Exploring the chemical space of post-translationally modified peptides in Streptomyces with machine learning

Kloosterman, A.M.

1

Omics-based strategies to discover novel classes of RiPP natural products

Alexander M. Kloosterman

Marnix H. Medema

Gilles P. van Wezel

The work described in this chapter is part of the following publication:

Kloosterman, et al., *Omics-based strategies to discover novel classes of RiPP natural products.* Curr. Opin. Biotechnol., 2020. **6**, p: 60-67.

1 The discovery of antibiotics and the rise of resistance

Every living organism has metabolic pathways to catabolize and produce organic molecules, collectively called natural products. A crude distinction made in 1891 by A. Kossel separates these molecules into two classes: those that are directly essential for growth, development and reproduction of the organism belong to primary metabolism, while the remainder belong to secondary metabolism [1]. Secondary metabolites are highly versatile and cover a large chemical space. Even closely related strains may produce a different repertoire of secondary metabolites [2-5]. While mostly not required for the growth of an organism in a pure culture, secondary metabolites are thought to confer advantages in the natural environment of the producing strains. They have been found to act as means of communication with other species, signal cellular differentiation and scavenge metal ions. Most importantly, many secondary metabolites act as competitive weapons against other species and can be used as anti-bacterial agents in the clinic [2].

Interest in secondary metabolites for medical applications arguably started in 1928, with the discovery of penicillin by Alexander Flemming [6]. Penicillin, first isolated from the fungus *Penicillium notatum*, has strong antimicrobial properties, making it a highly promising candidate for use as a medicine against bacterial infections. It would take another ten years before penicillin was isolated in a pure form and further investigated as a potential drug. When it appeared that the drug had low toxicity and proved suitable for human consumption, its use would become standard practice in the clinic [7].

The potential of secondary metabolites to be of practical use led to a surge of investigations towards the discovery of more of these molecules. Selman A. Waksman would follow up on the discovery of penicillin by screening soil samples for strains harboring biological activities. From these efforts, several more antibiotics were isolated, including actinomycin D, neomycin and streptomycin, all produced by filamentous bacteria belonging to the genus of Actinobacteria [8-10]. Similar screening methods were later executed on a much larger scale by the pharmaceutical industry. Between the 1940's and the 1970's, more than 20 different classes of antibiotics were discovered by high-

throughput screening (HTS) efforts, many of which are still being used in the clinic today [11]. Thousands of compounds with antibiotic activity are currently known, the majority of which were isolated from Actinobacteria [12]. By changing the parameters of the screening, additional secondary metabolites with different useful functions were identified, which could be used as insecticides, hypertension relievers or immunosuppressants [13]. Overall, this period is often considered a golden age of antibiotic discovery and biotechnology.

However, that golden age has since then been declining. A major problem is the rise of resistance against our current repertoire of antibiotics [14]. Although antibiotic resistance is not an unnatural phenomenon [14-16], it is generally accepted that the human over-, under- and misuse of antibiotics is the main cause for the spread of resistance. Genes conferring resistance, e.g. by encoding a copy of a household gene that is insensitive to the antibiotics, are thought to quickly swap between bacteria by horizontal gene transfer (HGT) events. As a result, infectious diseases involving multidrug-resistant (MDR) pathogenic strains, are one the rise, most notably those involving the six ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa* and *Enterobacter* species), and MDR *Mycobacterium tuberculosis* (TB) [17]. Several variants of TB have been identified which are extensively-drug-resistant (XDR) and even total-drug-resistant (TDR), setting us back to the time before antibiotics were discovered [18-20]. Resistance to antibiotics is considered a major threat to our health by the WHO, with casualty projections in 2050 exceeding those of any other disease [21].

At the same time, traditional methods for drug discovery have been on the decline. HTS approaches toward drug discovery suffer greatly from high rediscovery rates of known compounds, with massive screening efforts only finding a handful of potential leads [22-24]. These approaches have become unreliable investments for the pharmaceutical industry, further decreasing the overall output [25]. It appears that much of the low-hanging fruit that could be discovered by HTS has been found, and now, new methods are necessary to keep discovering novel antimicrobials.

