large portions of glycopeptide tailoring enzyme systems on single eDNA derived cosmid clones. The genes found on these clones would be useful for generating glycopeptides with novel functionalization patterns.

Due to the dearth of naturally occurring anionic glycopeptides characterized to date, and the promising improvement in ADME profile in telavancin by the addition of a group which somewhat mimics the sulfate seen in naturally derived anionic glycopeptides, the possibility of using the genes found on the six eDNA derived tailoring enzyme-rich clones to generate additional sulfonated glycopeptides using both *in vivo* and *in vitro* derivatization strategies was explored. For the *in vivo* derivatization studies, each tailoring enzyme-rich clone was initially retrofitted with the genetic elements necessary for transfer and integration into *Streptomyces* and then transferred by conjugation into *Streptomyces toyocaensis* (Matsushima and Baltz 1996). *S. toyocaensis* is the natural producer of the monosulfated glycopeptide congener A47934 (**28**) (Figure 12) (Boeck and Mertz 1986). With the exception of the sulfate found on the N-terminal HPG, A47934 is devoid of tailoring enzyme-derived functionality. Thus, A47934 provides numerous sites where new functionality could be added by eDNA derived tailoring

enzyme genes introduced into *S. toyocaensis*. While it is unlikely that all of the tailoring enzymes from any one clone would be functionally expressed using this strategy, the fact that each clone contains genes for multiple potential modification enzymes significantly increases the likelihood of new functionality being introduced onto A47934.

Cultures of both wild type *S. toyocaensis* as well as *S. toyocaensis* transformed with each of the six retrofitted eDNA cosmids were grown in *Streptomyces* antibiotic medium for 7 days and then crude glycopeptide extracts were generated from the resulting culture broths (Lamb, Patel et al. 2006). Reversed-phase HPLC analysis of the crude extracts indicated that five out the six *S. toyocaensis* transformants produced new A47934 derivatives. Each new compound was HPLC-purified and then analyzed by HRMS. Based on the molecular formulas predicted by HRMS, the introduction of either clone UTA15 (VEG pathway) or clone AZ205 leads to the addition of a new methyl substituent onto A47934, the introduction of clone UTD30 (TEG pathway) leads to the addition of a second sulfate onto A47934 and the introduction of either clone CA191 (CA37 pathway) or clone CA1627 (CA915 pathway) leads to the addition of a sugar onto A47934 (Figure 11). The position of each new functional group was determined by comparing 1D and 2D NMR spectra obtained from A47934 with those obtained from each derivative (Figures 33-48 in the Appendix).

Compound **29**, the C-terminal methyl ester of A47934, is produced by *S. toyocaensis* transformed with clone UTA15. The new oxygen substituted methyl (53.21 ppm in ^{1}H-^{13}C HMQC spectra) is visible as a deshielded singlet (3.85 ppm) in the ^{1}H spectra and an ^{1}H-^{13}C HMBC correlation from this new methyl to the C-terminal carbonyl carbon at 172.32 ppm, allowed us to position the methyl at the C-terminus of the heptapeptide (Table 17-18 in the Appendix and Figures 36-38 in the Appendix). Clone UTA15 contains three predicted methyltransferases (Table 7 in the Appendix), two of which are predicted to encode *N*-methyltransferases. The methyltransferase encode by the third methyltransferase gene found on this clone is therefore likely responsible for the addition of the methyl group to A47934.

The position of the second sulfate found on the disulfonated derivative of A47934 (compound **30**) that is produced by *S. toyocaensis* transformed with clone UTD30, was elucidated based on changes in ^{13}C chemical shifts observed for the HPG at position 4 in the heptapeptide (Figure 11, Tables 17-18 in the Appendix and Figures 39-41 in the Appendix). In compound **29**, the chemical shift for carbon 4d shifts up field by 4.70 ppm and the chemical shift for the carbons ortho to this carbon (4c and 4e) shift downfield by 5.12 and 4.76 ppm, compared to the chemical shifts for these same carbons in A47934. This pattern of chemical shift changes is predicted to arise from the sulfonation of an aromatic hydroxyl and mimics the changes we observed when the TEG14 sulfotransferase was used to sulfonate the teicoplanin

aglycone *in vitro* (Pretsch 2009). The *teg14* gene is found on clone UTD30 and TEG14 is likely responsible for this modification *in vivo* as well.

Cultures of *S. toyocaensis* transformed with either cosmid CA191 or cosmid CA1627 produce spectroscopically identical glycosylated derivatives of A47934 (**31**) (Figure 11). Based on the new chemical shifts and observed coupling constants seen in 1D and 2D NMR experiments this new sugar substituent was determined to be a β-linked D-glucose (Tables 17-19 in the Appendix and Figures 41-45 in the Appendix) (Pretsch 2009)**.** As is the case with the disulfonated derivative of A47934 described above, changes in ^{13}C chemical shifts for the carbons found in the HPG at position 4 in the heptapeptide allowed us to place the glucosyl-group on this amino acid. Both cosmid CA191 and cosmid CA1627 contain multiple genes encoding glycosyltransferases that could be responsible for appending the glucose moiety onto A47934 (Tables 9 and 11 in the Appendix, respectively).

Compound **32**, the N-terminal methyl amine of A47934, is produced by *S. toyocaensis* transformed with clone AZ205. As seen with compound **29**, the ^{1}H and ^{1}H-^{13}C HMQC NMR spectra from compound **32** indicate the presence of a new heteroatom-substituted methyl substituent (^{1}H 2.48 ppm, ^{13}C 31.34 ppm) (Figure 11, Tables 17-18 in the Appendix, and Figures 46-48 in the Appendix). In this case, however, the chemical shifts suggest that the methyl is bound to a nitrogen atom instead of an oxygen atom. An ^{1}H-^{13}C HMBC correlation from the new methyl to the alpha carbon (64.25 ppm) of the HPG

at position one in the heptapeptide confirmed the *N*-terminal methylation in compound **32**. ORF10 in clone AZ205 is predicted to encode an *N*-methyl transferase that is likely responsible for the addition of the methyl to the N-terminus of A47934 (Table 12 in the Appendix).

Many of the tailoring enzymes encoded on eDNA clones will not be expressed in a particular heterologous system and the activity of others is undoubtedly thwarted either by incompatibly with the available substrate or by the occupation or absence of the site to be functionalized. In A47934, for example, the hydroxyl on HPG1 is blocked by a sulfate and TYR2 is missing the beta carbon oxidation that is often seen in glycopeptides. Even with these limitations five out of six eDNA-derived tailoring enzyme-rich clones resulted in the production of new A47934 derivatives. The addition of clones in a combinatorial fashion or the cloning of individual eDNA derived tailoring enzymes under the control of model promoters would likely provide access to even more highly functionalized derivatives, as would the introduction of tailoring enzyme rich cosmid clones to an array of glycopeptide producers.

By coupling the *in vivo* approach used to generate compounds **29** - **32** with the *in vitro* approach used to generate compounds **20** - **26,** we were able to produce an additional set of sulfated A47934 analogs (Figure 11) (Banik and Brady 2008). In the presence of the co-substrate 3'-phosphoadenosine-5'-phosphosulfate (PAPS) and the teicoplanin aglycone, recombinant AZ205 sulfotransferase expressed as a HIS-tagged protein in *E. coli* was found to

produce a monosulfated teicoplanin aglycone that is spectroscopically identical to compound **21**, the sulfated product generated by Teg13. Both sulfotransferases from clone AZ205 and TEG13 therefore specifically sulfate the beta carbon hydroxyl on the beta-hydroxytyrosine (BHT) found at position six in glycopeptides. Sulfonation reactions using recombinant AR205 sulfotransferase were subsequently carried out on A47934 as well as on each of the A47934 derivatives that were generated *in vivo*. For compounds **28**, **30**, **31**, and **32**, we observed complete conversion of each metabolite to a new polysulfated derivative (**33**, **34**, **35**, and **36**) after 16 h. In each case HRMS confirmed that one additional sulfate had been added to the glycopeptide starting material. As expected from the addition of a sulfate, each beta carbon proton at position six is deshielded by ~0.4 ppm compared to the chemical shift for this same proton in the starting material (Figure 49 in the Appendix). No reaction was observed when compound **29** was used as a substrate. Methylation of the C-terminal carboxylic acid must disrupt a key interaction that is required for catalysis or substrate recognition by the AR205 sulfotransferase.

In total, fifteen new sulfated glycopeptides have been generated using the enzymes found on tailoring enzyme-rich clones recovered from the UT, CA and AZ eDNA mega-libraries (Figure 11). This represents a significant addition to the number of sulfonated glycopeptides that have been characterized thus far**.** Minimum inhibitory concentrations (MIC) for each

compound were determined using glycopeptide sensitive (*Staphylococcus aureus* ATCC 6538P, *Enterococcus faecalis* ATCC 47077, methicillin resistant (*Staphylococcus aureus* USA300 (Diep, Gill et al. 2006)) and glycopeptide VanA-type resistant (*Enterococcus faecalis* EF18 (Mato, de Lencastre et al. 1996)) bacteria (Table 2). As has been seen with other glycopeptide congeners, the more highly functionalized derivatives tend to show higher MICs. While highly functionalized glycopeptide congeners often show increased in vitro MICs, they have proved useful as therapeutics because some exhibit better pharmacological properties *in vivo*. The systematic introduction of tailoring enzyme-rich clones to easily cultured glycopeptide producers should prove to be a simple yet effective strategy for generating collections of glycopeptides with new functionalization patterns that can be examined for improved biological activities.

2.4 Discussion and Future Directions

The production of novel glycopeptides serves to increase the chemical diversity of this important class of molecules. The work contained in this chapter has more than doubled the number of sequenced glycopeptide biosynthetic gene clusters. The ability to combine both *in vivo* and *in vitro* approaches to molecule production and to identify novel glycopeptide clusters expands an already powerful molecular toolbox. It also appears likely that

the culture-independent methods used to identify the new glycopeptide pathways described herein may yield drastically different new molecules. Prior to this work, only two of the over 150 glycopeptides reported were sulfonated. However, two of the six pathways discovered as part of this work contain confirmed sulfotransferases, suggesting that perhaps sulfonation may be a more common functionalization in glycopeptide biosynthesis than previously estimated. It also appears very likely that, even in such an intensively researched group of small molecules, there is considerable chemical diversity remaining undiscovered.

The discovery of new glycopeptide pathways and the production of novel glycopeptide congeners, while an important proof of principle, would be monumentally more impactful and long-term beneficial if the molecules produced possessed novel characteristics. One hurdle to the examination of the glycopeptides produced herein is that, for those produced *in vitro*, the cost of production is a glaring impediment to the increase of scale to a level required for extensive assays. To that end, the transition of enzymatic processes carried out *in vitro* to *in vivo* reactions will significantly decrease the overall cost of production. The ability to produce a novel congener entirely *in vivo* will lead to the production of compound in quantities sufficient to truly investigate the therapeutic properties, and transition the field of eDNA-derived glycopeptides from novelty to limitless potential.

2.5 Materials and Methods

2.5.1 PCR screening of eDNA for oxyC sequences

Crude eDNA samples were prepared from soil samples collected in Pennsylvania (2 samples), New Jersey, Massachusetts (2 samples), Utah, Oregon, North Carolina, Tanzania (2 samples), and Costa Rica using standard eDNA isolation methodology (Brady 2007). In brief: A one to one mixture (w/v) of soil and lysis buffer (100 mM Tris-HCl, 100 mM Na EDTA, 1.5 M NaCl, 1% (w/v) CTAB, 2% (w/v) SDS, pH 8.0) was heated for 2h at 70°C, and then, the soil was removed by centrifugation (4,000 x g, 30 min). Crude environmental DNA was isopropanol precipitated from the supernatant with the addition of 0.6 volumes of isopropanol, collected by centrifugation (4,000 x g, 30 min), washed with 70% ethanol, and then resuspended in TE (10 mM Tris, 1 mM EDTA, pH 8.0). The remaining soil contamination present in the crude extract was removed by large-scale gel purification on a 1% agarose gel (16h at 20V), and purified high molecular weight eDNA was electroeluted from the gel. PCR ready eDNA was prepared from the gel purified eDNA using the QIAamp DNA Stool Mini Kit (Qiagen). *OxyC* sequences were amplified from these eDNA samples using the following nested PCR primers: 1st round forward: ATGCTSACSCCSGAGTTCACSGTVCGG, 1st round reverse: GCAGTRRTGGAYGCCGTGCCCGAA, 2nd round forward CTGTGYGARCTGCTCGGSRTCC, 2nd round reverse

CGACRCCRCCSAGGAKCAGC. 100 ng of eDNA was used as a template in all first round amplification reactions (cycling parameters: 30 rounds of PCR, denatured 95°C for 30 sec, 68°C for 30 sec, 72°C for 90 sec, 25 μL reaction conditions: 2.5 μL Thermo Pol Buffer (New England Biolabs), 1.25 μL DMSO, 0.625 μL of each 100mM oligonucleotide primer, 0.5 μL 10 mM dNTPs mix, 1 unit of Taq DNA Polymerase and water as needed). In second round amplification reactions 1 μL of the first round PCR reaction was used as a template (cycling parameters: 30 rounds of 95°C for 30 sec, 65°C for 30 s, and 72°C for 30 s, reaction conditions were identical to those used in the first round amplification reactions). Amplicons of the correct predicted size (245 bp) were gel purified and topo cloned into TOPO 2.1 (Invitrogen). Eight to ten unique clones obtained from each soil sample were sequenced.

2.5.2 Library construction and screening

Blunt ended (End-It Kit, Epicentre) eDNA was ligated into either pWEB or pWEB-TNC, packaged into lambda phage and transfected into *E. coli* EC100 (Epicentre). Three libraries, each containing over 2,000 unique 4,000-5,000 membered libraries were constructed from the transfected *E. coli*. For each unique library, a matching DNA miniprep (Qiagen) and glycerol stock set was archived. DNA from the individual 4,000-5,000 membered library pools was arrayed into 8X8 grids. Aliquots of DNA from each pool in a grid were then combined into unique sublibraries. DNA from each sublibrary pool was screened by PCR with two different sets of *oxyC* specific

degenerate primers: PS1FWD: 5'-ATGCTSACSCCSGAGTTCACSGTVCGG-3' and PS1REV: 5'-CGAGTRRTGGAYGCCGTGCCCGAA-3'; PS2FWD: 5'-CCGCAATTCASCMARKMGAARTCSG-3' and PS2REV: 5'-TGCCKGGCRAKGAGGTTGTC-3'. PS1 was designed based on the *oxyC* gene sequences found in five cultured glycopeptide producers: (*A. orientalis* which produces chloroeremomycin, *A. balhimyceticus* which produces balhimycin, *S. toycaensis* which produces A47934, *Nonomuraea* sp. ATCC 39727 which produces A40926, and *A. teichomyceticus* which produces teicoplanin). PS2 was designed to encompass the five *oxyC* sequences used in the design of PS1, as well as the two *oxyC* sequences previously cloned from eDNA. Each 25 μL PCR reaction contained 0.5 μL of cosmid DNA, 2.5 μL of ThermoPol Buffer (New England Biolabs), 1.25 μL of DMSO, 0.625 μL of each 100mM oligonucleotide primer, 0.5 μL of 10 mM dNTPs mix and 1 unit of Taq DNA Polymerase. For the PS1 primer set, the cycling parameters were as follows: 7 touch down cycles of 95°C for 30 s, 70-62°C (-1°C/cycle) for 30 s and 72°C for 90 s followed by 20 cycles of 95°C for 30 s, 62°C for 30 s, 72°C for 90 s. For the PS2 primer set the following cycling parameters were used: 30 cycles of 95°C for 30 s, 62°C for 30 s, 72°C for 30 s. PCR amplicons of the correct predicted size (700-800 bp for PS1 and 380-410BP for the PS2) were gel purified, A/T cloned into PCR2.1 (Invitrogen) and sequenced. DNA from the individual pools that were used to create each large sublibrary that yielded

an *oxyC* sequence of interest was then screened using the same set of PCR primers.

2.5.3 Clone recovery and sequencing

Once an individual sublibrary pool containing that sequence was identified, the cosmid of interest was recovered from the corresponding glycerol stock using successive rounds of serial dilution and PCR screening. Recovered cosmid clones were end-sequenced using vector specific primers: T7 Promoter and M13 Universal Fwd -40. End-sequencing data was used to design new sets of PCR primers that could be used to recover clones containing overlapping DNA sequences. This PCR screening process was repeated iteratively as needed for each gene cluster.

All cosmid clones were 454 sequenced (Roche) by the Memorial Sloan Kettering Cancer Center Genomics Core Laboratory. Raw reads were processed with Newbler (Roche) and Velvet ((Zerbino and Birney 2008)). Sequence manipulation and gene identification were carried out using MacVector and BLASTX, respectively. The specificity of each NRPS adenylation domain was determined using NRPS-PKS web-based analysis ((Ansari, Yadav et al. 2004)).

2.3.4 Cloning, Expression, and Purification of TEG11, 12, and 13 Sulfotransferase Genes

teg11, teg12, and *teg13* were amplified (30 cycles of 95°C for 30s, 60°C for 30s (65°C for AZ205 Sulf) and 72°C for 90s, FailSafe system from Epicentre) from clone D30 for *teg12-14* and AZ205 for *AZ205Sulf* using the following primers: TEG12FWD(BclI): **GCGCTGATCA**ATGAACGGAATTCGATGG, TEG12REV(HindIII): **GCGCAAGCTT**TCCTTAACCGGCATACCCGTA, TEG13FWD(BclI): **GCGCTGATCA**ATGAACGGCATTCGATGGATC, TEG13REV(HindIII): **GCGCAAGCTT**ATCTCTCCTCCCTCAGCCGGC, TEG14FWD(BclI): **GCGCTGATCA**ATGAACGGTATTCGATGGATC, TEG14REV(HindIII): **GCGCAAGCTT**ACAATCCGCCCGTTAGCCGGC, AZ205SulfFWD(*BamH*I) **GCGCGGATCC**ATGAACGGAATCCGATGG and AZ205SulfREV(*Hind*III) **GCGCAAGCTT**TCCCTAATCAGCGTACCCGTA. The restriction sites added for cloning purposes are shown in bold. Amplicons were doubly digested with either BclI/HindIII or BamHI/HindIII, ligated into the BamHI/HindIII digested pET28a, and then transformed into *E. coli* BL21 (DE3).

Fusion protein purification and sulfation reaction protocols were modified from published procedures used for StaL (Lamb, Patel et al. 2006). Overnight growths were used to inoculate (1:1000 dilution) 1 L cultures of LB, which were grown at 37°C until the OD_{600} reached 0.6. The temperature was then reduced to 20°C, and after 1h, the cultures were induced with IPTG

(0.5 mM). After 14-16h at 20°C, the cultures were harvested by centrifugation (3,200 x g for 30 minutes). The cell pellet was resuspended in 40 mL of lysis buffer (50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (v/v) glycerol, 20 mM imidazole, pH 8, 10 mM β-mercaptoethanol, and 0.5 % (v/v) Triton X-100), and the cells were lysed by sonication. Crude cell lysates were centrifuged at 25,000 x g for 30 minutes, and the supernatants were then incubated for 15 min (24°C) with 1 mL of Ni-NTA resin. After 15 min, this slurry was loaded onto a column, washed with 40 mL of lysis buffer, followed by 40 mL of wash buffer (50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (v/v) glycerol, 20 mM imidazole, pH 8, and 10 mM β-mercaptoethanol), and then the protein was eluted from the resin with 15 mL of elution buffer (50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (v/v) glycerol, 125 mM imidazole, pH 8, and 10 mM β-mercaptoethanol). The protein concentration was determined with the Bradford assay (yields: 74 mg/L TEG12, 31 mg/L TEG13, 88 mg/L TEG14, and 65 mg/L AZ205Sulf) and these samples were used in sulfation reactions without further purification. Two mililiter sulfonation reactions containing 1mg of each sulfotransferse, 1 mg teicoplanin aglycone and 0.5 mg 3'-phosphoadenosine 5'-phosphosulfate (PAPS) were setup in reaction buffer (250 mM HEPES, pH 7.5, 0.1 mM DTT) and incubated overnight at 30°C (For *in vitro* reactions with compounds isolated from culture extracts (compounds **28**, **30**, **31**, and **32**), reactions were scaled to 6 mL). Excess sulfotransferase was added to overcome the known inhibition of PAPS dependent

sulfotransferases by the sulfotransferase byproduct 3'-phosphoadenosine 5' phosphate (PAP) (Chapman, Best et al. 2004). The reactions were then placed in a boiling water bath for 10 minutes, placed on ice for 10 minutes, and centrifuged (21,000 x g at 4°C) for 10 minutes. The supernatant (and if necessary a water wash of the pellet) was evaporated to dryness and resuspended in 400 μL of a 50:50 DMSO:H_2O (1200 μL for 6 mL reactions). The analytical HPLC conditions were as follows: (C18 (4.6 x 150 mm), 20 mM ammonium acetate:acetonitrile, 95:5 to 70:30 over 20 min, 1.5 mL/min). Preparative HPLC was carried out using the same gradient conditions on a 10 x 150 mm column at a flow rate of 7 mL/min.

2.3.5 Cosmid retrofitting and conjugation into *S. toyocaensis*

Cosmids containing glycopeptide tailoring enzymes were each retrofitted to contain the genetic elements required for conjugation, integration and selection in *Streptomyces*. Each cosmid was initially digested with *PsiI* (New England Biolabs) and then ligated to the 6.7 kb DraI fragment from pOJ436 ((Matsushima, Broughton et al. 1994)). This DraI fragment contains the ϕC31 integration machinery, the RK2 *oriT* and the *aac(3)IV* apramycin resistance gene. Upon transformation into *E. coli*, retrofitted cosmids could be easily identified by selecting with both ampicillin and apramycin. Correctly retrofitted cosmids were transformed into *E. coli* S17.1 and conjugated into *Streptomyces toyocaensis* NRRL 15009 using previously described methods, with a few minor modifications ((Matsushima

and Baltz 1996)). For each conjugation, a 50 mL culture of *E. coli* S17.1 transformed with an appropriately retrofitted cosmid clone was grown to mid-logarithmic phase (LB with 50 μg/mL apramycin and 10 μg/mL trimethoprim). Cells were then pelleted by centrifugation (3500 x g, 30 min) and the resulting cell pellet was washed twice with 50 mL of antibiotic free LB. Washed *E. coli* were resuspended in 1 mL SOC (per liter: 20 gm Bacto Tryptone 5 gm Bacto Yeast Extract 10 mM NaCl, 2.5 mM Kcl, 10 mM $MgCl_2$ 10 mM $MgSO_4$ 20 mM glucose). 20 μL of a heat treated (50°C for 10 min) spore slurry (OD_{450}=40-80) was mixed with 50 μL of washed *E. coli* and this cell suspension was then plated onto R2 agar. After 16 hours at 30°C, the plates were flooded with 1 mL of sterile water containing 0.5 mg of apramycin and 0.5 mg of nalidixic acid and then returned to the incubator until exconjugants appeared (3-4 days). Exconjugants were restruck on modified Bennett's agar containing 25 μg/mL apramycin and 25 μg/mL nalidixic acid.

2.3.6 Glycopeptide congener production

S. toyocaensis harboring eDNA-derived glycopeptide tailoring enzymes was initially grown on modified Bennett's agar for 3 days at 30°C. Single colonies were inoculated into 50 mL of *Streptomyces* Vegetative Medium (SVM) and incubated at 30°C for 3 days (200 rpm orbital shaking)(Lamb, Patel et al. 2006). 50 mL of *Streptomyces* antibiotic medium containing 1 g/L yeast extract (SAM) was inoculated with a 1:1000 dilution of the 3-days-old

starter culture (Lamb, Patel et al. 2006). After 7 days (30°C, 200 rpm orbital shaking) the cell mass was collected by centrifugation (3500 x g, 10 min). The wet weight of the pellet was determined and the pellet was resuspended (vortex, 30s) in 0.75 mL of 2.5% (v/v) NH_4OH per gram of wet cell pellet. This basic cell suspension (final pH >11) was centrifuged (3500 x g, 10 min) and the supernatant collected. The pH of the supernatant was adjusted to 7.5 with 1 M HCl, and analyzed directly by HPLC [C_{18} (4.6 μm X 150 mm), 20 mM ammonium acetate:CH_3CN, from 95:5 to 70:30 over 20 min, 1.5 mL/min or C_{18} (10 μm X 150 mm), 20 mM ammonium acetate:CH_3CN, from 95:5 to 70:30 over 20 min, 7 mL/min].

2.3.7 Glycopeptide congener analysis

Hi-Res ESI-MS experiments were carried out either on a Thermo-Fisher LTQ-Obritrap mass spectrometer interfaced with a Dionex U3000 capillary/nano-HPLC system, or by direct injection onto a Waters LCT Premier XE mass spectrometer. NMR data was collected for all samples in a 3:1 mixture of D_2O:CD_3CN on a Bruker 600 MHz spectrometer and was processed using the MestreNova Software Package (MestreLab Research).

2.3.8 MIC determination

Stationary phase cultures of *Staphylococcus aureus* strains and *Enterococcus faecalis* strains, grown in tryptic soy broth and brain heart infusion broth, respectively, were diluted $1:10^6$ prior to addition of compounds. Individual wells in 96 well microtiter plates were filled with 100

μL of diluted culture. A stock solution of an individual compound dissolved in dimethylsulfoxide was diluted 1:100 into culture media and then 100 μL of this solution was added to the first well of a row in the filled microtiter plate. This well was then serially diluted 1:2 across the plate. Microtiter plates were incubated at 37°C with 100 rpm orbital shaking for 18 hours. MIC values are reported as the lowest concentration at which no bacterial growth was observed.

CHAPTER 3

3 Structural and Biochemical Analysis of eDNA-derived Glycopeptide Sulfotransferases

3.1 Chapter Summary

Prior to the work contained in chapter 2, there were only two known sulfonated glycopeptides (Boeck and Mertz 1986; Holdom 1988). The discovery of two new glycopeptide biosynthetic gene clusters containing glycopeptide-modifying sulfotransferases and the production of semi-synthetic glycopeptides displaying anionic modifications led to an increased desire to gain a broader understanding of the structural and catalytic properties of glycopeptide sulfotransferases. To that end, this chapter describes the structural and biochemical analysis of one of these sulfotransferases. In collaboration with Dr. Matthew J. Bick and Dr. Seth A. Darst of the Darst Laboratory at the Rockefeller University, four separate X-ray crystal structures have been solved, three of TEG12 and one of TEG14 (Bick, Banik et al. 2010; Bick, Banik et al. 2010). The three TEG12 structures are an apo-structure, a binary structure containing teicoplanin aglycone bound to TEG 12, and a ternary structure, containing both 3'-phosphoadenosine-5'phosphate (PAP) and the teicoplanin aglycone bound to TEG12. The TEG14 structure is an apo-structure, which has provided

structural information, but limited biochemical data, and therefore has largely been omitted from this chapter. The discussion in this chapter will therefore be focused on the various TEG12 structures. Upon solution of the various TEG12 structures, a number of amino acids appeared to be fundamentally important in substrate recruitment, substrate specificity, and/or catalytic activity. These key residues were mutated and observations were made as to the effects on the enzyme catalytic reactivity. Taken together, the structural and kinetic data provide a much better understanding of these enzymes and have provided general insight into key regions of these proteins, which have allowed three sulfotransferases with in excess of 80% amino acid identity to sulfonate the same glycopeptide substrate at three different sites. Only a brief description of the details of crystal structure solution will be given for the purposes of archiving, as this work was carried out exclusively by Dr. Matthew J. Bick and Dr. Seth A. Darst, and is outside the expertise of this author.

3.2 Introduction

Sulfonation is one of the rarest modifications seen in glycopeptide biosynthesis. Prior to the metagenomics efforts described in chapter 2 of this thesis, there were only two known sulfonated glycopeptides, the biosynthesis of which was characterized for only one, A47934 (**28**) (Pootoolal, Thomas et al. 2002). The sulfotransferase, which acts during the biosynthesis of A47934, StaL, has been well studied, with publications detailing both significant

structural and biochemical studies (Lamb, Patel et al. 2006; Shi, Lamb et al. 2007). However, with only one characterized glycopeptide sulfotransferase, it was difficult to gain much more than anecdotal, pathway-specific information about this enzyme. This dearth of knowledge became even more pressing after the generation of a semi-synthetic derivative of vancomycin, telavancin (**27**), which is modified from its natural product precursor by the addition of a long hydrophobic acyl chain and a phosphonic acid moiety to amino acid seven of the heptapeptide backbone. This phosphonic acid moiety is structurally and chemically rather similar to the naturally occurring sulfate seen in A47934 (**28**). The discovery of four novel sulfotransferases through the course of the culture-independent glycopeptide pathway discovery efforts described in chapter 2 provided an opportunity for a more in-depth investigation into these understudied enzymes. One pathway in particular, the TEG pathway, contained three novel sulfotransferases, which displayed over 80% amino acid identity to each other. These very closely related sulfotransferases sulfonated the same predicted substrate, the teicoplanin aglycone, at three different sites. It was observed that, for all intents and purposes, there were only three regions that varied within the three enzymes. It was therefore hypothesized that a structural analysis of these enzymes, coupled with biochemical data, would provide insight into this potentially very useful class of proteins.

3.3 Results

3.3.1 TEG12 Structural Overview

A ClustalW alignment of the three sulfotransferases found in the TEG pathway (TEG12-14) displays the high level of amino acid identity among the three enzymes, with three main regions of divergence (Figure 13). These three regions, termed variable loops 1-3 (V1,V2, and V3) correspond in TEG12 to Ile37-Thr43, Gly127-137, and Thr204-Asp250, respectively. As all three TEG sulfotransferases can use the teicoplanin aglycone as a substrate, this small molecule was used, along with the cofactor byproduct 3'-phosphoadenosine-5'phosphate (PAP), in various combinations for co-crystallization experiments. Vmax and Km were calculated for TEG12 with the teicoplanin aglyonce as substrate. These values were found to be 215.3 +/- 25.2 nmol/min-mg and 559.6 +/- 1.1 μM, respectively.

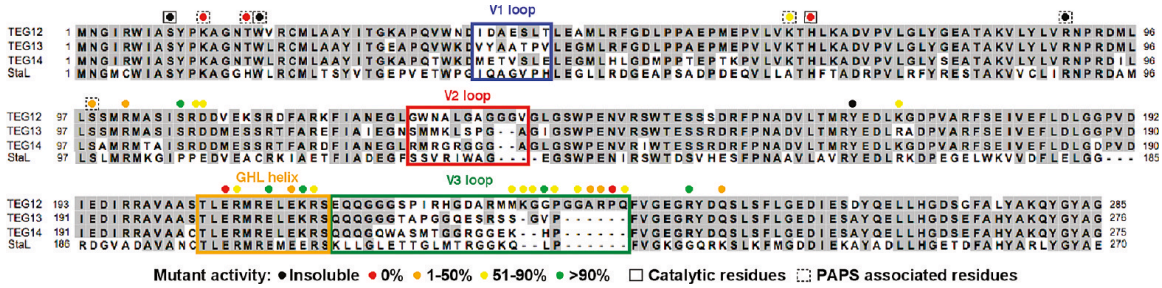


Figure 13: ClustalW alignment of glycopeptide sulfotransferases.
Highly variable sequences (V1-V3) and the flexible part of the GHL helix appear in colored boxes that match the coloration seen in panels B and C of Figure 13. Results from alanine exchange mutagenesis experiments are color-coded by percent activity. Results for catalytic and PAPS-associated residues are shown in solid and dashed boxes, respectively. All other mutated residues are thought to interact with the glycopeptides. The H67A mutant was insoluble, and therefore, H67Q data are reported.

