Supporting Information

Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining

Michael A. Skinnider[a,1], Chad W. Johnston[a,1], Robyn E. Edgar[a], Chris A. Dejong[a], Nishanth J. Merwin[a], Philip N. Rees[a], and Nathan A. Magarvey[a,2]

[a] Departments of Biochemistry and Biomedical Sciences and Chemistry and Chemical Biology, Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, L8S 4K1, Canada

[1] M.A.S. and C.W.J. contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail: magarv@mcmaster.ca

Figure S1

Average maximum Tanimoto coefficient between predicted structure libraries and true RiPP structures for 21 RiPP families. Error bars represent standard deviation.

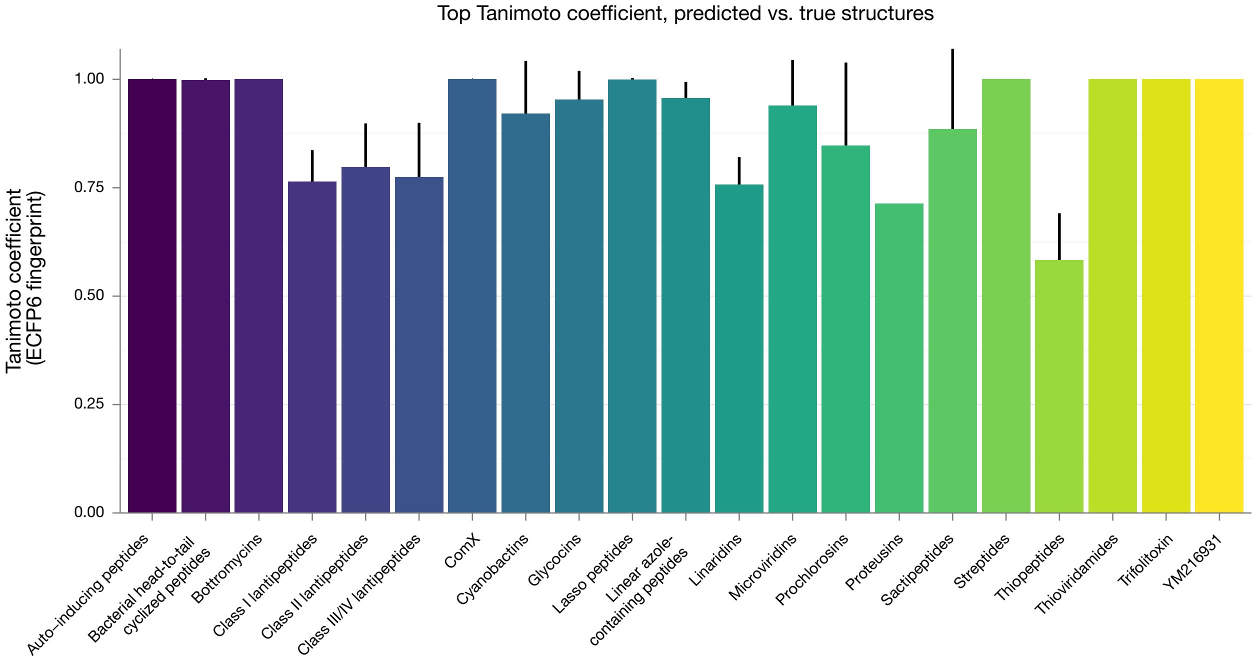


Figure S2

Effect of systematically withholding all sequences derived from a given biosynthetic gene cluster on structure prediction for that cluster product. (*A*) Overall median Tanimoto coefficient between predicted structure libraries and true RiPP structures before and after withholding cluster sequences. (*B*) Median Tanimoto coefficient between predicted structure libraries and true RiPP structures before and after withholding cluster sequences for each RiPP family. (*C*) Overall maximum Tanimoto coefficient between predicted structure libraries and true RiPP structures before and after withholding cluster sequences. (*D*) Maximum Tanimoto coefficient between predicted structure libraries and true RiPP structures before and after withholding cluster sequences for each RiPP family. Error bars represent standard deviation.

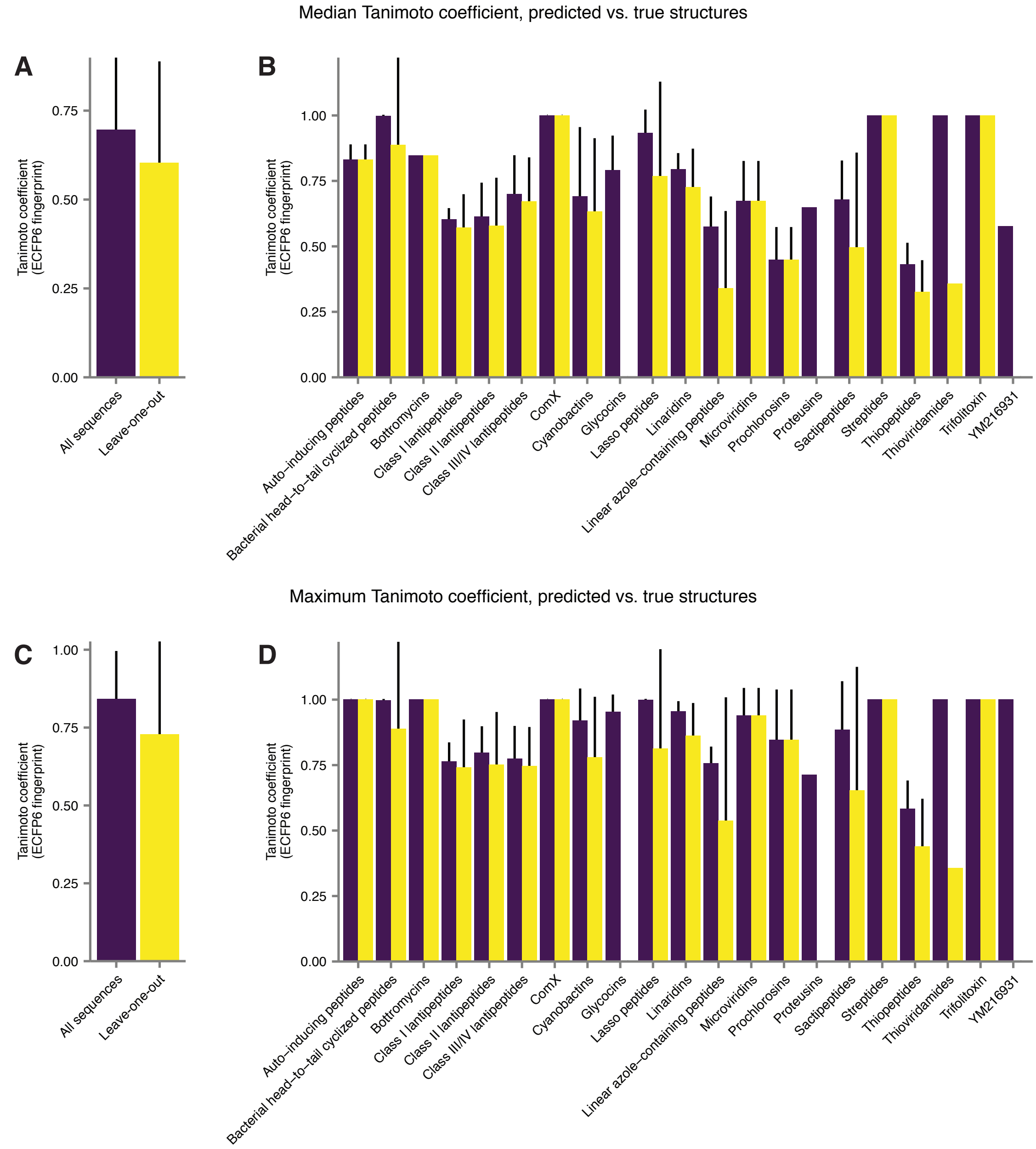


Figure S3

Examples of noteworthy RiPP biosynthetic gene clusters detected during RiPP-PRISM analysis of microbial genomes discussed in the main text.

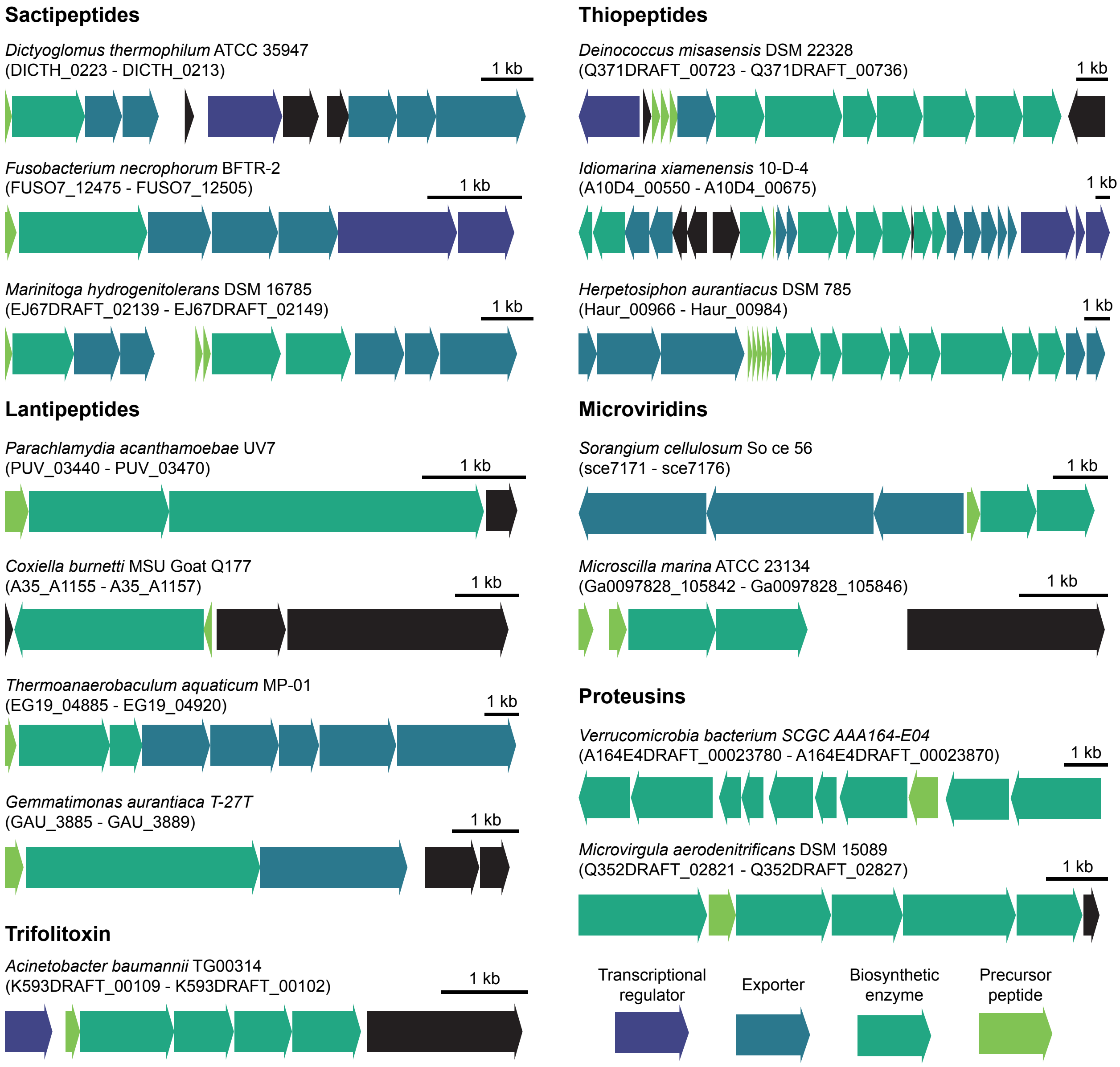

Sactipeptides
Dictyoglomus thermophilum ATCC 35947
(DICTH_0223 - DICTH_0213)
1 kb
Fusobacterium necrophorum BFTR-2
(FUSO7_12475 - FUSO7_12505)
1 kb
Marinitoga hydrogenitolerans DSM 16785
(EJ67DRAFT_02139 - EJ67DRAFT_02149)
1 kb
Lantipeptides
Parachlamydia acanthamoebae UV7
(PUV_03440 - PUV_03470)
1 kb
Coxiella burnetti MSU Goat Q177
(A35_A1155 - A35_A1157)
1 kb
Thermoanaerobaculum aquaticum MP-01
(EG19_04885 - EG19_04920)
1 kb
Gemmatimonas aurantiaca T-27T
(GAU_3885 - GAU_3889)
1 kb
Trifolitoxin
Acinetobacter baumannii TG00314
(K593DRAFT_00109 - K593DRAFT_00102)
1 kb
Thiopeptides
Deinococcus misasensis DSM 22328
(Q371DRAFT_00723 - Q371DRAFT_00736)
1 kb
Idiomarina xiamenensis 10-D-4
(A10D4_00550 - A10D4_00675)
1 kb
Herpetosiphon aurantiacus DSM 785
(Haur_00966 - Haur_00984)
1 kb
Microviridins
Sorangium cellulosum So ce 56
(sce7171 - sce7176)
1 kb
Microscilla marina ATCC 23134
(Ga0097828_105842 - Ga0097828_105846)
1 kb
Proteusins
Verrucomicrobia bacterium SCGC AAA164-E04
(A164E4DRAFT_00023780 - A164E4DRAFT_00023870)
1 kb
Microvirgula aerodenitrificans DSM 15089
(Q352DRAFT_02821 - Q352DRAFT_02827)
1 kb
Transcriptional regulator
Exporter
Biosynthetic enzyme
Precursor peptide

Figure S4
Graphical output of RiPP-PRISM for the aurantizolicin biosynthetic gene cluster and renderings of each SMILES contained within the library of predicted structures.