1 Genome mining for antibiotics

The advent of next generation sequencing (NGS) in the past two decades has made sequencing of entire genomes feasible and this has given the discovery of natural products new momentum. Due to newly developed techniques like Illumina, Nanopore and PacBio sequencing, it is now cheaper than ever to analyze the genomes of the bacteria and fungi that produce natural products. Conveniently, the genes encoding the enzymes capable of synthesizing natural products are clustered together into biosynthetic gene clusters (BGCs). By using biosynthesis genes characterized from known BGCs as queries, new, homologous BGCs can quickly be identified [26]. The first study investigating genome sequences of *Streptomyces coelicolor*, a model organism of the prolific antibiotic producers, the streptomycetes, discovered more than 20 BGCs, while the strain was known to only produce three compounds [27]. It quickly became clear that only a fraction of identified BGCs could be linked to a natural product. At time of writing, more than a million BGCs can be detected in publicly available genomic and metagenomic sequences [28], while only about 2,000 compounds have been directly linked to a BGC [29]. The BGCs without a known product, called cryptic BGCs, represent a vast reservoir of novel natural products that could be clinically relevant.

The existence of large amounts of uncharacterized BGCs raises the question about why their products have not been previously detected. A reasonable explanation is that secondary metabolite production is a costly process for a microorganism. Production levels must therefore be finely tuned in order to gain a competitive edge while not wasting resources, likely as a response to chemical signals in the environment [30, 31]. In Actinobacteria, this finetuning is accomplished by extensive regulatory networks that govern natural product production [30, 32]. A large number of two-component regulators and as many as 60 different sigma factors per strain make an organism capable of responding to many signals and scenario's.

In an attempt to exploit this, strategies have been developed to activate silent BGCs work by giving cultures specific signals they might encounter in their natural environment. These strategies include the use of molecular elicitors in

high-throughput elicitor screening (HiTES), screening with various carbon sources, applying stress conditions like starvation, and co-culturing strains together [33, 34]. Besides the novel natural products that are isolated in these studies, they also provide insight into what role the secondary metabolite might fulfill, knowing under what conditions it is activated. Alternatively, silent BGCs can be activated by genetically refactoring the BGC. Genes and operons can be rearranged and promoters can be replaced with strong, constitutive promoters to find optimal production conditions [35]. These efforts are generally more labor-intensive than using general chemical signals, requiring a large amount of genetic engineering to isolate the product or products of a single BGC. However, they do allow one to target a single BGC at a time, rather than evoke a more general response from an organism with potentially unwanted side effects. Combination of both approaches are required at different stages of investigation to completely understand the roles and products of novel BGCs.

The tradeoff between confidence and novelty in genome mining

Natural products are divided into classes or families based on their chemical makeup and biochemical origins. These products are built up from primary metabolites, such as amino acids and acetyl- or malonyl moieties. Natural products belonging to the same class share a common biosynthetic logic and homologous enzymes carrying out the reactions. For example, two major classes, called the non-ribosomal peptides (NRPs) and the type I polyketides (PKs), are synthesized by assembly-line machineries [36]. These are large enzyme complexes, that can be divided into modular units, each of which attaches one precursor molecule to a growing chain. The diversity of the resulting secondary metabolites is achieved either by using a large variety of precursors, or by applying additional tailoring to the products, such as glycosylation or cyclization. Other classes, like terpenes and type II PKs, rely on other enzymes to convert precursors into the final product, but these enzymes are still encoded by conserved genes [37, 38].

The classification of natural products provides a suitable framework for the purpose of identifying BGCs from genomic information. For each class of natural products, a set of enzymes can usually be identified that determines a family of natural products, and which can be used as “bait” to identify more

BGCs of that family. Tools like antiSMASH [39] and PRISM [40] use rule-based identification of these BGCs by targeting specific genes conserved among specific natural product classes. The surrounding region of hits found is the scanned for genes encoding additional tailoring enzymes, transporters, regulators, and immunity proteins, thereby identifying the BGC.

While these methods for genome mining have identified large amounts of BGCs, only BGCs with some relation to previously characterized ones will be detected. Methods extrapolating from known BGCs may give high-confidence output, but depending on the exact ruleset used, the BGCs detected will lack novelty [26]. Given the large diversity of natural product classes, it is tempting to speculate about the existence of completely novel classes and chemical scaffolds. To identify BGCs of these classes using only bioinformatics is a difficult challenge, as there should be at least some criteria to identify them. One tool has been described primarily for this purpose, called ClusterFinder [41]. ClusterFinder uses two collections of protein domains: one with those that are frequently associated with BGCs, and one with those that are present in other parts of the genome. It then uses a Hidden Markov Model (HMM) to identify regions of the genome that are enriched in domains indicating BGCs. Combined with large comparative genomics, a novel class of BGCs was identified and characterized. In general, using less restricted search criteria will increase the amount of candidate BGCs found, such as was the case with ClusterFinder. Among these may be more novel BGCs, but at the same time the amount of false positives will increase. This tradeoff between confidence and novelty is one that everyone undertaking of genome mining must consider carefully [26].