Multiple crystal structures of TEG12 were solved, all as homo-dimers: an apo-structure solved to 2.91 Å, a binary structure of TEG12 complexed with the teicoplanin aglycone, solved to 2.27 Å, and a ternary structure of TEG12 containing in both monomers of the dimer a molecule of PAP, and in one of the monomers two molecules of the teicoplanin aglycone, solved to 2.05 Å (Figure 14). All protein used for crystallization was N-terminally His_6-tagged and purified by affinity chromatography on Ni-NTA resin. There was no additional purification, nor was there any additional processing to remove the 34 vector-derived amino acids, which were later observed to play a possible role in crystal packing in the TEG12 structures. The TEG12 crystal structures described herein are the first examples of a glycopeptide sulfotransferase with substrate bound. The previous crystal structure of StaL, the sulfotransferase from the A47934 pathway, described a modeling of

the substrate docked with the enzyme, as their various structures contained no electron density for the glycopeptide (Shi, Lamb et al. 2007).

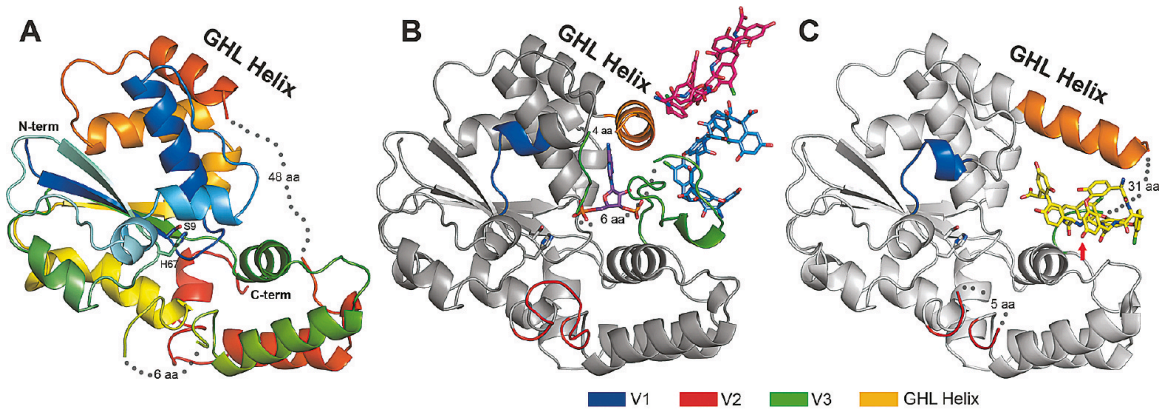


Figure 14: TEG12 crystal structures.

A): TEG12-apo is colored using a rainbow scheme, from blue (N-terminus) to red (C-terminus). Regions of disorder are connected by gray dots, with the number of disordered residues indicated. The flexible GHL helix is colored orange. Side chains for the proposed active site residues His67 and Ser9 are also shown. B): TEG12-ternary complex containing PAP and the teicoplanin aglycone. The protein is colored gray to accentuate regions of the structure that differ from TEG12-apo. Specifically, these regions are variable loops V1-V3, colored blue, red, and green, respectively. Also shown are a molecule of PAP in the active site, a molecule of teicoplanin aglycone that interacts with the V3 loop (sky blue), and an additional molecule of teicoplanin aglycone involved in crystal packing interactions (hot pink). C): TEG12-binary structure in a complex with teicoplanin aglycone in the active site cavity. Again, the protein is colored gray, with V1-V3 colored as in the ternary structure. TEG12 sulfates the hydroxyl of residue 3 of the teicoplanin aglycone, which has been denoted with a red arrow. Omitted from the binary structure is a second molecule of teicoplanin aglycone bound to the outside of the protein on the opposite side of the GHL helix.

3.3.2 Co-factor-binding Residues

TEG12, like many other PAPS-dependent (3'-phosphoadenosine-5'-phosphosulfate) sulfotransferases utilize conserved structural motifs to bind

this sulfate donating co-factor. A loop in TEG12 comprised of Pro11-Thr16 represents one of these common motifs, known as the phosphosulfate-binding (PSB) loop, due to its coordination of the 5'-phosphosulfate. In the ternary structure, Lys12 and Thr16 form hydrogen-bonding contacts with the 5'-phosphate of PAP and Trp17, which sits just outside the PSB loop, forms a parallel stack with the adenine base of PAP (Figure 15). Another conserved structural motif, the phosphate-binding loop (Campbell, Singh et al.), is comprised of Val89-Ser99, with Ser98 observed to form a hydrogen-bond with the 3'-phosphate of PAP in the ternary structure (Campbell, Singh et al.).

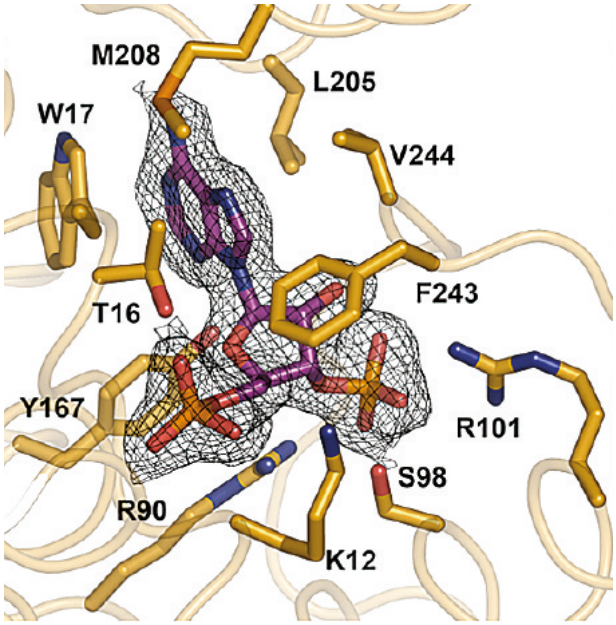


Figure 15: PAP bound in theTeg12-ternary complex.
Leu205 and Met208 from the GHL helix and Phe243 and Val244 fromthe V3 loop make significant van der Waals contacts with PAP.

The hydrogen bonds formed between Thr16 and the 5'-phosphate mimic hydrogen bonds seen in the StaL crystal structure, although in StaL,

there is a histidine at position 16. Nonetheless, the hydrogen bonds, which appeared to be important in coordinating PAPS in the StaL structure, are present in the TEG12 structure as well. To determine whether these hypothesized PAPS coordinating hydrogen bonds were important for catalytic activity, alanine replacement mutants were constructed, using a previously reported protocol and PCR primers listed (Table 3) and activity experiments were carried out (Sarkar 1990). T16A was unable to catalyze the sulfonation reaction, as was K12A, confirming the necessity of both residues. Alanine replacement mutants of three other residues implicated in coordinating PAPS, Trp17, Lys65, and Ser98, were constructed as well. W17A did not express as soluble protein, confirming its role in protein stability, but leaving its necessity for PAPS binding as solely an inferred role. K65A and S98A produced enzymes with only partial activity (65% and 27% relative to wild-type, respectively). Additional alanine replacement mutations were constructed for Arg 90 and Tyr167, which are each seen to interact with both PAP and the aglycone in the ternary structure. Unfortunately, neither R90A nor Y167A expressed as soluble protein, suggesting an additional possible structural stability role.

Table 3: Oligonucleotide primers used for the generation of site-directed mutants in TEG12

Mutant	Primer (5'-3')
S9A	GTTTCCAGCCTTTGGATA**CGC**TGCGATCCATCGAATTCC
K12A	CCACGTGTTTCCAGC**CGC**TGGATACGATGC
T16A	CACCTGACCCA**CGC**GTTTCCAGCCTTTGG
W17A	CAACATGCACCTGAC**CGC**CGTGTTTCCAGC
K65A	CATCGGCCTTGAGGTGCGT**CGC**CACCAGCACCGGTTC
H67A	CACATCGGCCTTGAG**GGC**CGTCTTCACCAG
R90A	CATATCCCGCGGGTT**CGC**CACGAGATAGAG
S98A	GGCCATGCGATCGA**GGC**GAGCAGCATATC
R101A	TATCGAGGCCAT**GGC**CATCGAGCTGAG
S106A	CTACGTCGTCGCG**CGC**TATCGAGGCCATGCG
R107A	GCTTTTTTCTACGTCGTC**GGC**CGATATCGAGGC
D108A	GCTTTTTTCTACGTC**GGC**GCGCGATATCGAGGCC
Y167A	GTGCTGACGATGCGT**GCT**GAGGACCTGAAGGGC
K171A	CGTTATGAGGACCTG**GCG**GGCGATCCGGTCGCACGG
E206A	GCTGCCTCCACGCTG**GCG**CGGATGCGTGAACTG
R207A	GCTGCCTCCACGCTGGAG**GCG**ATGCGTGAACTG
E210A	GAGCGGATGCGT**GCA**CTGGAGAAACGGAG
E212A	CGGATGCGTGAACTG**GCG**AAACGG
K213A	CGGATGCGTGAACTGGAG**GCA**CGGAGCGAGCAGCAG
R214A	CGGATGCGTGAACTGGAGAAA**GCG**AGCGAGCAGCAG
M232A	GGTGATGCGAGAATG**GCG**AAAGGG
K233A	GATGCGAGAATGATG**GCA**GGGGG
G234A	GCGAGAATGATGAAA**GCG**GGACC
G235A	AGAATGATGAAAGGG**GCA**CCTGGTG
P236A	ATGATGAAAGGGGGA**GCT**GGTGG
G238A	AAAGGGGGACCTGGT**GCC**GCGAGG
A239R	AAAGGGGGACCTGGTGGC**CGG**AGG
R240A	GGACCTGGTGGCGCG**GCG**CCCCAG
P241A	CCTGGTGGCGCGAGG**GCC**CAGTTC
Q242A	GGTGGCGCGAGGCCC**GCG**TTCGTG
R248A	CAGTTCGTGGGCGAGGGC**GCG**TACGACCAGTCCCTG
Q251A	GAGGGCAGGTACGAC**GCG**TCCCTGTCCTTCTTG

3.3.3 Residues Involved in Catalysis

It had been previously hypothesized that PAPS-dependent sulfotransferases, like TEG12, utilize conserved histidines, which act as general bases to activate the hydroxyl or amino groups in the sulfate exchange from PAPS to substrate. Of the three histidines in TEG12, only His67 is located near the molecule of PAP in the ternary structure, in fact modeling PAPS in for PAP in the ternary structure places the 5'-sulfate directly adjacent to His67 (Figure 16). Mutation of His67 to alanine yieled insoluble protein, as did the H67E mutant. However, mutation of His67 to glutamine yielded a soluble, but fully inactive, enzyme. Both the histidine at position 67 and a serine at position 9 are conserved between StaL and TEG12. Ser9 is thought to hydrogen bond to the imidazole ring of the histidine and aid in catalytic activation. Unfortunately, an alanine replacement mutant of Ser9 yielded insoluble protein, so as with Trp17, the role played by this amino acid can only be inferred from the crystal structure.

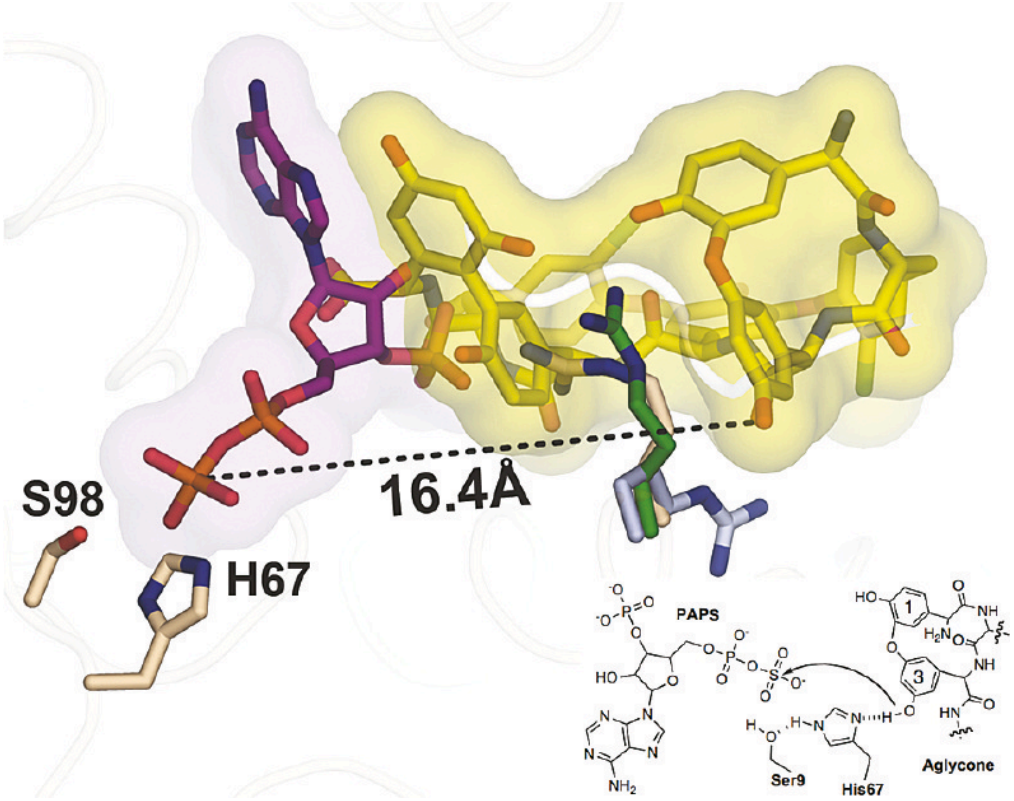


Figure 16: Composite of the active site from the Teg12-binary and-ternary structures.

PAPS was modeled into the ternary structure in place of PAP. In this aligned view, PAPS and teicoplanin aglycone would not be able to occupy their respective positions simultaneously, as there would be strong clashes between the two substrates. Residue 3 of the teicoplanin aglycone, the site of sulfation, is more than 16 Å from the sulfate of PAPS when the ternary and binary structures are overlapped. A substantial rearrangement of teicoplanin aglycone within the active site would have to occur for sulfation to proceed by the proposed in-line attack mechanism. The positions of Arg101 in the binary (green) and ternary (tan) structures appear to preclude the movement of teicoplanin aglycone toward His67. Arg101's position in the context of the apo structure creates a more open active site, where the aglycone could pivot toward His67 more easily.

3.3.4 The Glycopeptide Helix Loop (GHL)

There are a few regions, which vary in their level of disorder amongst the three TEG12 structures (Figure 14). In the apo-structure, the regions

from Gly127-Val137 and from Thr204-Asp250 are almost completely disordered. However, in the binary structure, the latter of those two regions, which corresponds to a region containing variable loop 3 (V3), becomes slightly more ordered. In the ternary structure, this region is almost fully ordered. The stabilization in this region is due largely to interactions between the protein and the glycopeptide substrate that is bound in these co-crystal structures. This region is also the largest region of variability amongst the three TEG sulfotransferases (TEG12-TEG14). Due to the protein-substrate interactions, the region slightly upstream of V3 and V3 has been collectively labeled the glycopeptide helix loop (GHL) (Figure 14). Based on the significant number of contacts made between the protein and the substrate, as well as the variability in the TEG sulfotransferases in V3, it seemed quite likely that the GHL could play a very significant role in this enzyme, perhaps in substrate recruitment, and shuttling of substrate to the active site.

In the binary structure, the GHL helix and a helix, which extends from Ile193-Ser203 form a single helix that runs along the top of the glycopeptide substrate, while in the ternary structure, this single contiguous helix is kinked at ~90°, which is similar to the orientation adopted by these helices in the StaL structure (Figure 14). The position occupied by the GHL and PAP in the ternary structure overlaps the position of the teicoplanin aglycone substrate in the binary structure, indicating the great degree of movement and flexibility found in this region of the protein. The displaced, disordered

loop that contacts PAP in the ternary structure is located behind the bound glycopeptide substrate in the binary structure. Other than the movement seen in this loop, the binary structure differs little from either the apo- or the ternary structures. In addition to the interactions with the GHL, the glycopeptide substrate also forms interactions with the PSB and PB loops. The PSB loop directly interacts with the C-terminus of the glycopeptide substrate, while the PB loop runs underneath the glycopeptide in the binary structure.

There are two molecules of the teicoplanin aglycone found in one of the monomers of the ternary structure, although neither of these molecules was located in the active site of the enzyme. There are numerous contacts made between one of these molecules and the GHL loop (Figure 17). A number of residues within the GHL were mutated to determine the effects of changes within this loop would alter the sulfotransferase activity. Glu212, which does not directly interact with the glycopeptide in the binary structure, when mutated to alanine, displays a vastly reduced level of enzyme activity (26% relative to wild type). Additional alanine replacement mutations in the GHL led to varied effects on the activity of TEG12. Alanine replacement mutations of three residues seen to make hydrogen bonding contacts in the binary structure, Glu206, Arg207, and Glu 210 led to complete loss of activity, only slight (79% relative to wild type) loss of activity, and negligible (97% relative to wild type) change in activity, respectively. Three other GHL residues were

mutated to alanine and assessed for relative activity. E212A, K213A, and R214A displayed activities relative to wild type of 26, 99, and 67%, respectively.

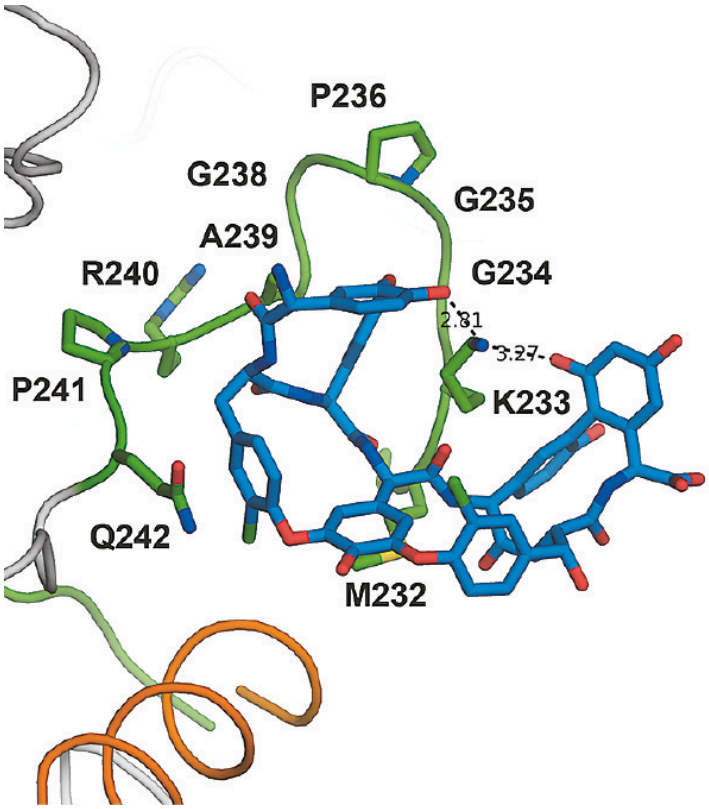


Figure 17: Close-up of the GHL loop aglycone complex from the Teg12-ternary structure.

Side chains for all alanine replacement experiments are shown. Lys233 intercalates into the glycopeptide and makes hydrogen bond contacts with residues 1 and 7 of the aglycone.

Another large stretch of the GHL (Met232-Pro236 and Gly238-Gln242) interacts directly with the teicoplanin aglycone in the ternary structure. Alanine replacement mutations of these residues, which directly interact with the aglycone, have been constructed and kinetics data recorded (Table 4). Not surprisingly, all but the mutation of Gly235 to alanine led to a

decrease in enzyme activity. Two additional residues outside the GHL, which are not directly observed to make contact with the substrate or cofactor in either the binary or ternary structures, but instead are seen to define the contour of the active site, Arg248 and Gln251 were also mutated to alanine. Interestingly, while Q251A resulted in a marked decrease in enzyme activity, R248A resulted in an almost two-fold increase in enzyme activity. It seems likely, in the case of R248A, that the removal of the steric bulk of the arginine side chain creates a more flexible active site, which is more amenable to accommodating the sizable glycopeptide substrate. This hypothesis is further supported by the substitution in StaL of a glycine for the arginine at position 248, which is conserved in all of the TEG sulfotransferases.

Table 4: Relative activity of TEG12 mutant sulfotransferases

Catalytic or PAP binding	Activity (%)	V3 aglycone binding	Activity (%)	Non-V3 aglycone binding	Activity (%)
S9A	N/A	M232A	74	R101A	29
K12A	0	K233A	79	S106A	92
T16A	0	G234A	82	R107A	82
W17A	N/A	G235A	100	D108A	82
K65A	61	P236A	73	Y167A	N/A
H67A	N/A	G238A	56	K171A	69
H67E	N/A	A239R	37	E206A	0
H67Q	0	R240A	48	R207A	79
R90A	N/A	P241A	0	E210A	97
S98A	27	Q242A	64	E212A	26
				K213A	99
				R214A	67
				R248A	196
				Q251A	41

3.3.5 Sulfotransferase Intercalation in the Glycopeptide Substrate

In both the binary and ternary structures, there are individual amino acids (Arg101 in the binary and Lys233 in the ternary structure) that actually interacalate into the inner surface of the teicoplanin aglycone (Figure 16). All other substrate-protein interactions involve the external surface of the glycopeptide. These long, positively charged side chains intercalate between amino acids 3 and 5, forming a number of hydrogen bonding contacts. Although the binding observed for Arg101 and Lys233 is very similar in its nature, the effect of mutation of these residues to alanine is drastically different. R101A possesses only a small fraction of the activity

of wild-type TEG12 (27%), while K233A is only marginally affected (79% relative to wild-type).

3.3.6 Second Molecule of the Teicoplanin Aglycone in the Ternary Structure

As earlier mentioned, there is electron density corresponding to two molecules of the teicoplanin aglycone in the ternary TEG12 structure. The second of these two molecules, which is bound to the enzyme considerably farther away from the active site than the other molecule of aglycone, appears to simply be forming crystal contacts between adjacent symmetry partners. However, there is a region of electron density near this second molecule of aglycone, which is most likely hydrogen bonding of vector-derived amino acids. The region occupied by this density is the same as that observed for the binding of the D-ala-D-ala tail of peptidoglycan, which is the traditional target for glycopeptide antibiotics (Nicolaou 1999). Although it is quite likely that these interactions are of no biological importance to the sulfotransferase, they are nonetheless reminiscent of the true role of the glycopeptide cleft.

3.4 Discussion and Future Directions

There are a number of residues that, based on alanine replacement mutations, are essential for catalytic conversion of the teicoplanin aglycone to sulfo-teicoplanin aglycone by TEG12. TEG12 sulfates the teicoplanin

aglycone at the DPG at position 3. The conserved histidine, His67, has been confirmed in its role in the sulfonation reaction. All but two of the predicted catalytic or substrate binding residues (K65A and S98A) were found to be either essential to catalysis or to play an essential structure stabilizing role, based on the expression of insoluble protein. All but one of the alanine replacement mutants generated for residues in the GHL led to a diminished capacity for substrate conversion.

It is impossible from the current structural and kinetics data to determine which of two possible reaction mechanisms take place. It is possible that the sulfate group is first covalently attached to the catalytic histidine, forming a sulfo-TEG intermediate. Alternatively, the catalytic histidine could act as a general base, as is depicted in Figure 17. Both strategies have been documented in various prokaryotic and eukaryotic sulfotransferases (Chapman, Best et al. 2004). Even in very distantly related sulfotransferases such as TEG12 and the human cytosolic sulfotransferase SULT1A1, which only display 20% amino acid identity, the catalytic histidine is conserved, as is the serine which is predicted to aid in subsrate activation. Possible experiments that could lead to the determination of which of the two mechanisms is correct include low-energy ionization mass spectrometry to attempt to identify the sulfo-TEG intermediate, or kinetic experiments where the effect of PAPS concentration on the initial velocity of the enzyme. These

and other experiments could lead to the elucidation of which reaction mechanism is correct.

Although the interactions of the substrate teicoplanin aglycone and the protein vary between the binary and ternary structures, the observation of a long, positively-charged side chain intercalating into the middle of the glycopeptide is common in both structures. Arg101, based on its position in both structures, appears to play an important role in the active site dynamics. Arg101 appears to block the glycopeptide from entering the active site and gaining access to the catalytic histidine (His67). In the ternary structure, Arg101 interacts directly with the 3'-phosphate of PAP and does not appear to interact with the glycopeptide whatsoever. However, in the apo-structure, Arg101 adopts a confirmation that would open up the active site to accommodate both the glycopeptide substrate and the PAPS cofactor. It seems likely then that Arg101 plays an important role in modulating active site dynamic changes, which are required for catalysis.

The portion of the GHL loop and the loop directly preceding it, as mentioned earlier, form a single, contiguous loop in the binary structure, while they are two distinct loops, with a 90° kink in between. The distance between the ends of the straight, longer loop and the kinked loop is approximately 16 Å. This is also the distance between the site of sulfonation of the glycopeptide substrate and His67 in the binary structure. The flexibility of this loop seems very likely a significant factor in the ability of

this protein to act on such a large substrate, much larger than most sulfotransferase substrates. The interactions seen in the binary structure between the substrate and this loop, based on the diminished enzyme activity of alanine replacement mutations, appear to represent a snapshot on the path the substrate takes from outside the enzyme to the active site. The residues that interact in the GHL with the glycopeptide in the binary structure are seen to point in an outward direction in the ternary structure, suggesting that their role is more for shuttling the substrate to the active site, and that other contacts likely moderate the unique substrate positioning in the active site. The conservation of the GHL in the three TEG sulfotransferases (and largely in StaL as well) lends further credence to this hypothesis.

The role of substrate positioning is likely played by the V3 residues, which would account for the variation in the site of sulfonation in the three TEG sulfotransferases. The varied binding of the substrate, and the presentation of different faces of the glycopeptide to the catalytic residues in the active site are likely responsible for this diversity. One possibility for generating variation in these sulfotransferases is to simply exchange the GHL regions from one TEG sulfotransferase for another. Exchange of the entire GHL region in TEG12 for the GHL region from TEG13 or TEG14 led to the expression of soluble protein at quantities similar to wild-type. However, in both cases (TEG12 GHL exchanged fro TEG13 GHL and TEG12 GHL

exchanged for TEG14 GHL), all catalytic activity was abolished. The enzymes neither retained their original activity nor gained new activity. It therefore seems likely that each individual V1-V3 collection is required for each specific sulfonation reaction. Therefore, the roles played by V1 and V2 cannot be discounted. These regions shift slightly between the apo-, binary, and ternary structures, presumably to accommodate the glycopeptide at various stages of shuttling into the active site. It is therefore most likely that the production of novel sites of sulfonation in the glycopeptide substrate would be possible, but would require alteration of not only V3, but also changes in V1 and V2 as well. However, based on the observed interactions in the TEG12 structures, the possibility of generating a novel sulfotransferase by specific alterations in the variable loops seems an exciting and distinct possibility.

3.5 Materials and Methods

3.5.1 TEG12 Expression and Purification

TEG12 was cloned and expressed as previously described (Banik and Brady 2008). Briefly, *teg12* was amplified (30 cycles of 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 90 s; FailSafe system from Epicentre) from eDNA cosmid clone D30 using the following primers: TEG12FWD(BclI):GCGC**TGATCA**ATGAACGGAATTCGATGG, TEG12REV(HindIII):GCGC**AAGCTT**TCCTTAACCGGCATACCCGTA. Restriction enzyme sites used for cloning are shown in bold. The resulting product was doubly digested with BclI and HindIII and subsequently ligated

into pET28a, which had been BamHI/HindIII doubly digested. The resulting construct was then transformed into *E. coli* BL21(DE3) for protein expression. Expression cultures were grown to OD_{600}=0.6, followed by IPTG induction, and overnight growth at 20 °C. The culture was pelleted by centrifugation (3,200 x g for 30 min), the supernatant was discarded, and the cell pellet was resuspended in 40 mL lysis buffer [50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (vol/vol) glycerol, 20 mM imidazole, pH 8, 10 mM β-mercaptoethanol and 0.5% (vol/vol) Triton X-100]. The resuspended cell pellet was lysed by sonication, and the insoluble portion was removed by centrifugation (15,000 x g for 30 min). The cleared cell lysate was incubated with 1 mL Ni-NTA resin for 15 min. The slurry was loaded onto a column, allowed to empty by gravity flow, washed with 40 mL lysis buffer, and finally washed with 40 mL wash buffer [50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (vol/vol) glycerol, 20 mM imidazole, pH 8.0, and 10 mM β-mercaptoethanol]. The protein was eluted by the addition of 15 mL of elution buffer [50 mM HEPES, pH 7.5, 0.5 M NaCl, 5% (vol/vol) glycerol, 125 mM imidazole, pH 8, and 10 mM β-mercaptoethanol]. No attempt was made to remove the vector derived 6X-histidine tag, resulting in a TEG12 protein plus 34 additional residues N-terminal to the start methionine. Protein was concentrated using Vivascience Vivaspin 30,000 MWCO ultrafiltration concentrators, and was buffer exchanged 3 times into protein buffer [200 mM NaCl, 20 mM HEPES, pH 7.5, 5% glycerol, and 1 mM DTT].

3.5.2 TEG12 Crystallization

Concentrated protein was centrifuged at 14,000 rpm (4 °C, 30 min) in a microcentrifuge to remove any insoluble material prior to crystallization. All crystals were grown using the hanging drop vapor diffusion method. Initial TEG12-apo crystals were obtained by mixing 1 μl of protein (7.5 mg/ml in protein buffer) with 1 μl of reservoir solution (1.0 M sodium citrate, 0.1 M sodium cacodylate, pH 6.5, JCSG core III-48, Qiagen) over a 500 μl reservoir. Blade-like crystals grew overnight at 22 °C and reached a maximal size of 400 μm X 50 μm X 20 μm in approximately one week. To improve crystal thickness, TEG12-apo crystals were optimized by microseeding, in addition to mixing 1 μl of protein at 7.5 mg/ml with 0.5 μl of reservoir and 0.5 μl of Silver Bullet reagent 29 (Hampton Research). A component of the Silver Bullet screen, aspartame, was modeled into the PAPS binding site of one of the monomers of the TEG12 dimer. Crystals were soaked in a 20 μl drop containing reservoir solution plus 10% ethylene glycol. The drop was allowed to dehydrate by exposure to open air at room temperature for approximately 5 hours before flash cooling the crystals in liquid ethane.

TEG12-ternary crystals were co-crystallized at 4 °C in the presence of 2 mM PAP and 1 mM teicoplanin aglycone. TEG12 was first concentrated to 20 mg/ml in protein buffer. The protein was then diluted 1:1 with 50 mM CHES, pH 9.1, 2 mM teicoplanin aglycone, 4 mM PAP, achieving a final concentration of 10 mg/ml TEG12 in 0.5X protein buffer, 25 mM CHES, pH

9.1, 1 mM teicoplanin aglycone, 2 mM PAP. 1 μl of protein was mixed 1:1 with reservoir solution (0.2 M ammonium acetate and 20% w/v PEG 3350, JCSG, core I-25, Qiagen). Crystals appeared in 2-3 days and grew to a maximal size of 100 μm X 50 μm X 50 μm in approximately 1 week. These crystals were of an irregular chunk-like morphology and had cracks throughout. Crystals were cryo-protected by quickly dunking in reservoir solution plus 15% ethylene glycol and were flash cooled in liquid nitrogen.