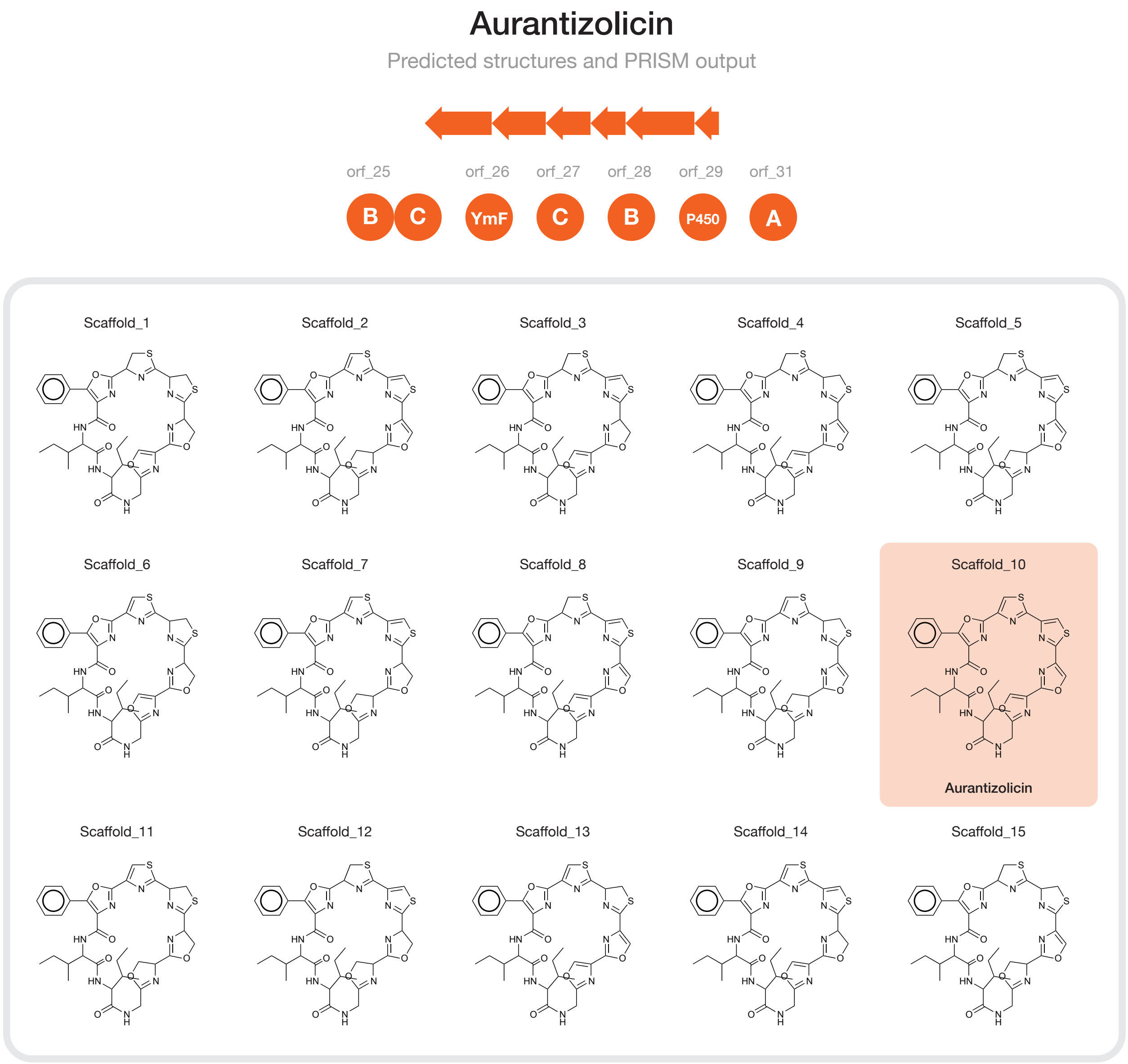


Figure S5

Incorporation of deuterium-labelled phenylalanine (d_8-Phe) into aurantizolicin. (*A*) Structure of aurantizolicin, corresponding to #10 from a library of 15 structures aurantizolicin predicted by PRISM. (*B*) Incorporation of d_8-Phe into aurantizolicin reveals the loss of three deuteriums, resulting from the formation of the phenyloxazole.

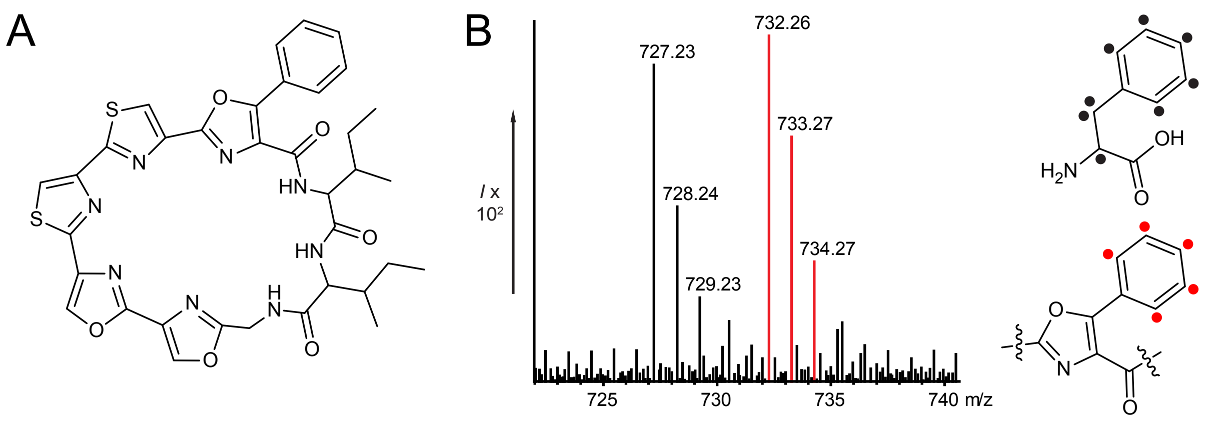


Figure S6

Comparison of RiPP biosynthetic gene clusters identified by RiPP-PRISM, antiSMASH, and ClusterFinder in a sample of 5,049 complete bacterial genomes from NCBI. (*A*) Mode of detection for each of 2,258 identified clusters. (*B*) Mode of detection for each cluster by producer taxonomy, as a count. (*C*) Mode of detection for each cluster by producer taxonomy, as a percentage of all identified clusters for that taxonomic category. "Other" includes taxonomic classes with fewer than 20 RiPP clusters. (*D*) Mode of detection for each cluster by RiPP family, as a count. (*E*) Mode of detection for each cluster by RiPP family, as a percentage of all identified clusters of that family. (*F*) Direct comparison of RiPP-PRISM and ClusterFinder by RiPP family. (*G*) Direct comparison of RiPP-PRISM and ClusterFinder by producer taxonomy. (*H*) Direct comparison of RiPP-PRISM and antiSMASH 3.0 by RiPP family. (*I*) Direct comparison of RiPP-PRISM and antiSMASH 3.0 by producer taxonomy.

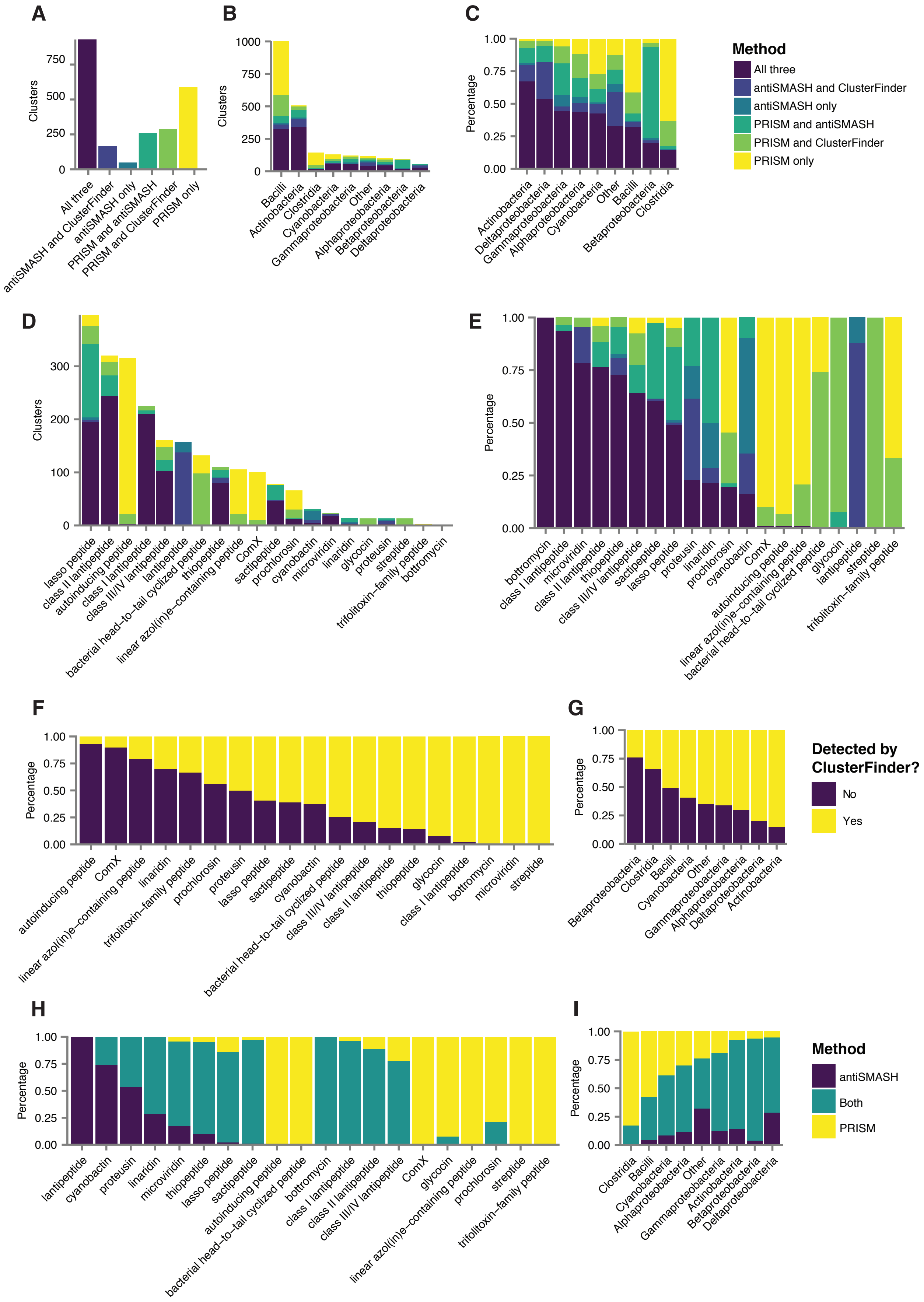

A
Clusters
0
250
500
750
All three
antiSMASH and ClusterFinder
antiSMASH only
PRISM and antiSMASH
PRISM and ClusterFinder
PRISM only
B
Clusters
0
250
500
750
1000
Bacilli
Actinobacteria
Clostridia
Cyanobacteria
Gammaproteobacteria
Other
Alphaproteobacteria
Betaproteobacteria
Deltaproteobacteria
C
Percentage
0.00
0.25
0.50
0.75
1.00
Actinobacteria
Deltaproteobacteria
Gammaproteobacteria
Alphaproteobacteria
Cyanobacteria
Other
Bacilli
Betaproteobacteria
Clostridia
Method
All three
antiSMASH and ClusterFinder
antiSMASH only
PRISM and antiSMASH
PRISM and ClusterFinder
PRISM only
D
Clusters
0
100
200
300
lasso peptide
class II lantipeptide
autoinducing peptide
class I lantipeptide
class III/IV lantipeptide
lantipeptide
bacterial head–to–tail cyclized peptide
thiopeptide
linear azol(in)e–containing peptide
ComX
sactipeptide
prochlorosin
cyanobactin
microviridin
linaridin
glycocin
proteusin
streptide
trifolitoxin–family peptide
bottromycin
E
Percentage
0.00
0.25
0.50
0.75
1.00
bottromycin
class I lantipeptide
microviridin
class II lantipeptide
thiopeptide
class III/IV lantipeptide
sactipeptide
lasso peptide
proteusin
linaridin
prochlorosin
cyanobactin
ComX
autoinducing peptide
linear azol(in)e–containing peptide
bacterial head–to–tail cyclized peptide
glycocin
lantipeptide
streptide
trifolitoxin–family peptide
F
Percentage
0.00
0.25
0.50
0.75
1.00
autoinducing peptide
ComX
linear azol(in)e–containing peptide
linaridin
trifolitoxin–family peptide
prochlorosin
proteusin
lasso peptide
sactipeptide
cyanobactin
bacterial head–to–tail cyclized peptide
class III/IV lantipeptide
class II lantipeptide
thiopeptide
glycocin
class I lantipeptide
bottromycin
microviridin
streptide
G
Percentage
0.00
0.25
0.50
0.75
1.00
Betaproteobacteria
Clostridia
Bacilli
Cyanobacteria
Other
Gammaproteobacteria
Alphaproteobacteria
Deltaproteobacteria
Actinobacteria
Detected by ClusterFinder?
No
Yes
H
Percentage
0.00
0.25
0.50
0.75
1.00
lantipeptide
cyanobactin
proteusin
linaridin
microviridin
thiopeptide
lasso peptide
sactipeptide
autoinducing peptide
bacterial head–to–tail cyclized peptide
bottromycin
class I lantipeptide
class II lantipeptide
class III/IV lantipeptide
ComX
glycocin
linear azol(in)e–containing peptide
prochlorosin
streptide
trifolitoxin–family peptide
I
Percentage
0.00
0.25
0.50
0.75
1.00
Clostridia
Bacilli
Cyanobacteria
Alphaproteobacteria
Other
Gammaproteobacteria
Actinobacteria
Betaproteobacteria
Deltaproteobacteria
Method
antiSMASH
Both
PRISM

Figure S7

Biosynthetic gene clusters for which RiPP-PRISM did not generate a predicted structure by RiPP family (most abundant first) and taxonomy of producing organism. “Other” includes all classes with fewer than 25 RiPP clusters.