RiPPs form a diverse group of peptidic natural products

The ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important and diverse group of natural products, produced by all three branches of life. The unifying theme among all RiPPs is their biosynthetic logic: a precursor gene is translated into a precursor peptide, usually no longer than 100 amino acids. The precursor peptide is then extensively modified by a set of RiPP Tailoring Enzymes (RTE). RTEs usually recognize the peptide by binding a recognition sequence located outside the region that is modified.

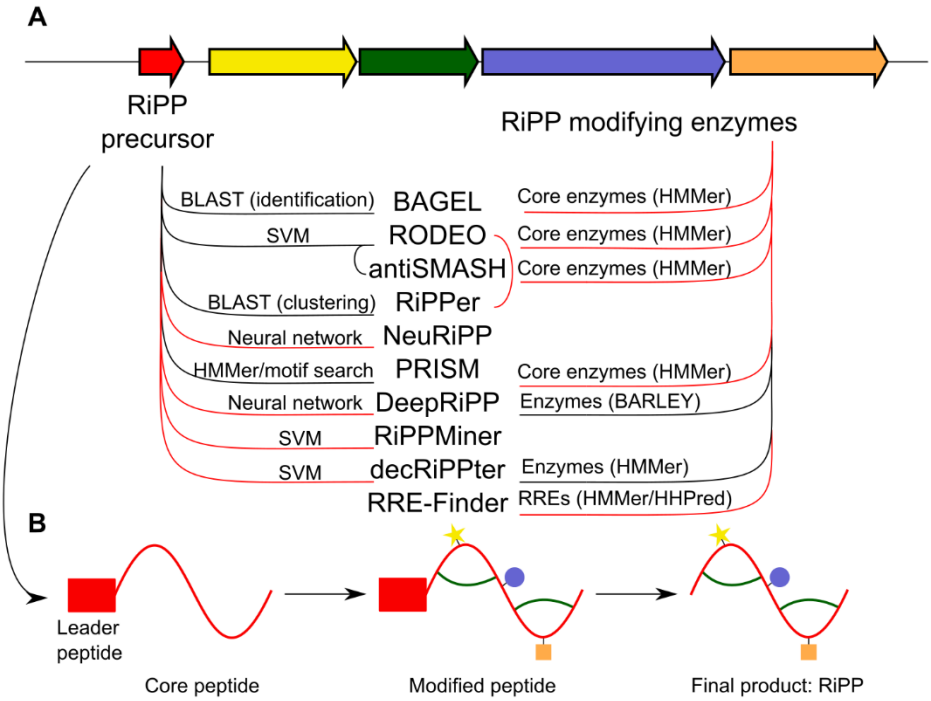

Figure 1. The conserved features of RiPP BGCs provides a framework for genome mining tools. A) A RiPP BGC typically has a gene for a precursor peptide (red) in an operon-like arrangement with genes for RTEs. These genes are used as targets for genome mining tools (see also Table 2). A red line indicates that the detection method is the primary detection method, while black lines indicate additional annotation. B) After translation, a precursor peptide is modified by RTEs, which use the leader peptide as a handle for peptide recognition. After modifications are applied, the leader peptide is cleaved off, resulting in the final product, the RiPP.

This recognition sequence is called the leader peptide if N-terminal, or follower peptide if C-terminal. Afterwards, leader and follower peptides are cleaved off, resulting in the mature RiPP [42-44] (Figure 1).

RiPPs are classified into subclasses or families, all of which share the biosynthetic logic with wildly different results depending on the precursor sequence and the enzymes involved. Every subclass of RiPPs typically has one characteristic modification. For example, all lasso peptides have at least a single crosslink, forming a small loop through which the amino acid chain is threaded [45]. Thiopeptides are all macrocyclized as the result of a 4+2 cycloaddition

between two dehydrated serine residues [46]. Lanthipeptides contain a thioether bridge between a non-C-terminal cysteine and a dehydrated serine or threonine residue [47]. Asides from core modifications, RiPPs may be tailored with several accessory modifications like disulfide bridges, acetylation, methylation and glycosylation. The combination of different precursor sequences and different possible combinations of modifications applied to them creates a wide diversity of different natural product (Figure 2).

As a rapidly expanding class of natural products, RiPPs represent an excellent candidate for further exploration to discover novel chemical scaffolds and antimicrobial leads. The last comprehensive review from 2013 [42] lists more than 20 different RiPP classes. Since then, several dozen RiPPs have been identified with modifications and genetic markers that set them apart from known RiPPs and thus form new RiPP families. An updated review published in 2020 expands this list to more than 40 different candidates [48] (Table 1). Given the number of novel RiPP classes discovered in the past decade, it is likely that many more exist. After all, it takes only a few evolutionary steps for a RiPP BGC to take shape. While the BGCs encoding the assembly lines of NRPS and PKS can be more than 100 kbp long, a RiPP BGC only needs to encode a small precursor and a single modifying enzyme for it to be a RiPP, and some are no larger than that [49]. RTEs could arise out of primary metabolism enzymes and protein modification enzymes with relatively few mutations and genetic arrangements. These BGCs would not need to be directly homologous to known RiPP classes, and would therefore be missed by the current methods of RiPP genome mining.