TEG12-binary crystals were co-crystallized at 4 °C in the presence of 1 mM teicoplanin aglycone. Similar to TEG12-ternary crystallization, the protein was first concentrated to 20 mg/ml in protein buffer, then diluted to 10 mg/ml with 50 mM CHES, pH 9.1, 2 mM aglycone (final 0.5X protein buffer, 25 mM CHES, 1 mM teicoplanin aglycone). 1 μl of protein solution was mixed with 1 μl of reservoir solution (2.0 M sodium formate, 0.1 M sodium acetate, pH 4.6, JCSG core III-85, Qiagen). Cubic crystals grew between 2 and 3 weeks, and were approximately 50 μm X 50 μm X 50 μm. Crystals were soaked in 6.0 M sodium formate, 0.1 M sodium acetate, pH 4.6, 1 mM teicoplanin aglycone overnight, prior to flash cooling in liquid nitrogen.

3.5.3 Data Collection and Structure Solving

All data sets were reduced and scaled using the HKL2000 package (Otwinowski and Minor 1997). Data for TEG12-apo crystals were collected at the NSLS, beamline X29A. All but one of the crystals screened diffracted

poorly to approximately 4 Å resolution. The crystal from which the 2.91 Å dataset was collected rotated briefly out of the cryostream, and thereby had gone through a room temperature annealing cycle of several seconds. Diffraction from this crystal was dramatically improved compared with other crystals taken from the same drop. Data for TEG12-apo was reduced and scaled in space group $C222_1$. Phase information was obtained by molecular replacement using the program Phaser and StaL (GenBank accession number AAM80529, PDB code 2OV8), devoid of all flexible loops, as the search model (McCoy, Grosse-Kunstleve et al. 2007). The initial molecular replacement model was refined against the TEG12-apo dataset using rigid body refinement in Refmac (Murshudov, Vagin et al. 1997). Additional features of the map were enhanced through density modification, and 2-fold ncs averaging in CNS (Brunger, Adams et al. 1998; Brunger 2007). The model was rebuilt manually using the program Coot (Emsley and Cowtan 2004). Full restrained refinement was carried out using the translation/libration/screw model in Refmac, with the addition of hydrogen atoms, converging to a final R_{work} and R_{free} of 21.70 and 27.05, respectively (Winn, Isupov et al. 2001). NCS restraints were not used during refinement. The final model comprises residues 1-129, 136-203, and 251-285 for monomer A, and 1-27, 42-128, 137-210, 247-285 for monomoer B. The TEG12-apo model was used as a molecular replacement model for all subsequent structures.

TEG12-ternary and TEG12-binary data sets were collected at the APS, microfocus beamline 24-IDE. TEG12-ternary data was reduced and scaled in space group $P2_12_12_1$. TEG12-binary was scaled in space group $I2_12_12_1$. The crystal structure of glycopeptide aglycone A-40926 was used as a starting point to generate a restraint definition file for teicoplanin aglycone using the program Phenix Elbow (Schafer, Schneider et al. 1996).Geometry optimization was achieved using the semi-empirical quantum mechanical AM1 method. TEG12-binary and TEG12-ternary models were refined using the translation/libration/screw model in Phenix Refine to a final R_{work} and R_{free} of 17.30 and 22.61, and 17.12 and 22.47, respectively (Winn, Isupov et al. 2001). The final TEG12-binary model comprises residues 1-129, 135-216, and 247-285. The final TEG12-ternary model comprises residues 1-215, 220-224, and 231-285 for monomer A, and residues 1-129, 136-224, and 240-285. All structures were validated using the Molprobity server from the Richardson laboratory at Duke University (Davis, Leaver-Fay et al. 2007).

3.5.4 Site-Directed Mutagenesis

TEG12 point mutants were generated using the "megaprimer" method, with slight modifications (Sarkar and Sommer 1990). Oligonucleotide primers were designed for each mutant (Table 3), and a megaprimer was generated by PCR amplification from the TEG12/pET28a construct using the Pfx Accuprime System (Invitrogen), the relevant mutant oligonucleotide primer, and either the T7 promoter (for mutations at residues 9-108) or the

T7 terminator (for mutations at residues 167-251) as the second oligonucleotide primer, (30 rounds of amplification: 95 °C for 30 s, 55°C for 30 s, 68 °C for 30 s). The full length mutant TEG12 gene was amplified from the TEG12/pET28a construct, using the megaprimer, which then contained the bases that code for the specific mutant residue, and either the T7 terminator (for mutations at residues 9-108) or the T7 promoter (for mutations at residues 167-251) as the second oligonucleotide primer (30 rounds of amplification: 95 °C for 30 s, 55 °C for 30 s, 68 °C for 80 s). Full-length mutant amplicons were then sequentially digested with BamHI and HindIII, and subsequently ligated into BamHI/HindIII doubly digested pET28a. Ligated constructs were transformed into *E. coli* EC100 (Epicentre), and sequenced to identify successfully mutated constructs. Mutant constructs containing the desired point mutation were then transformed into *E. coli* BL21 (DE3) for protein expression.

3.5.5 Mutant TEG12 Expression and Purification

Mutant proteins were expressed and purified in a manner similar to the native TEG12, except on a reduced scale. 100 mL overnight expression cultures were pelleted and resuspended in 4 mL lysis buffer. After sonication to lyse the cells, the crude lysates were centrifuged to remove insoluble material (10 min at 15,000 x g). The cleared lysates were incubated with 100 μl Ni-NTA resin for 15 min. The slurry was then loaded onto a column, allowed to empty by gravity flow, washed with 4 mL lysis buffer, followed by

a second wash with 4 mL wash buffer. The protein was eluted by the addition of 1.5 mL elution buffer. All TEG12 mutants used in activity assays appeared to be homogeneous by polyacrylamide gel electrophoresis

3.5.6 TEG12 Activity Assays

All soluble TEG12 mutants were assayed for activity using the teicoplanin aglycone as a substrate. 50 μL reactions were run in duplicate, as follows: 15 mM HEPES, pH 7.5, 1 mM 3'-phosphoadenosine-5'-phosphosulfate (PAPS), 0.1 mM DTT, 1.2 mM teicoplanin aglycone (in DMSO), and 500 ng purified protein in elution buffer. Reactions were carried out at 30 °C for each of the four time points (10, 15, 20, 25 min), followed by heat inactivation at 99 °C for 10 min, and a further 10 min in an ice water bath. V_{max} and K_m values were determined under the same reaction conditions using the teicoplanin aglycone as substrate (5 μM to 100 μM). 25 μL of each reaction was run on a Waters analytical HPLC system (C_{18} (4.6 x 150 mm)). A linear gradient (1.5 ml/min) was run from an initial condition of 95:5 20 mM ammonium acetate:acetonitrile to 70:30 20 mM ammonium acetate:acetonitrile over twenty minutes. The area under the UV peak (Diode Array, 240 nm-400 nm) was determined for both the monosulfated product and the teicoplanin aglycone substrate at each time point. The percent substrate conversion for duplicate time points was averaged. The slope of the graph derived from the four time points for TEG12 and each

mutant was then determined. Relative activity of each mutant is reported as a percent of the slope for wild-type TEG12.

CHAPTER 4

4 Selective Enrichment of eDNA Libraries for Clones Rich in Secondary Metabolism Genes

4.1 Chapter Summary

Chapter 4 outlines the process by which the AB and AZ eDNA-derived cosmid mega-libraries, discussed in chapter 2, were enriched for genes associated with secondary metabolite biosynthetic gene clusters. The selection-based strategy used to achieve this enrichment involved the eDNA-mediated complementation of the *E. coli* enterobactin biosynthetic pathway (*entA-F),* which is responsible for siderophore production and is required for growth under iron-limiting conditions. The enterobactin pathway gene *entD* encodes for a 4'-phosphopantetheinyltransferase (PPtase) enzyme that carries out the post-translational modification of EntB and EntF, the non-ribosomal peptide synthetases (NRPSs) that biosynthesize the enterobactin backbone. When *entD* was knocked out of *E. coli*, the resulting mutant strain was unable to grow on iron-deficient media (modified M9 media). However, complementation of this mutant strain by exogenous PPtases encoded on eDNA-derived cosmids was able to successfully rescue *E. coli* by restoring enterobactin biosynthesis. As is the case with *entD* and the enterobactin pathway from *E. coli,* PPtases from Gram-negative organisms are frequently

clustered with the biosynthetic gene clusters upon which they act. Consequently, this selection strategy is able to enrich eDNA megalibraries for clones harboring secondary metabolite biosynthetic genes in addition to the functional PPTase activities through which such clones are selected. Next-generation (454) pyrosequencing of eDNA library clones selected for in this manner was then used to compare the "secondary metabolism" contents of enriched versus non-enriched library clones. This analysis demonstrated an increase in the secondary metabolism content of enriched library clones compared to randomly selected (non-enriched) library clones. This increase corresponded to over an order of magnitude rise in the percentage of nucleotide bases derived from NRPS, Polyketide synthase (PKS), or PPtase genes found within the enriched library. This preliminary demonstration of the efficacy of this enrichment strategy is a promising indication that such methods might be used in the future to quickly and easily obtain novel secondary metabolites from eDNA mega-libraries of over ten million unique cosmid clones.

4.2 Introduction

eDNA-derived cosmid mega-libraries possess enormous biosynthetic potential. However, bulk eDNA samples represent pools of bacterial genomic DNA and are therefore undoubtedly dominated by genes encoding proteins involved in primary metabolic functions. Given the enormous amount of genetic information contained in a 10+ million-member eDNA-derived cosmid

library (~100,000 bacterial genomes), the identification of novel biosynthetic information from the milieu possess a rather arduous task. If, however, the cosmid clones containing genes likely to be involved in secondary metabolite biosynthesis could be isolated from the rest, a more systematic, exhaustive examination of the remaining clones could be carried out. The strategy described in this chapter utilizes a complementation strategy to enrich an eDNA cosmid mega-library for clones likely to be rich in secondary metabolite biosynthetic enzymes.

This enrichment strategy focuses on the enrichment of libraries for clones containing two of the more common classes of secondary metabolite biosynthetic enzymes, non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs). These enzymes are post-translationally modified by the action of 4'-phosphopantetheinyltransferases (PPtases), whose action is essential to the maturation of NRPSs and PKSs, producing active enzyme via the covalent attachment of a 4'-phosphopantetheinyl group to a conserved serine residue in the peptidyl carrier protein (PCP) of NRPSs or the acyl carrier (ACP) of PKS proteins (Macpherson, Manning et al.). NRPSs and PKSs are capable of biosynthesizing complex small molecules from simple starting materials. The genes encoding PPtases are often found either internal to or in close proximity to the NRPS and PKS systems they modify. This proximal linkage was essential to the selection of PPtases as a

target for our complementation strategy to enrich our metagenomic libraries for NRPS and PKS enzymes.

Enterobactin (**37**) is an NRPS-derived siderophore produced by *E. coli* (Figure 18) (Liu, Duncan et al. 1989). In addition to the NRPS biosynthetic genes *entE, entB, and entF* contained in the enterobactin biosynthetic gene cluster, there is a cluster-specific PPtase, *entD*. The removal of *entD* from *E. coli* strain EC100, which produces enterobactin as its sole siderophore, abolishes the ability of the *entD*-minus strain to produce this siderophore and results in the inability to grow on iron deficient media. By targeting *entD*, it becomes possible to select for eDNA-derived cosmid clones capable of complementing the enterobactin pathway, generally through the expression of an active, exogenous PPtase (Figure 18). The selection of clones based on the presence of an expressed PPtase allows us to take advantage of the commonly observed proximal linkage of genes encoding PPtases and genes encoding NRPSs and PKSs to enrich a library for two of the most common classes of enzymes seen in secondary metabolite biosynthesis.

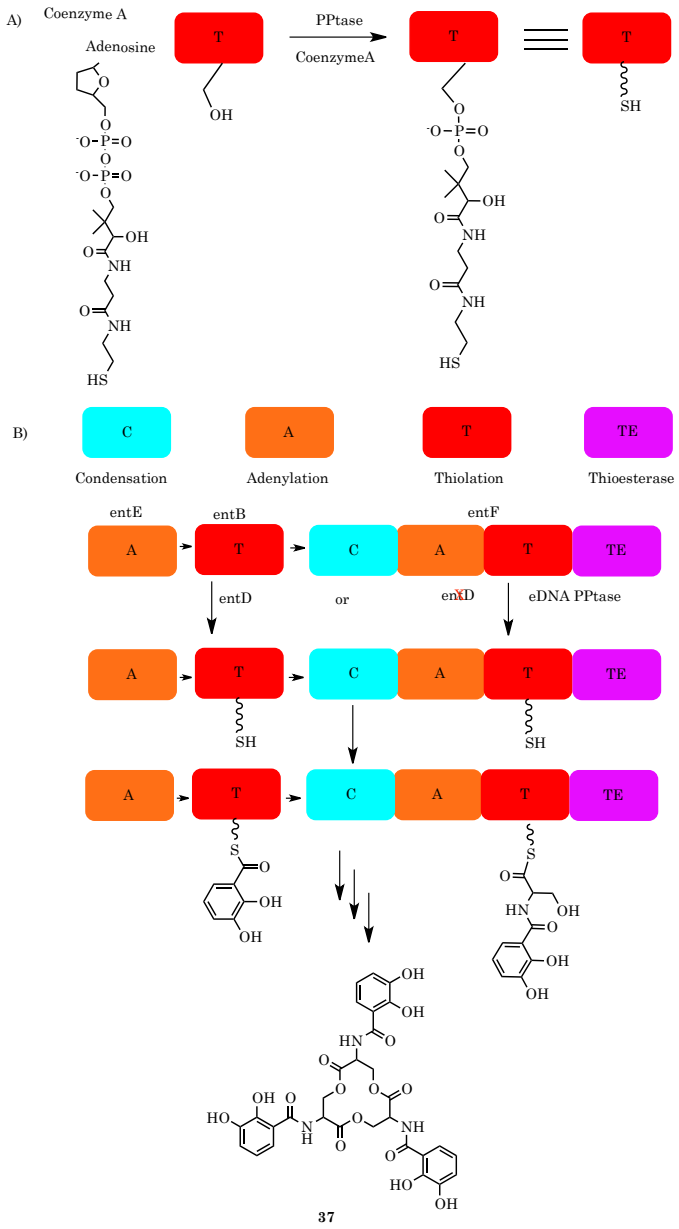


Figure 18: Enterobactin biosynthesis and complementation strategy

A): 4'-phosphopantetheinylation of PKSs and NRPSs (on a conserved serine on the ACP or PCP, respectively) leads to the activation of these enzymes. B): Enterobactin (**37**) is an NRPS derived siderophore, utilized by *E. coli* in iron scavenging under low iron conditions. C): Production of enterobactin in *E. coli* is disrupted by knocking out the native PPtase, entD. entD is responsible for the maturation of the NRPS system, which biosynthesizes enterobactin.

4.3 Results

4.3.1 Construction of an *entD*$^-$ *E. coli* strain for Complementation and Enrichment of an eDNA-derived Cosmid Mega-library

A PPtase-deficient *E. coli* strain (*E. coli* EC100 Δ*entD*) was generated using RecET recombination (Zhang, Muyrers et al. 2000). *E. coli* EC100 Δ*entD* (hereafter referred to as EC-Δ*entD*) does not produce enterobactin, and is thus incapable of growth on iron deficient media. EC-Δ*entD* was then used to enrich an eDNA-derived cosmid megalibrary (constructed in either of two cosmid vectors (pWEB or pWEB-TNC, Epicentre) for clones with the ability to grow on iron-deficient "*entD* complementation selection" media (see Materials and Methods).

Two cosmid mega-libraries, each containing in excess of 10,000,000 clones, were subjected to the PPtase-based enrichment strategy described above. Both libraries were transformed into EC-Δ*entD* and plated directly onto iron-deficient "entD complementation selection" media plates. Approximately 1 of every 2,500 library clones grew on the low-iron selection media. After a second round of selection, approximately 1 in every 20 library members was capable of growth on the selection media, representing a 125-fold enrichment over the first round of selection. An additional round of selection on low iron media produced a collection of clones from which 50-90% were independently proven to be capable of complementing the entD knockout mutation.

4.3.2 Enriched libraries contain an abundance of NRPS and PKS biosynthetic genes

To examine the relative success of the *entD* complementation enrichment strategy, an approximate (4.5 Mb) genome equivalent of cosmid clones was sequenced from the enriched library along with a similar equivalent from randomly selected cosmid clones from the original eDNA library used in the enrichment process. Cosmids were sequenced (454 pyrosequencing, Roche), and assembled contigs greater than 30 kb in length were submitted to the RAST online server for automated annotation (Aziz, Bartels et al. 2008). These two data sets were then analyzed for the presence of putative NRPSs and PKSs, as well as for genes encoding PPtases (Figure 19). A simple qualitative analysis, such as that displayed in Figure 19, indicated that this enrichment strategy had been successful at recovering eDNA clones rich in secondary metabolite biosynthetic enzymes. All predicted NRPS and PKS enzymes over 3 kb were then submitted as BLASTP search queries against all non-redundant protein sequences, and bacterial phyla predictions were made based on the consensus of top scoring BLAST hits. The results of this analysis are summarized in Table 5.

Figure 19: ORF prediction maps for Non-enriched (RANDOM) and Enriched library clones.

Predicted ORF maps from both sequencing runs; Non-enriched (i.e., random) (left), Enriched (right). Genes filled red are putative NRPSs and/or PKSs that are longer than 3,000 nucleotides. Genes filled yellow are putative 4'-phosphopantetheinyltransferases.

Table 5: A comparison of the fractional biosynthetic contents (% total nucleotides belonging to NRPS/PKS genes) of enriched library clones, non-enriched library clones and select bacterial genomes.

Sample	Total NTs	% NRPS/PKS
Enriched	4544371	27.31
Non-Enriched	4688540	0.88
Escherichia coli UTI89	5065741	0.54
Bacillus cereus G9842	5387334	0.14
Pseudomonas aeruginosa LESB58	6601757	0.34
Salinispora tropica CNB-440	5183331	0.53
Bacteroides fragilis YCH46	5277274	0.02
Streptomyces coelicolor A3(2)	9023530	2.31
Prochlorococcus marinus MIT 9303	2682675	0.28

In order to quantitatively assess the success of this enrichment strategy, the percentage of nucleotides contained in predicted NRPS and PKS enzymes (i.e., the "biosynthetic content") was calculated for both the enriched and non-enriched data sets and several sequenced genomes from a range of different bacteria (Table 5). As can be seen in Table 5, the percentage of nucleotides from most sequenced bacterial genomes corresponding to NRPS and PKS genes is below 1%, with the notable exception being the *Streptomyces coelicolor* A3(2) genome, which encodes the PKS derived antibiotic actinorhodin and contains over a dozen cryptic secondary metabolite biosynthetic gene clusters. Remarkably, the genomic equivalent from our enriched library has a biosynthetic content that is over an order of magnitude greater than the *Streptomyces coelicolor* A3(2) genome. Furthermore, compared to most sequenced bacterial genomes and to our

randomly selected (non-enriched) genome equivalent, the disparity in biosynthetic content is closer to two orders of magnitude. This significant level of enrichment should allow for a much more extensive examination of the select group of enriched library clones than would ever be feasible with eDNA megalibraries containing 10+ million clones.

4.3.3 Cultivation of Clones from Enriched Library

To evaluate the potential of enriched eDNA library clones to heterologously produce clone specific secondary metabolites in *E. coli*, 1152 unique clones from the enriched library were cultured in iron-deficient *entD* complementation selection medium to ensure that clone-specific PPtases were actively expressed as functional enzymes. It is logical that such conditions should increase the potential for expression of clone-specific small molecule biosynthetic gene clusters. A subset (576) of these clones was also cultured in LB medium. Individual cell cultures of these 576 clones were extracted with ethyl acetate under both neutral (unadjusted) and acidic (pH 3-4) conditions, while cultures of the remaining 576 clones were extracted only at neutral (unadjusted) pH. Analysis of these 1152 *E. coli* based heterologous expression cultures did not yield clone-specific metabolites as determined by TLC and LCMS.

4.3.4 Phenotypic Screening of Enriched Libraries

Enriched libraries were plated onto both LB and *entD* complementation selection medium to attempt to identify clones, which

possessed the ability to produce clone-specific metabolites, demonstrated by the generation of a high-yield phenotype (pigmentation, altered colony morphology, or antibiosis). Clones were plated at a titer, which would yield distinct, well-spaced colonies (300-500 clones on a 150 mm diameter plate). These plates were incubated for four days, at which point they were examined for clones displaying pigmentation or altered colony morphology. To identify clones displaying antibiosis, the plates were overlayed with one of three assay strains. After incubation for an additional two days, the plates were examined for zones of growth inhibition in the various assay strains. No clones were identified from the *entD* complementation selection medium screen that produced clone-specific metabolites. One clone, which was identified from a screen on LB media, produced a clone-specific metabolite. This clone, which did not grow on the *entD* complementation selection medium was, based on HPLC-MS, determined to be producing N-acyl tyrosines. One major hurdle, which was not overcome, was the generation of a small zone of growth inhibition by every clone, which grew on the *entD* complementation selection medium when grown in the presence of *B. subtilis*. This impediment is one that must be overcome in future work, as this bacterium is very commonly used for phenotypic screening due to its high sensitivity to many test compounds.

4.4 Discussion and Future Directions

The eDNA-derived cosmid mega-libraries constructed for the metagenomics efforts described in chapter 2 contain too many individual members to exhaustively examine each for the production of novel, clone-specific small molecules. The work presented in this chapter outlines a means by which 10+ million cosmid clones can be sorted (i.e., "selected for" or "enriched in") to identify several thousand with a very high likelihood of containing secondary metabolite biosynthetic machinery. Although the examination of several thousand individual clones in pure culture is not a trivial task, it is logistically much more possible than a similar examination of over ten million clones.

The enrichment strategy described in this chapter has obvious limitations. First and foremost, the selective-enrichment process is carried out in *E. coli*, which is predicted to be limited in terms of the expression of eDNA-derived genes (Gabor, Alkema et al. 2004). While being one of the most widely used heterologous expression hosts in the majority of functional metagenomic efforts, the development of additional heterologous hosts that are better suited for the production of secondary metabolites is likely to result in more favorable outcomes.

The bacterial phyla predictions that were made for each NRPS/PKS clone sequenced from the enriched and non-enriched libraries indicate a

major need for more taxonomically diverse hosts for the selective enrichment process. The majority of predicted NRPS and PKS enzymes from the enriched set of clones belong to either cyanobacterial or proteobacterial sources (Table 6). However, as *E. coli* is a Gram-negative member of the *γ-Proteobacteria*, it is perhaps not surprising that the majority of the NRPSs and PKSs selected for by this *E. coli*-based process derive from *Proteobacteria* or members of the Gram-negative *Cyanobacteria*. Similar complementation strategies in bacteria that are taxonomically distant from the *Proteobacteria* would presumably result in a vastly different yield of recovered NRPSs and PKSs.

Table 6: Phylum-level source predictions for eDNA-derived biosynthetic genes.

Phyla	PPtases ENRICHED	PPtases RANDOM	NRPS/PKS ENRICHED	NRPS/PKS RANDOM
Cyanobacteria	78		131	3
Proteobacteria	16	1	71	1
Verrucomicrobia	7			
Actinobacteria	6		2	
Bacteroidetes	1		3	
Chloroflexi	1			
Acidobacteria	1			
Firmicutes			4	

One additional observation that must contribute heavily to future endeavors with this system is that many of the possible pathways identified using this specific enrichment strategy will undoubtedly be truncated due to the use of cosmid-based libraries housing 35-40 kb eDNA inserts. One possible solution to this problem is the recovery of overlapping cosmid clones,

as was done for the recovery of intact glycopeptide biosynthetic gene clusters in chapter 2. Alternatively, one could imagine creating libraries with significantly larger eDNA inserts. Larger insert libraries, such as those constructed using Bacterial Artificial Chromosomes (BAC) are notoriously difficult to build, however addressing this issue may be essential to the expansion and eventual success of this strategy.

4.5 Materials and Methods

4.5.1 Library Enrichment

EC-Δ*entD* was constructed using Rec/ET recombineering (Zhang, Muyrers et al. 2000). For library enrichment, pools of two separate eDNA-derived cosmid mega-libraries, each containing in excess of ten million individual clones, were transformed into EC-Δ*entD*. These mega-libraries were constructed using two cosmid vectors: pWEB and pWEB-TNC (Epicentre). Both libraries were first transformed into the *entD- E. coli* strain, outgrown for 1 h at 37°C in SOC, and then plated onto an LB medium plate containing either 30 μg/mL kanamycin (pWEB) or 12.5 μg/mL chloramphenicol (pWEB-TNC). These plates were then scraped after the addition of liquid LB and stored away as a glycerol stock. This glycerol stock was then used as an inoculum for a 50 mL LB media culture, which was grown to an OD_{600} ~ 0.5. The culture was then pelleted, washed three times with sterile M9 media, and plated directly onto iron-deficient *entD*

complementation selection media (M9 media supplemented with 1 g/L casamino acids, 10 μM thiamine-HCl, 100 μM 2,2-dipyridyl, and either 30 μg/mL kanamycin (pWEB) or 12.5 μg/mL chloramphenicol (pWEB-TNC)) with 15 g/L agar and antibiotic selection, as well as onto LB agar with antibiotic selection to determine the titer of the enrichment. Plates were kept at 37°C, and colonies appeared after 12-16 h. Sterile M9 media was added to the selection plates after 36-48 hours, and the plates were scraped and used to inoculate an overnight LB culture containing appropriate antibiotic selection. This culture was grown overnight at 37°C and then used as an inoculum for a 50 mL LB culture, which was grown to an OD_{600} ~ 0.5, washed three times with sterile M9, and plated yet again onto iron-deficient *entD* complementation selection media containing 15 g/L agar. The plating on *entD* complementation selection media, scraping, growing overnight, growing to mid-log phase, washing, and re-plating represented one round of enrichment. A total of three enrichment rounds were carried out.

4.5.2 Sequencing of Clones from Enriched and Non-enriched Libraries

Individual colonies from the selective enrichment were chosen for sequencing following the final round of the enrichment process. Colonies were picked from plates after incubation at 37°C for 36 hours. These colonies were cultured overnight in LB medium, and mini-prepped (Qiagen). eDNA-derived cosmids for 500 individual clones were sequenced using 454 pyrosequencing

(Roche). These colonies were considered representative of the enriched library. For sequencing of the un-enriched library, pools from the AB and AZ libraries were plated out at a sufficient titer to yield well-spaced individual colonies. Colonies were cultured overnight in LB medium, and mini-prepped (Qiagen). 500 individual eDNA-derived cosmids were sequenced using 454 pyrosequencing (Roche), representing the non-enriched (original) library.

4.5.3 Bioinformatics Analysis of Cosmid Clones

126 contigs from both the enriched and the non-enriched 454 sequencing data sets, each in excess of 30 kb, were chosen for bioinformatic analysis. This quantity of sequencing information was roughly equivalent to a small-medium sized bacterial genome. Following an Open Reading Frame (ORF) prediction using the RAST server (Aziz, Bartels et al. 2008), all predicted PKSs and/or NRPSs in excess of 3 kilobases were assigned as "PKSs or NRPSs" and are depicted in Figure 19. Predicted PPtases are also shown in Figure 20. A calculation was then made to express the fraction of nucleotides from each 126 contig data set that were contained in predicted PKSs, NRPSs, and PPtases as a percentage of the total nucleotides sequenced from each data set. A similar calculation was made for a few select bacterial genomes, representing species found across various major phyla. This data is reported in Table 5.

A phylogenetic analysis of the PPtases identified from the enriched library was then performed using a BlastP analysis. Probable phylum-level

source predictions were made for each individual PPtase based on origins of the top Blast hit(s). A tabulation of this analysis is found in Table 6.

4.5.4 Culture of Individual Cosmid Clones

Individual *E. coli* clones from the final round of library enrichment were cultured and extracted with organic solvent to determine if this enrichment technique was identifying clones with the immediate potential to produce clone-specific secondary metabolites. Following the final stage of library enrichment, 1152 colonies were picked and used to inoculate 96 well culture plates containing LB medium. These plates were incubated overnight at 37°C. An equal volume of 30% sterile glycerol was added to each well, mixed, and stored at -80°C. 25 mL of culture media (either iron-deficient *entD* complementation selection medium or LB medium) was added to 50 mL conical tubes. Individual tubes were inoculated with a single colony from the stored glycerol stocks (their caps attached loosely for air exchange) and placed at 30°C and 200 rpm orbital shaking for 4 days. After 4 days, tubes were removed from the incubator and mixed with an equal volume (~ 20 mL) of ethyl acetate.

The first 576 cultures were grown both in LB and in iron-deficient *entD* complementation selection media, and were extracted both under neutral (unadjusted) pH and under acidic (adjusted to 3-4 using 10% HCl) pH. A negative control of *E. coli* EC100 transformed with pWEB-TNC was added to each round of cultures. Extractions were centrifuged briefly (3 min) at ~3,000

x g and the ethyl acetate fraction was decanted into glass tubes for vacuum drying. Dried samples were resuspended in methanol (~0.5 mL) and transferred to 1 dram glass vials to concentrate the sample for TLC. Dried samples were resuspended in 40 μL of methanol (acidified for adjusted samples) and 5 μL was loaded onto a TLC plate. These plates were run in 90:10 chloroform:methanol and assayed under UV-light (365 nm) and by crystal iodine staining. *E. coli* EC100 transformed with pWEB-TNC was used as a control culture.

Cultures displaying possible clone-specific metabolite production were subjected to LCMS using a Waters analytical HPLC system (C_{18} (4.6 x 150 mm)). Initial conditions were 90:10 water containing 0.1% formic acid:methanol, followed by a linear gradient from 90:10 to 0:100 over 20 minutes, followed by 5 minutes at 0:100. Based on the observed background for the LB cultures and cultures extracted under acidic conditions, the second 576 cultures were grown exclusively in iron-deficient *entD* complementation selection media and extracted under neutral (unadjusted) pH.

4.5.5 Phenotypic Screening of Enriched Libraries

A glycerol stock of each enriched library was used as an inoculum for a 50 mL LB media culture. This culture was grown at 37°C with 200 rpm shaking until the OD_{600} reached 0.5. The culture was then washed three times with 50 mL of sterile M9 media and plated. A titer of 300-500 clones on a 150 mm diameter plate was achieved by plating 5, 10, and 50 μL of a 10^{-4}

serial dilution. Cultures were plated on solid *entD* complementation selection media with antibiotic selection (M9 media supplemented with 1 g/L casamino acids, 10 μM thiamine-HCl, 100 μM 2,2-dipyridyl, 15 g/L agar) and 100 μg/mL ampicillin and LB agar with 100 μlgmL ampicillin. These plates were incubated overnight at 30°C, removed to room temperature, where they remained for a following three days. The plates were then examined for clones displaying pigmentation or altered colony morphology. To identify clones displaying antibiosis, the plates were overlayed with *E. coli* BAS849, *B. subtilis* 1E9, or *Saccharomyces cerevisiae* W303. 100 plates each were overlayed for each strain. Overlay was carried out using top agar (LB or *entD* complementation selection media with 5 g/L agar) inoculated with a 10^{-4} dilution of an OD_{600} = 0.5 of each assay strain. After incubation for an additional two days at room temperature, the plates were examined for zones of growth inhibition in the various assay strains.