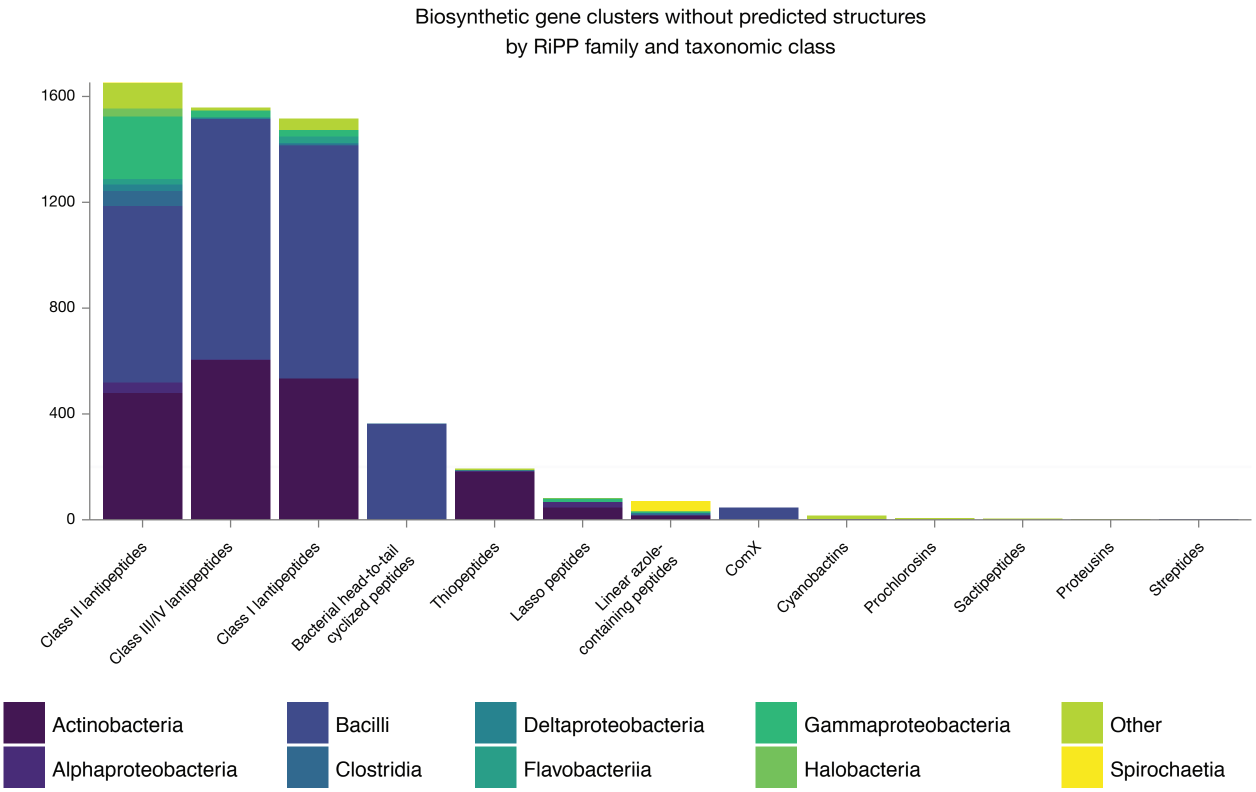


Table S1
Rules for RiPP biosynthetic gene cluster detection in PRISM.

RiPP class	Hidden Markov model hits
Autoinducing peptide	AgrB *and* AgrD
Bacterial head-to-tail cyclized peptide	DUF95 *and* HTT_precursor
Bottromycin	BotA *and* BotC
ComX	ComQ *and* ComX
Cyanobactin	PatA *and* PatE *and* (PatG *or* PatG_ox)
Glycocin	SunA *and* SunS
Class I lantipeptide	LanB *and* LanC
Class II lantipeptide	LanM
Class III/IV lantipeptide	LanKC
Lasso peptide	transglutaminases *and* asparagine_synthases
Linaridin	(CypA *or* LegA) *and* (CypH *or* LegH) *and* CypL
Linear azole-containing peptide	(McbB *and* (McbC *or* McbD)) *or* GodG
Microviridin	MdnA *and* (MdnB *or* MdnC)
Prochlorosin	ProcA
Proteusin	PoyA *and* PoyD
Sactipeptide	SboA *and* AlbA
Streptide	StrA *and* StrB *and* StrC
Thiopeptide	LazA *and* LazB *and* LazC
Trifolitoxin	TfxA *and* TfxB *and* TfxB
Thioviridamide	TvaA *and* TvaH
YM-216391	YmA *and* YmF

SI Text

Cluster identification, precursor cleavage, and tailoring of RiPPs in PRISM. We developed 58 motifs, 154 new hidden Markov models, and 97 new virtual tailoring reactions in order to comprehensively predict the chemical structures of genetically encoded RiPPs. Here, we detail the implementation of algorithms for biosynthetic gene cluster identification, precursor peptide cleavage, and virtual tailoring reaction execution in PRISM.

Autoinducing peptides. Autoinducing peptides (AIPs) are quorum-sensing peptides produced by firmicutes, and are distinguished from linear regulatory peptides by the presence of a cyclic ester or thioester (1,2). Their biosynthesis is encoded in the accessory gene regulator (agr) locus (1). Conserved features of this locus include the AgrD and AgrB genes, which encode the precursor peptide and the enzyme responsible for leader peptide cleavage and macrocyclization, respectively. Putative AIP clusters are identified in PRISM if they contain homologs of both AgrD and AgrB. Analysis of AgrD sequences revealed a single conserved motif with C-terminal leader peptide cleavage four residues from the end of the motif. However, the relationship between the beginning of the motif and N-terminal cleavage was less clear, with cleavage known to occur at a distance of three, four, or five amino acids before the start of the motif. Consequently, all three possibilities are considered within PRISM. AgrB is assumed to catalyze ester or thioester formation between the C terminus of the cleaved leader peptide and the side chain of a conserved serine or cysteine residues four amino acids from the C terminus.

Bacterial head-to-tail cyclized peptides. Bacterial head-to-tail cyclized peptides are relatively large bacterial peptides which contain a peptide bond between the C and N termini (2,3). They are distinguished from other cyclic RiPPs both by their size and by the mechanism of macrocyclization. A conserved DUF95-family enzyme is present in all known clusters, and may be associated with transport or macrolactam formation. Putative clusters are identified if they contain a precursor peptide and a DUF95-family enzyme. Three conserved motifs were identified within a dataset of 21 known and putative precursor peptides. Cleavage generally occurred five amino acids before the start of the first motif, or three amino acids before the start of the second motif (six when the third motif was also present). The DUF95 enzyme is assumed to catalyze macrolactam formation between the C and N termini.

Bottromycins. The bottromycin family of natural products includes several closely related compounds with unique structural features, including a decarboxylated C-terminal thiazole, a macrocyclic amidine, and multiple C-methylated amino acids (2). The precursor peptide is notable for conserved C-terminal leader cleavage. Putative clusters are identified if they contain the bottromycin precursor and a homolog of BotC, an enzyme with weak homology to the YcaO family of proteins which has been speculated to catalyze macrocyclodehydration (4). A conserved motif was identified at the C-terminal 'follower' sequence, analogous to the N-terminal leader sequence in lantipeptides. Cleavage occurs a single amino acid before the start of this motif in all known bottromycins. Seven virtual tailoring reactions were developed for bottromycin-family RiPPs. BotC is assumed to catalyze macrocyclodehydration and amidine formation between the N-terminal residue and the 5th residue, while the related YcaO family protein BotCD is assumed to catalyze cyclodehydration of the C-terminal cysteine residue to form a thiazole (4). Although no known bottromycin precursor sequences contain serine or threonine residues, these are also considered potential substrates of BotCD in the event that oxazole or 5-methyloxazole-containing bottromycins are discovered in the future. Homologs of the bottromycin O-methyltransferase BotOMT are assumed to catalyze aspartate side chain O-methylation, while removal of the N-terminal methionine is assumed to be catalyzed by the homologs of the leucyl aminopeptidase BotP. Finally, BotRMT3 catalyzes proline β-C-methylation, BotRMT1 catalyzes phenylalanine β-C-methylation, and BotRMT2 catalyzes β-C-methylation of any subset of valine residues.

ComX. ComX is a quorum-sensing peptide produced by bacilli, which is characterized by a unique isoprenylation and cyclization reaction at a conserved tryptophan residue (2,5). Putative clusters are identified if they contain homologs of both the ComX precursor peptide and the isoprenyltransferase ComQ. Cleavage is assumed to occur at 11, 9, or 6 residues after the end of the conserved motif; the largest which produces a peptide with at least five amino acids, although the limited number of annotated precursor peptide sequences and variable length of the precursor peptides limits the accuracy of cleavage prediction. ComQ is assumed to catalyze formation of the unique geranyl-derived tricyclic ring structure of ComX at any tryptophan (5,6).

Cyanobactins. Cyanobactins are a large family of RiPPs produced by cyanobacteria, with common modifications including N-to-C macrocyclization, cyclodehydration to form thiazolines and oxazolines, heterocycle oxidation to form thiazoles and oxazoles, and prenylation (2). Macrocyclization was until recently thought to be a defining feature of this family of RiPPs, but linear cyanobactins have also been reported (7). Biosynthesis involves both N- and C-terminal cleavage of the PatE precursor peptide; several precursor peptides encode multiple core peptides, which are each flanked by conserved N- and C-terminal recognition sequences. The PatA protease catalyzes N-terminal recognition sequence cleavage, while the PatG protease catalyzes C-terminal recognition sequence cleavage and macrocyclization. Several PatG homologs also contain an oxidative domain responsible for azoline oxidation. Other conserved biosynthetic machinery includes the prenyltransferase PatF, which may or may not be active, the cyclodehydratase PatD, and two proteins of unknown function (PatB and PatC).

Putative cyanobactin clusters are identified within PRISM if they contain a cyanobactin precursor and homologs of both the PatA and PatG proteases; the PatG protease may or may not have an oxidative domain. A single propeptide is assumed to be produced from the cyanobactin precursor. Two general patterns of precursor cleavage were observed. One large group of cyanobactins had a single occurrence of a distinct C-terminal motif, with one or two N-terminal motifs. Cleavage occurred at the end of the second N-terminal motif, if present, or twelve residues after the end of the first, with C-terminal cleavage seven residues before the start of the C-terminal motif. The second large group of cyanobactins also had two N-terminal motifs and one of two repeating motifs. N-terminal cleavage occurred four residues after the second N-terminal motif, with C-terminal cleavage four residues before the end of the repeating motif.

Four tailoring reactions were developed for cyanobactin post-translational modifications, including macrocyclization, cyclodehydration, azoline oxidation, and prenylation. PatD homologs are assumed to catalyze cyclodehydration of any subset of serine, threonine, and cysteine residues without *bis*-azoline formation. Distinct hidden Markov models were constructed for PatG sequences with and without oxidation domains. PatG homologs are assumed to catalyze N-to-C macrocyclization, and those with an oxidative domain are additionally assumed to catalyze azoline oxidation. Finally, PatF domains are assumed to each catalyze prenylation of either one or two residues, with possible sites of prenylation including the N-terminus, side chain hydroxyls of threonines, serines, and tyrosines, and arginine guanidino groups. A hidden Markov model was also constructed with inactive PatF sequences to predict the presence of inactive prenylating domains. A hidden Markov model was also constructed specific to the unique aeruginosamide putative bifunctional methyltransferase/prenyltransferase AgeF1, which is assumed to catalyze C-terminal methylation and N-terminal prenylation.

Glycocins. Glycocins are a small family of RiPPs with only two characterized members (sublancin 168 and glycocin F). Their biosynthesis is characterized by the formation of two alpha helices by the introduction of two disulfide bonds, with O- or S-glycosylation of a serine or cysteine residue. Putative clusters are detected if they contain both a glycocin precursor and a glycosyltransferase. The paucity of known glycocin sequences complicates accurate cleavage prediction. However, within PRISM, cleavage is assumed to occur either zero or two residues after the lone identified motif. Although the sublancin 168 cluster contains two thiol-disulfide oxidoreductases (8), the mechanism of disulfide bond formation is not clear, and therefore is associated with the precursor itself within PRISM. Only combinations of disulfide bonds which result in the characteristic ‘hairpin’ secondary structure of glycocins are considered. Homologs of the SunS enzyme are assumed to catalyze glycosylation of cysteine and serine residues. If the tail is glycosylated, as in glycocin F, the hairpin is also assumed to be glycosylated.

Lantipeptides. Lantipeptides are a large class of RiPPs characterized by the presence of the amino acids lanthionine and methyl-lanthionine. They are divided into four classes based on the enzymatic mechanism by which these groups are installed.

In class I lantipeptides, a dehydratase catalyzes the dehydration of serine and threonine residues, and a cyclase subsequently catalyzes cyclization of cysteine residues on the dehydro amino acids. Putative class I lantipeptides are identified by the presence of a dehydratase (LanB) and cyclase (LanC). LanB is assumed to catalyze dehydration of any subset of serine and threonine residues, with a minimum size of two when there are two or more serine and threonine residues. N-terminal dehydroalanine and dehydrobutyric acid residues are tautomerized to pyruvate and 2-oxobutyrate, respectively. LanC is assumed to catalyze cyclization between all possible permutations of cysteines and dehydrated serine or threonine residues.

Hidden Markov models were also developed for several tailoring enzymes specific to individual families of class I lantipeptides. LanD enzymes catalyze oxidative decarboxylation of C-terminal residues to form aminovinylcysteine, and are assumed to react at any combination of a C-terminal cysteine and a dehydroalanine residue (2). LanJ enzymes, such as that found in the biosynthetic gene cluster of the class I lantipeptide lacticin 3147 (9), catalyze the reduction of dehydroalanine to D-alanine, and are assumed to react at any dehydroalanine residue. Homologs of the epilancin short-chain dehydrogenase ElxO are assumed to reduce an N-terminal pyruvate residue to lactate (10). Homologs of the microbisporicin halogenase MibH are assumed to catalyze C-5 chlorination of a tryptophan residue, while homologs of the cytochrome P450 MibO are assumed to catalyze 3,4-dihydroxylation of a proline residue (11). Homologs of the paenibacillin acetyltranferase PaeN are assumed to catalyze N-terminal acetylation (12).