The diversity of RiPP BGCs reflects the diversity on the chemical level, and as such, there is no single method that effectively identifies all RiPP BGCs. Rather, new genome mining tools are still being developed and new approaches are still being experimented with. Uniquely, the small precursor genes provide a handhold suitable for all RiPP families. In the following section, we review the available tools and the way they differ in their approaches (Figure 1, Table 2), and highlight a few concentrated genome mining efforts has dramatically expanded the number of members of known RiPP families.

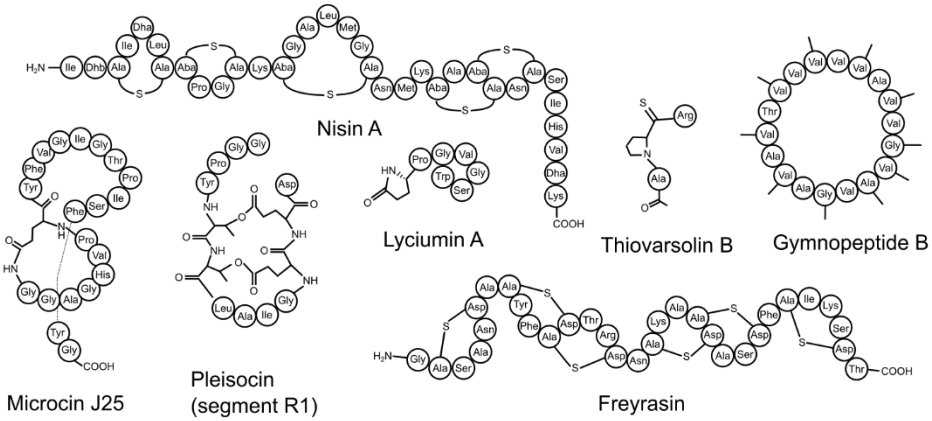

Figure 2. Examples of the rich chemical diversity of RiPPs. RiPP precursors are highly diverse in sequence and can be modified in many ways. Several old and new examples are shown here: nisin A (lanthipeptide) [50], lyciumin A (lyciumin) [51], thiovarsolin B (thioamitide) [52], gymnopeptide B (borosin) [51], microcin J25 (lasso peptide) [53], plesiocin segment R1 (omega-ester containing peptide/graspetide) [54] and freyrasin (ranthipeptide) [55].

Bioinformatic tools for homology-based genome mining of known RiPP families

A widely used approach for RiPP genome mining is to target the modifying enzymes encoded by the RiPP BGC. Most RiPP families contain conserved genes that encode the enzymes responsible the post-translational modification (PTM) characteristic of that RiPP family. The earliest methods of RiPP genome mining used simple BLAST or PSI-BLAST [56, 57] searches to identify homologs of genes encoding RTEs as a starting point for novel RiPP BGCs. These searches lead to the discovery of novel RiPPs with similar modifications as known RiPPs, like haloduracin, trichamide and capistruin [58-61]. BAGEL [62, 63] developed in 2006, was the first tool for automated genome mining of RiPPs. Since then, the more general algorithms antiSMASH [39] and PRISM [40, 64] have been developed, which allow genome mining for BGCs of any class of natural products, including RiPPs. All of these tools use profile Hidden Markov Models (pHMMs [65]) rather than BLAST queries, which are built from protein alignments of many members of a protein domain family. RiPP BGCs are identified based on manually set rules determining which pHMMs detected in

Table 1. All currently classified RiPP families. Adapted from [48].