CHAPTER 5

5. Future Directions

5.1 BAC Libraries

One of the largest obstacles to the recovery of intact secondary metabolite biosynthetic gene clusters from eDNA-derived cosmid libraries is the low probability (for medium sized gene clusters) or complete inability (for very large gene clusters) of many pathways to be contained on an individual cosmid due to size constraints. Additionally, even if one recovers a complete pathway contained on overlapping cosmid clones, the reconstruction of a contiguous pathway on a single linear vector has been, until recent developments with transformation associated recombination (TAR) in yeast, technically very difficult (Kim, Feng et al. 2010). An obvious solution to this problem is to simply construct larger insert libraries. If individual clones contained over 100 kilobases of contiguous eDNA, many of the difficulties associated with cosmid-based heterologous expression studies would be overcome. One strategy that has been used to construct large insert gDNA libraries involves the construction of bacterial artificial chromosomes or BACs, which have been reported to stably maintain over 300 kilobases of DNA (Tao and Zhang 1998).

One problem with the construction of BAC libraries for metagenomics research has been the technical inability to produce large collections of BAC clones. Initial studies yielded libraries with at most tens of thousands of clones (Rondon, August et al. 2000). Thus, it was hypothesized that one must sacrifice overall metagenomic coverage to gain access to larger contiguous pieces of DNA. Recent advancements in BAC cloning have led to the production of libraries containing in excess of one million clones (Magbanua, Ozkan et al. 2011). Additional internal observations in the Brady lab have determined that using desert soils, which contain less of the organic substances (humates, etc.) found in richer soils collected in eastern temperate environments, cosmid cloning efficiency can be increased by orders of magnitude. Therefore, it seems likely that the production of multi-million member eDNA-derived BAC libraries from soil bacteria is an inevitablilty. Such advancements will increase the rate with which large pathways are identified and used successfully in downstream heterologous expression experiments.

5.2 eDNA-derived Cosmid Libraries Hosted in *Streptomyces sp.*

Much of the heterologous expression of eDNA-derived secondary metabolite biosynthetic gene clusters has been carried out in *Streptomyces sp.* (Banik and Brady 2008; King, Bauer et al. 2009; Bauer, King et al. 2010; Feng, Kim et al. 2010). *Streptomyces* produces many of the secondary

metabolites that have been identified using culture-based methods (Buckingham 1994). Creating an eDNA-derived cosmid library and transferring the entire library to various *Streptomyces sp.* would open up a number of different avenues in terms of metagenomics research. First, with the possible expansion of enrichment strategies such as described in chapter 4 to Gram-positive species, it would seem only natural to transfer a library enriched in a Gram-positive host into representatives of this prolific small molecule producing genera of bacteria. Additionally, functional metagenomics experiments could be carried out, where colonies from multiple million-member libraries could be assayed for phenotypes commonly associated with the production of clone-specific metabolites. Much smaller metagenomic libraries have already been hosted in *Streptomyces*, with some success (Wang, Graziani et al. 2000). However, the expansion from just over a thousand member library to a greater than ten million member library will undoubtedly increase the productivity of a strategy that is the natural crossroads of multiple avenues.

5.3 Large-scale Sequencing Efforts

The discovery of bioactive small molecules using metagenomic methods will undoubtedly benefit greatly from future advances in sequencing technology that allow for the comprehensive sequencing of complex microbiomes (Chan, Hsu et al. 2008; Morozova and Marra 2008; Glass, Wilkening et al. 2010; Yang, Peng et al. 2010), as well as from increasing our

understanding of the expression barriers encountered by foreign DNA in model laboratory grown bacterial hosts (Wang, Isaacs et al. 2009; Komatsu, Uchiyama et al. 2010). The TerraGenome project was established in 2008 in an effort to bring together sufficient sequencing power to sequence the first complete soil microbiome (Vogel, Simonet et al. 2009). Although only a small number of compounds have been characterized to date using culture-independent methods, these initial studies indicate that as yet uncultured bacteria are likely to be a rich source of previously unknown biologically active small molecules.

APPENDIX

Table 7: Predicted ORFs for VEG pathway

Protein length is given in amino acids (AA).

VEG ORFs ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA% ID (*Organism*)	Accession #
1	319	Transcriptional Regulator	Transcriptional Regulator	92%(*A. orientalis*)	CAB45047.1
2	104	Transcriptional Regulator	Transcriptional Regulator	72%(*A. orientalis*)	CAB45048.1
3	415	Integral Membrane Antiporter	Integral Membrane Antiporter	91%(*A. orientalis*)	CAB45049.1
4	356	Oxidoreductase	Oxidoreductase	87%(*A. orientalis*)	CAB45050.1
5	693	ABC Transporter	ABC Transporter	78%(*A. orientalis*)	CAB45051.1
6	2069	Nonribosomal Peptide Synthetase Modules 1-2	NRPS	75%(*A. orientalis*)	CAB45052.1
7	1057	Nonribosomal Peptide Synthetase Module 3	NRPS	73%(*A. teichomyceticus*)	CAE53351.1
8	4002	Nonribosomal Peptide Synthetase Modules 4-6	NRPS	82%(*A. orientalis*)	CAA11795.1
9	1866	Nonribosomal Peptide Synthetase Module 7	NRPS	87%(*A. orientalis*)	CAA11796.1
10	382	P450 Monooxygenase OxyA	OxyA	80%(*A. orientalis*)	AEI58868.1
11	356	P450 Monooxygenase OxyB	OxyB	92%(*A. balhimycina*)	CAA76548.1
12	430	P450 Monooxygenase OxyC	OxyC	87%(*A. orientalis*)	AEI58870.1
13	492	Halogenase	Halogenase	94%(*A. orienatlis*)	CAA11780.1
14	390	Glycosyltransferase	GtfA	69%(*A. orientalis*)	CAA11774.1
15	409	Glycosyltransferase	GtfE	78%(*A. orientalis*)	AAB49299.1
16	351	Glycosyltransferase	GtfC	68%(*A. orientalis*)	AAB49294.1
17	408	Methyltransferase	Methyltransferase	87%(*A. orientalis*)	CAA11777.1
18	273	Methyltransferase	None	56%(*Frankia sp.*)	ECF83210.1
19	379	Glycosyltransferase	GtfA	67%(*A. teichomyceticus*)	CAE53349.1
20	268	Deacetylase	Deacetylase	67%(*A .teichomyceticus*)	CAE53355.1
21	578	Mannosyltransferase	Mannosyltransferase	84%(*A. teichomyceticus*)	CAE53356.1
22	604	Mannosyltransferase	Mannosyltransferase	80%(*A. teichomyceticus*)	CAE53356.1
23	407	Transposase	None	67%(*Frankia sp.*)	EFC78751.1
24	283	N-methyltransferase	N-methyltransferase	69%(*A. orientalis*)	CAA11779.1
25	280	N-methyltransferase	N-methyltransferase	55%(*A. orientalis*)	CAA11779.1
26	433	Transcriptional Regulator	Transcriptional Regulator	92%(*A. orientalis*)	CAA11790.1
27	277	BHT Perhydrolase	BhtA	96%(*A. orientalis*)	CAA11784.1
28	578	BHT Peptide Synthetase	BhtB	94%(*A. orientalis*)	CAA11773.1
29	397	BHT Oxygenase	BhtC	91%(*A. orientalis*)	CAA11772.1
30	358	Hydroxymandelate Synthase	HmaS	91%(*A. orientalis*)	CAA11761.1
31	370	Hydroxymandelate Oxidase	Hmo	90%(*A. orientalis*)	CAA11762.1
32	472	NDP-Hexose-2,3,-Dehydratase	EvaA	90%(*A. orientalis*)	CAA11763.1
33	326	NDP-Hexose-4-Ketoreductase	EvaD	86%(*A. orientalis*)	CAA11764.1
34	370	NDP-Hexose-3-Aminotransferase	EvaB	93%(*A. orientalis*)	CAA11782.1
35	203	NDP-Hexose-3,5-Epimerase	EvaE	82%(*A. orientalis*)	CAA11781.1
36	379	Glycosyltransferase	None	48%(*M. echinospora*)	AF505622_8
37	373	Dihydroxyphenylacetic Acid Synthase	DpgA	95%(*A. orientalis*)	CAA11765.1
38	218	Enoyl-CoA Hydratase/Isomerase	DpgB	86%(*A. orientalis*)	CAA11766.1

39	431	Dihydroxyphenylacetyl-CoA Oxygenase	DpgC	91%(*A. orientalis*)	CAA11787.1
40	268	Enoyl-CoA Hydratase/Isomerase	DpgD	91%(*A. orientalis*)	CAA11767.1
41	357	Aldolase	Aldolase	88%(*A. orientalis*)	CAA11768.1

Table 8: Predicted ORFs for TEG Pathway

Protein length is given in amino acids (AA).

TEG ORFs ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA% ID (*Organism*)	Accession #
1	322	Transcriptional Regulator	Transcriptional Regulator	86%(*A. orientalis*)	CAB45047.1
2	122	Chorismate Mutase	None	80%(*T.fusca YX)*	AAZ55245.1
3	363	Prephenate Dehydrogenase	Prephenate Dehydrogenase	89%(*A. orientalis*)	CAA11792.1
4	655	ABC Transporter	ABC Transporter	79%(*A. teichomyceticus*)	CAE53357.1
5	2078	Nonribosomal Peptide Synthetase Modules 1-2	NRPS	79%(*A. orientalis*)	CAA11794.1
6	1070	Nonribosomal Peptide Synthetase Module 3	NRPS	81%(*A. teichomyceticus*)	CAE53351.1
7	4133	Nonribosomal Peptide Synthetase Modules 4-6	NRPS	82%(*A. orientalis*)	CAA11795.1
8	1865	Nonribosomal Peptide Synthetase Module 7	NRPS	83%(*A. orientalis*)	CAA11796.1
9	392	P450 Monooxygenase OxyA	OxyA	81%(*A. teichomyceticus*)	CAE53359.1
10	386	P450 Monooxygenase OxyD	OxyD	79% (*S. toyocaensis*)	AAM80534
11	399	P450 Monooxygenase OxyB	OxyB	83%(*A. teichomyceticus*)	CAE53361.1
12	286	Sulfotransferase	StaL	49%(*S. toyocaensis*)	AAM80529.1
13	277	Sulfotransferase	StaL	54%(*S. toyocaensis*)	AAM80529.1
14	276	Sulfotransferase	StaL	51%(*S. toyocaensis*)	AAM80529.1
15	411	P450 Monooxygenase OxyC	OxyC	84%(*A. teichomyceticus*)	CAG15021.1
16	494	Halogenase	Halogenase	86%(*A. teichomyceticus*)	CAG15020.1
17	444	Dyp-Type Peroxidase	None	42%(*M. nodulans*)	ACL56652.1
18	583	Pyrrolo-Quinoline Quinone	None	24%(*M. maripaludis*)	ABX01476.1
19	393	Arylsulfotransferase	None	38% (*S. sp. SN-1061M*)	ADC96650.1
20	435	Hydroxyphenylglycine Transaminase	HpgT	89% (*A. orientalis*)	CAC48367.1
21	277	BHT Perhydrolase	BhtA	91% (*A. orientalis*)	CAA11784.1
22	588	BHT Peptide Synthetase	BhtB	89% (*A. orientalis*)	CAA11773.1

Table 9: Predicted ORFs for AB37 Pathway

Protein length is given in amino acids (AA).

AB 37 ORFs ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA% ID (*Organism*)	Accession #
1	202	TetR Family Transcriptional Regulator	None	62%(*K. flavida*)	ZP_03865120
2	229	Short Chain Oxidoreductase	None	82%(*K. flavida*)	ZP_03863529
3	291	Phytanoyl CoA Dioxygenase	None	43%(*M. carbonacea*)	ZP_04605278
4	217	D-ala-D-ala Peptidase	VanX	85%(*A. teichomyceticus*)	CAE53345
5	346	D-ala-D-lac Ligase	VanA	86%(*A. teichomyceticus*)	CAE53344
6	347	D-lactate Dehydrogenase	VanH	85%(*A. teichomyceticus*)	CAE53343
7	430	D-ala-D-ala Carboxypeptidase	DalaDala Carboxypeptidase	65%(*A. balhimycina*)	CAG25753
8	74	Two-Component Regulatory System	VanR	63% (*A. balhimycina*)	CAG25752
9	387	Two-Component Regulatory System	VanS	72%(*A. teichomyceticus*)	CAE53347
10	146	N-acetyltransferase	None	38%(*B. cepacia*)	YP_368498
11	448	UDP-N-acetylmuramoyltripeptide D-ala-Dala Ligase	MurF	78%(*A. teichomyceticus*)	CAE53342
12	369	DalaDala Ligase	None	62%(*T. fusca YX*)	YP_289698
13	370	UDP-glucosaminyltransferase	None	88%(*K. flavida*)	ZP_03862660
14	396	Ferredoxin Reductase	None	71%(*S. avermitilis*)	NP_822785
15	348	C/O-Methyltransferase	None	48%(*S. ghanaensis*)	ZP_04690473
16	2085	Nonribosomal Peptide Synthetase, modules 1-2	NRPS 1+ 2	79%(*A. teichomyceticus*)	CAE53350
17	1482	Nonribosomal Peptide Synthetase, module 3	NRPS3	83%(*A. teichomyceticus*)	CAE53351
18	4075	Nonribosomal Peptide Synthetase, modules 4-6	NRPS 4-6	82%(*A. teichomyceticus*)	CAE53352
19	1850	Nonribosomal Peptide Synthetase, module 7	NRPS 7	86%(*A. teichomyceticus*)	CAE53353
20	491	NDP-Hexose-2,3-Dehydratase	EvaA	56%(*A. balhimycina*)	CAC48374
21	357	Glucose Thymidyltransferase	Glucose Thymidyltransferase	69%(*S. sviceus*)	YP_002205669
22	378	Glycosyltransferase	GtfA	65%(*A. teichomyceticus*)	CAE53349
23	407	Glycosyltransferase	GtfB	72%(*A. teichomyceticus*)	CAE53364
24	572	Mannosyltransferase	MtfA	77%(*A. teichomyceticus*)	CAE53356
25	629	ABC Transporter	ABC transporter	83%(*A. teichomyceticus*)	CAE53357
26	392	P450 Monooxygenase OxyA	OxyA	88%(*A. teichomyceticus*)	CAE53359
27	385	P450 Monooxygenase OxyB	OxyB	86%(*A. teichomyceticus*)	CAE53360
28	399	P450 Monooxygenase OxyD	OxyD	86%(*A. teichomyceticus*)	CAE53361
29	411	P450 Monooxygenase OxyC	OxyC	87%(*A. teichomyceticus*)	CAE53363
30	441	Halogenase	Halogenase	69%(*A. teichomyceticus*)	CAE53362
31	507	Non-heme Iron Dioxygenase	Non-heme Iron dioxygenase	93%(*A. teichomyceticus*)	CAE53366
32	356	DAHP Synthase	DAHP synthase	90%(*A. teichomyceticus*)	CAE53368
33	331	Str Family Transcriptional Regulator	Transcriptional Regulator	83%(*A. teichomyceticus*)	CAE53369
34	230	Response Regulator	Res. Regulator (2 comp)	88%(*A. teichomyceticus*)	CAE53348
35	836	LuxR Regulator	LuxR (2 comp)	78%(*A. teichomyceticus*)	CAE53370
36	368	Dihydroxyphenylacetic Acid Synthase	DpgA	93%(*A. teichomyceticus*)	CAE53371
37	222	Enoyl-CoA Hydratase/Isomerase	DpgB	84%(*A. teichomyceticus*)	CAE53372
38	434	Dihydroxyphenylacetyl-CoA Oxygenase	DpgC	88%(*A. teichomyceticus*)	CAE53373
39	239	Enoyl-CoA Hydratase/Isomerase	DpgD	91%(*A. teichomyceticus*)	CAE53374

40	199	GTP Cyclohydrolase	GTP Cyclohydrolase	91%(*A. teichomyceticus*)	CAE53376
41	420	Hydroxyphenylglycine Transaminase	HpgT	89%(*A. teichomyceticus*)	CAE53377
42	374	Prephenate Dehydrogenase	Prephenate Dehydrogenase	83%(*A. teichomyceticus*)	CAG15036
43	301	Transcriptional Regulator	None	50%(*S. kasugaensis*)	BAC53615
44	329	Transcriptional Regulator	None	50%(*S. griseus*)	YP_001827443
45	522	Non-heme Iron Dioxygenase	Non-heme Iron Dioxygenase	70%(*A. teichomyceticus*)	CAG15037
46	352	Hydroxymandelate Synthase	HmaS	92%(*A. teichomyceticus*)	CAE53378
47	365	Hydroxymandelate Oxygenase	Hmo	87%(*A. teichomyceticus*)	CAE53379

Table 10: Predicted ORFs for AB878 Pathway

Protein length is given in amino acids (AA).

AB878 ORFs					
ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA %ID (*Organism*)	Accession #
1	350	D-lactate Dehydrogenase	VanH	74% (*A. teichomyceticus)*	CAG15002
2	346	D-ala-D-lac Ligase	VanA	77% (*A. teichomyceticus)*	CAE53344
3	203	D-ala-D-ala Dipeptidase	VanX	81% (*A. teichomyceticus)*	CAE53345
4	344	Two-Component Regulatory System	VanR	69% (*S. toyocaensis*)	AAM80542
5	224	Two-Component Regulatory System	VanS	90% (*S. toyocaensis*)	AAM80541
6	161	D-ala-D-ala Carboxypeptidase	VanY	73% (*A. balhimycina*)	CAG25753
7	322	Str Family Transcriptional Regulator	Trans Reg	85% (*A. balhimycina*)	CAG25754
8	358	Prephenate Dehydrogenase	Prephenate Dehydrogenase	83% (*A. balhimycina*)	CAG25755
9	709	ABC Transporter	ABC	81% (*A. teichomyceticus*)	CAE53357
10	1858	Nonribosomal Peptide Synthetase, modules 1-2	NRPS 1-2	77% (*A. orientalis*)	CAA11794
11	1058	Nonribosomal Peptide Synthetase, modules 3	NRPS 3	80% (*A. teichomyceticus*)	CAE53351
12	4078	Nonribosomal Peptide Synthetase, modules 4-6	NRPS 4-6	84% (*A. orientalis*)	CAA11795
13	1835	Nonribosomal Peptide Synthetase, modules 7	NRPS 7	87% (*A. orientalis*)	CAA11796
14	364	P450 Monooxygenase OxyA	OxyA	81% (*A. teichomyceticus*)	CAE53359
15	386	P450 Monooxygenase OxyB	OxyB	78% (*A. teichomyceticus*)	CAE53360
16	356	P450 Monooxygenase OxyD	OxyD	86% (*A. orientalis*)	CAA11798
17	346	DAHP Synthase	DAHP Synthase	72% (*A. teichomyceticus*)	CAE53368
18	411	P450 Monooxygenase OxyC	OxyC	81% (*A. teichomyceticus*)	CAG15021
19	492	Halogenase	Halogenase	93% (*A. orientalis*)	CAA11780
20	388	Glycosyltransferase	GtfA	70% (*A. orientalis*)	AAB49292
21	409	Glycosyltransferase	GtfB	76% (*A. orientalis*)	AAB49293
22	407	Glycosyltransferase	GtfC	72% (*A. orientalis*)	AAB49294
23	278	Methyltransferase	None	54% (*S. coelicolor)*	NP_628799
24	385	Glycosyltransferase	GtfA	69% (*A. teichomyceticus*)	CAE53349
25	286	Deacetylase	Deacetylase	71% (*A. teichomyceticus*)	CAE53355
26	577	Mannosyltransferase	Mannosyltransferase	79% (*A. teichomyceticus*)	CAE53356
27	347	C/O-Methyltransferase	None	51% (*S. ghanaensis*)	ZP_04690473
28	435	Phenylglycine Aminotransferase	HpgT	89% (*A. balhimycina*)	CAC48367
29	277	BHT Perhydrolase	BhtA	92% (*A. orientalis*)	CAA11784
30	581	BHT Peptide Synthetase	BhtB	92% (*A. orientalis*)	CAA11773
31	397	BHT Oxygenase	BhtC	89% (*A. orientalis*)	CAA11772
32	354	Hydroxymandelate Synthase	HmaS	82% (*A. orientalis*)	CAA11761
33	359	Hydroxymandelate Oxidase	Hmo	88% (*A. orientalis*)	CAA11762
34	456	Na+/H+ Antiporter	Putative Antiporter	85% (*A. balhimycina*)	CAC48373
35	477	NDP-Hexose-2,3-Dehydratase	EvaA	90% (*A. balhimycina*)	CAC48374
36	326	NDP-Hexose-4-Ketoreductase	EvaD	86% (*A. orientalis*)	CAA11764
37	370	NDP-Hexose-3-Aminotransferase	EvaB	91% (*A. orientalis*)	CAA11782
38	206	NDP-Hexose-3,5-Epimerase	EvaE	82% (*A. orientalis*)	CAA11781
39	608	Mannosyltransferase	None	64% (*S. avermitilis*)	NP_822291
40	373	Dihydroxyphenylacetic Acid Synthase	DpgA	94% (*A. orientalis*)	CAA11765
41	219	Enoyl-CoA Hydratase/Isomerase	DpgB	85% (*A. balhimycina*)	CAC48379
42	379	Dihydroxyphenylacetyl-CoA Oxygenase	DpgC	87% (*A. orientalis*)	CAA11787
43	268	Enoyl-CoA Hydratase/Isomerase	DpgD	89% (*A. balhimycina*)	CAC48381

Table 11: Predicted ORFs for AB915 Pathway

Protein length is given in amino acids (AA).

AB915 ORFs

ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA %ID (*Organism*)	Accession #
1	330	D-ala-D-ala Ligase	None	61% (T. fusca YX)	YP_289698
2	214	D-ala-D-ala Dipeptidase	VanX	81% (A. teichomyceticus)	CAE53345
3	347	D-ala-D-lac-Ligase	VanA	79% (A. teichomyceticus)	CAE53344
4	419	Betalactamase	None	51% (A. mirum)	ZP_03818051
5	277	Phytanoyl CoA Dioxygenase	None	46%(M. carbonacea)	ZP_04605278
6	448	UDP-N-acetylmuramoyltripeptide-D-ala-D-ala Ligase	MurF	75% (A. teichomyceticus)	CAE53342
7	333	D-lactate Dehydrogenase	VanH	70% (A. teichomyceticus)	CAG15002
8	399	Two-component Regulatory System	VanS	67% (Frankia sp. EAN1pec)	YP_001511361
9	230	Two-component Regulatory System	VanR	87% (S. toyocaensis)	AAM80541
10	206	D-ala-D-ala Carboxypeptidase	VanY	54% (S. nassauensis)	ZP_04486505
11	368	C/O-Methyltransferase	None	50% (S. ghanaensis)	ZP_04690473
12	1054	Nonribosomal Peptide Synthetase Module 1	NRPS1	75% (S. toyocaensis)	AAM80539
13	1106	Nonribosomal Peptide Synthetase Module 2	NRPS2	76% (A. teichomyceticus)	CAE53350
14	1060	Nonribosomal Peptide Synthetase Module 3	NRPS3	77% (A. teichomyceticus)	CAE53351
15	4068	Nonribosomal Peptide Synthetase Modules 4-6	NRPS4-6	78% (A. teichomyceticus)	CAE53352
16	1872	Nonribosomal Peptide Synthetase Module 7	NRPS7	83% (A. teichomyceticus)	CAE53353
17	501	NDP-Hexose-2,3-Dehydratase	EvaA	66% (S. cyanogenus)	AAD13549
18	357	Glucose Thymidylyltransferase	Glucose Thymidylyltransferase	70% (S. sviceus)	YP_002205669
19	375	Glycosyltransferase	GtfA (glycosyltransferase)	61% (A. orientalis)	AAB49292
20	408	Glycosyltransferase	GtfB (glycosyltransferase)	70% (Nonomuraea sp.)	CAD91204
22	407	Glycosyltransferase	GtfC (glycosyltransferase)	67% (A. orientalis)	AAB49294
21	274	Methyltransferase Type 11	None	57% (Frankia sp. EAN1pec)	YP_001508433
23	381	Glycosyltransferase	GtfA (glycosyltransferase)	63% (A. teichomyceticus)	AAB49294
24	374	NDP-Hexose-3-Aminotransferase	EvaB	80% (S. arenicola)	YP_001537186
25	201	NDP-Hexose-3,5-Epimerase	EvaD	68% (S. arenicola)	YP_001537187
26	334	NDP-Hexose-4-Ketoreductase	EvaE	69% (S. arenicola)	YP_001537188
27	238	Methyltransferase Type 12	None	74% (S. arenicola)	YP_001539299
28	579	Mannosyltransferase	Mannosyltransferase	71% (A. teichomyceticus)	CAE53356
29	659	ABC Transporter	ABC Transporter	81% (A. teichomyceticus)	CAE53357
30	231	Str Family Transcriptional Regulator	None	42% (S. griseus)	YP_001822146
31	392	P450 Monooxygenase, OxyA	OxyA	84% (A. teichomyceticus)	CAE53359
32	385	P450 Monooxygenase, OxyD	OxyD	80% (A. teichomyceticus)	CAE53360
33	399	P450 Monooxygenase, OxyB	OxyB	78% (A. teichomyceticus)	CAE53361
34	411	P450 Monooxygenase, OxyC	OxyC	86% (A. teichomyceticus)	CAE53363
35	412	Halogenase	Halogenase	66% (A. teichomyceticus)	CAE53362
36	531	Non-Heme Iron Dioxygenase	Non-Heme Iron Dioxygenase	84% (A. teichomyceticus)	CAE53366
37	352	DAHP Synthase	DAHP Synthase	87% (A. teichomcyeticus)	CAE53368
38	339	Str Family Transcriptional Regulator	Str Family Trans Reg	77% (A. teichomyceticus)	CAE53369
39	513	Non-Heme Iron Dioxygenase	iron dioxygenase	56% (S. toyocaensis)	AAM80528
40	389	Hydroxymandelate Oxygenase	Hmo	70% (S. toyocaensis)	AAM80552
41	806	Transcriptional Regulator	trans regulator	58% (A. teichomyceticus)	CAE53370
42	199	GTP Cylcohydrolase	GTP Cyclohydrolase	92% (A. teichomyceticus)	CAE53376
43	420	Hydroxyphenylglycine Transaminase	HpgT	88% (A. teichomyceticus)	CAG15035
44	122	Chorismate Mutase	None	68% (N. dassonvillei)	ZP_04335576
45	354	Hydroxymandelate Synthetase	HmaS	64% (Frankia sp. CcI3)	YP_481550
46	365	Prephenate Dehydrogenase	Prephenate Dehydrogenase	73% (A. teichomyceticus)	CAG15036
47	369	Dihydroxyphenylacetic Acid Synthase	DpgA	84% (A. teichomyceticus)	CAE53371
48	221	Enoyl-CoA Hydratase/Isomerase	DpgB	67% (A. teichomyceticus)	CAE53372
49	434	Dihydroxyphenylacetyl-CoA Oxygenase	DpgC	78% (A. teichomyceticus)	CAE53373

50	239	Enoyl-CoA Hydratase/Isomerase	DpgD	83% (A. teichomyceticus)	CAE53374
51	395	UDP muramoyl pentapeptide transferase	MurG	60% (S. arenicola)	YP_001535200

Table 12: Predicted ORFs for the AZ205 Pathway

Protein length is given in amino acids (AA).

AZ205 ORFs

ORF	Length (AA)	Predicted Gene Product	Glycopeptide Homolog	AA% ID (*Organism*)	Accession #
1	184	VanY-Carboxypeptidase	VanY	78%(*Nonomuraea sp.*)	CAD91202.1
2	232	Two-Component Regulatory System	VanR	91%(*Nocardia farcinica*)	BAD55538.1
3	374	Two-Component Regulatory System	VanS	77%(*Nocardia farcinica*)	BAD55539.1
4	446	UDP-N-acetylmuramoyltripeptide-D-ala-D-ala Ligase	MurF	80% (*A. teichomyceticus*)	CAE53342.1
5	169	Transposase	None	57% (*S. erythraea*)	CAM04227.1
6	344	D-lactate Dehydrogenase	VanH	76% (*A. orientalis*)	AEI58859.1
7	347	D-ala-D-lac-Ligase	VanA	76% (*A. teichomyceticus*)	CAE53344.1
8	203	D-ala-D-ala Dipeptidase	VanX	79% (*A. teichomyceticus*)	CAE53345.1
9	226	Hypothetical Protein	None	43% (*C. michiganensis*)	CAN02774.1
10	222	Transposase	None	54% (*Frankia alni*)	CAJ61201.1
11	384	Methyltransferase	None	43% (*Nodularia spumigena*)	EAW44464.1
12	653	ABC Transporter	ABC Transporter	94% (*A. orientalis*)	AEI58864.1
13	3151	Nonribosomal Peptide Synthetase Module 1-3	NRPS1-3	71% (*A. orientalis*)	CAA11794.1
14	4112	Nonribosomal Peptide Synthetase Modules 4-6	NRPS4-6	81% (*A. orientalis*)	CAA11795.1
15	1864	Nonribosomal Peptide Synthetase Module 7	NRPS7	84% (*A. orientalis*)	CAA11796.1
16	494	P450 Monooxygenase, OxyA	OxyA	74% (*A. orientalis*)	AEI58868.1
17	356	P450 Monooxygenase, OxyB	OxyB	89% (*A. orientalis*)	CAA11798.1
18	278	Sulfotransferase	StaL	51% (*S. toyocaensis*)	AAM80529.1
19	442	P450 Monooxygenase, OxyC	OxyC	80% (*A. balhimycina*)	CAA76549.1
20	493	Halogenase	Halogenase	89% (*A. balhimycina*)	CAA76550.1
21	409	Glycosyltransferase	Glycosyltransferase	81% (*A. balhimycina*)	CAA76552.1
22	275	Deacetylase	Deacetylase	63% (*A. teichomyceticus*)	CAE53355.1
23	274	N-methyltransferase	N-methyltransferase	70% (*A. orientalis*)	CAA11779.1
24	429	Hydroxyphenylglycine Transaminase	HpgT	89% (*A. orientalis*)	CAA11790.1
25	277	BHT Perhydrolase	BhtA	87% (*A. orientalis*)	CAA11784.1
26	579	BHT Peptide Synthetase	BhtB	84% (*A. orientalis*)	CAA11773.1
27	400	BHT Oxygenase	BhtC	88% (*A. orientalis*)	CAA11772.1
28	361	Hydroxymandelate Synthase	HmaS	80% (*A. balhimycina*)	CAC48371.1
29	359	Hydroxymandelate Oxidase	Hmo	81% (*A. orientalis*)	CAA11762.1
30	373	Dihydroxyphenylacetic Acid Synthase	DpgA	92% (*A. orientalis*)	CAA11765.1
31	225	Enoyl-CoA Hydratase/Isomerase	DpgB	84% (*A. balhimycina*)	CAC48379.1
32	433	Dihydroxyphenylacetyl-CoA Oxygenase	DpgC	70% (*A. balhimycina*)	CAC48380.1
33	447	Antiporter	None	43% (*B. thuringiensis*)	ADY24350.1
34	389	femAB protein	StaO	88% (*S. toyocaensis*)	AAM80555.1
35	469	Prephenate Dehydrogenase	Prephenate Dehydrogenase	53% (*A. orientalis*)	AEI58863.1
36	259	Thioesterase	None	66% (*Streptomyces sp. CS*)	ADM46371.1
37	239	Methyltransferase	None	66% (*S. flavogriseus*)	ADW05446.1

Figure 20: ESI-MS/MS fragmentation data for sulfo-teicoplanin aglycone A (20).