In class II lantipeptides, a fused cyclase/dehydratase (LanM) catalyzes both steps of lanthionine formation. Putative class II lantipeptide clusters are identified within PRISM based on the presence of a LanM enzyme. Since class II lantipeptides may also contain uncyclized dehydroalanine or dehydrobutyric acid residues, LanM is assumed within PRISM to catalyze both cyclization between all possible permutations of cysteines and serine or threonine residues, and dehydration of any subset of remaining serine or threonine residues. Hidden Markov models were also constructed for tailoring enzymes specific to individual families of class II lantipeptides, including homologs of the cinnamycin biosynthesis proteins cinorf7, which is assumed to catalyze cross-linking of a lysine residue and a dehydroalanine residue, and CinX, an α-ketoglutarate/iron(II)-dependent hydroxylase assumed to catalyze β-hydroxylation of an aspartate residue (13). A hidden Markov model was also constructed for homologs of the actagardine luciferase-like monooxygenase GarO, which is assumed to catalyze oxidation of a lanthionine to form a sulfoxide group (14).

In class III and IV lantipeptides, a peptide with lyase, kinase, and cyclase domains catalyzes a range of reactions, including dehydration of serine and threonine residues, lanthionine and methyl-lanthionine formation, as well as the formation of the modified amino acid labionin. Putative class III/IV lantipeptide clusters are identified within PRISM based on the presence of a LanKC/LanL enzyme. Within PRISM, labionin formation occurs between any ordered triad of two serine residues and a cysteine residue such that the size of the first ring is three amino acids, and the size of the second ring is between three and six amino acids. All potential labionins are formed, where the number of possible labionins is defined as the smaller of the number of cysteine residues or half the number of serine residues. For each permutation of labionins, all combinations of dehydroalanines, dehydrobutyrates, and lanthionines are additionally considered.

Prochlorosins represent a distinct group of class II lantipeptides, whose biosynthesis is characterized by the action of a single, highly promiscuous LanM-type enzyme on a diverse series of precursor peptides with a conserved leader peptide, but little core peptide sequence identity (15). A distinct model was constructed for prochlorosin precursors, generically termed ProcA. Because the genomes of many prochlorosin producers contain only a single LanM-type enzyme, but encode precursor peptides at multiple different loci, the presence of a ProcA precursor alone defines a prochlorosin cluster. The LanM annotator and reaction classes from class II lantipeptides are reused for prochlorosins, but with the assumption that this reaction occurs regardless of the presence of a LanM-type enzyme within the same cluster.

With the exception of prochlorosins, precursor identification within PRISM is independent of lantipeptide class. Thirteen hidden Markov models were developed for lantipeptide precursors. Three TIGRFAM models are also included to detect unusual precursors related to the enzymes nitrile hydratase and Nif11 (16), for either lantipeptides or linear azol(in)e-containing peptides (see below). In lantipeptide clusters which still lack a precursor, a heuristic strategy is applied to identify putative precursors: open reading frames between 30 and 80 amino acids containing two cysteines are considered as potential lantipeptide precursors.

Highly conserved motifs were identified for three large clades of lantipeptides. The common GG/GA motif is used preferentially to cleave the precursor peptide, with cleavage at the end of the motif, even if other motifs are present. Two conserved motifs were also identified within two other large clades of lantipeptides, with cleavage seven residues after the start of the first motif, and four residues after the end of the second. A smaller clade of two-component lantipeptides, including cytolysin, cerecidin, and lichenicidin, also had a conserved motif with cleavage four residues after the end. However, more specific motifs were also identified for specific families of lantipeptides, including SapB/AmfS-type lantipeptides (cleavage four residues before the end of the motif), pep5/epilancin-type lantipeptides (cleavage nine residues after the end of the motif), actagardine/michiganin-type lantipeptides (cleavage eight residues after the start of the motif), type IV lantipeptides (cleavage four residues after the end of the motif), labyrinthopeptins (cleavage nine residues before the end of the motif), pinensins (cleavage two residues after the end of the motif), SmbA/BhtA2-type precursors (seven residues before the end of the motif) and SmbB/BhtA1-type precursors (18 residues after the end of the motif), and staphylococcin C55-type lantipeptides (three residues before the end of the motif).

Lasso peptides. Lasso peptides are a family of RiPPs characterized by their unique secondary structure, with an isopeptide bond between the N-terminus and a side chain carboxylic acid forming a loop through which the tail of the peptide is threaded (2). Identification of putative lasso peptide clusters requires the presence of two conserved lasso peptide biosynthesis proteins with homology to transglutaminases and asparagine synthases. Putative lasso peptide precursors are identified with a hidden Markov model or, in lasso peptide clusters where no precursor can be identified, by adapting a previously published heuristic for lasso peptide precursor discovery (17). We relaxed this heuristic to consider all open reading frames less than 80 amino acids in length, with a Tx(G/C)x_{6-8}(D/E) motif at a distance of 0-50 amino acids from the start and more than 5 amino acids from the end. For lasso peptide precursors identified with an HMM, cleavage is predicted to occur three residues before the end of the leader peptide motif. However, for heuristically identified precursors, cleavage is predicted to occur two residues after the start of the motif. Asparagine synthase homologs are assumed to catalyze formation of the lasso fold by condensing the N-terminal amine with the side chain of a glutamate or aspartate residue between positions 7 and 10. Finally, when one or more disulfide bonds are possible, all possible permutations of one, two, or three disulfide bonds are considered.

Linaridins. Linaridins are a small, linear family of natural products characterized by the presence of dehydrated amino acids with aminovinylcysteine moieties installed by a biosynthetic route which diverges from that of lantipeptides. Putative linaridin clusters require a precursor peptide and homologs of the unique CypH and CypL proteins. Separate hidden Markov models were constructed for cypemycin and grisemycin-type precursors and for legonaridin-type precursors, which represent a distinct subfamily of linaridins (18). Distinct motifs were likewise identified for cypemycin and grisemycin-type precursors and for legonaridin-type precursors, with leader cleavage occurring three residues before the end of each. The cypemycin genes CypL and CypH encode proteins with no functional homologs required for cypemycin production, which are proposed to catalyze dehydration of threonine residues (19). Although no dehydroalanine-containing legonaridins are known, these proteins are also considered within PRISM to potentially catalyze dehydration of serine residues, in the event that dehydroalanine-containing legonaridins are isolated in the future. The CypH and CypL enzymes are arbitrarily assigned to threonine and serine dehydration respectively within PRISM, although these enzymes are not found separately (19). A separate hidden Markov model was developed for the distinct legonaridin protein LegH (18). Homologs of the lantipeptide aminovinylcysteine biosynthesis flavoprotein in linaridin clusters are assumed to react at a C-terminal cysteine residue and any other cysteine. Homologs of the cypemycin methyltransferase CypM are assumed to catalyze N,N-dimethylation of the N-terminal residue.

Linear azol(in)e-containing peptides. Linear azol(in)e-containing peptides (LAPs) are a family of RiPPs defined by the presence of thiazole and (methyl)oxazole heterocycles in an otherwise linear core peptide. Identification of putative LAP clusters requires the presence of either the dehydrogenase McbB and one of the potentially fused cyclodehydratases McbC or McbD, or the goadsporin enzyme GodG. Separate hidden Markov models were developed for goadporin, microchip B17, streptolysin, and plantazolicin-type precursor sequences. When no precursor is identified, a heuristic strategy is adapted for precursor identification (20). Open reading frames between 30-70 amino acids which contain a sequence of 7 amino acids of which all 7 are serine, threonine, or cysteine, or a sequence of 8 amino acids of which at least 7 are serine, threonine, or cysteine, are considered potential LAP precursors. Due to the low homology of known LAP precursors, we identified four separate leader peptide motifs for each, with predicted cleavage occurring three residues after the end of the goadsporin-type motif, five residues after the end of the microcin B17-type motif, one residue after the end of the streptolysin-type motif, and five residues before the end of the plantazolicin-type motif. Dehydratases (McbB) and cyclodehydratases (McbC and McbD) are assumed to react at identical substrates to analogous enzymes in lantipeptides, with the exception that a maximum of four serine or threonine residues are left uncyclized in LAPs. All cysteine residues are assumed to undergo heterocyclization. In homologs of the goadsporin cluster, both GodF and GodG are required for dehydroalanine formation (21), while GodD and GodF are arbitrarily associated with cyclodehydration and oxidation, respectively, in azole formation (22). Finally, homologs of the GodH acetyltransferase catalyze N-terminal acetylation, while homologs of the plantazolicin methyltransferase PznL catalyze N,N-dimethylation of the N-terminus.

Microviridins. Microviridins are a family of N-acetylated cyanobacterial peptides containing intramolecular amide and ester bonds, which are installed by the ATP-grasp enzymes MdnB and MdnC respectively (23). Putative microviridin clusters are identified if they contain both a precursor peptide and at least one ATP-grasp enzyme. A conserved motif was identified within the core peptide, with N-terminal cleavage predicted to occur three residues before its start. If the C-terminus of the peptide is more than one residue from the end of the open reading frame, C-terminal cleavage is also predicted to occur. The ATP-grasp enzyme MdnB is assumed to catalyze amide bond formation between lysine residues and aspartate or glutamate side-chain carboxylic acids, while MdnC is assumed to catalyze ester bond formation between any subset of two serine and threonine residues and the side chain carboxylic acids of any subset of two aspartate and glutamate residues. Finally, the acetyltransferase MdnD is assumed to catalyze acetylation of the N-terminus.

Proteusins. Polytheonamides are the lone member of the proteusin family of RiPPs, a family of large D-amino acid-containing peptides which form unimolecular membrane channels with femtomolar affinity (24). Their biosynthesis is notable for the installation of 48 post-translational modifications by six enzymes. Putative proteusin clusters are identified by the presence of a precursor and the PoyD epimerase. A conserved motif was identified within the leader peptide of putative proteusins, with cleavage predicted to occur immediately after this motif. Polytheonamide biosynthesis involves a set of iterative enzymes with a high degree of substrate specificity. However, the absence of other members of the proteusin class limits the specificity with which the activities of homologs of these enzymes can be predicted within PRISM. Thus, for instance, the SAM-dependent methyltransferase PoyE is assumed to catalyze side chain N-methylation of 6 to 9 asparagine residues, while either of the closely related radical SAM proteins PoyB and PoyC is assumed to catalyze β-methylation of 10 to 15 isoleucine, valine, threonine, glutamine, or methionine carbons. The Fe(II)/α-ketoglutarate oxidoreductase PoyI is assumed to catalyze β-hydroxylation of 3 to 5 asparagine and valine residues. PoyF, which has homology to the dehydratase domain of LanM lantibiotic synthetases, is assumed to catalyze dehydration of the polytheonamide N-terminal threonine, with subsequent leader peptide cleavage causing tautomerization and α-ketone formation. It is assumed to react at a N-terminal serine or threonine residue. The activity of the proteusin epimerase PoyD is not predicted because PRISM does not generate chiral structure predictions.

Sactipeptides. Sactipeptides are a family of peptides defined by the presence of one or more bonds between a cysteine sulfur and the α-carbon of another residue. Identification of putative sactipeptide clusters requires the presence of a sactipeptide precursor and a homolog of the subtilosin radical SAM enzyme AlbA. We identified five conserved motifs. Precursors with an SkfA-type motif in the leader peptide have predicted cleavage one residue before the end of the motif. Precursors with an SboA-type motif spanning the leader and core peptide have predicted cleavage five residues after the start of the motif. Precursors with a thuricin-type motif have predicted cleavage two residues before the end of the leader peptide motif. Precursors with a thurincin-type motif have predicted cleavage nine residues after the start of a leader and core peptide motif. Finally, cleavage for a clade of putative sactipeptide precursors with high homology to known precursors is predicted to occur seven residues before the end of the leader peptide motif. Predicting the reaction sites of the radical SAM protein AlbA represented a computational challenge, as its specificity is poorly understood. A recursive algorithm was implemented within PRISM to identify all potential sites of sulfhydryl to α-carbon bond formation such that the final molecule has the characteristic hairpin secondary structure of sactipeptides, at least two residues separate each cysteine-alpha carbon bond, and the final cysteine-alpha carbon bond is at least two residues from the end. When the cluster contains a homolog of the sporulation killing factor thioredoxin SkfH, only combinations with a single cysteine are considered, and the remaining two cysteines are linked by a disulfide bond (25). Otherwise, all combinations of three cysteines are identified. If present, homologs of either the sporulation killing factor membrane-bound protease SkfC (25) or the split zinc-dependent protease AlbE/AlbF (26) are assumed to catalyze head-to-tail macrocyclization.

Streptide. Streptide is the founding member of a unique class of macrocyclic peptides characterized by the covalent linkage of inactivated carbons from the side chains of lysine and tryptophan residues (27). Identification of putative streptide clusters requires the presence of the streptide precursor (StrA), a SPASM-domain-containing radical SAM protein (StrB), and the streptide transporter (StrC). Conserved N- and C-terminal motifs were identified based on a published alignment of streptide precursors (27). When only the N-terminal motif is identified, cleavage is predicted to occur 13 residues after its start. When the C-terminal motif is also identified, and cleavage one residue after its start produces a peptide of at least five amino acids, C-terminal cleavage is also predicted to occur. The SPASM-domain-containing radical SAM protein StrB catalyzes cross-linking between the β-carbon of a lysine residue and the C7 position of a tryptophan residue.

Thiopeptides. Thiopeptides are a large and complex family of RiPPs characterized by the presence of multiple thiazoles and a central six-membered nitrogenous ring (pyridine, dehydropiperidine, piperidine). Thiopeptides of the nosiheptide and thiostrepton families are further derivatized by the addition of an indolic acid or quinaldic acid moiety to form a second macrocycle. Putative thiopeptide clusters are identified by the presence of a dehydratase and a cycloaddition enzyme. Precursors are identified by a hidden Markov model, or by a heuristic strategy when a thiopeptide cluster is identified without any hits to the precursor hidden Markov model. This heuristic considers open reading frames of length 30-90 amino acids as potential thiopeptide precursors when they contain a sequence of 12 to 20 amino acids composed of at least 40% serines, threonines, and cysteines, and containing at least two serines.