Class	Example	Class-defining PTM(s)
Amatoxins/phallotoxins	Phalloidin	N-to-C cyclization, Cys-Trp crosslink
Amidinotides	Pheganomycin	Amidino amino acid containing peptides (ATP-grasp)
Atropitides	Tryptorubin	Aromatic amino acids crosslinked to give a noncanonical atropisomer
Autoinducing peptides	AIP-I	Cyclic ester or thioester
Bacterial head-to-tail cyclized peptides	Enterocin AS-48	N-to-C cyclization (DUF95 & ATP-grasp)
Borosins	Omphalotin	Amide backbone N-methylation (N-MT), N-to-C cyclization (POP)
Bottromycins	Bottromycin A1	Macrolactamidine (YcaO)
ComX	ComX168	Indole cyclization and prenylation
Conopeptides	Conantokin G	Peptides produced by cone snails
Crocagins	Crocagin A	Indole-backbone cyclization
Cyanobactins	Patellamides	N-terminal proteolysis (PatA protease)
Cyclotides	Kalata B1	N-to-C cyclization, disulfide(s) (AEP)
Dikaritins	Ustiloxin	Tyr-Xxx ether crosslink (UstY)
Epipeptides	YydF	D-amino acids (rSAM)
Glycocins	Sublancin 168	S, O, or N-glycosylation of Ser/Cys
Graspetides	Microviridin J	Macrolactones/lactams (ATP-grasp)
Lanthipeptides	Nisin	(Methyl)lanthionine, labionin
Lasso peptides	Microcin J25	Macrolactam with threaded C-terminal tail (Asn synthetase homolog)
Linaridins	Cypemycin	Dhb, no lanthionines
Linear azol(in)e containing peptides (LAPs)	Microcin B17	Cys, Ser, or Thr derived azol(in)es (YcaO)
Lipolanthines	Microvionin	C-terminal labionin/avionin containing peptide and N-terminal FAS/PKS segment
Lyciumins	Lyciuman A	Pyroglutamate, Trp-Gly crosslink
Methanobactins	Methanobactin	Oxazolones (DUF692)
Microcin C	Microcin C	Aminoacyl adenylate or cytidylate with a phosphoramidate linkage (ubiquitin E1 homolog)
Mycofactocin	Mycofactocin	Val-Tyr crosslink (rSAM)
Orbitides	Cyclolinopeptide A	N-to-C cyclization; no disulfides
Pantocins	Pantocin A	Glu-Glu crosslink (PaaA)
Pearlins	Thiaglutamate	aa-tRNA derived (PEARL)
Proteusins	Polytheonamide	Nitrile hydratase LP
Pyrroloquinoline quinones	PQQ	Glu-Tyr crosslink (rSAM)
Ranthipeptides	Freyrasin	Sulfur-to-non-Cα thioether crosslink (rSAM)
Rotapeptides	TQQ	Oxygen-to-α-carbon crosslink
Ryptides	RRR	Arg-Tyr crosslink (rSAM)
Sactipeptides	Subtilosin	Thioether crosslink to alpha-carbon (rSAM)
Spliceotides	PlpA	β-amino acids (rSAM)
Streptides	Streptide	Trp-Lys crosslink
Sulfatyrotides	RaxX	Tyrosine sulfation
Thioamitides	Thioviridamide	Backbone thioamide (YcaO)
Thiopeptides	Thiostrepton	[4+2] Cycloaddition of two Dha
Thyroid hormones	Triiodothyronin	Triiodothyronin

close proximity of one another form a RiPP BGC. For example, type I lanthipeptide BGCs can be identified by targeting the protein domains present in the modifying enzymes LanB (PF04738/PF14028) and LanC (PF05147), both of which should be be found encoded by genes located near one another on the genome.

These methods excel at the detection of known RiPP families, for which the RTEs responsible for the hallmark modifications have been identified. Completely novel RiPP families which lack these modifications cannot be detected, however. Nevertheless, these BGCs may still specify RiPPs that are novel because they encode different precursor. Examples of studies investigating these are numerous, and only a handful are mentioned here. For example, antiSMASH-based genome mining led to the discovery of streptocollin, a type IV lanthipeptide [66]. A study investigating the RiPP BGCs of 629 actinobacterial genomes using BAGEL3 detected 477 different RiPP BGCs[67]. Most of these contained unique precursor peptides (e.g. lanthipeptides: 276 out of 301 unique, lasso peptides: 62 out of 67 unique, LAPs: 43 out of 48 unique). A more thorough investigation into only lanthipeptide-like BGCs in Actinobacteria detected 1,163 in 830 genomes. These were further grouped into 100 gene cluster families (GCFs) based on sequence and RTEs encoded. Interestingly, several GCFs encoded RTEs not previously associated with lanthipeptide BGCs, like O-methyltransferases, NRPSs and PKSs [68].

Although detection of RTEs is relatively straightforward, the challenge in the automated detection of RiPP BGCs lies in the correct annotation of the genes encoding precursor peptides. Gene finding algorithms such as Glimmer [69, 70] and Prodigal [71] frequently miss the open reading frames (ORFs) that encode precursor peptides, as they can be as small as 15 nucleotides [49]. BAGEL4, the latest of version of BAGEL, takes additional steps to increase the number of precursor genes detected [62]. In a genomic area that contains genes encoding RTEs, all intergenic small ORFS (<= 72 nt) are extracted, translated and BLASTed against a database of the core sections of known RiPP precursor peptides. This method provides a more detailed annotation of precursor genes. However, since detection is based on known core peptides, completely novel precursor peptides will not be detected by this method.