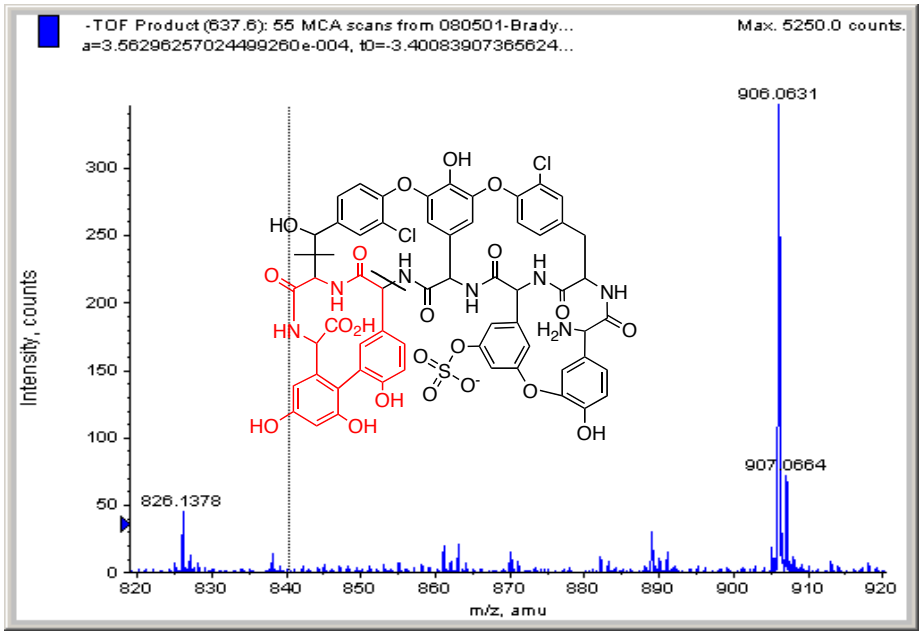


Figure 21: ^{1}H NMR of sulfo-teicoplanin aglycone A (20) in d_6-DMF at 323 K

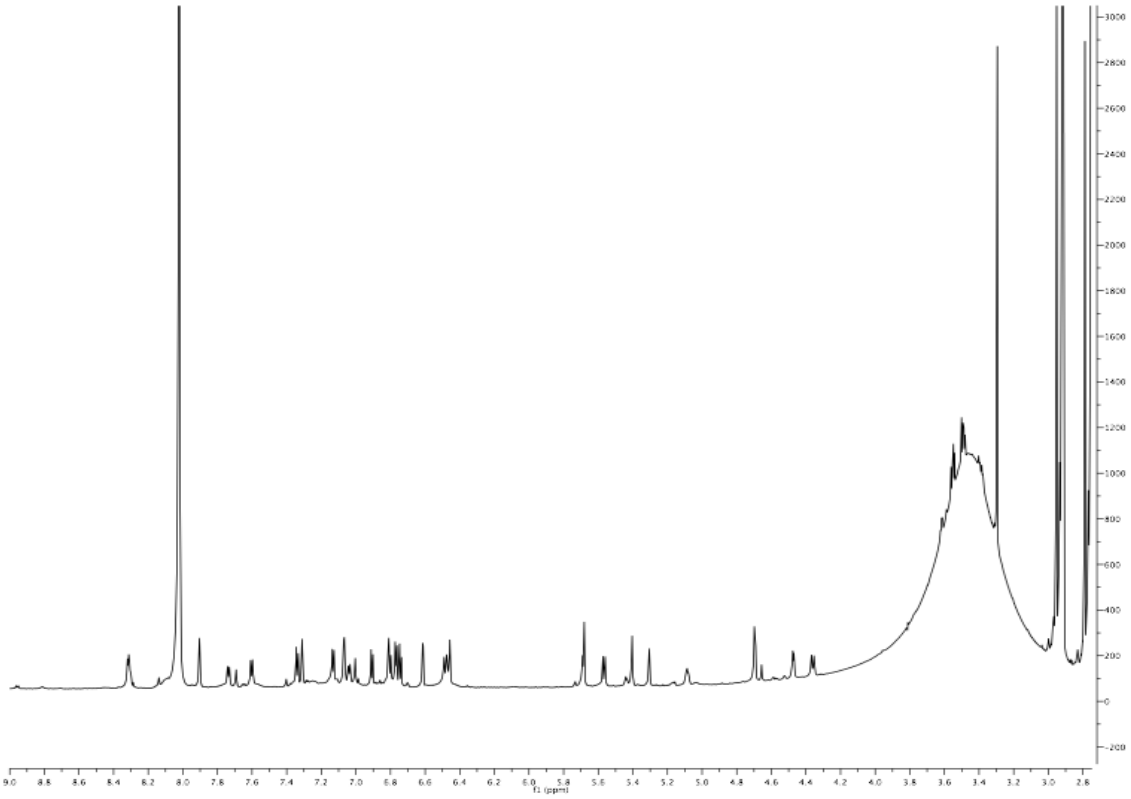


Figure 22: ^{1}H - ^{1}H Correlation Spectroscopy (COSY) NMR of sulfo-teicoplanin aglycone A (20) in d_6-DMF at 323 K

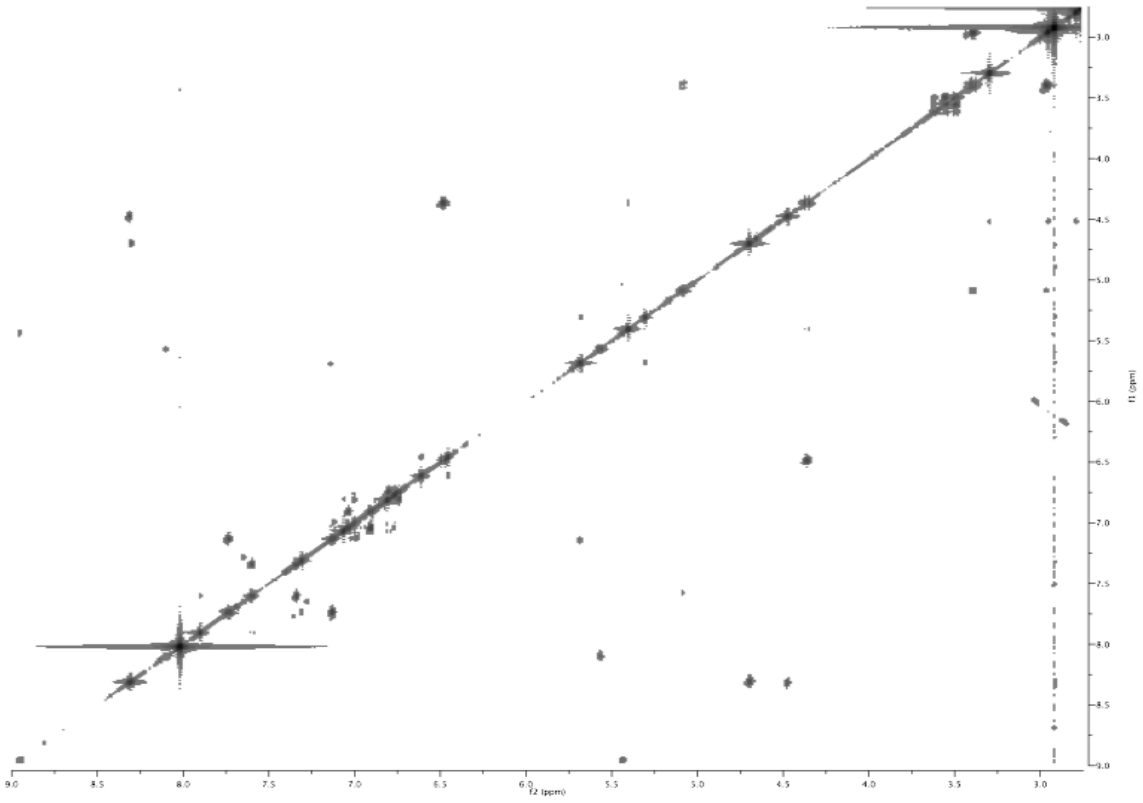


Figure 23: ESI-MS/MS fragmentation data for sulfo-teicoplanin aglycone B (21)

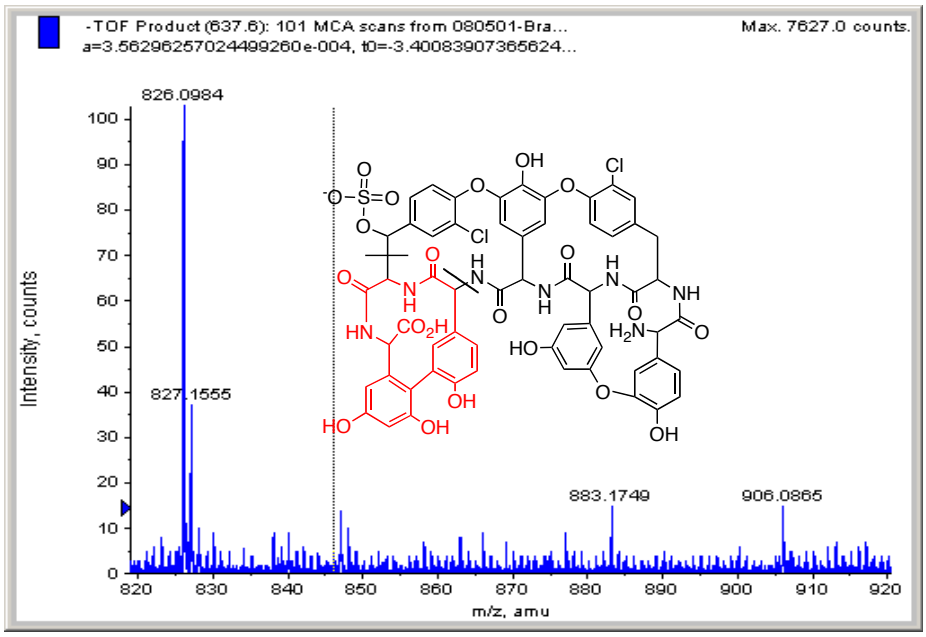


Figure 24: ^{1}H NMR of sulfo-teicoplanin aglycone B (21) in d_6-DMF at 323 K

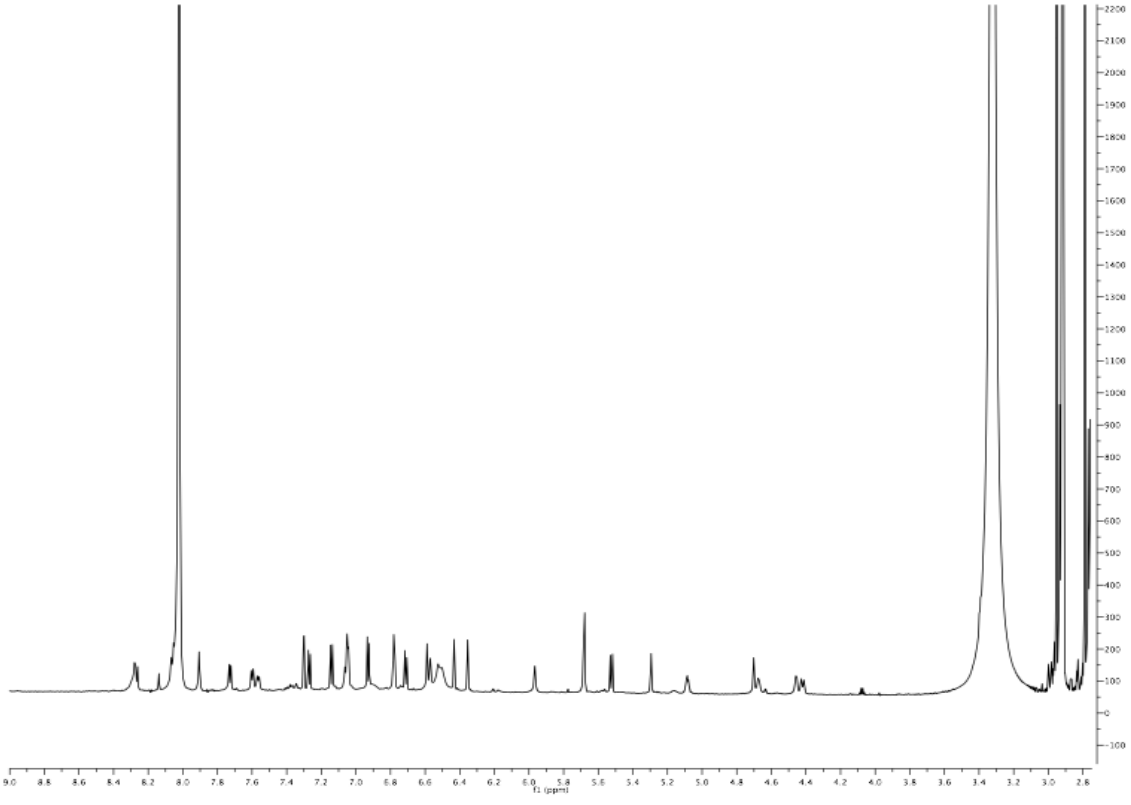


Figure 25: ^{1}H - ^{1}H Correlation Spectroscopy (COSY) NMR of sulfo-teicoplanin aglycone B (21) in d_6-DMF at 323 K

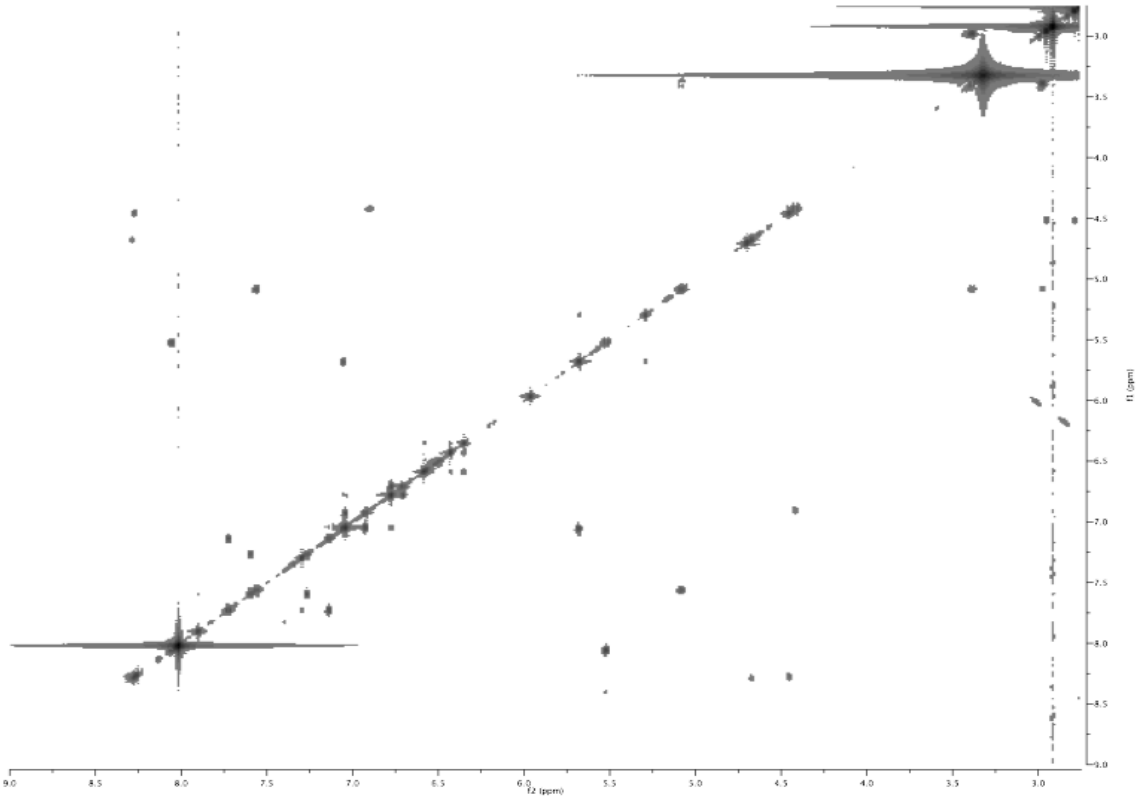


Figure 26: ESI-MS/MS fragmentation data for sulfo-teicoplanin aglycone C (22)

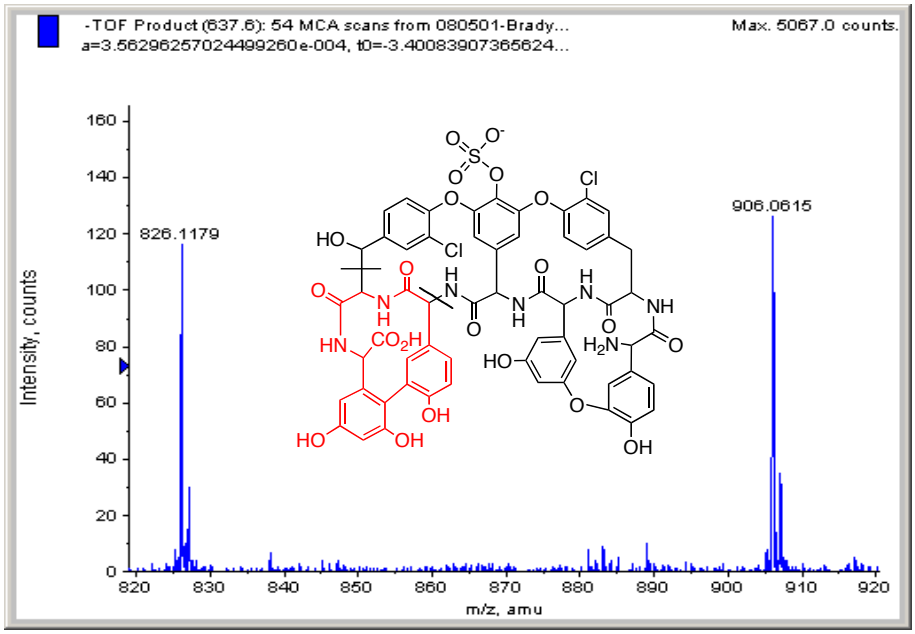


Figure 27: ^{1}H NMR of sulfo-teicoplanin aglycone C (22) in d_6-DMF at 323 K

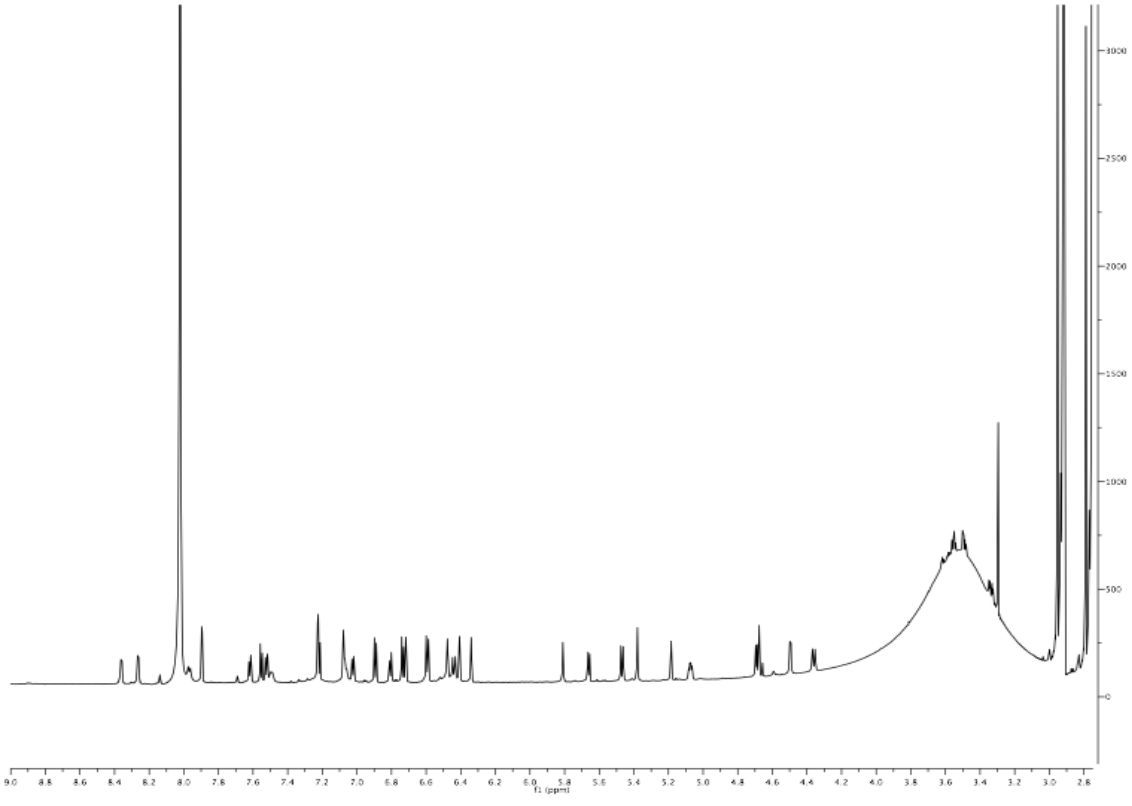


Figure 28: ^{1}H - ^{1}H Correlation Spectroscopy (COSY) NMR of sulfo-teicoplanin aglycone C (22) in d_6-DMF at 323 K

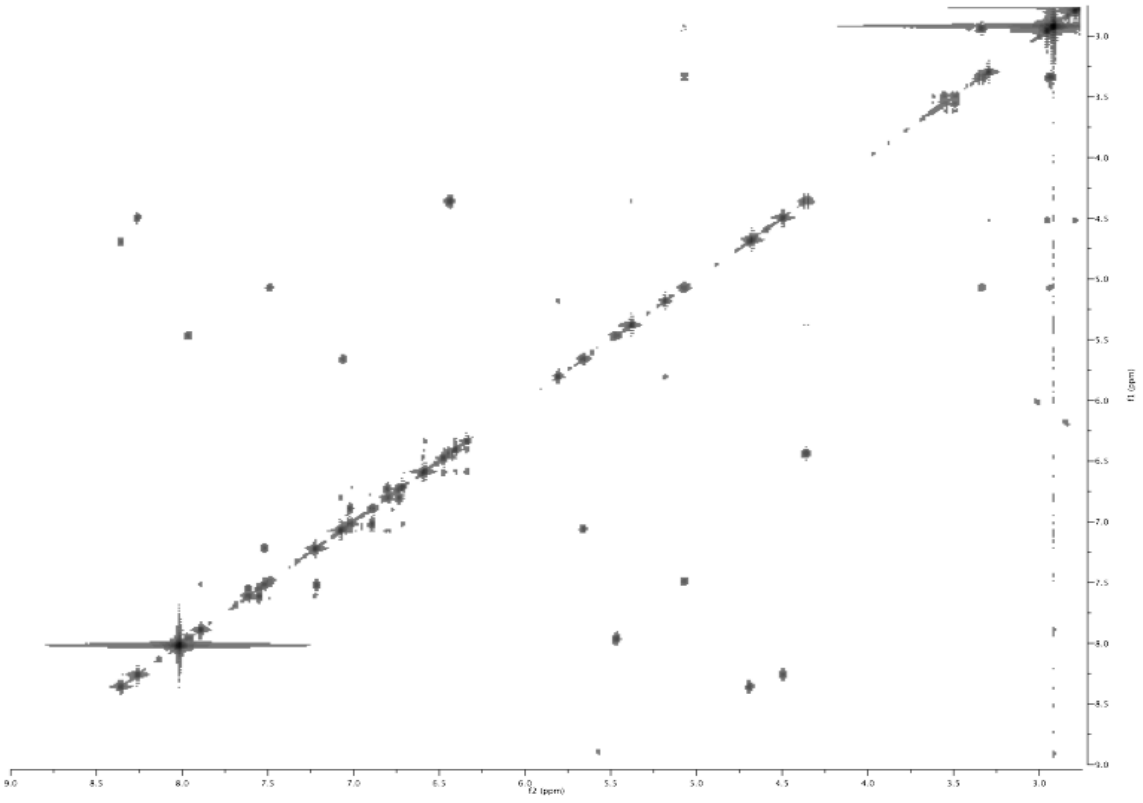


Figure 29: ESI-MS/MS fragmentation data for teicoplanin aglycone

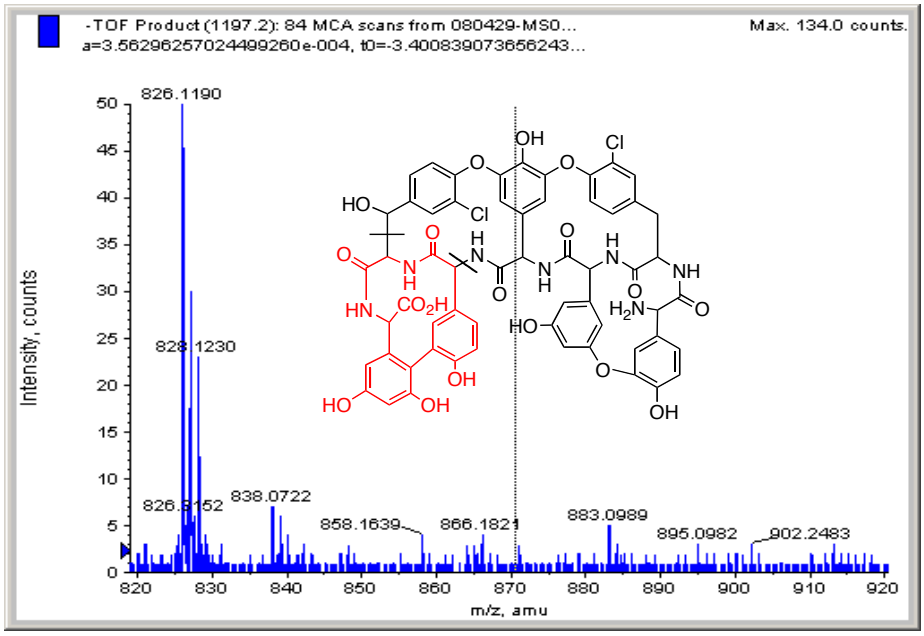


Figure 30: ^{1}H NMR of teicoplanin aglycone in d_6-DMF at 323 K

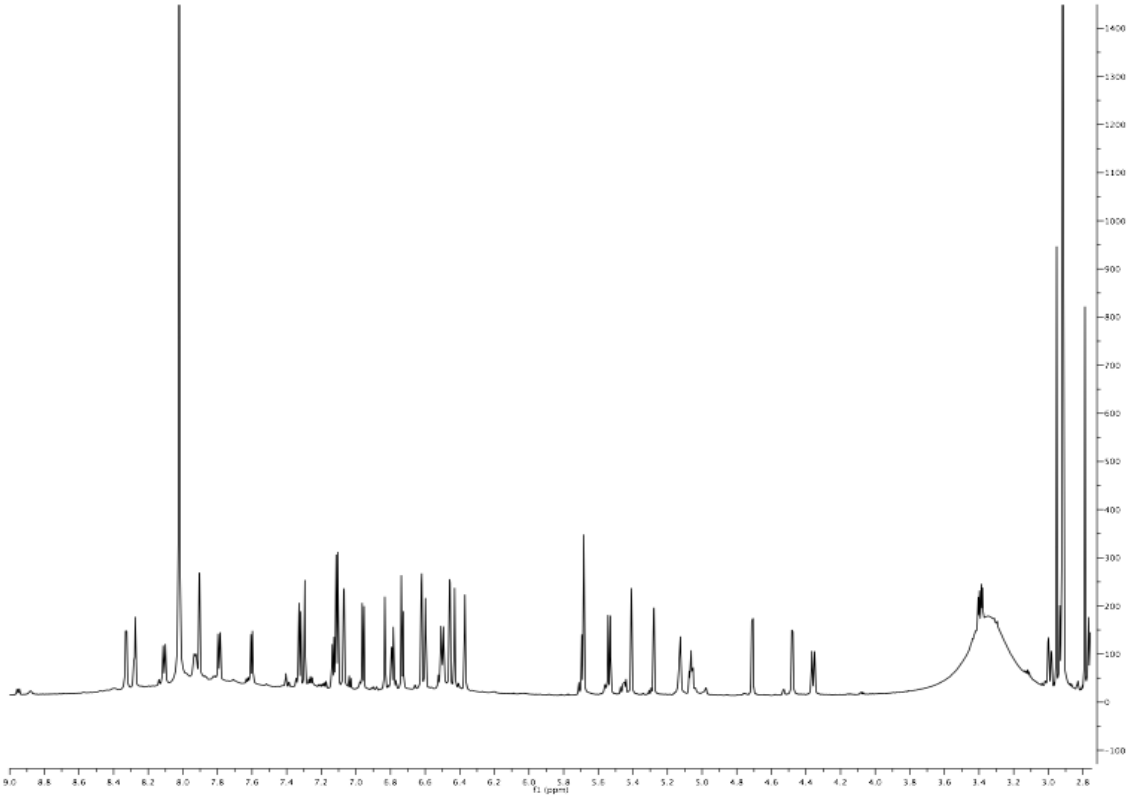


Figure 31: ^{1}H - ^{1}H Correlation Spectroscopy (COSY) NMR of teicoplanin aglycone in d_6-DMF at 323 K