Five conserved motifs were identified to predict leader peptide cleavage. Two conserved motifs corresponded to large clades of thiopeptides, with predicted cleavage three residues before the end of the first and four residues after the start of the second. More specific motifs were identified for thiostrepton and siomycin-type thiopeptides, with cleavage 32 residues after the start of the motif, and for nosiheptide and nocathiacin-type thiopeptides, with cleavage 33 residues after the start of the motif. For thiopeptides with C-terminal cleavage, such as thiomuracin and GE2270, N-terminal cleavage is predicted to occur 29 residues after the start of the motif, with C-terminal cleavage three residues after the end of the motif.

The biosynthesis of thiopeptide core scaffolds is characterized by a diverse set of conserved post-translational modifications, including dehydration of serine and threonine residues, heterocyclization and oxidation of serine, threonine, and cysteine residues, and cycloaddition of two dehydroalanine residues to form a six-membered nitrogenous ring. Using nomenclature from the simplest cluster, lactazole (28), we developed Hidden Markov models for core thiopeptide biosynthesis enzymes LazB, LazC, LazE, and LazF. As with the lantipeptide dehydratase LanB, homologs of the thiopeptide dehydratase LazB are assumed to catalyze the dehydration of any subset of serine and threonine residues to dehydroalanine and dehydrobutyric acid, respectively, with tautomerization of N-terminal dehydrated residues to pyruvate and 2-oxobutyrate. When at least two serine or threonine residues are present in the cleaved precursor peptide, a minimum of two dehydrations are assumed to occur. LazC is assumed to catalyze pyridine formation between an N-terminal dehydroalanine and any other dehydroalanine residue resulting in formation of a macrocycle of size 8 to 13 amino acids. Substituents at positions 1 and 2 are assumed to be fully heterocyclized and oxidized (i.e., to form thiazoles, oxazoles, or methyloxazoles) when possible. A separate hidden Markov model was developed for the distinct thiostrepton-type LazC enzyme, which is assumed to catalyze dehydropiperidine formation (29) between any pair of dehydroalanine residues resulting in 8 to 13 amino acid macrocycle formation. The cyclodehydratase LazE is assumed to catalyze heterocyclization of any subset of serine, threonine, and cysteine residues, and the oxidase LazF is assumed to catalyze azole formation at any overlapping subset of azolines.

A number of hidden Markov models were also constructed for tailoring enzymes specific to individual families of thiopeptides. Homologs of the berninamycin cytochrome P450 BerH catalyze β-hydroxylation of a valine residue (30). Homologs of the cyclothiazomycin protein CltM catalyze tertiary thioether formation between a cysteine residue and a dehydroalanine residue (31). Homologs of the GE37468 cytochrome P450 GetJ catalyze the conversion of an isoleucine residue to β-methyl-δ-hydroxyproline (32). NocQ is a putative SAM-dependent methyltransferase from the nocathiacin cluster proposed to catalyze the methylation of a hydroxylated dehydrothreonine residue; since the enzyme responsible for dehydrothreonine hydroxylation is not known, homologs of this unique enzyme are assumed to catalyze both reactions (33). The nosiheptide protein NosA is a cofactor-independent enzyme whose homologs catalyze C-terminal amine formation by dealkylation of a terminal dehydroalanine residue (34). Homologs of the nosiheptide cytochrome P450s NosB and NosC catalyze γ-hydroxylation of a glutamate residue and C-5 hydroxylation of a pyridine moiety, respectively (35). Hidden Markov models were constructed for four enzymes proposed to be involved in the biosynthesis of the (NosI, NosK, NosL, and NosN). However, because the exact function of each enzyme is unclear (35), the presence of the NosI acyl-CoA ligase alone is assumed to catalyze the esterification of the modified indolic acid observed in the nosiheptide structure. The carboxylic acid is assumed to react with a free cysteine or serine residue, while the hydroxyl group is assumed to react with the side chain carboxylic acid of a glutamate or aspartate residue. Homologs of the thiocillin nonheme iron-dependent hydroxylase TclD catalyze β-hydroxylation of a valine residue, while homologs of the SAM-dependent methyltransferase TclO catalyze O-methylation of a threonine residue (36). Homologs of the 4-hydroxybutyrate dehydrogenase TpaJ, from the TP-1161 cluster, are assumed to catalyze conversion of a C-terminal threonine residue to an amino acetone group (37). TpdJ1 and TpdJ2 are cytochrome P450 monooxygenases from the thiomuracin cluster, which contains two cryptic oxidations (phenylalanine β-hydroxylation and isoleucine epoxidation) (38). Since it is not known which cytochrome P450 catalyzes which oxidation, homologs of either of these two enzymes are assumed to catalyze either of the two possible reactions. Homologs of the GE2270 cytochrome P450 TpdQ catalyze β-hydroxylation of a phenylalanine residue (39). Homologs of the thiomuracin radical SAM protein TpdI and the GE2270 radical SAM protein TpdL catalyze thiazole 5-methylation, while the GE2270 radical SAM protein TpdM is assumed to hydroxymethylation of a methylthiazole moiety (38). Homologs of the GE2270 N-methyltransferase TpdT catalyze side chain N-methylation of an asparagine residue (38). Homologs of the thiostrepton amidotransferase protein TsrC catalyze the conversion of a C-terminal carboxylic acid to an amide, while homologs of the cytochrome P450 TsrR catalyze the oxidation of an isoleucine residue to form β- and γ-hydroxyl groups (29). As in the nosiheptide cluster, hidden Markov models were constructed for seven proteins implicated in the biosynthesis of the modified quinaldic acid moiety found in thiostrepton (TsrA, TsrB, TsrD, TsrE, TsrI, TsrT, and TsrU), but the presence of the TsrI esterase alone is assumed within PRISM to catalyze esterification of the quinaldate carboxylic acid at a free serine and the attachment of C-7 to the precursor peptide N-terminal amine (40).

Thioviridamide. Thioviridamide is a ribosomally synthesized natural product notable for its distinct repertoire of post-translational modifications, including the installation of five thioamide bonds and a unique serine-derived N-terminal acyl group. Identification of putative thioviridamide clusters requires the presence of the thioviridamide precursor and the putative thioamide-forming enzyme TvaH. Leader cleavage is predicted to occur immediately after the conserved N-terminal motif. Within PRISM, homologs of the TvaF decarboxylase catalyze aminovinylcysteine formation between a C-terminal cysteine and a serine or threonine residue; homologs of the TvaJ oxygenase catalyze histidine β-hydroxylation; and homologs of the TvaG methyltransferase catalyze histidine N1,N3-dimethylation. The candidate thioamide-forming enzyme TvaH is assumed to catalyze thioamide between the second and sixth residue. One of TvaC, TvaD, or TvaE is proposed to accomplish the transformation of the N-terminal serine residue to the unique acyl unit found in the structure of thioviridamide, with the presence of all three required to execute the virtual reaction.

Trifolitoxin. Trifolitoxin is a unique RiPP with a UV-absorbing chromophore and potent antibiotic activity (41). The presence of the trifolitoxin biosynthesis enzymes TfxB and TfxC is sufficient to define the presence of a cluster, and to catalyze the conversion of a xQGC tetrapeptide within the cleaved precursor (TfxA) to the putative trifolitoxin chromophore. Cleavage is predicted to occur five residues before the start of the conserved motif in the core peptide.

YM-216391. YM-216931 is a cyclic RiPP characterized by a polyoxazole-thiazole moiety (2). Its biosynthesis is notable for a unique mechanism of heterocyclization (42). Putative YM-216391 family clusters are identified if they contain a precursor peptide and the distinct YmF macrocyclase. The identification of both N- and C-terminal conserved motifs is required to execute leader peptide cleavage, which is predicted to occur immediately after the end of the N-terminal leader peptide motif and eight residues before the start of the C-terminal motif. Within PRISM, cyclodehydration and oxidation of serine, threonine, and cysteine residues to form azolines and azoles is assumed to be catalyzed by the N- and C-terminal domains of YmBC, respectively. All possible cyclodehydrations are assumed to occur, but any subset of heterocycle oxidations is permitted. Homologs of the putative protease YmF catalyze head-to-tail macrocyclization. Homologs of the cytochrome P450 YmE catalyze β-hydroxylation of a phenylalanine residue. Finally, homologs of the YmB1 and YmC1 heterocyclase and oxidase catalyze formation of the unique phenyloxazole moiety from β-hydroxyphenylalanine.

Comparative analysis of RiPP cluster identification. ClusterFinder source code was accessed from http://github.com/petercim/ClusterFinder (last updated November 26, 2013). Version 30.0 of Pfam-A and version 2.6.3 of Prodigal were used to satisfy ClusterFinder dependencies. antiSMASH version 3.0.5 was obtained from http://bitbucket.org/antismash/antismash. glimmer 3.0.2, GlimmerHMM 3.0.2, hmmer 2.3.2 and 3.1b1, fasttree 2.1.7, diamond 0.7.11, muscle 3.8.31, Prodigal 2.6.1, and NCBI blast+ 2.2.30 were installed to satisfy antiSMASH 3 dependencies. All families of RiPPs detected by antiSMASH 3 (lantipeptides, thiopeptides, linaridins, cyanobactins, glycocins, linear azol(in)e-containing peptides, lasso peptides, sactipeptides, bottromycins, bacterial head-to-tail cyclized peptides, microviridins, and proteusins) were included in the comparative analysis. Non-posttranslationally modified bacteriocins and microcins were not considered in the comparison. Because ClusterFinder relies on antiSMASH to annotate biosynthetic gene cluster types, we were unable to evaluate the ability of ClusterFinder to identify RiPP clusters not detected by either RiPP-PRISM or antiSMASH without extensive manual annotation. Consequently, our analysis of ClusterFinder was restricted to its ability to identify RiPP clusters also identified by RiPP-PRISM. In the interest of providing the most stringent possible comparison for RiPP-PRISM, we considered all clusters detected by ClusterFinder, regardless of whether they were also classified as RiPPs by antiSMASH.

Structure confirmation of aurantizolicin. The structure of aurantizolicin that was predicted by PRISM and identified in LCMS chromatograms was highly similar to the known structures urukthapelstatin and YM-216391, whose NMR assignments had been made previously in d_6-DMSO. As aurantizolicin was produced in extremely limited quantities (60 L of optimized production culture provided <1 mg of aurantizolicin) the structure was confirmed from a mixed NMR sample. Chemical shifts of aurantizolicin detected from the NMR sample were nearly identical to reported shifts for urukthapelstatin and YM-216391.

^{1}H proton and ^{1}H-^{13}C HSQC experiments confirmed the presence of four well conserved azole protons that were nearly identical to those from urukthapelstatin, which possesses the same azole ring system (avg. ^{1}H Δppm of 0.01, avg. ^{13}C Δppm of 0.47). ^{1}H-^{13}C HMBC experiments from these protons revealed correlations to aromatic carbons that again closely matched those from the identical urukthapelstatin ring system, as well as from the glycine carbonyl-derived carbon of the first oxazole of YM-216391 (avg. ^{13}C Δppm of 0.09).

Chemical shifts for the phenyloxazole which had been indicated in the structure prediction and isotope incorporation data were most highly similar to those observed in urukthapelstatin. ^{1}H proton signal integration and coupling constants, ^{1}H-^{1}H COSY and TOCSY experiments, and ^{1}H-^{13}C HSQC experiments confirmed the presence of the aromatic phenyl ring, with signals nearly identical to those reported for urukthapelstatin (avg. ^{1}H Δppm of 0.01, avg. ^{13}C Δppm of 0.19). ^{1}H-^{13}C HMBC experiments confirmed HSQC correlations and provided correlations to aromatic carbon 1 and to the β-carbon found in the oxazole.

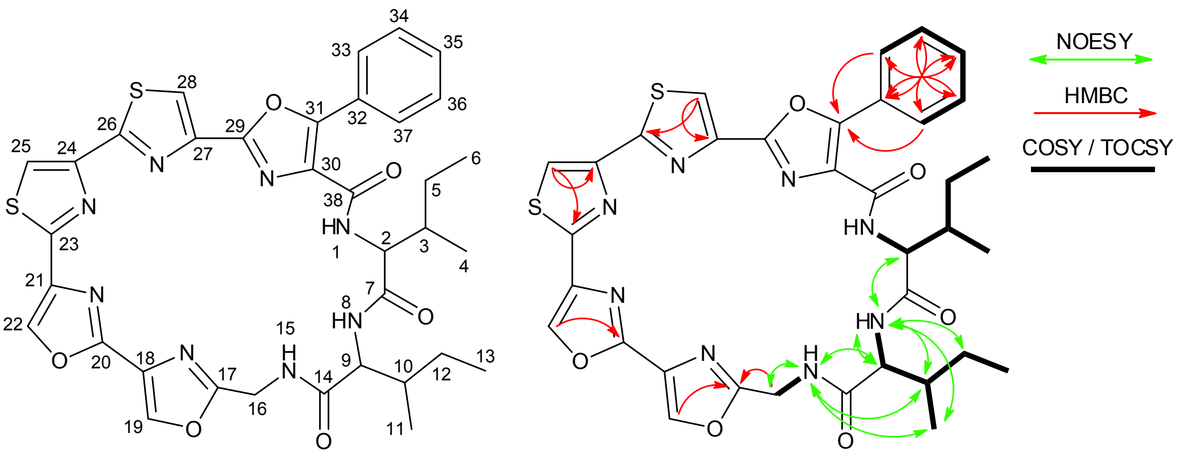
^{1}H-^{1}H 2D-TOCSY, COSY, and NOESY experiments revealed three systems related to the predicted amino acids. Glycine was identified via an NH doublet signal at 8.80 ppm, which was found to possess correlations to a pair of CH_2 protons at 5.02 ppm and 4.17 ppm. These protons shared a carbon signal at 34.85 ppm and possessed a ^{1}H-^{13}C HMBC correlation to the adjacent aromatic oxazole carbon. Chemical shifts for this glycine were nearly identical to those reported in the YM-216391 structure (avg. ^{1}H Δppm of 0.06, ^{13}C Δppm of 0.35). The isoleucine adjacent to the phenyloxazole was identified via an NH doublet signal at 8.13 ppm which possessed a correlation to a doublet-of-doublets at 4.82 ppm. This putative alpha-proton signal had correlations to another proton at 1.98 ppm, which was then connected to a fourth proton at 0.91 ppm. ^{1}H-^{1}H 1D TOCSY experiments irradiating either the distinct amide NH at 8.13 ppm or the alpha proton at 4.82 ppm were used to reveal the entire isoleucine spin system, including CH_2 protons at 1.62 ppm and 1.06 ppm, as well as the terminal CH_3 at 0.87 ppm. ^{1}H-^{13}C HSQC experiments were used to reveal associated carbon chemical shifts, although CH_3 signals could not be definitively assigned due to extensive overlap. Chemical shifts for this isoleucine were nearly identical to those reported in the YM-216391 structure (avg. ^{1}H Δppm of 0.05, ^{13}C Δppm of 0.46). The unique aurantizolicin isoleucine (replacing the valine observed in YM-216391 and alanine observed in urukthapelstatin) was identified via an NH doublet signal at 8.60 ppm that possessed a correlation to a triplet at 4.39 ppm. This alpha proton possessed a correlation to a CH proton at 1.88 ppm, which was then connected to a CH_3 at 0.90 ppm. A ^{1}H-^{1}H NOESY experiment revealed CH_2 protons at 1.51 ppm and 1.25 ppm, and confirmed the order of the observed amino acids. ^{1}A ^{1}H-^{1}H 1D TOCSY experiment irradiating the distinct amide NH at 8.60 ppm was used to reveal the entire isoleucine spin system, including the terminal CH_3 at 0.87 ppm. ^{1}H-^{13}C HSQC experiments were used to reveal associated carbon chemical shifts, although carbon signals corresponding to the CH_3 groups could not be definitively assigned due to extensive overlap.