A more sophisticated approach for precursor detection is taken by RODEO [45, 55, 72-74]. RODEO allows a user to analyze the genomic context of any gene matching a query domain on NCBI. Given its accession number, genes in the context of a query gene are annotated with Pfam and TIGRFAM [75, 76]. The tool was first used to mine genomes for lasso peptide BGCs, using a rule-based system based on detected protein domains. To better detect precursor genes, a machine learning classifier called a Support Vector Machine (SVM) was trained to distinguish between lasso precursor peptides and other peptides. This SVM was trained on several hundreds of features, such as frequency of specific amino acids or amino acid pairs, charge and hydrophobicity. The prediction of this SVM was combined with heuristic scoring of a given small ORF to effectively detect precursor genes. The same model for precursor detection was integrated into antiSMASH, as of version 4.0 [77].

The prerequisite of both a precursor peptide and a specific protein domain has been used to mine for thiopeptides [72], sactipeptides and ranthipeptides [55], lasso peptides [45], lanthipeptides[73] and linaridins [74]. These genome mining efforts have expanded the list of candidate BGCs belonging to each family, and led to the discovery of novel RiPPs, such as citrulassin. In theory, the same process could be applied to any RiPP family, as long as sufficient precursor sequences are available to train an SVM. This method therefore lends itself mostly to well-characterized classes. Interestingly, like in the study described above, the BGCs detected contain a wide assortment of different putative modifying enzymes, which occasionally co-occur within the core RTEs (1-25%). These are predicted to encode for e.g. acetyltransferases, glycosyltransferases, FAD oxidoreductases or methyltransferases. The existence of RiPPs with additional tailoring is not without precedence, exemplified by reports of acetylated lasso peptides [78], glycosylated lanthipeptides [79] and lipidated lanthipeptides [80]. Characterization of these secondary tailoring enzymes provides interesting opportunities to further chart the chemical landscape covered by RiPPs. In addition, given that many RTEs recognize via the leader peptide, these enzymes may be capable of modifying other RiPPs as well, allowing one to further tweak their properties with synthetic biology [81].

1 Explorative domain-based genome mining expands and defines novel RiPP families

The rule-based genome mining used by high-confidence RiPP genome mining tools described above is an effective way to expand known RiPP families, for which a conserved set of genes has been identified. However, for many newly discovered RiPP families, sometimes only a single example BGC is known. Highly homologous BGCs can easily be identified, but more interesting is perhaps the identification of a minimal set of genes that is required for a specific modification unique to the RiPP. Identification of these in novel contexts could lead to the discovery of novel RiPPs that belong to the same or related families. In the following section, we will describe several studies aimed at genome mining of novel RiPP families as well as the discovery of related RiPP families by shared modifications.

Discovery and expansion of omega-ester peptides

The first member of the omega-ester peptides of RiPPs was microviridin, a cytotoxic RiPP with three intramolecular omega-ester or omega-amide crosslinks, which was isolated in 2008 [82]. While initial studies focused on identifying microviridins from the cyanobacterial genus *Mycrocystis* [83], the characterization of two homologous BGCs from *Plesiocystis pacifica* and *Bacillus thuringiensis serovar huazhongensis* led to the identification of plesiocin and the thuringinin group, respectively [54, 84]. Like microviridin, these RiPPs also contained omega-ester and omega-amide crosslinks, although the number of crosslinks and the overall topology of the products were different. As all BGCs encoded a homologous ATP-grasp ligase, these proteins could be used as a query for genome mining of novel BGCs of the same type [85]. This search resulted in 5,276 homologous proteins. Inspection of context of the encoding genes for possible precursor peptides resulted in the identification of 12 groups of new omega-ester containing peptides. This is a sizable increase in the number of candidate BGCs of this family, especially considering that only four ATP-grasp ligases were used as a query. However, the authors note that ~3,200 protein hits could not be assigned a specific precursor, which could be false positives.

Table 2. Tools available for RiPP genome mining.