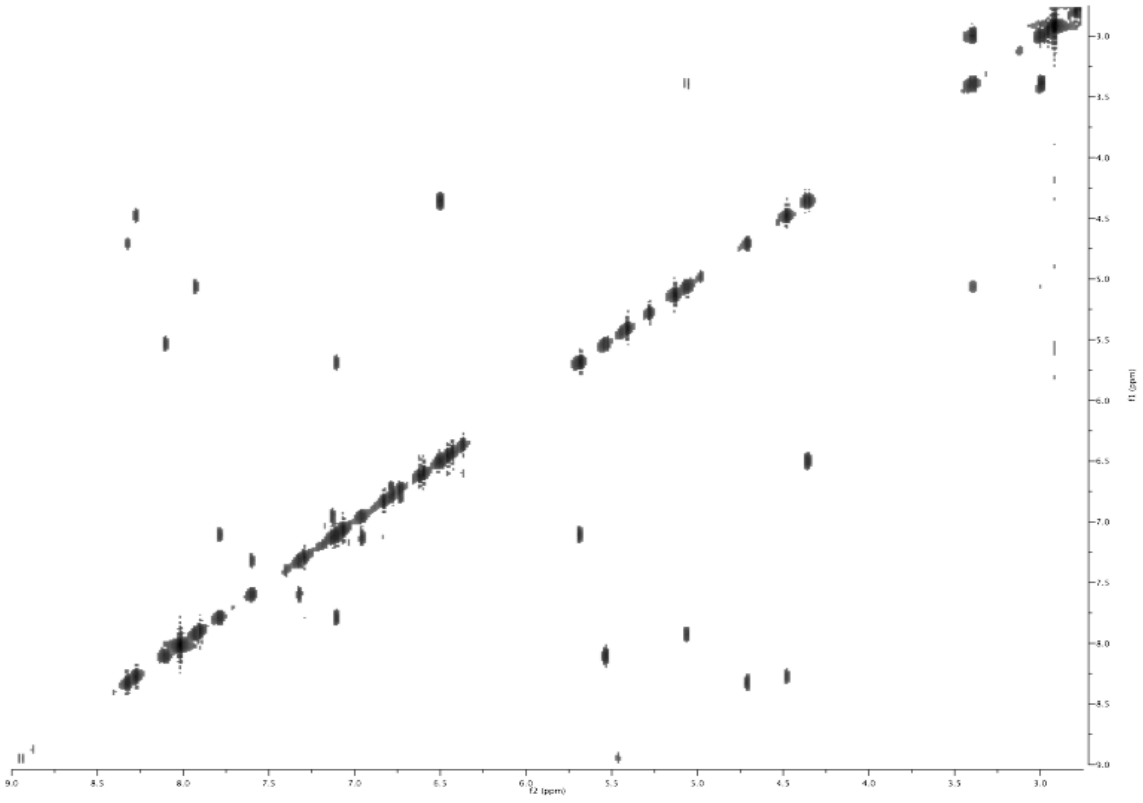
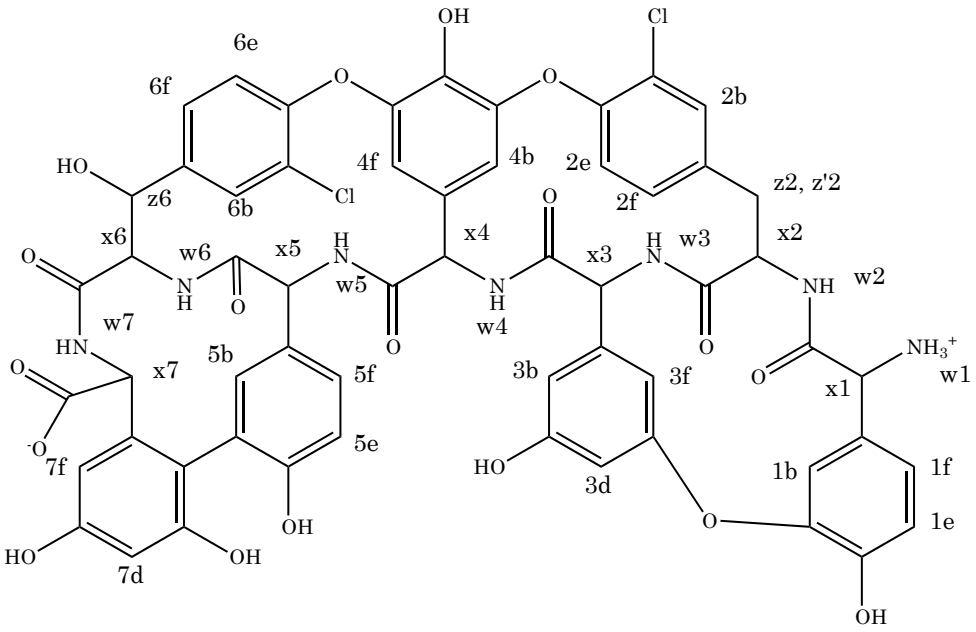


Figure 32: Numbering scheme for teicoplanin aglycone and derivatives



Teicoplanin Aglycone

Table 13: ^{1}H NMR chemical shift data for sulfo-teicoplanin aglycone A (20)

	Sulfo-teicoplanin aglycone A 323K in d6-DMF			
	δ (ppm)	Multiplicity, J Coupling	COSY (strong)	COSY (weak)
1b	6.8	d, J = 2.7 Hz		
1e	6.91	d, J = 9.1 Hz	7.04	
1f	7.04	m	6.91	
2b	7.31	d, J = 1.7 Hz		7.78
2e	7.13	d, J = 8.2 Hz	7.78	
2f	7.78	dd, J = 1.7, 8.3 Hz	7.13	7.31
3b	7	t, J = 2.1 Hz		6.77, 6.81
3d	6.81	bs		6.81
3f	6.77	m		7
4b	5.68	m		5.3
4f	5.3	s		5.68
5b	7.07	m		6.74
5e	6.74	d, J = 8.5 Hz		
5f	6.77	bs		
6b	7.9	d, J = 1.9 Hz		7.6
6e	7.34	d, J = 8.3 Hz	7.6	
6f	7.6	m	7.34	
7d	6.61	d, J = 2.5 Hz		6.47
7f	6.47	m		6.61
w1	ND			
w2	7.6	m	5.09	
w3	8.1	m	5.56	
w4	7.13	d, J = 8.2 Hz	5.68	
w5	8.31	m	4.47	
w6	6.47	m	4.36	
w7	8.31	m	4.69	
x1	4.69	m		
x2	5.09	m	7.6	2.95, 3.39
x3	5.56	d, J = 10.5 Hz	8.1	
x4	5.68	m	7.13	
x5	4.47	d, J = 5.1 Hz	8.31	
x6	4.36	d, J = 12.1 Hz	6.47	
x7	4.69	m	8.31	
z2	2.95	Under Solvent	3.39	5.09
z2'	3.39	dd, J = 5.4, 13.8 Hz	2.95	5.09
z6	5.4	s	4.36	

Table 14: ^{1}H NMR chemical shift data for sulfo-teicoplanin aglycone B (21)

	Sulfo-teicoplanin aglycone B 323K in d6-DMF			
	δ(ppm)	Multiplicity, J Coupling	COSY (strong)	COSY (weak)
1b	6.77	m		
1e	6.92	d, J = 8.3 Hz	7.05	
1f	7.05	m	6.92	
2b	7.3	d, J = 1.7 Hz		
2e	7.14	d, J = 8.3 Hz	7.73	
2f	7.73	dd, J = 1.8, 8.4 Hz	7.14	
3b	6.52	m		
3d	6.35	t, J = 2.1 Hz		6.58
3f	6.58	bs		6.35
4b	5.68	m		5.29
4f	5.29	bs		5.68
5b	7.05	m		
5e	6.71	d, J = 8.5 Hz	6.77	
5f	6.77	m	6.71	
6b	7.9	s		
6e	7.26	d, J = 8.5 Hz	7.6	
6f	7.6	dd, J = 1.5, 8.5 Hz	7.26	
7d	6.57	bs		6.43
7f	6.43	s		6.57
w1	ND			
w2	7.56	d, J = 8.9 Hz	5.08	
w3	8.07	d, J = 10.4 Hz	5.52	
w4	7.05	m	5.68	
w5	8.28	m	4.47	
w6	6.52	m	4.41	
w7	8.28	m	4.68	
x1	4.7	s		
x2	5.08	m	7.56	2.98, 3.39
x3	5.52	d, J = 10.5 Hz	8.07	
x4	5.68	m	7.05	
x5	4.47	d, J = 4.7 Hz	8.28	
x6	4.41	d, J = 12.1 Hz	6.92	
x7	4.68	bs	8.28	
z2	2.98	Under Solvent	3.39	5.08
z2'	3.39	dd, J = 5.5, 13.9 Hz	2.98	5.08
z6	5.97	s		

Table 15: ^{1}H NMR chemical shift data for sulfo-teicoplanin aglycone C (22)

	Sulfo-teicoplanin aglycone C 323K in d6-DMF			
	δ(ppm)	Multiplicity, J Coupling	COSY (strong)	COSY (weak)
1b	6.72	d, J = 1.8 Hz		7.02
1e	6.89	d, J = 8.4 Hz	7.02	
1f	7.02	dd, J = 1.8, 8.4 Hz	6.89	7.02
2b	7.22	m		7.62
2e	7.55	d, J = 8.6 Hz	7.62	
2f	7.62	dd, J = 1.8, 8.6 Hz	7.55	7.22
3b	6.41	bs		6.34
3d	6.34	t, J = 2.2 Hz		6.41, 6.60
3f	6.6	d, J = 2.2 Hz		6.34
4b	5.81	d, J = 1.4 Hz		
4f	5.18	bs		
5b	7.07	m		6.81
5e	6.74	d, J = 8.4 Hz	6.81	
5f	6.81	dd, J = 2.2, 8.4 Hz	6.74	7.07
6b	7.89	d, J = 2.0 Hz		7.52
6e	7.22	m		
6f	7.52	dd, J = 1.5, 8.6 Hz	7.89	
7d	6.59	bs		6.48
7f	6.48	d, J = 2.1 Hz		6.59
w1	ND			
w2	7.49	d, J = 8.9 Hz	5.07	
w3	7.97	d, J = 11.1 Hz	5.46	
w4	7.07	m	5.66	
w5	8.26	d, J = 5.0 Hz	4.5	
w6	6.44	d, J = 12.2 Hz	4.36	
w7	8.36	d, J = 6.0 Hz	4.68	
x1	4.68	m		
x2	5.07	m	7.49	2.98, 3.38
x3	5.46	d, J = 10.6 Hz	7.97	
x4	5.66	d, J = 8.1 Hz	7.07	
x5	4.5	d, J = 5.2 Hz	8.26	
x6	4.36	d, J = 12.2 Hz	6.44	
x7	4.68	m	8.36	
z2	2.98	Under Solvent	3.38	5.07
z2'	3.38	dd, J = 5.5, 13.9 Hz	2.98	5.07
z6	5.38	bs		

Table 16: ^{1}H NMR chemical shift data for teicoplanin aglycone

	Teicoplanin aglycone 323K in d6-DMF			
	δ (ppm)	Multiplicity, J Coupling	COSY (strong)	COSY (weak)
1b	6.83	d, J = 1.7 Hz		
1e	6.96	d, J = 8.3 Hz	7.12	
1f	7.12	m	6.96	
2b	7.29	d, J = 1.9 Hz		
2e	7.12	m	7.79	
2f	7.79	dd, J = 1.6, 8.4 Hz	7.12	
3b	6.43	bs		
3d	6.3	t, J = 2.1 Hz		
3f	6.62	d, J = 2.4 Hz		
4b	5.68	m		5.28
4f	5.28	d, J = 1.4 Hz		5.68
5b	7.07	d, J = 2.3 Hz		
5e	6.73	d, J = 8.5 Hz		
5f	6.79	dd, J 2.3, 8.3 Hz		
6b	7.9	d, J = 2.0 Hz		7.6
6e	7.32	d, J = 8.5 Hz	7.6	
6f	7.6	dd, J = 1.9, 8.3 Hz	7.32	7.9
7d	6.6	s		
7f	6.46	d, J = 2.4 Hz		
w1	ND			
w2	7.92	d, J = 9.6 Hz	5.06	
w3	8.11	d, J = 10.9 Hz	5.53	
w4	7.11	m	5.68	
w5	8.27	d, J = 5.2 Hz	4.48	
w6	6.5	m	4.36	
w7	8.32	d, J = 6.1 Hz	4.71	
x1	5.13	d, J = 4.5 Hz		
x2	5.06	m	7.92	2.98, 3.40
x3	5.53	d, J = 10.3 Hz	8.11	
x4	5.68	m	7.11	
x5	4.48	d, J = 5.3 Hz	8.27	
x6	4.36	d, J = 12.1 Hz	6.5	
x7	4.71	d, J = 6.0 Hz	8.32	
z2	2.98	dd, J = 3.2, 14.2 Hz	3.4	5.06
z2'	3.4	dd, J = 5.4, 14.2 Hz	2.98	5.06
z6	5.41	bs		

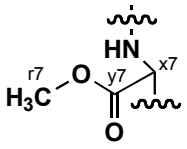
Table 17: ^{1}H NMR data for A47934 and derivatives produced *in vivo*

	δ (ppm) Compound 28	**δ (ppm) Compound 29**	**δ (ppm) Compound 30**	**δ (ppm) Compound 31**	**δ (ppm) Compound 32**
1b	7.01	7.02	7.00	7.02	6.97
1e	7.61	7.60	7.65	7.61	7.61
1f	7.38	7.35	7.38	7.38	7.33
2b	7.32	7.31	7.32	7.33	7.31
2e	7.21	7.19	7.35	7.22	7.20
2f	7.58	7.57	7.58	7.63	7.57
3b	6.49	6.49	6.49	6.50	6.46
3d	6.62	6.61	6.63	6.62	6.61
3f	6.24	6.27	6.22	6.26	6.21
4b	5.27	5.26	5.27	5.30	5.27
4f	5.68	5.68	5.82	5.70	5.68
5b	6.96	6.97	6.98	6.97	6.95
5f	7.01	6.98	7.02	7.01	7.01
6b	7.60	7.59	7.61	7.60	7.59
6e	7.21	7.18	7.21	7.21	7.20
6f	7.47	7.47	7.51	7.47	7.49
7d	6.52	6.58	6.53	6.52	6.51
7f	6.44	6.23	6.45	6.44	6.43
x1	5.40	5.40	5.40	5.40	4.94
x2	4.94	4.98	4.94	4.94	4.87
x3	5.49	5.49	5.49	5.49	5.43
x4	5.49	5.52	5.49	5.49	5.49
x5	4.35	4.35	4.36	4.36	4.35
x6	4.00	4.03	4.02	4.02	4.00
x7	4.53	4.72	4.53	4.53	4.53
z2	3.35	3.35	3.36	3.34	3.35
z2'	3.10	3.09	3.11	3.11	3.10
z6	5.47	5.43	5.48	5.47	5.47

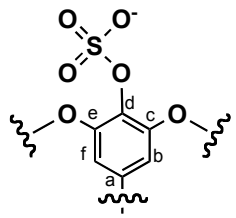
Table 18: Compound specific ^{1}H and ^{13}C assignments for compounds 29-32

The ^{1}H and ^{13}C chemical shifts reported in this table detail the specific chemical shifts used to make the structural assignments of compounds 29-32. For compound 29, these are the chemical shifts for both the parent compound (Compound 28) and the new compound at the C-terminus of the glycopeptide. For compounds 30 and 31, the chemical shifts reported are the key shifts vital to the assigning the position of the compound specific modification (sulfonation and glycosylation, respectively) of the hydroxyl at the top of ring 4. For compound 32, the shifts reported are those new signals and the observed changes in the ^{13}C at the C-α, x1, at the N-terminus.

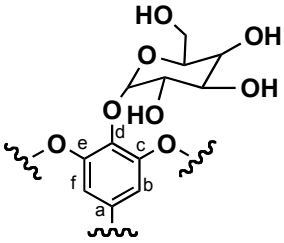
	Compound 28	Compound 29
r7 (1H)	-	3.85
r7 (13C)	-	53.21
y7 (13C)	176.79	172.32
x7 (13C)	59.09	57.11



	Compound 28	Compound 30
4a (13C)	135.27	138.22
4b (13C)	107.5	108.93
4c (13C)	146.53	151.65
4d (13C)	132.98	128.41
4e (13C)	147.08	151.84
4f (13C)	104.16	104.22



	Compound 28	Compound 31
4a (13C)	135.27	134.29
4b (13C)	107.5	107.56
4c (13C)	146.53	151.38
4d (13C)	132.98	130.95
4e (13C)	147.08	151.78
4f (13C)	104.16	104.11



	Compound 28	Compound 32
4a (13C)	135.27	134.29
4b (13C)	107.5	107.56
4c (13C)	146.53	151.38
4d (13C)	132.98	130.95
4e (13C)	147.08	151.78
4f (13C)	104.16	104.11

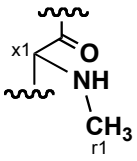
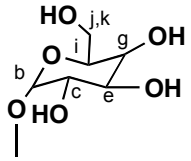


Table 19: Numbering scheme and ^{1}H and ^{13}C assignments for compound 31 - glucose



	1H (ppm)	13C (ppm)	Observed 3J (Guzman-Martinez, Lamer et al.)
b	5.33	104.49	7.8
c	3.75	73.2	8.6
e	3.68	75.28	9.5
g	3.63	68.69	9.5
i	3.56	76.06	-
j	3.88	59.98	5.18
k	3.8	59.98	-