High-resolution mass spectrometry of aurantizolicin

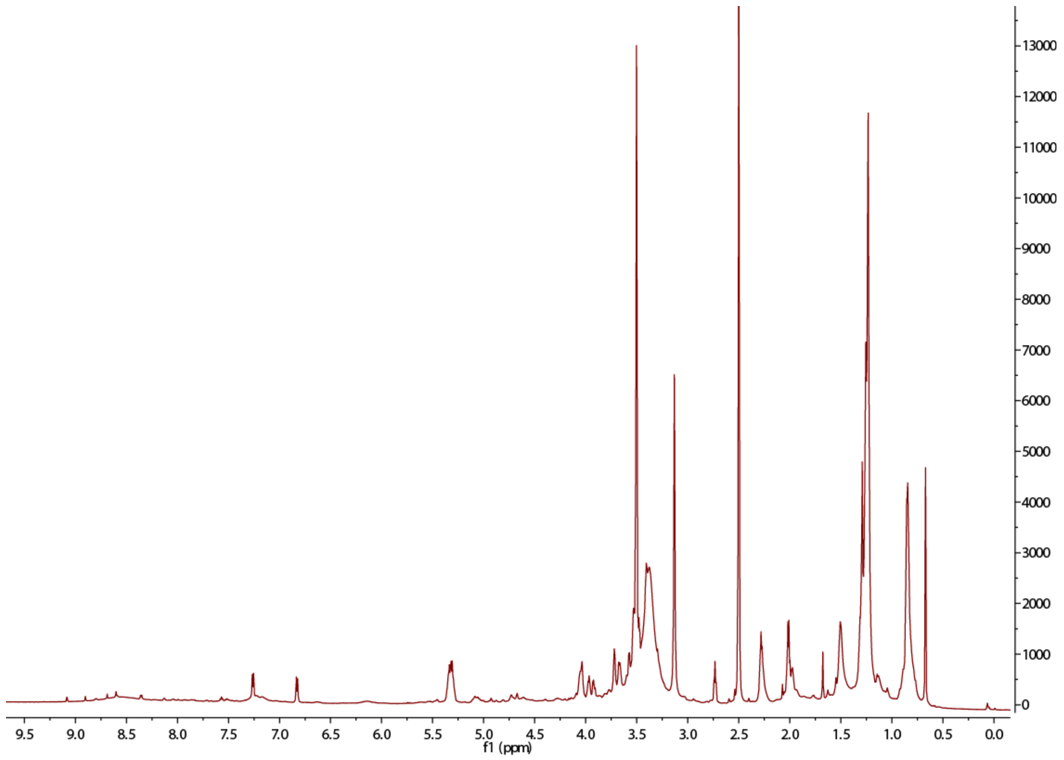
Compound	**Formula**	**Calc.**	**Obs.**	**Δppm**
Aurantizolicin	$C_{35}H_{35}N_8O_6S_2$ [M+H]	727.21155	727.21119	0.495

NMR spectroscopic data for aurantizolicin (700 MHz, DMSO-d_6)

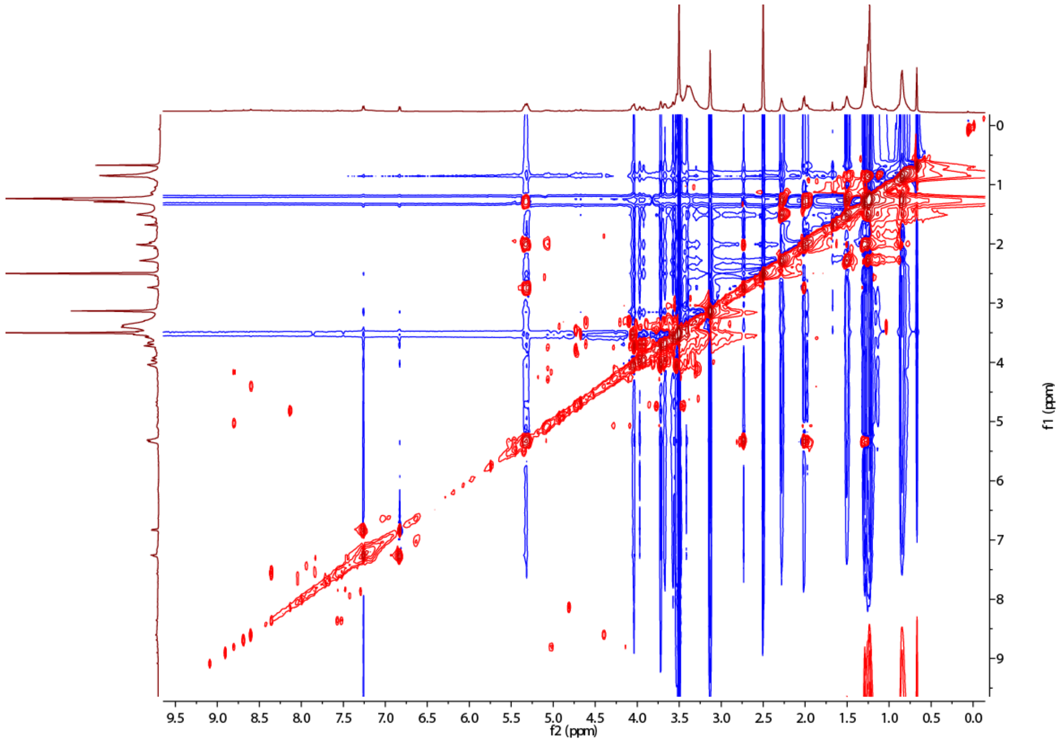
Position	δH (mult.)	δC	Position	δH (mult.)	δC
1	8.13 (*d*, 7)	-	20	-	155.30
2	4.82 (*m*)	56.11	21	-	-
3	1.98 (*ov.*)	38.66	22	9.08 (*s*)	139.00
4	0.91 (*ov.*)	-	23	-	157.06
5	1.62, 1.05 (*ov.*)	25.34	24	-	147.79
6	0.87 (*ov.*)	-	25	8.60 (*s*)	120.70
7	-	-	26	-	161.11
8	8.60 (*d*, 7)	-	27	-	141.93
9	4.39 (*t*, 7, 14)	56.88	28	8.69 (*s*)	122.59
10	1.88 (*ov.*)	36.75	29	-	-
11	0.90 (*ov.*)	-	30	-	-
12	1.52, 1.27 (*ov.*)	28.21	31	-	150.87
13	0.86 (*ov.*)	-	32	-	126.66
14	-	-	33	8.35 (*d*, 7)	127.37
15	8.80 (*d*, 7)	-	34	7.57 (*t*, 7, 14)	128.37
16	5.02, 4.17 (*m*)	34.85	35	7.51 (*t*, 7, 14)	129.84
17	-	162.78	36	7.57 (*t*, 7, 14)	128.37
18	-	-	37	8.35 (*d*, 7)	127.37
19	8.90 (*s*)	139.48	38	-	-



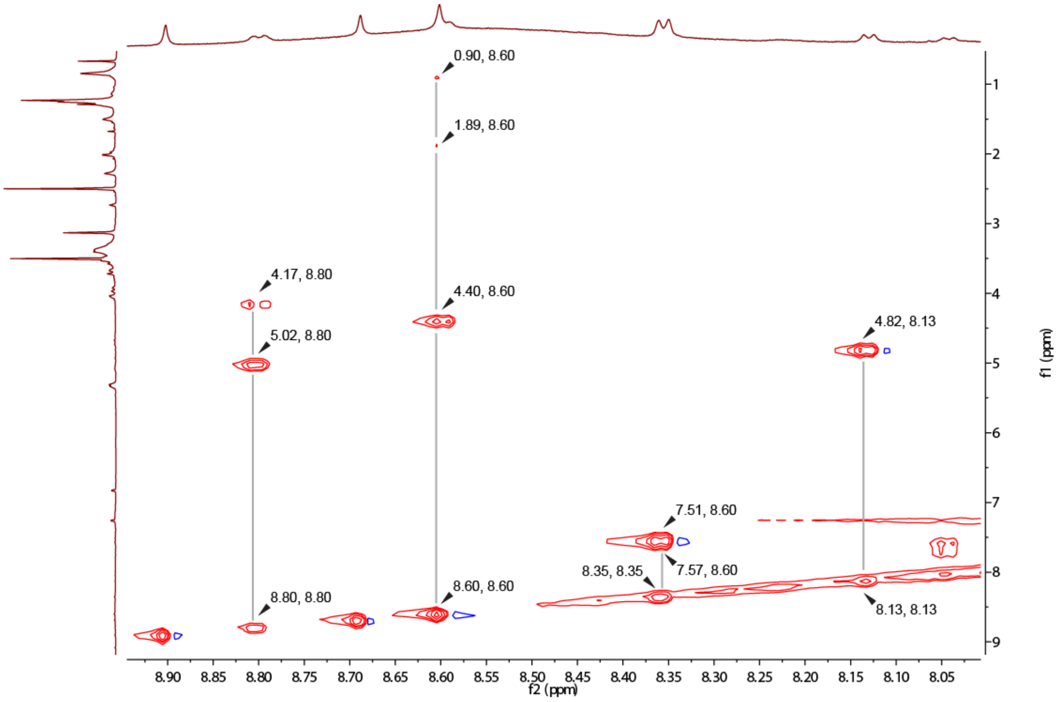
^{1}H NMR spectrum of aurantizolicin in DMSO.



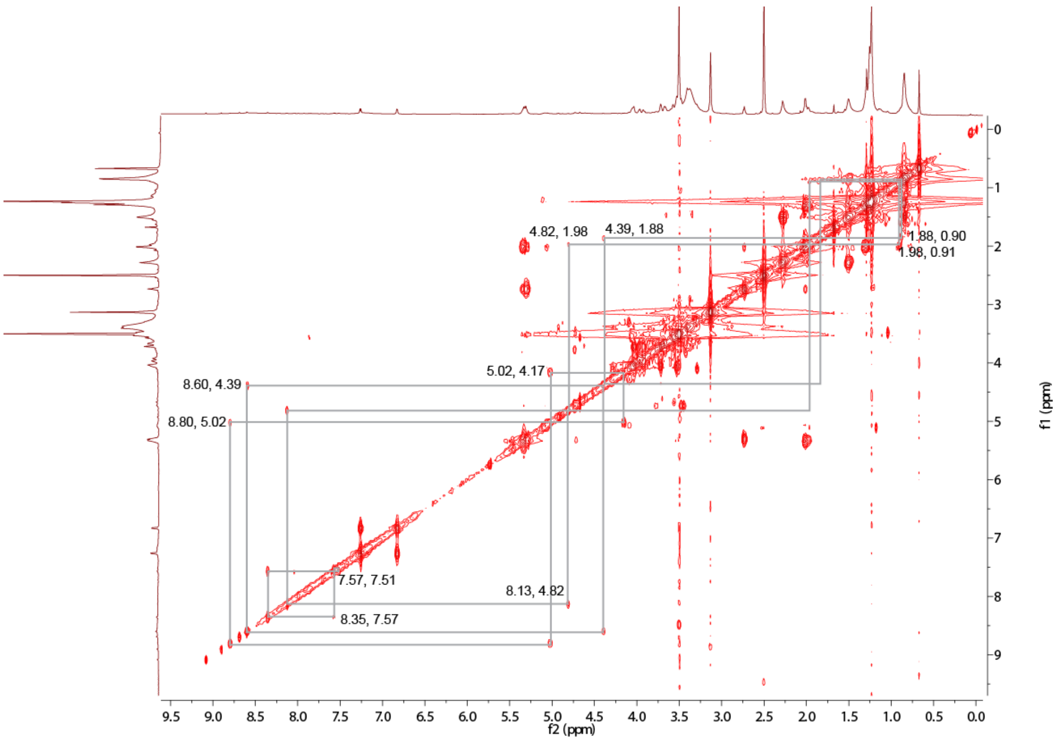
^{1}H-^{1}H TOCSY spectrum of aurantizolicin in DMSO.