Name	BGC identification target	Method description	Reference
antiSMASH	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with RODEO's SVMs.	Blin *et al*. [39]
BAGEL	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with BLAST and a known precursor database.	Van Heel *et al*. [62]
RiPP-PRISM	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with HMMer and a motif search.	Sknnider *et al*. [40, 64]
RODEO	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identification of precursor peptides with SVMs.	Tietz *et al* [45], Schwalen *et al* [72], Hudson *et al* [55], DiCaprio *et al* [86], Walker *et al* [73], Georgiou *et al* [74].
RiPPer	Any enzyme	Identifies RiPP BGCs with any query enzyme. Prioritizes candidate precursor peptides with prodigal-short and BLAST-based clustering.	Santos-Aberturas *et al*. [52]
RiPPMiner	Precursor peptides	Identifies and classifies precursors with a single SVM.	Agrawal *et al*. [87]
NeuRiPP	Precursor peptides	Identifies precursors with a neural network.	De Los Santos. [88]

Table 2 (continued).

DeepRiPP	Precursor peptides	Identifies and classifies precursors and BGCs with a neural network (NLPPrecursor). Predicts products and estimates novelty based on genetic context and known modifications (BARLEY). Compares metabolomics and matches MS/MS spectra to predicted products (CLAMS).	Merwin *et al*. [89]
DEREPLICATOR	NA	Clusters peptide natural products based on MS/MS spectra.	Mohimani *et al*. [90]
VarQuest	NA	Matches peptide natural products to their variants with unknown modifications based on MS/MS spectra.	Gurevich *et al*. [91]
MetaMiner	Core enzymes	Identifies RiPP BGCs with antiSMASH. Predicts products based on genetic context and known modifications. Matches predicted products to MS/MS spectra.	Cao *et al*. [92]

Novel thioamidated RiPPs found by a bait-based approach combined with precursor clustering

The above example highlights how the combination of a putative precursor and a single RTE of interest as a query allows identification of new types of RiPP BGCs. RiPPer was developed to generalize this procedure for any type of RTE [52]. The search starts with a query RTE, which is used find the genes encoding their homologs within a given database. To identify possible precursor genes, the surrounding region (+- 8 kbp) of each hit is reannotated with an adapted version of the genefinding software Prodigal called prodigal-short. The adapted version has a lower cut-off point for the minimum size of a gene (60 nt instead of 90), allowing it to more effectively identify RiPP precursors. All short genes

(between 60 and 360 nt) are scored by the prodigal score, which is increased if it is on the same strand as the query RTE. This approach does not take into account the sequence of the precursors, but was still able to detect 94.1% and 96.7% of two test sets of precursor peptides from microviridin and lasso peptide genome mining studies [45, 83]. However, because multiple candidate precursor peptides are reported per BGC, the total number of precursor peptides identified by this method was several times higher than the training set. To increase the specificity of detected precursors, the authors clustered the precursor peptides detected based on sequence similarity. Large groups of conserved peptides are more likely to be encoded by real ORFs, and indeed, the largest group of peptides was found to overlap with previously identified precursor peptides. Peptides encoded by spurious ORFs are less likely show significant similarity to one another, and therefore small groups of precursor peptides can be discarded as false positives.

The authors used the *tfua* gene as a query RTE, which encodes a protein thought to be involved in the formation of thioamidated RiPPs, like thioviridamide [93, 94]. The nearby candidate precursors were clustered, which resulted in thirty networks, two of which were encoded by thioviridamide-like BGCs. Of each of the networks, the genomic context of the genes encoding these precursor peptides for each of the networks was manually analyzed. Based on this, the authors highlighted 12 different candidate RiPP families. Experimental characterization of one of these families led to the discovery of the thiovarsiolins. These RiPPs share the small thioamidated amino acids backbones, but otherwise share no similarity with thioviridamide-like RiPPs. While a single RTE like the one encoded by the *tfua* gene may not be a reliable marker, in conjunction with a group of predicted precursor peptides, it leads to many promising and yet to be uncovered RiPPs.

Radical SAM enzymes as a versatile RiPP marker

Genes that are shared between different RiPP families make interesting targets for the discovery of novel RiPP families. Examples include the cyclase domain, which is shared between all different lanthipeptides subtypes [47], and the YcaO protein, which is involved in the maturation of bottromycins, LAPs, thiopeptides and thioamitides, and can catalyze three different reactions [95]. A protein that

is involved in the maturation of a wide number of RiPPs is the radical S-adenosyl methionine (rSAM) enzyme. The rSAM enzyme uses a radical S-adenosyl methionine (SAM) intermediate to catalyze a range of different reactions [96], ranging from simple methyltransferase reactions in bottromycins [97] to conversions of L-D stereochemistry in proteusins [98] to crosslinks in sactipeptides [99]. Radical SAMs can be involved in primary metabolism as well as secondary metabolism, and are highly diverse, which makes them somewhat unreliable for RiPP genome mining [100]. In conjunction with a precursor peptide, however, or by targeting a specific clade more closely associated with RiPP metabolism, many more interesting RiPP BGCs could be uncovered.