Figure 33: ^{1}H of Compound 28, 298K in 3:1 D_2O:CD_3CN

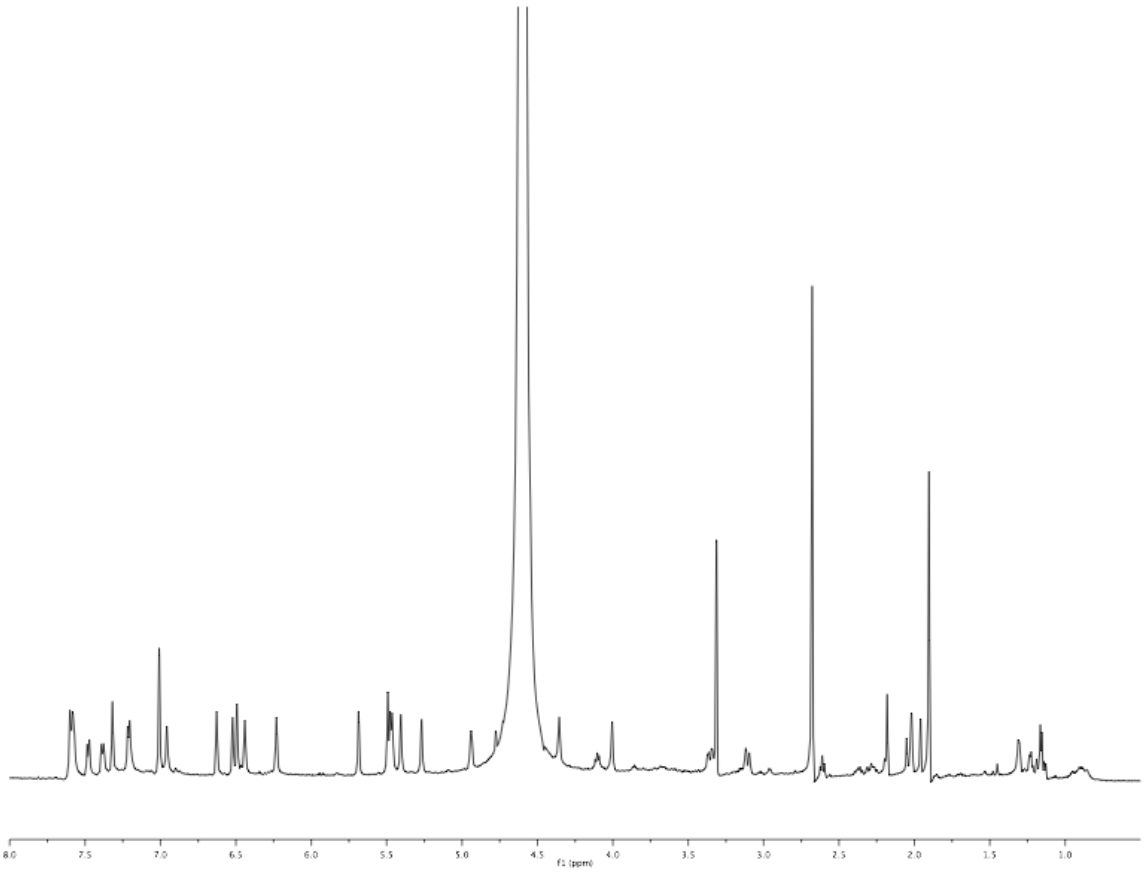


Figure 34: ^{1}H-^{13}C HMQC of Compound 28, 298K in 3:1 D_2O:CD_3CN

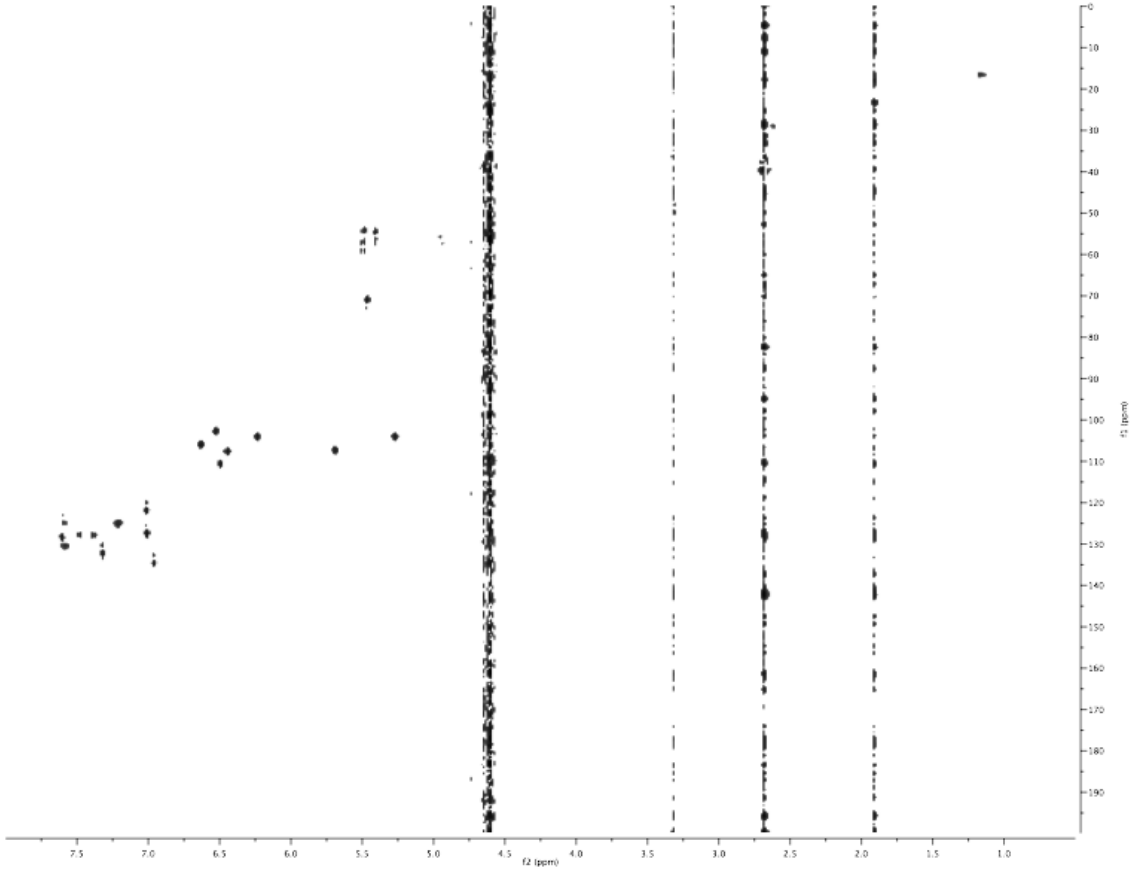


Figure 35: ^{1}H-^{13}C HMBC of Compound 28, 298K in 3:1 D_2O:CD_3CN

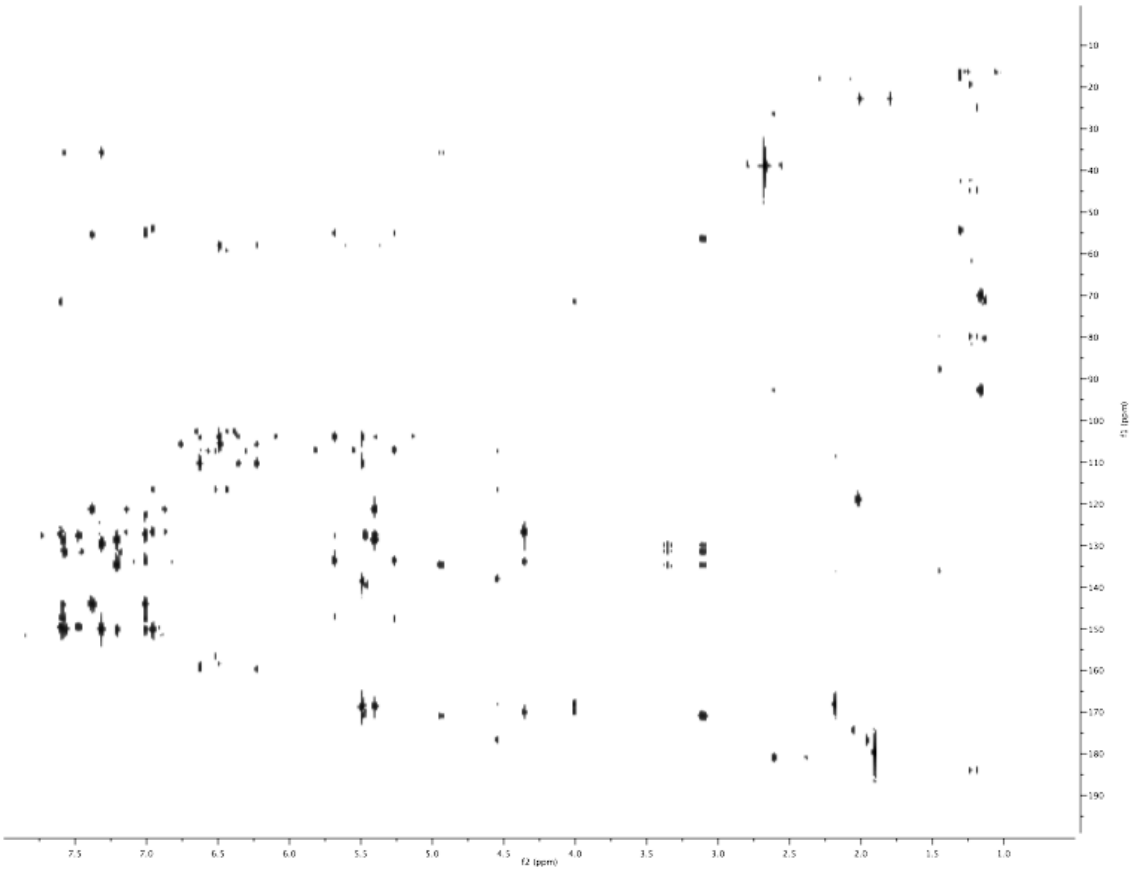


Figure 36: ^{1}H of Compound 29, 298K in 3:1 D_2O:CD_3CN

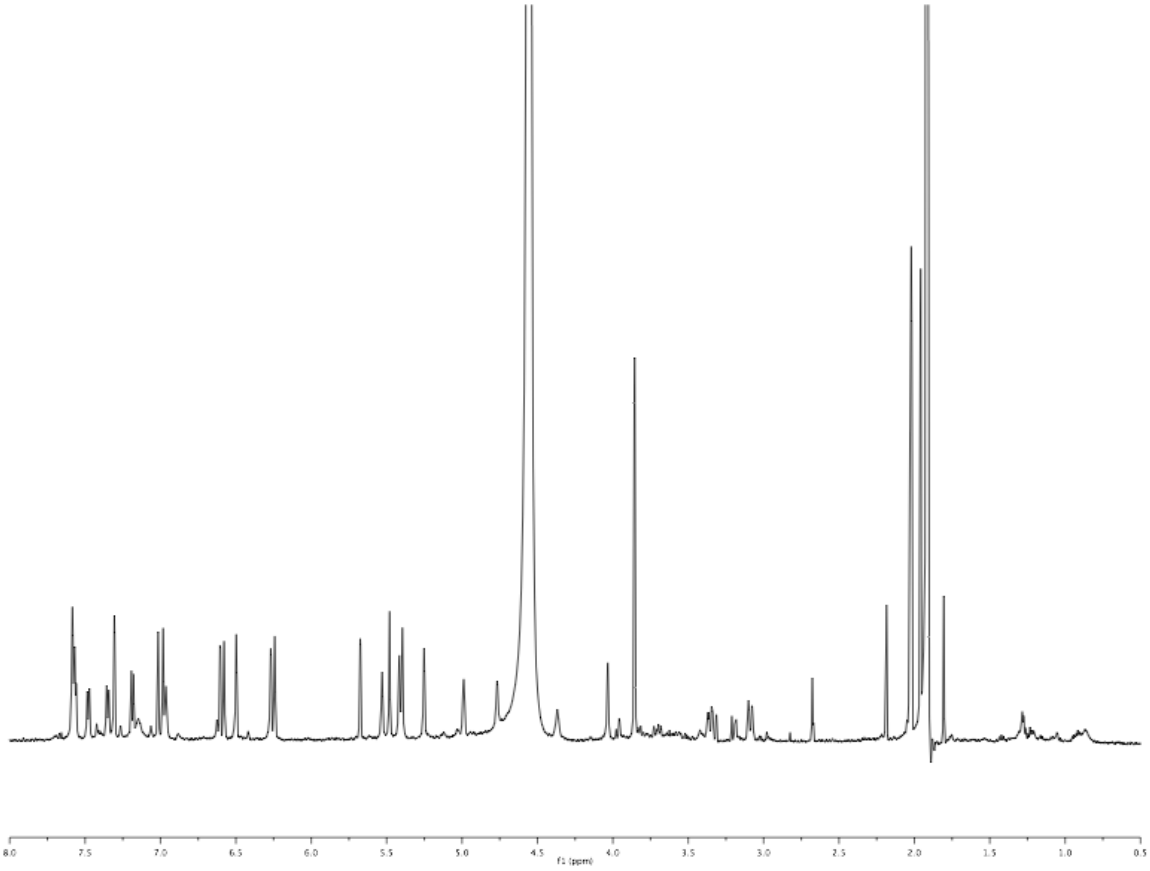


Figure 37: ^{1}H-^{13}C HMQC of Compound 29, 298K in 3:1 D_2O:CD_3CN

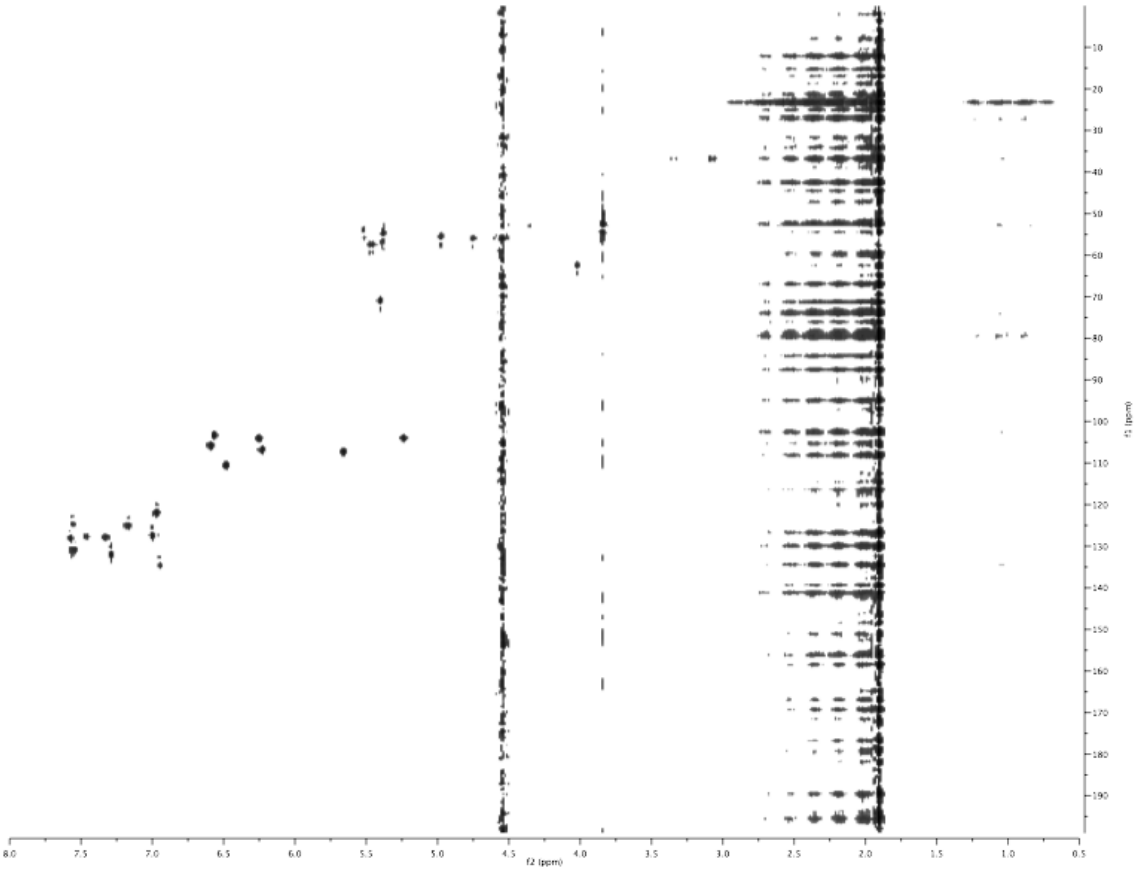


Figure 38: ^{1}H-^{13}C HMBC of Compound 29, 298K in 3:1 D_2O:CD_3CN

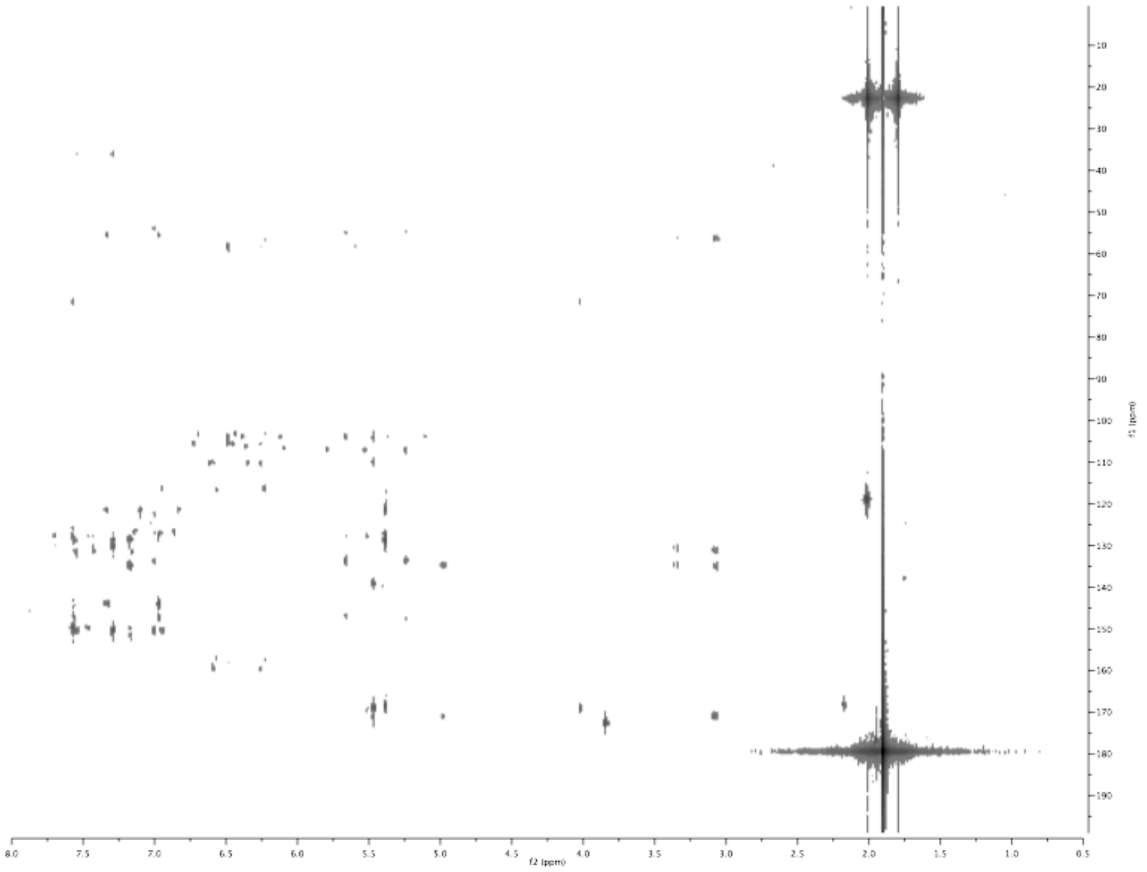


Figure 39: ^{1}H of Compound 30, 298K in 3:1 D_2O:CD_3CN

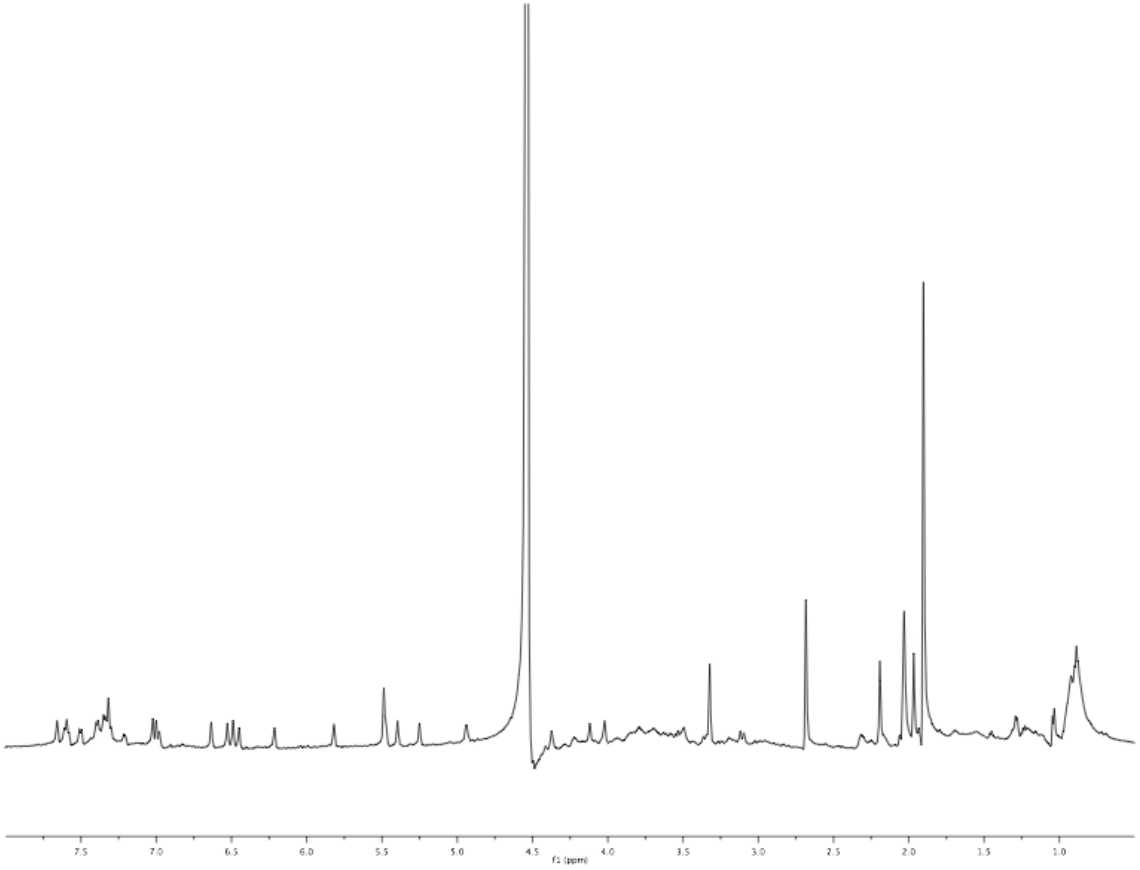


Figure 40: ^{1}H-^{13}C HMQC of Compound 30, 298K in 3:1 D_2O:CD_3CN

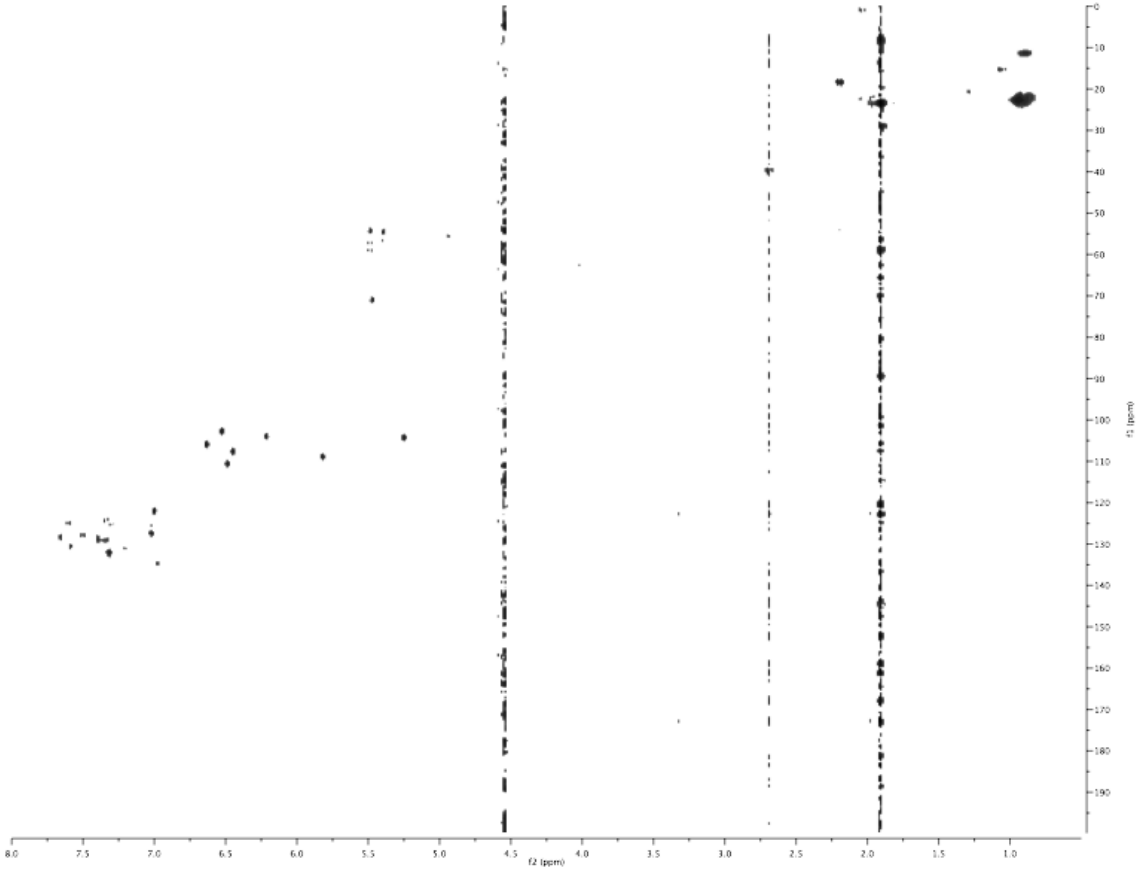


Figure 41: ^{1}H-^{13}C HMBC of Compound 30, 298K in 3:1 D_2O:CD_3CN

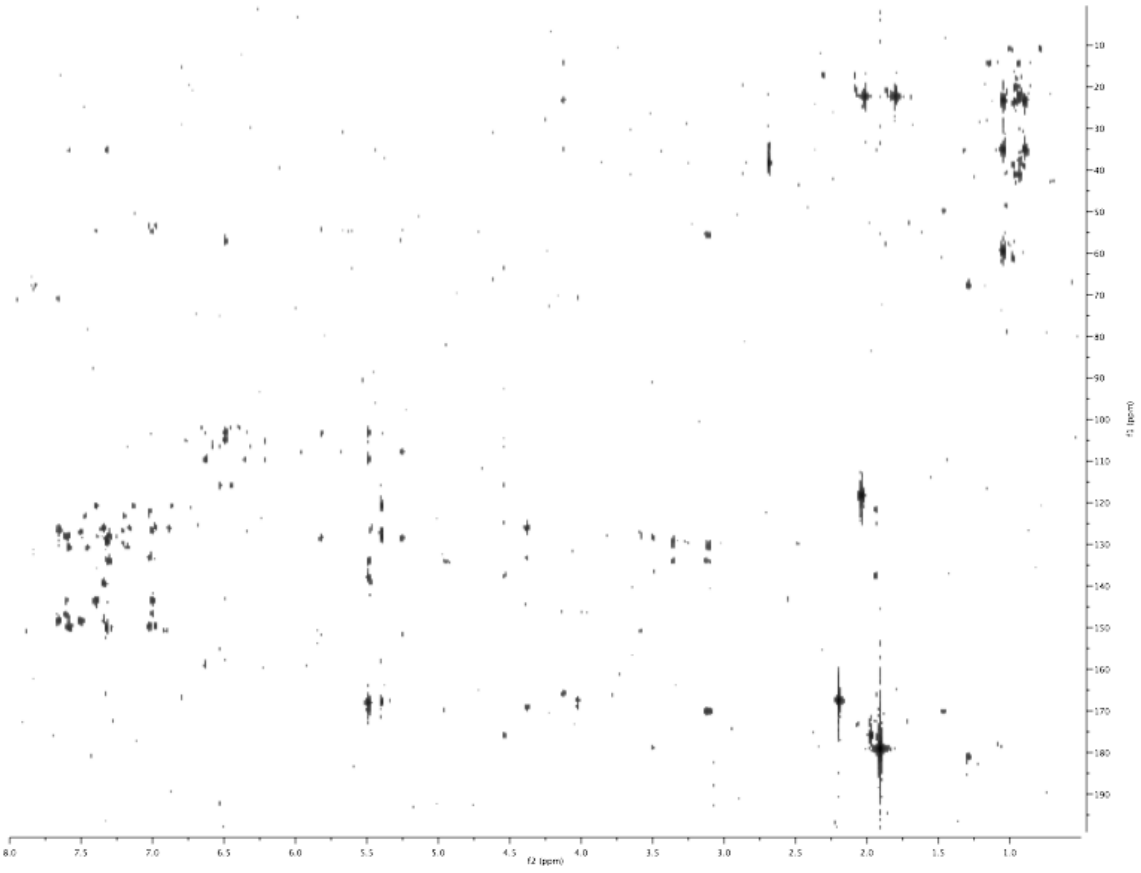


Figure 42: ^{1}H of Compound 31, 298K in 3:1 D_2O:CD_3CN

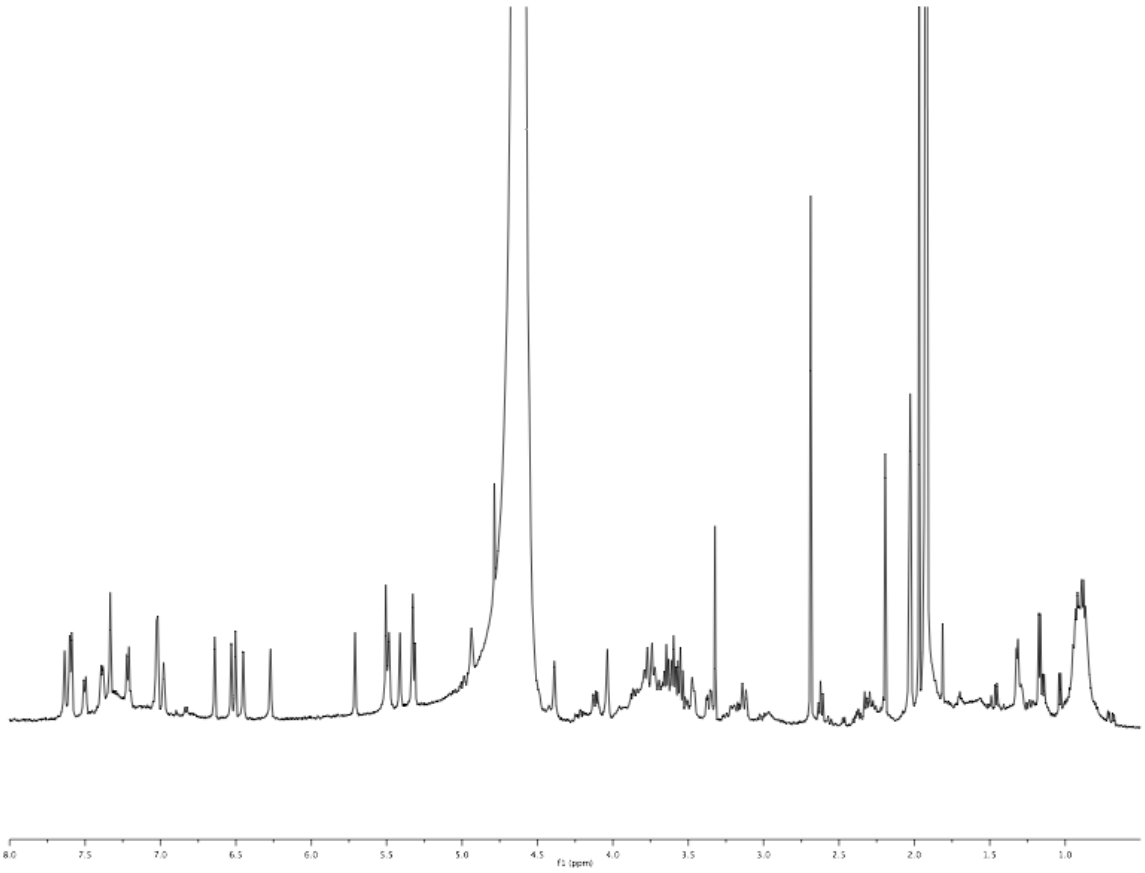


Figure 43: ^{1}H-^{13}C HMQC of Compound 31, 298K in 3:1 D_2O:CD_3CN

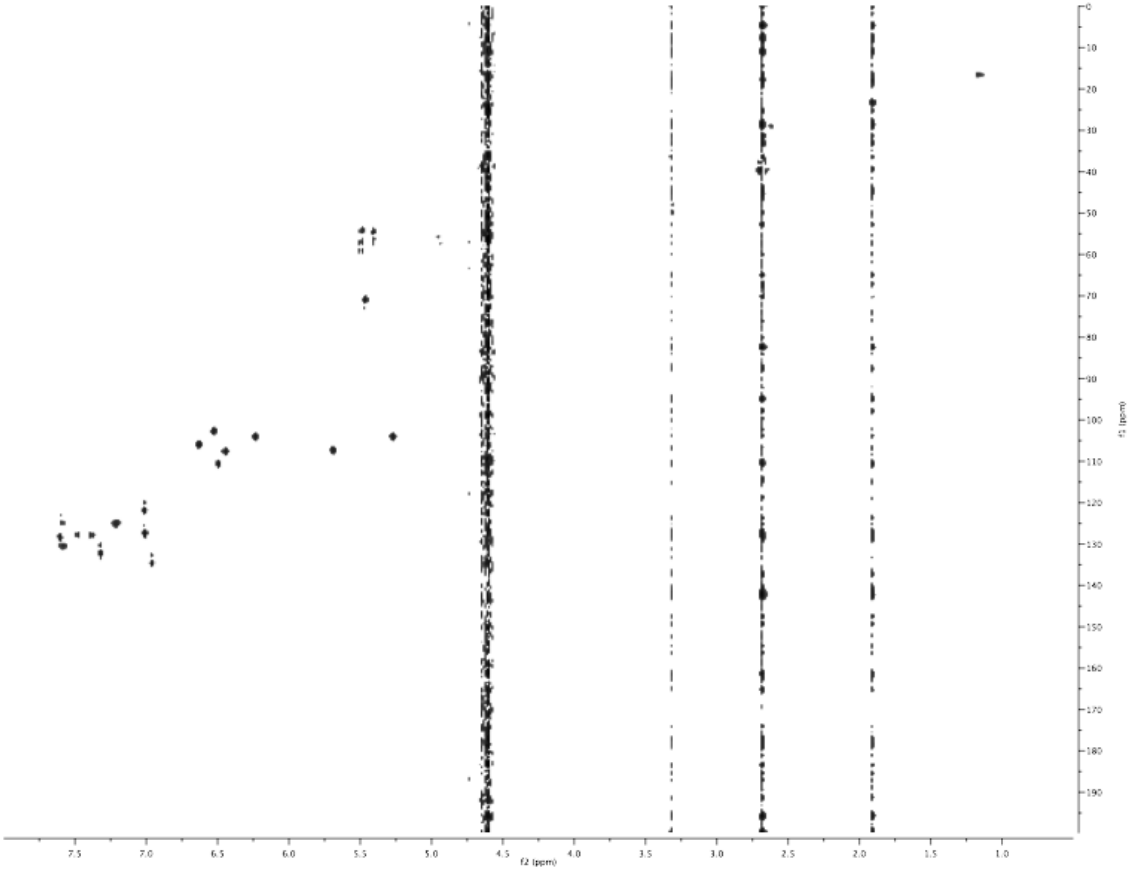


Figure 44: ^{1}H-^{13}C HMBC of Compound 31, 298K in 3:1 D_2O:CD_3CN

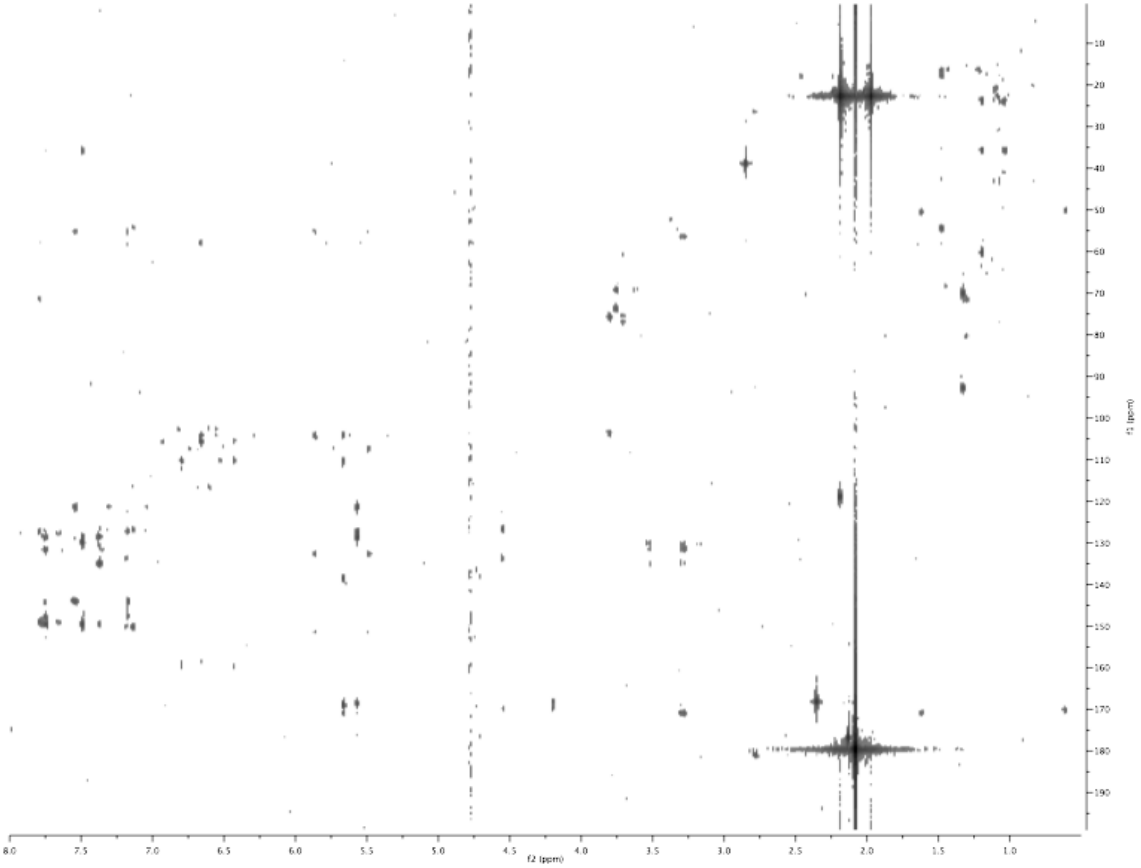


Figure 45: ^{1}H of Compound 31, 313K in 3:1 D_2O:CD_3CN

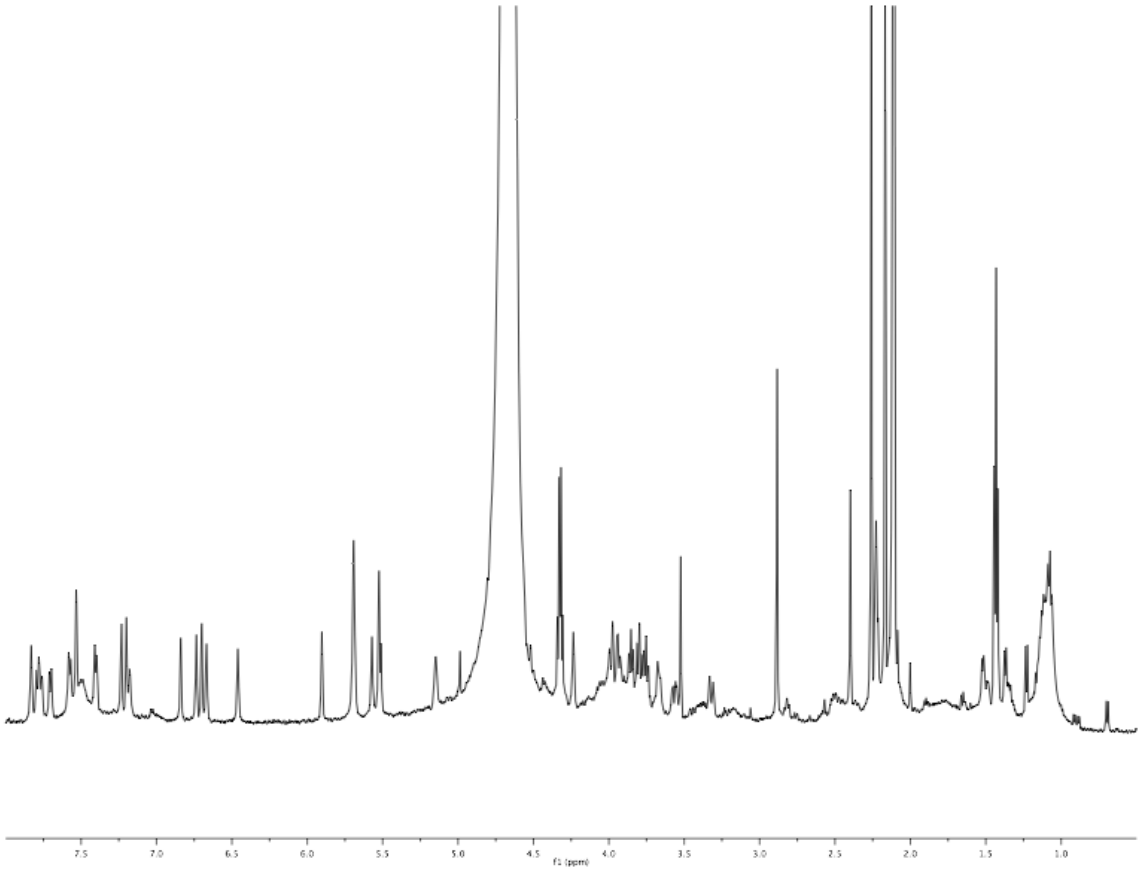


Figure 46: ^{1}H of Compound 32, 298K in 3:1 D_2O:CD_3CN

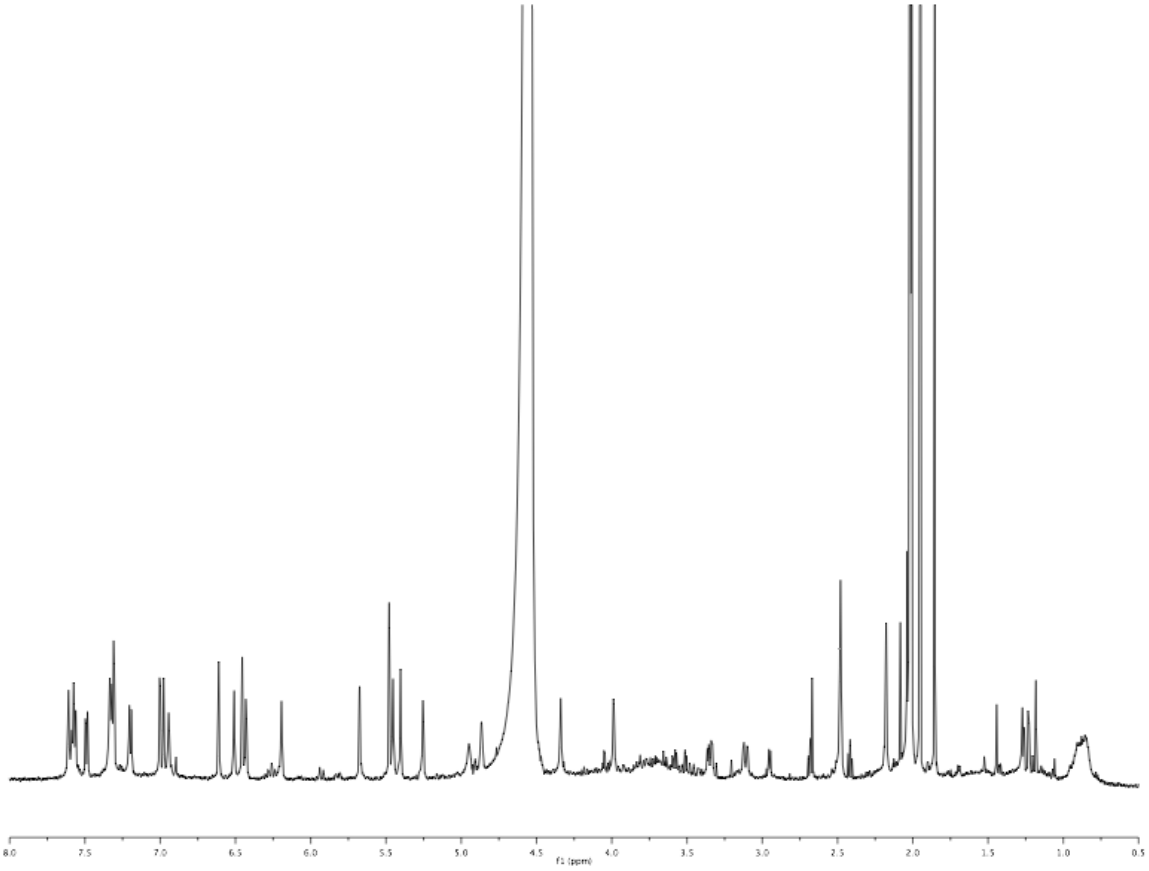


Figure 47: ^{1}H-^{13}C HMQC of Compound 32, 298K in 3:1 D_2O:CD_3CN

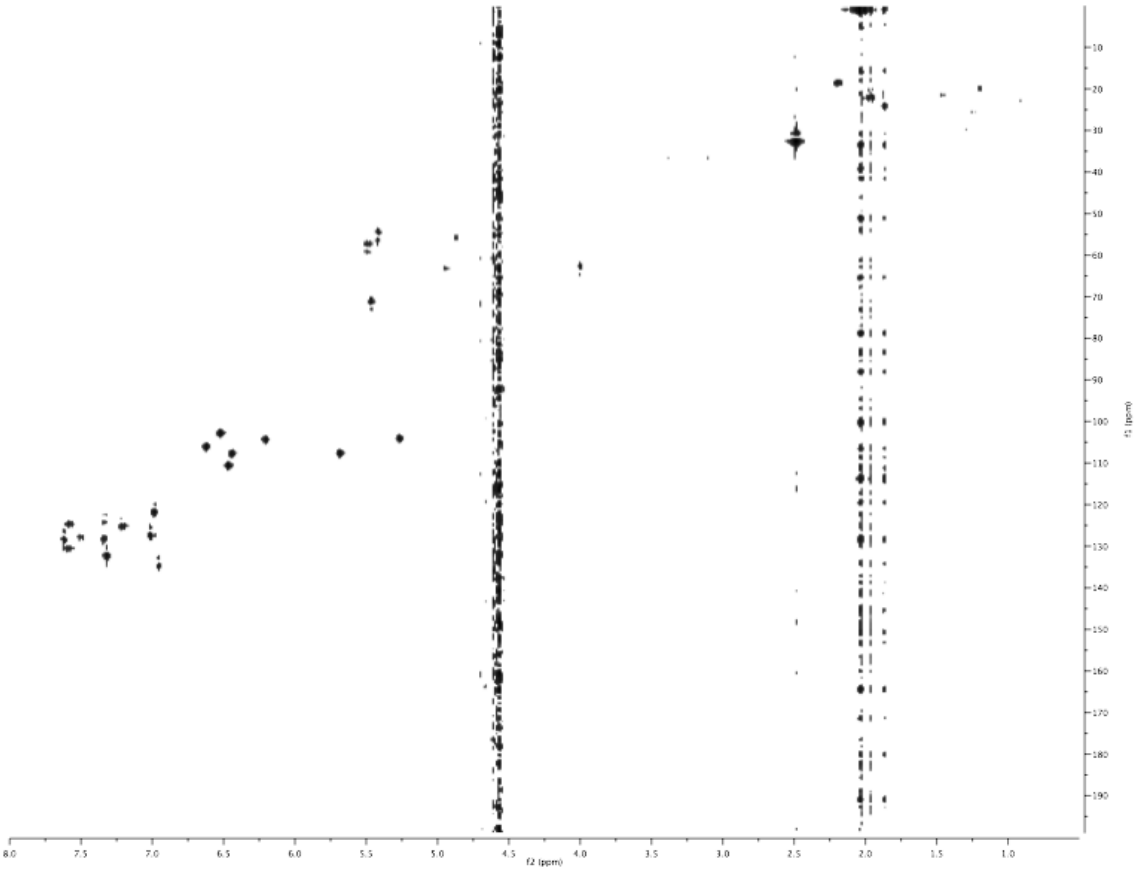


Figure 48: ^{1}H-^{13}C HMBC of Compound 32, 298K in 3:1 D_2O:CD_3CN

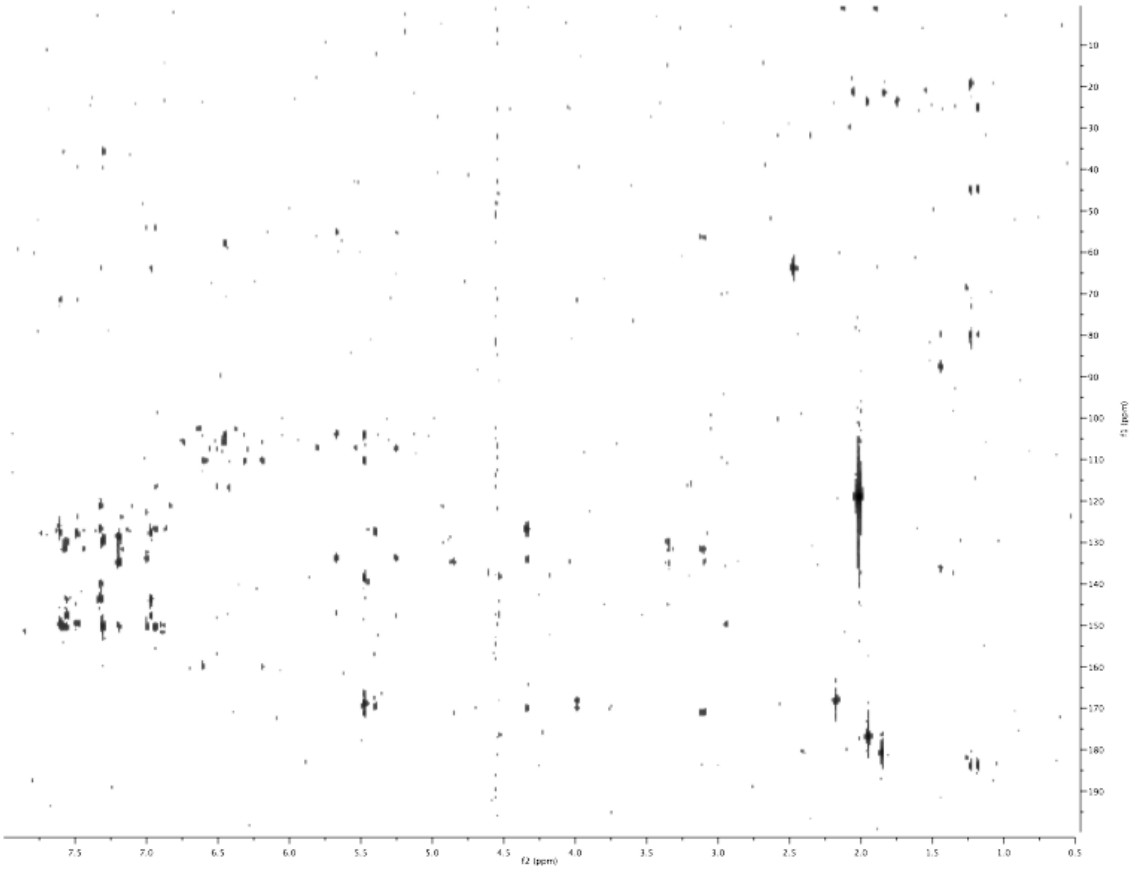
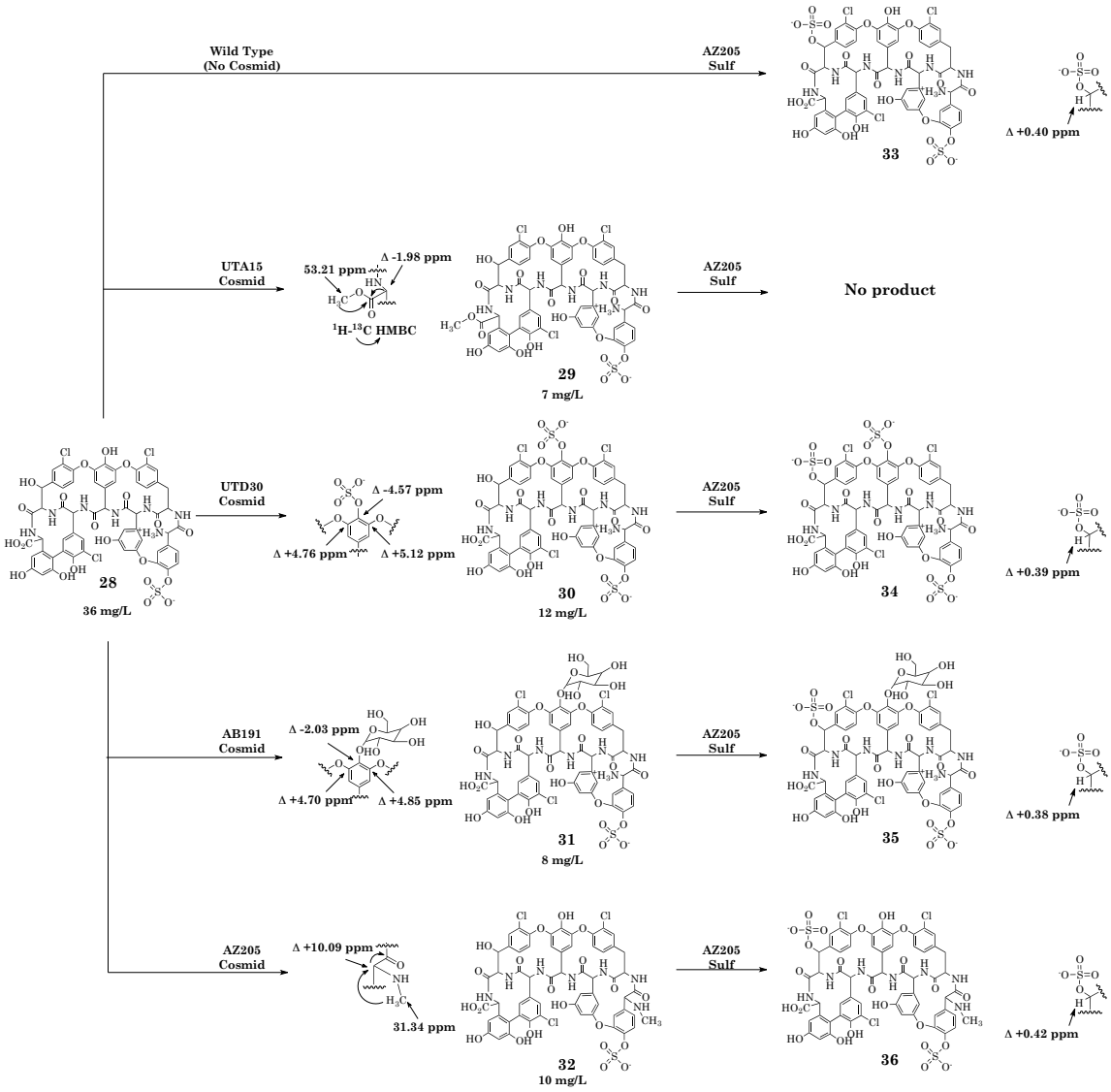


Figure 49: New glycopeptide congeners produced using eDNA-derived tailoring enzymes



28	29	30	31	32	33	34	35	36
HRMS: $[M+H]^+$	HRMS: $[M+H]^+$	HRMS: $[M+H]^+$	HRMS: $[M+H]^+$	HRMS: $[M-H]^-$	HRMS: $[M-H]^-$	MS: $[M+H]^+$	HRMS: $[M-H]^-$	HRMS: $[M-H]^-$
$C_{58}H_{45}Cl_3N_7O_{21}S$	$C_{59}H_{47}Cl_3N_7O_{21}S$	$C_{58}H_{45}Cl_3N_7O_{24}S_2$	$C_{64}H_{55}Cl_3N_7O_{26}S$	$C_{59}H_{45}Cl_3N_7O_{21}S$	$C_{58}H_{43}Cl_3N_7O_{24}S_2$	$C_{58}H_{45}Cl_3N_7O_{27}S_3$	$C_{64}H_{53}Cl_3N_7O_{29}S_2$	$C_{59}H_{45}Cl_3N_7O_{24}S_2$
Calc: 1312.1455	Calc: 1326.1611	Calc: 1392.1023	Calc: 1474.1983	Calc: 1324.1455	Calc: 1390.0866	Calc: 1472.06	Calc: 1552.1395	Calc: 1404.1023
Obs: 1312.1454	Obs: 1326.1681	Obs: 1392.1104	Obs: 1474.1949	Obs: 1324.1517	Obs: 1390.0934	Obs: 1472.09	Obs: 1552.1267	Obs: 1404.1329
	1 + CH_2	**1** + SO_3	**1** + $C_6H_{10}O_2$	**1** + CH_2	**1** + SO_3	**3** + SO_3	**4** + SO_3	**5** + SO_3

REFERENCES

Aakvik, T., K. F. Degnes, et al. (2009). "A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species." FEMS Microbiol Lett **296**(2): 149-158.

Ansari, M. Z., G. Yadav, et al. (2004). "NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases." Nucleic Acids Res **32**(Web Server issue): W405-413.

Aziz, R. K., D. Bartels, et al. (2008). "The RAST Server: rapid annotations using subsystems technology." BMC genomics **9**: 75.

Banik, J. J. and S. F. Brady (2008). "Cloning and characterization of new glycopeptide gene clusters found in an environmental DNA megalibrary." Proc Natl Acad Sci U S A **105**(45): 17273-17277.

Bauer, J. D., R. W. King, et al. (2010). "Utahmycins a and B, azaquinones produced by an environmental DNA clone." J Nat Prod **73**(5): 976-979.

Bick, M. J., J. J. Banik, et al. (2010). "The 2.7 A resolution structure of the glycopeptide sulfotransferase Teg14." Acta crystallographica. Section D, Biological crystallography **66**(Pt 12): 1278-1286.

Bick, M. J., J. J. Banik, et al. (2010). "Crystal Structures of the Glycopeptide Sulfotransferase Teg12 in a Complex with the Teicoplanin Aglycone." Biochemistry **49**(19): 4159-4168.

Boeck, L. D. and F. P. Mertz (1986). "A47934, a novel glycopeptide-aglycone antibiotic produced by a strain of Streptomyces toyocaensis taxonomy and fermentation studies." J Antibiot (Tokyo) **39**(11): 1533-1540.

Boger, D. L., J.-H. Weng, et al. (2000). "Thermal Atropisomerism of Teicoplanin Aglycon Derivatives: Preparation of the P,P,P and M,P,P

Atropisomers of the Teicoplanin Aglycon via Selective Equilibration of the DE Ring System." J Am Chem Soc **122**(41): 10047 - 10055.

Brady, S. F. (2007). "Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules." Nat Protoc **2**(5): 1297-1305.

Brady, S. F. (2007). "Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules." Nature Protocols **2**: 1297-1305.

Brady, S. F., C. J. Chao, et al. (2001). "Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA." Org Lett **3**(13): 1981-1984.

Brady, S. F. and J. Clardy (2000). "Long-Chain N-Acyl Amino Acid Antibiotics Isolated from Heterologously Expressed Environmental DNA." Journal of the American Chemical Society **122**(51): 12903-12904.

Brady, S. F. and J. Clardy (2005). "Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA." Angew Chem Int Ed Engl **44**(43): 7063-7065.

Brunger, A. T. (2007). "Version 1.2 of the Crystallography and NMR system." Nat Protoc **2**(11): 2728-2733.

Brunger, A. T., P. D. Adams, et al. (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta crystallographica. Section D, Biological crystallography **54**(Pt 5): 905-921.

Buckingham, J. (1994). Dictionary of natural products. London ; New York, Chapman & Hall.

Campbell, J., A. K. Singh, et al. (2010). "Synthetic lethal compound combinations reveal a fundamental connection between wall teichoic

acid and peptidoglycan biosyntheses in Staphylococcus aureus." ACS Chem Biol **6**(1): 106-116.

Chan, C. K., A. L. Hsu, et al. (2008). "Binning sequences using very sparse labels within a metagenome." BMC Bioinformatics **9**: 215.

Chapman, E., M. D. Best, et al. (2004). "Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility." Angew Chem Int Ed Engl **43**(27): 3526-3548.

Chen, H., M. G. Thomas, et al. (2000). "Deoxysugars in glycopeptide antibiotics: enzymatic synthesis of TDP-L-epivancosamine in chloroeremomycin biosynthesis." Proc Natl Acad Sci U S A **97**(22): 11942-11947.

Chen, H., C. C. Tseng, et al. (2001). "Glycopeptide antibiotic biosynthesis: enzymatic assembly of the dedicated amino acid monomer (S)-3,5-dihydroxyphenylglycine." Proc Natl Acad Sci U S A **98**(26): 14901-14906.

Chew, Y. V. and A. J. Holmes (2009). "Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest." Journal of Microbiological Methods **78**(2): 136-143.

Choroba, O. W., Williams, D. H., and Spencer, J. B. (2000). "Biosynthesis of the Vancomycin Group of Antibiotics: Involvement of an Unusual Dioxygenase in the Pathway to (S)-4-Hydroxyphenylglycine." J. Am. Chem. Soc. **122**: 5389-5390.

Courtois, S., C. M. Cappellano, et al. (2003). "Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products." Appl Environ Microbiol **69**(1): 49-55.

Craig, J. W., F. Y. Chang, et al. (2009). "Natural products from environmental DNA hosted in Ralstonia metallidurans." ACS Chem Biol **4**(1): 23-28.

Craig, J. W., F. Y. Chang, et al. (2010). "Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria." Appl Environ Microbiol **76**(5): 1633-1641.

Davis, I. W., A. Leaver-Fay, et al. (2007). "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids." Nucleic Acids Res **35**(Web Server issue): W375-383.

Debono, M., K. E. Merkel, et al. (1984). "Actaplanin, new glycopeptide antibiotics produced by Actinoplanes missouriensis. The isolation and preliminary chemical characterization of actaplanin." J Antibiot (Tokyo) **37**(2): 85-95.

Diep, B. A., S. R. Gill, et al. (2006). "Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus." The Lancet **367**(9512): 731-739.

Diep, B. A., S. R. Gill, et al. (2006). "Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus." Lancet **367**(9512): 731-739.

Donia, M. S., B. J. Hathaway, et al. (2006). "Natural combinatorial peptide libraries in cyanobacterial symbionts of marine ascidians." Nat Chem Biol **2**(12): 729-735.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-2132.

Fehlner, J. R., R. E. Hutchinson, et al. (1972). "Structure of ristocetin A." Proc Natl Acad Sci U S A **69**(9): 2420-2421.

Feng, Z., J. H. Kim, et al. (2010). "Fluostatins Produced by the Heterologous Expression of a TAR Reassembled Environmental DNA Derived Type II PKS Gene Cluster." J. Am. Chem. Soc. **132**: 11902-11903.

Fisch, K. M., C. Gurgui, et al. (2009). "Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting." Nat Chem Biol **5**(7): 494-501.

Fox, J. L. (2006). "The business of developing antibacterials." Nat Biotechnol **24**(12): 1521-1528.

Gabor, E. M., W. B. Alkema, et al. (2004). "Quantifying the accessibility of the metagenome by random expression cloning techniques." Environ Microbiol **6**(9): 879-886.

Gillespie, D. E., S. F. Brady, et al. (2002). "Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA." Appl Environ Microbiol **68**(9): 4301-4306.

Glass, E. M., J. Wilkening, et al. (2010). "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." Cold Spring Harb Protoc **2010**(1): pdb prot5368.

Goldstein, B. P., E. Selva, et al. (1987). "A40926, a new glycopeptide antibiotic with anti-Neisseria activity." Antimicrob Agents Chemother **31**(12): 1961-1966.

Gray, K. A., T. H. Richardson, et al. (2003). Soil-Based Gene Discovery: A New Technology to Accelerate and Broaden Biocatalytic Applications. Advances in Applied Microbiology, Academic Press. **Volume 52:** 1-27.

Guzman-Martinez, A., R. Lamer, et al. (2007). "Total synthesis of lysobactin." J Am Chem Soc **129**(18): 6017-6021.

Haeder, S., R. Wirth, et al. (2009). "Candicidin-producing Streptomyces support leaf-cutting ants to protect their fungus garden against the pathogenic fungus Escovopsis." Proc Natl Acad Sci U S A **106**(12): 4742-4746.

Handelsman, J., M. R. Rondon, et al. (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." Chemistry & Biology **5**(10): R245-R249.

Heald, S. L., L. Mueller, et al. (1987). "Actinoidins A and A2: structure determination using 2D NMR methods." J Antibiot (Tokyo) **40**(5): 630-645.

Higgins, D. L., R. Chang, et al. (2005). "Telavancin, a multifunctional lipoglycopeptide, disrupts both cell wall synthesis and cell membrane integrity in methicillin-resistant Staphylococcus aureus." Antimicrob Agents Chemother **49**(3): 1127-1134.

Holdom, K. S., Maeda, H., Ruddock, J.C, and Tone, J. (Pfizer, Ltd.) (1988). EP-B **265143**.

Hubbard, B. K., M. G. Thomas, et al. (2000). "Biosynthesis of L-p-hydroxyphenylglycine, a non-proteinogenic amino acid constituent of peptide antibiotics." Chem Biol **7**(12): 931-942.

Hugenholtz, P., B. M. Goebel, et al. (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J Bacteriol **180**(18): 4765-4774.

Kalyuzhnaya, M. G., R. Zabinsky, et al. (2006). "Fluorescence In Situ Hybridization-Flow Cytometry-Cell Sorting-Based Method for Separation and Enrichment of Type I and Type II Methanotroph Populations." Applied and Environmental Microbiology **72**(6): 4293-4301.

Kieser, T., Bibb, MJ, Buttner MJ, Chater, KF, Hopwood, DA. (2000). "Practical Streptomyces Genetics."

Kim, J. H., Z. Feng, et al. (2010). "Cloning large natural product gene clusters from the environment: Piecing environmental DNA gene clusters back together with TAR." Biopolymers **93**(9): 833-844.