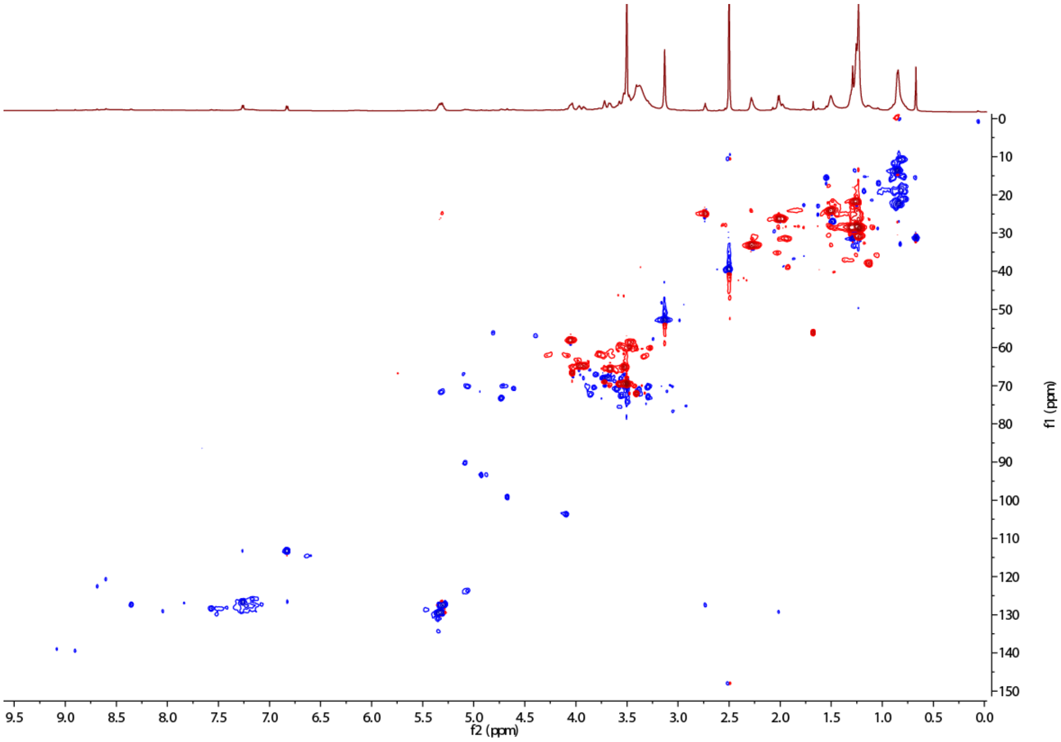
^{1}H-^{1}H TOCSY spectrum of aurantizolicin in DMSO.



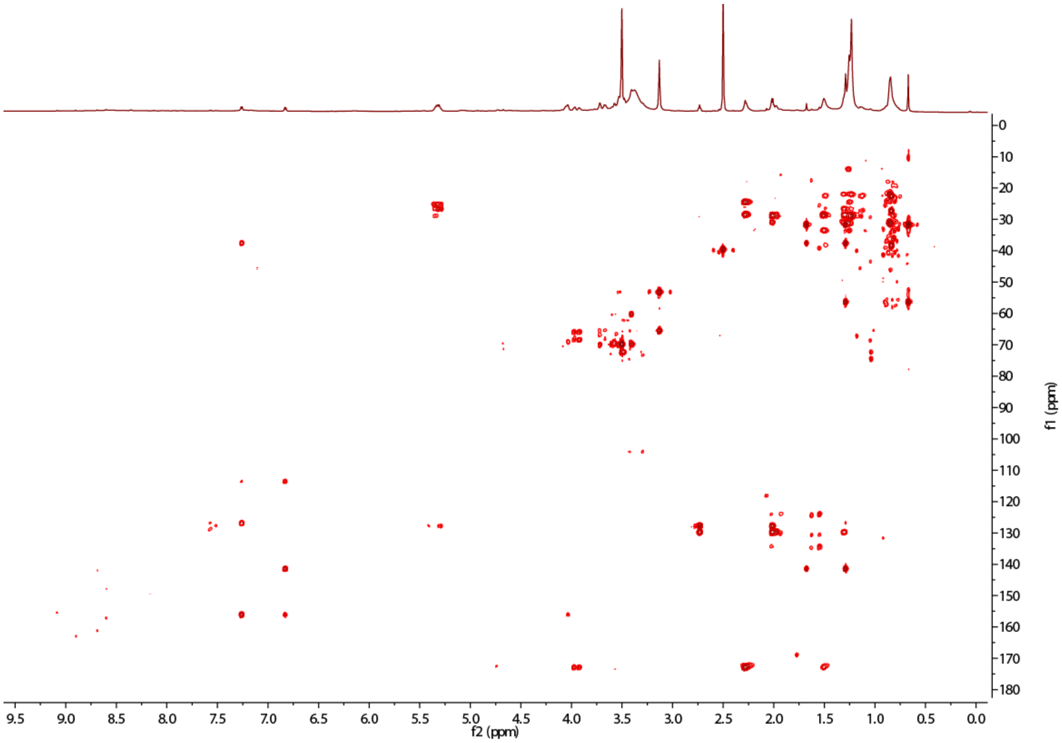
^{1}H-^{1}H COSY spectrum of aurantizolicin in DMSO.



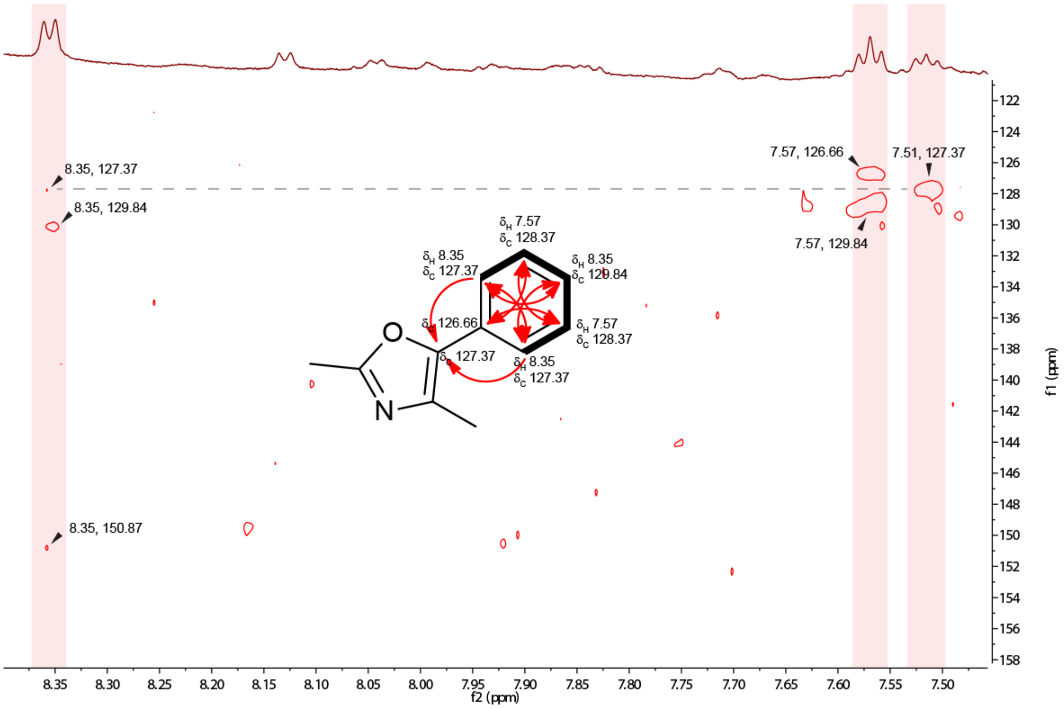
^1H-^{13}C HSQC spectrum of aurantizolicin in DMSO.



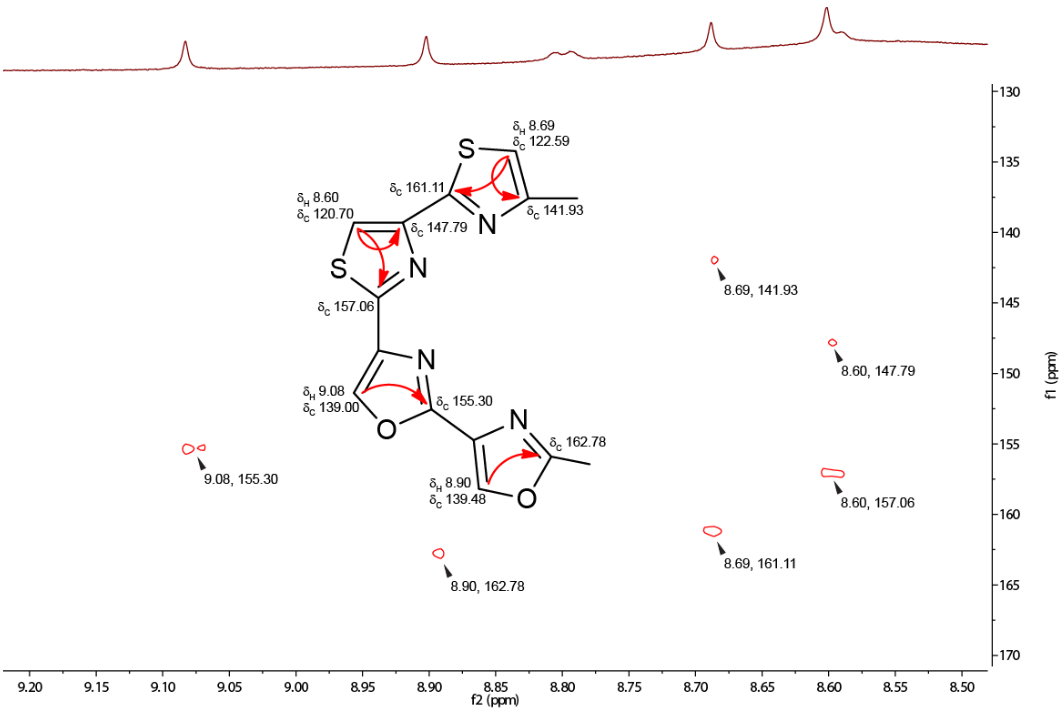
^1H-^{13}C HMBC spectrum of aurantizolicin in DMSO.



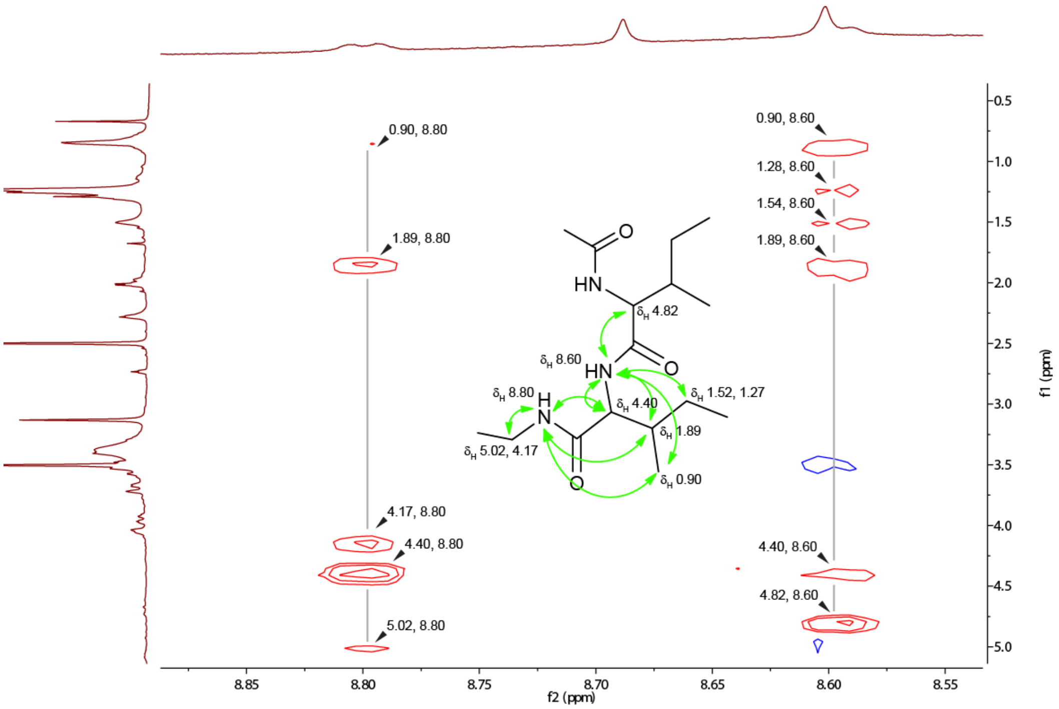
^{1}H-^{13}C HMBC spectrum of aurantizolicin in DMSO.



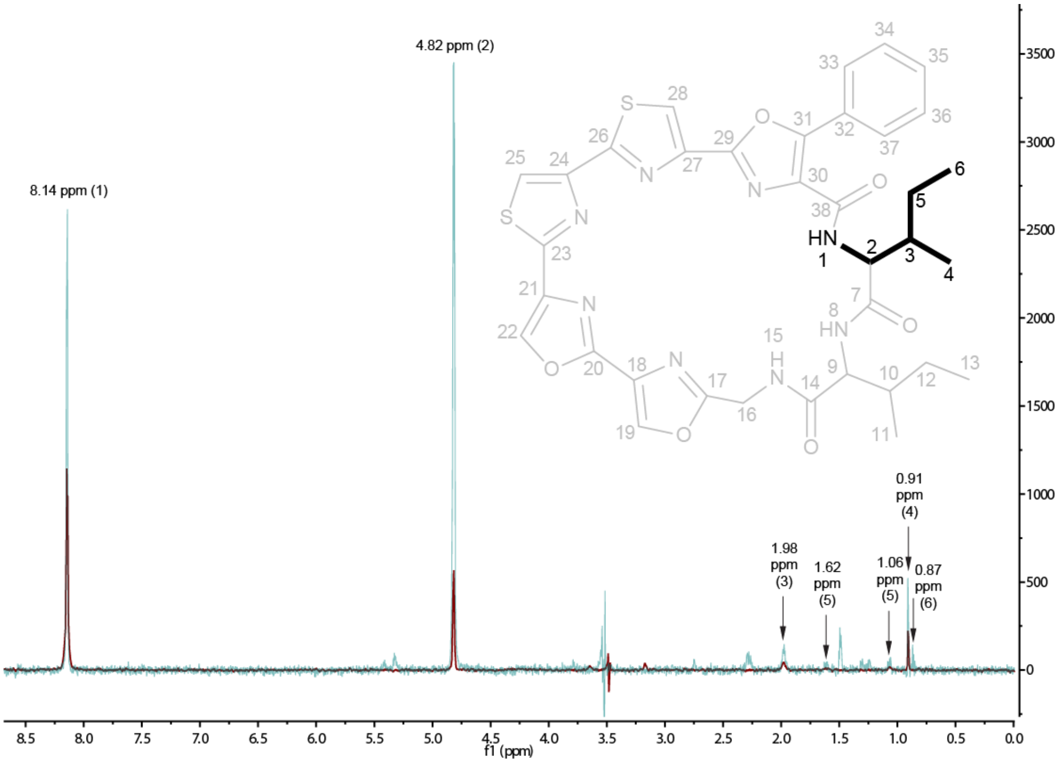
^{1}H-^{13}C HMBC spectrum of aurantizolicin in DMSO.