King, R. W., J. D. Bauer, et al. (2009). "An environmental DNA-derived type II polyketide biosynthetic pathway encodes the biosynthesis of the pentacyclic polyketide erdacin." Angew Chem Int Ed Engl **48**(34): 6257-6261.

Komatsu, M., T. Uchiyama, et al. (2010). "Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism." Proc Natl Acad Sci U S A **107**(6): 2646-2651.

Kovach, M. E., P. H. Elzer, et al. (1995). "Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes." Gene **166**(1): 175-176.

Lamb, S. S., T. Patel, et al. (2006). "Biosynthesis of sulfated glycopeptide antibiotics by using the sulfotransferase StaL." Chem Biol **13**(2): 171-181.

Li, B., D. Sher, et al. (2010). "Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria." Proc Natl Acad Sci U S A **107**(23): 10430-10435.

Lim, H. K., E. J. Chung, et al. (2005). "Characterization of a forest soil metagenome clone that confers indirubin and indigo production on Escherichia coli." Appl Environ Microbiol **71**(12): 7768-7777.

Liu, J., K. Duncan, et al. (1989). "Nucleotide sequence of a cluster of Escherichia coli enterobactin biosynthesis genes: identification of entA and purification of its product 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase." J Bacteriol **171**(2): 791-798.

Long, P. F., W. C. Dunlap, et al. (2005). "Shotgun cloning and heterologous expression of the patellamide gene cluster as a strategy to achieving sustained metabolite production." Chembiochem **6**(10): 1760-1765.

MacNeil, I. A., C. L. Tiong, et al. (2001). "Expression and isolation of antimicrobial small molecules from soil DNA libraries." J Mol Microbiol Biotechnol **3**(2): 301-308.

Macpherson, D. F., P. A. Manning, et al. (1994). "Characterization of the dTDP-rhamnose biosynthetic genes encoded in the rfb locus of Shigella flexneri." Mol Microbiol **11**(2): 281-292.

Magbanua, Z. V., S. Ozkan, et al. (2011). "Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine." PloS one **6**(1): e16214.

Malabarba, A., P. Ferrari, et al. (1986). "Teicoplanin, antibiotics from Actinoplanes teichomyceticus nov. sp. VII. Preparation and NMR characteristics of the aglycone of teicoplanin." J Antibiot (Tokyo) **39**(10): 1430-1442.

Maplestone, R. A., M. J. Stone, et al. (1992). "The evolutionary role of secondary metabolites -- a review." Gene **115**(1-2): 151-157.

Marahiel, M. A., T. Stachelhaus, et al. (1997). "Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis." Chem Rev **97**(7): 2651-2674.

Martinez, A., S. J. Kolvek, et al. (2004). "Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts." Appl Environ Microbiol **70**(4): 2452-2463.

Mato, R., H. de Lencastre, et al. (1996). "Multiplicity of genetic backgrounds among vancomycin-resistant Enterococcus faecium isolates recovered from an outbreak in a New York City hospital." Microb Drug Resist **2**(3): 309-317.

Matsushima, P. and R. H. Baltz (1996). "A gene cloning system for 'Streptomyces toyocaensis'." Microbiology **142 (Pt 2)**: 261-267.

Matsushima, P., M. C. Broughton, et al. (1994). "Conjugal transfer of cosmid DNA from Escherichia coli to Saccharopolyspora spinosa: effects of chromosomal insertions on macrolide A83543 production." Gene **146**(1): 39-45.

McCoy, A. J., R. W. Grosse-Kunstleve, et al. (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

Morimoto, S. and T. Fujii (2009). "A new approach to retrieve full lengths of functional genes from soil by PCR-DGGE and metagenome walking." Appl Microbiol Biotechnol **83**(2): 389-396.

Morozova, O. and M. A. Marra (2008). "Applications of next-generation sequencing technologies in functional genomics." Genomics **92**(5): 255-264.

Murshudov, G. N., A. A. Vagin, et al. (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-255.

Newman, D. J. and G. M. Cragg (2007). "Natural products as sources of new drugs over the last 25 years." J Nat Prod **70**(3): 461-477.

Nicolaou, K. C., Boddy, C.N. C., Brase, S., and Winssinger, N. (1999). "Chemistry, Biology, and Medicine of the Glycopeptide Antibiotics." Angew. Chem. Int. Ed. **38**: 2096-2152.

Otwinowski, Z. and W. Minor (1997). "Processing of X-ray Diffraction Data Collected in Oscillation Mode." Methods in Enzymology **276A**: 307-326.

Pelzer, S., R. Sussmuth, et al. (1999). "Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in Amycolatopsis mediterranei DSM5908." Antimicrob Agents Chemother **43**(7): 1565-1573.

Peraud, O., J. S. Biggs, et al. (2009). "Microhabitats within venomous cone snails contain diverse actinobacteria." Appl Environ Microbiol **75**(21): 6820-6826.

Piel, J. (2002). "A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles." Proc Natl Acad Sci U S A **99**(22): 14002-14007.

Piel, J., D. Hui, et al. (2004). "Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei." Proc Natl Acad Sci U S A **101**(46): 16222-16227.

Pootoolal, J., M. G. Thomas, et al. (2002). "Assembling the glycopeptide antibiotic scaffold: The biosynthesis of A47934 from Streptomyces toyocaensis NRRL15009." Proc Natl Acad Sci U S A **99**(13): 8962-8967.

Pretsch, E. (2009). Structure determination of organic compounds : tables of spectral data. New York, Springer.

Pretsch, E., P. Bühlmann, et al. (2000). Structure determination of organic compounds : tables of spectral data. Berlin ; New York, Springer.

Puk, O., D. Bischoff, et al. (2004). "Biosynthesis of chloro-beta-hydroxytyrosine, a nonproteinogenic amino acid of the peptidic backbone of glycopeptide antibiotics." J Bacteriol **186**(18): 6093-6100.

Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Quinton, C. M., G. B. Stephanie, et al. (2007). "Subtractive hybridization magnetic bead capture: A new technique for the recovery of full-length ORFs from the metagenome." Biotechnology Journal **2**(1): 36-40.

Rappe, M. S. and S. J. Giovannoni (2003). "The uncultured microbial majority." Annu Rev Microbiol **57**: 369-394.

Rondon, M. R., P. R. August, et al. (2000). "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms." Appl Environ Microbiol **66**(6): 2541-2547.

Sarkar, G., and Sommer, S.S. (1990). "The "megaprimer" method of site-directed mutagenesis." BioTechniques **8**: 404-407.

Sarkar, G. and S. S. Sommer (1990). "The "megaprimer" method of site-directed mutagenesis." Biotechniques **8**(4): 404-407.

Schafer, M., T. R. Schneider, et al. (1996). "Crystal structure of vancomycin." Structure **4**(12): 1509-1515.

Schipper, C., C. Hornung, et al. (2009). "Metagenome-derived clones encoding two novel lactonase family proteins involved in biofilm inhibition in Pseudomonas aeruginosa." Appl Environ Microbiol **75**(1): 224-233.

Schmidt, E. W. and M. S. Donia (2009). "Chapter 23. Cyanobactin ribosomally synthesized peptides--a case of deep metagenome mining." Methods Enzymol **458**: 575-596.

Schmidt, E. W., J. T. Nelson, et al. (2005). "Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella." Proc Natl Acad Sci U S A **102**(20): 7315-7320.

Schmitz, J. E., A. Daniel, et al. (2008). "Rapid DNA library construction for functional genomic and metagenomic screening." Appl Environ Microbiol **74**(5): 1649-1652.

Seow, K. T., G. Meurer, et al. (1997). "A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms." J Bacteriol **179**(23): 7360-7368.

Shi, R., S. S. Lamb, et al. (2007). "Crystal structure of StaL, a glycopeptide antibiotic sulfotransferase from Streptomyces toyocaensis." J Biol Chem **282**(17): 13073-13086.

Sieradzki, K., T. Leski, et al. (2003). "Evolution of a vancomycin-intermediate Staphylococcus aureus strain in vivo: multiple changes in the antibiotic resistance phenotypes of a single lineage of methicillin-resistant S. aureus under the impact of antibiotics administered for chemotherapy." J Clin Microbiol **41**(4): 1687-1693.

Soldati, M., Fioretti, A., and Ghione, M. (1966). "Cytotoxicity of pederin and some of its derivatives on cultured mammalian cells." Experientia **22**(3): 176-178.

Sosio, M., H. Kloosterman, et al. (2004). "Organization of the teicoplanin gene cluster in Actinoplanes teichomyceticus." Microbiology **150**(Pt 1): 95-102.

Sri, M., E. Ellen, et al. (2009). "The Thioesterase Bhp is Involved in the Formation of beta-Hydroxytyrosine during Balhimycin Biosynthesis in <I>Amycolatopsis balhimycina</I>." ChemBioChem **11**(2): 266-271.

Stachelhaus, T., H. D. Mootz, et al. (1999). "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases." Chem Biol **6**(8): 493-505.

Stinchi, S., L. Carrano, et al. (2006). "A derivative of the glycopeptide A40926 produced by inactivation of the beta-hydroxylase gene in Nonomuraea sp. ATCC39727." FEMS Microbiol Lett **256**(2): 229-235.

Tao, Q. and H. B. Zhang (1998). "Cloning and stable maintenance of DNA fragments over 300 kb in Escherichia coli with conventional plasmid-based vectors." Nucleic Acids Res **26**(21): 4901-4909.

Torsvik, V., F. L. Daae, et al. (1998). "Novel techniques for analysing microbial diversity in natural and perturbed environments." J Biotechnol **64**(1): 53-62.

Torsvik, V., J. Goksoyr, et al. (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-787.

Torsvik, V. and L. Ovreas (2002). "Microbial diversity and function in soil: from genes to ecosystems." Curr Opin Microbiol **5**(3): 240-245.

Torsvik, V., K. Salte, et al. (1990). "Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria." Appl Environ Microbiol **56**(3): 776-781.

van Wageningen, A. M., P. N. Kirkpatrick, et al. (1998). "Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic." Chem Biol **5**(3): 155-162.

Vogel, T. M., P. Simonet, et al. (2009). "TerraGenome: a consortium for the sequencing of a soil metagenome." Nat Rev Micro **7**(4): 252-252.

Wang, G. Y., E. Graziani, et al. (2000). "Novel natural products from soil DNA libraries in a streptomycete host." Org Lett **2**(16): 2401-2404.

Wang, H. H., F. J. Isaacs, et al. (2009). "Programming cells by multiplex genome engineering and accelerated evolution." Nature **460**(7257): 894-898.

Winn, M. D., M. N. Isupov, et al. (2001). "Use of TLS parameters to model anisotropic displacements in macromolecular refinement." Acta crystallographica. Section D, Biological crystallography **57**(Pt 1): 122-133.

Yang, B., Y. Peng, et al. (2010). "Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers." BMC Bioinformatics **11 Suppl 2**: S5.

Yates, E. A., F. Santini, et al. (2000). "Effect of substitution pattern on 1H, 13C NMR chemical shifts and 1J(CH) coupling constants in heparin derivatives." Carbohydr Res **329**(1): 239-247.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zhang, K., J. He, et al. (2009). "Identifying natural product biosynthetic genes from a soil metagenome by using T7 phage selection." Chembiochem **10**(16): 2599-2606.

Zhang, Y., J. P. Muyrers, et al. (2000). "DNA cloning by homologous recombination in Escherichia coli." Nat Biotechnol **18**(12): 1314-1317.

Ziemert, N., K. Ishida, et al. (2010). "Exploiting the natural diversity of microviridin gene clusters for discovery of novel tricyclic depsipeptides." Appl Environ Microbiol **76**(11): 3568-3574.

Zimmermann, K., M. Engeser, et al. (2009). "Pederin-type pathways of uncultivated bacterial symbionts: analysis of o-methyltransferases and generation of a biosynthetic hybrid." J Am Chem Soc **131**(8): 2780-2781.