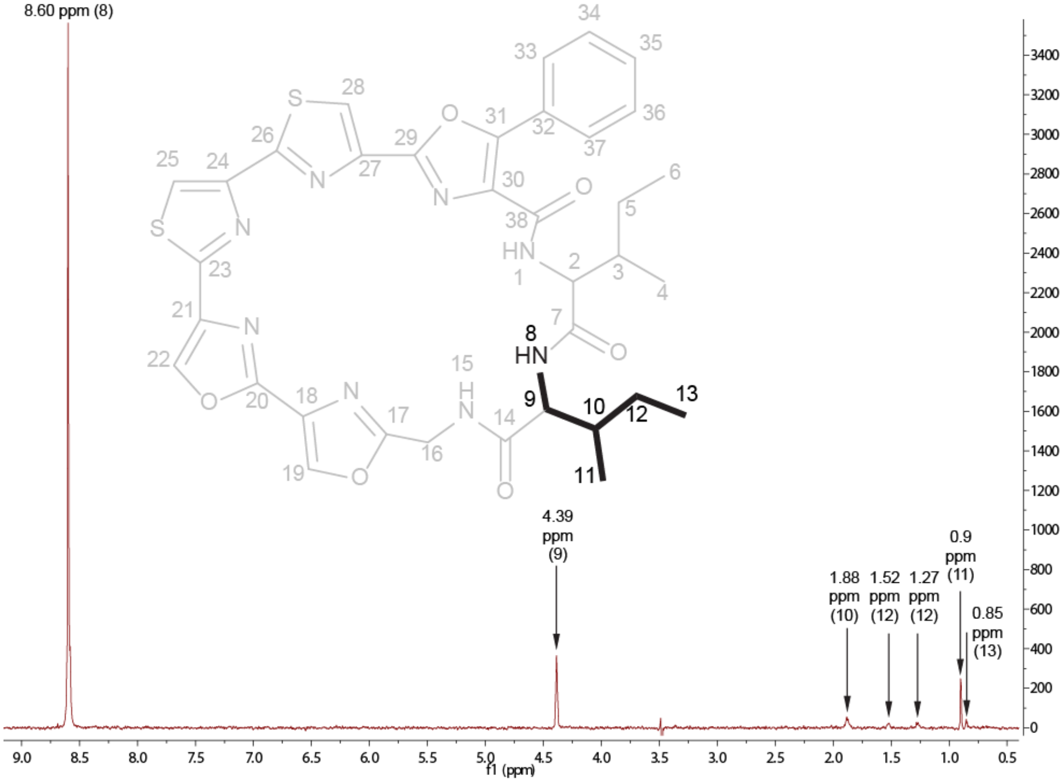
^1H-^1H NOESY spectrum of aurantizolicin in DMSO.



^1H-^1H 1D TOCSY spectrum of aurantizolicin in DMSO. Overlay of spectra from irradiation of 8.14 ppm and irradiation of 4.82 ppm.



^{1}H-^{1}H 1D TOCSY spectrum of aurantizolicin in DMSO. Irradiation of 8.60 ppm.



References

1. Wuster, A. and Babu, M.M. (2008) Conservation and evolutionary dynamics of the agr cell-to-cell communication system across firmicutes. J Bacteriol, 190, 743-746.
2. Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J. et al. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep, 30, 108-160.
3. van Belkum, M.J., Martin-Visscher, L.A. and Vederas, J.C. (2011) Structure and genetics of circular bacteriocins. Trends Microbiol, 19, 411-418.
4. Huo, L., Rachid, S., Stadler, M., Wenzel, S.C. and Muller, R. (2012) Synthetic biotechnology to study and engineer ribosomal bottromycin biosynthesis. Chem Biol, 19, 1278-1287.
5. Okada, M., Sato, I., Cho, S.J., Iwata, H., Nishio, T., Dubnau, D. and Sakagami, Y. (2005) Structure of the Bacillus subtilis quorum-sensing peptide pheromone ComX. Nat Chem Biol, 1, 23-24.
6. Bacon Schneider, K., Palmer, T.M. and Grossman, A.D. (2002) Characterization of comQ and comX, two genes required for production of ComX pheromone in Bacillus subtilis. J Bacteriol, 184, 410-419.
7. Leikoski, N., Liu, L., Jokela, J., Wahlsten, M., Gugger, M., Calteau, A., Permi, P., Kerfeld, C.A., Sivonen, K. and Fewer, D.P. (2013) Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides. Chem Biol, 20, 1033-1043.
8. Garcia De Gonzalo, C.V., Zhu, L., Oman, T.J. and van der Donk, W.A. (2014) NMR structure of the S-linked glycopeptide sublancin 168. ACS Chem Biol, 9, 796-801.
9. Cotter, P.D., O'Connor, P.M., Draper, L.A., Lawton, E.M., Deegan, L.H., Hill, C. and Ross, R.P. (2005) Posttranslational conversion of L-serines to D-alanines is vital for optimal production and activity of the lantibiotic lacticin 3147. Proc Natl Acad Sci U S A, 102, 18584-18589.
10. Ortega, M.A., Velasquez, J.E., Garg, N., Zhang, Q., Joyce, R.E., Nair, S.K. and van der Donk, W.A. (2014) Substrate specificity of the lanthipeptide peptidase ElxP and the oxidoreductase ElxO. ACS Chem Biol, 9, 1718-1725.
11. Foulston, L.C. and Bibb, M.J. (2010) Microbisporicin gene cluster reveals unusual features of lantibiotic biosynthesis in actinomycetes. Proc Natl Acad Sci U S A, 107, 13461-13466.
12. Huang, E. and Yousef, A.E. (2015) Biosynthesis of paenibacillin, a lantibiotic with N-terminal acetylation, by Paenibacillus polymyxa. Microbiol Res, 181, 15-21.
13. Okesli, A., Cooper, L.E., Fogle, E.J. and van der Donk, W.A. (2011) Nine post-translational modifications during the biosynthesis of cinnamycin. J Am Chem Soc, 133, 13753-13760.
14. Boakes, S., Cortes, J., Appleyard, A.N., Rudd, B.A. and Dawson, M.J. (2009) Organization of the genes encoding the biosynthesis of actagardine and engineering of a variant generation system. Mol Microbiol, 72, 1126-1136.
15. Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W. and van der Donk, W.A. (2010) Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. Proc Natl Acad Sci U S A, 107, 10430-10435.
16. Haft, D.H., Basu, M.K. and Mitchell, D.A. (2010) Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. BMC Biol, 8, 70.
17. Maksimov, M.O., Pelczer, I. and Link, A.J. (2012) Precursor-centric genome-mining approach for lasso peptide discovery. Proc Natl Acad Sci U S A, 109, 15223-15228.
18. Rateb, M.E., Zhai, Y., Ehrner, E., Rath, C.M., Wang, X., Tabudravu, J., Ebel, R., Bibb, M., Kyeremeh, K., Dorrestein, P.C. et al. (2015) Legonaridin, a new member of linaridin RiPP from a Ghanaian Streptomyces isolate. Org Biomol Chem, 13, 9585-9592.

19. Claesen, J. and Bibb, M. (2010) Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. Proc Natl Acad Sci U S A, 107, 16297-16302.

20. Cox, C.L., Doroghazi, J.R. and Mitchell, D.A. (2015) The genomic landscape of ribosomal peptides containing thiazole and oxazole heterocycles. BMC Genomics, 16, 778.

21. Ozaki, T., Kurokawa, Y., Hayashi, S., Oku, N., Asamizu, S., Igarashi, Y. and Onaka, H. (2016) Insights into the Biosynthesis of Dehydroalanines in Goadsporin. Chembiochem, 17, 218-223.

22. Onaka, H., Nakaho, M., Hayashi, K., Igarashi, Y. and Furumai, T. (2005) Cloning and characterization of the goadsporin biosynthetic gene cluster from Streptomyces sp. TP-A0584. Microbiology, 151, 3923-3933.

23. Ziemert, N., Ishida, K., Liaimer, A., Hertweck, C. and Dittmann, E. (2008) Ribosomal synthesis of tricyclic depsipeptides in bloom-forming cyanobacteria. Angew Chem Int Ed Engl, 47, 7756-7759.

24. Freeman, M.F., Gurgui, C., Helf, M.J., Morinaka, B.I., Uria, A.R., Oldham, N.J., Sahl, H.G., Matsunaga, S. and Piel, J. (2012) Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science, 338, 387-390.

25. Fluhe, L., Burghaus, O., Wieckowski, B.M., Giessen, T.W., Linne, U. and Marahiel, M.A. (2013) Two [4Fe-4S] clusters containing radical SAM enzyme SkfB catalyze thioether bond formation during the maturation of the sporulation killing factor. J Am Chem Soc, 135, 959-962.

26. Fluhe, L., Knappe, T.A., Gattner, M.J., Schafer, A., Burghaus, O., Linne, U. and Marahiel, M.A. (2012) The radical SAM enzyme AlbA catalyzes thioether bond formation in subtilosin A. Nat Chem Biol, 8, 350-357.

27. Schramma, K.R., Bushin, L.B. and Seyedsayamdost, M.R. (2015) Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. Nat Chem, 7, 431-437.

28. Hayashi, S., Ozaki, T., Asamizu, S., Ikeda, H., Omura, S., Oku, N., Igarashi, Y., Tomoda, H. and Onaka, H. (2014) Genome mining reveals a minimum gene set for the biosynthesis of 32-membered macrocyclic thiopeptides lactazoles. Chem Biol, 21, 679-688.

29. Liao, R., Duan, L., Lei, C., Pan, H., Ding, Y., Zhang, Q., Chen, D., Shen, B., Yu, Y. and Liu, W. (2009) Thiopeptide biosynthesis featuring ribosomally synthesized precursor peptides and conserved posttranslational modifications. Chem Biol, 16, 141-147.

30. Malcolmson, S.J., Young, T.S., Ruby, J.G., Skewes-Cox, P. and Walsh, C.T. (2013) The posttranslational modification cascade to the thiopeptide berninamycin generates linear forms and altered macrocyclic scaffolds. Proc Natl Acad Sci U S A, 110, 8483-8488.

31. Wang, J., Yu, Y., Tang, K., Liu, W., He, X., Huang, X. and Deng, Z. (2010) Identification and analysis of the biosynthetic gene cluster encoding the thiopeptide antibiotic cyclothiazomycin in Streptomyces hygroscopicus 10-22. Appl Environ Microbiol, 76, 2335-2344.

32. Young, T.S. and Walsh, C.T. (2011) Identification of the thiazolyl peptide GE37468 gene cluster from Streptomyces ATCC 55365 and heterologous expression in Streptomyces lividans. Proc Natl Acad Sci U S A, 108, 13053-13058.

33. Ding, Y., Yu, Y., Pan, H., Guo, H., Li, Y. and Liu, W. (2010) Moving posttranslational modifications forward to biosynthesize the glycosylated thiopeptide nocathiacin I in Nocardia sp. ATCC202099. Mol Biosyst, 6, 1180-1185.

34. Yu, Y., Guo, H., Zhang, Q., Duan, L., Ding, Y., Liao, R., Lei, C., Shen, B. and Liu, W. (2010) NosA catalyzing carboxyl-terminal amide formation in nosiheptide maturation via an enamine dealkylation on the serine-extended precursor peptide. J Am Chem Soc, 132, 16324-16326.

35. Yu, Y., Duan, L., Zhang, Q., Liao, R., Ding, Y., Pan, H., Wendt-Pienkowski, E., Tang, G., Shen, B. and Liu, W. (2009) Nosiheptide biosynthesis featuring a unique indole side ring formation on the characteristic thiopeptide framework. ACS Chem Biol, 4, 855-864.

36. Wieland Brown, L.C., Acker, M.G., Clardy, J., Walsh, C.T. and Fischbach, M.A. (2009) Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. Proc Natl Acad Sci U S A, 106, 2549-2553.
37. Engelhardt, K., Degnes, K.F. and Zotchev, S.B. (2010) Isolation and characterization of the gene cluster for biosynthesis of the thiopeptide antibiotic TP-1161. Appl Environ Microbiol, 76, 7093-7101.
38. Morris, R.P., Leeds, J.A., Naegeli, H.U., Oberer, L., Memmert, K., Weber, E., LaMarche, M.J., Parker, C.N., Burrer, N., Esterow, S. et al. (2009) Ribosomally synthesized thiopeptide antibiotics targeting elongation factor Tu. J Am Chem Soc, 131, 5946-5955.
39. Tocchetti, A., Maffioli, S., Iorio, M., Alt, S., Mazzei, E., Brunati, C., Sosio, M. and Donadio, S. (2013) Capturing linear intermediates and C-terminal variants during maturation of the thiopeptide GE2270. Chem Biol, 20, 1067-1077.
40. Duan, L., Wang, S., Liao, R. and Liu, W. (2012) Insights into quinaldic acid moiety formation in thiostrepton biosynthesis facilitating fluorinated thiopeptide generation. Chem Biol, 19, 443-448.
41. Breil, B.T., Ludden, P.W. and Triplett, E.W. (1993) DNA sequence and mutational analysis of genes involved in the production and resistance of the antibiotic peptide trifolitoxin. J Bacteriol, 175, 3693-3702.
42. Jian, X.H., Pan, H.X., Ning, T.T., Shi, Y.Y., Chen, Y.S., Li, Y., Zeng, X.W., Xu, J. and Tang, G.L. (2012) Analysis of YM-216391 biosynthetic gene cluster and improvement of the cyclopeptide production in a heterologous host. ACS Chem Biol, 7, 646-651.