rSAMs themselves are as diverse as their modifications, but pHMMs have been developed for rSAMs in specific niches, such as those responsible for the maturation of sporulation killing factor A (TIGR04403) [101]. Identification of a specific subclade of rSAMs and using this as a query has previously led to the discovery of mycofactocin [102]. The authors in this study identify a subclade of without any assigned function, and use partial phylogenetic profiling to discover protein families associated with this subclade. A group of proteins was found strongly associated with this particular clade, which led to the discovery of the new RiPP mycofactocin. Expanded genome mining using the rSAMs associated with sactipeptides as a query, in conjunction with SVM-based precursor identification, has led to the discovery of ranthipeptides [55]. In addition, rSAMs were identified in a RiPP BGC in *Pleurocapsa sp. PCC 7319*, which was found to specify α-keto-β-amino acid-containing RiPPs, although the exact genome mining method was not reported [103].

Last of all, by building a query for the detection of genes encoding rSAMs and quorum-sensing type regulators, a study reported the presence of these enzymes in many genomes [104]. The rules for detection were based on the BGC of streptide [105]. The production of streptide is regulated by quorum sensing, and its BGC contains two genes encoding a two-component system that mediate this regulation. A search in Streptococcal genomes revealed that the three genes encoding the regulatory system and a rSAM appeared to co-occur frequently. One of these BGCs was experimentally characterized, leading to the discovery of WGK, rotapeptides and ryptides [104, 106-108].

1

The RiPP recognition element as a guide for novel RiPP family discovery

The core element that is associated with most RiPP families is the RiPP Recognition Element (RRE). RREs were first discovered as short domains of roughly 100 amino acids that showed high structural similarity to PqqD [109]. Experimental characterization showed that the element could bind precursor peptides, and in many cases was required for RiPP maturation. Thorough analysis of all known RiPP classes revealed that the element is present in roughly half of all RiPP classes discovered to date. This domain may either be present in a small, stand-alone protein, or be fused to another enzymatic domain. In many cases, the domain is essential for RiPP maturation, even for stand-alone RREs, suggesting they act as a guide for other modifying enzymes to aid in precursor peptide recognition. In some cases, though, the domain is vestigial [110]. This similarity was discovered with HHPred [111], an algorithm for the comparison of pHMMs and secondary structure predictions made by PSIPRED [112]. As of now the only known method for the reliable detection of RREs is via HHPred, which is a time-consuming algorithm taking several minutes per query. Nevertheless, as this element promises to be highly specific towards RiPPs, but is still independent of any specific RiPP family, it would make an excellent target for RiPP genome mining.

Shared enzymology between RiPPs and non-RiPPs leads to discovery of 3-thiaglutamate

In a few unusual cases, genome mining for RiPP BGCs can lead to the discovery of non-RiPP BGCs. For example, a recent genome mining effort aimed at finding lanthipeptide-like gene clusters in Actinobacteria identified several genes for LanB, without a nearby gene encoding LanC [68]. LanB is involved in the dehydration of serine and threonine residues in type I lanthipeptides. This reaction takes place in two steps, where first the hydroxyl group is glutamylated using tRNA-glutamate as the donor, and then eliminated by a separate protein domain [47]. Strikingly, these LanB homologs did not contain an elimination domain. Characterization of the BGC showed that the precursor peptide is used catalytically to produce 3-thiaglutamate [113]. A cysteine was attached to the C-terminus of a precursor-like peptide by the LanB homolog, converted to 3-thiaglutamate and then excised. The precursor peptide could still be used as a

recognition site for the other enzymes, three of which even contained RREs, but was otherwise not consumed for the production. The LanB homologue was renamed as a peptide aminoacyl-tRNA ligases, and these products were collectively called pearlins.

Precursor-centric genome mining of RiPP BGCs

The one thing that all RiPP BGCs have in common is the gene encoding a precursor peptide, and this is a vital element in all RiPP genome mining strategies. Usually, the precursor gene is in operon with the genes encoding the modifying enzymes, so detection of a precursor gene will result in detection of a new BGC. The precursor peptides show a high amount of variability, however, which limits precursor-centric genome mining using similarity-based methods, even within a single RiPP family. In addition, the precursor genes may be very small (<= 100 nt), and therefore missed by automatic gene annotation programs. Nevertheless, identification of the precursor peptide is highly valuable as it greatly speeds up experimental characterization.

The reliability of machine-learning methods to detect encoded precursor peptides could re-invigorate precursor-centric genome mining. In general, any such method should have a low false positive rate in order to be useful for precursor-based genome mining. After all, the number of small open reading frames that are precursor candidates far exceeds the number of expected RiPP precursors. A *Streptomyces* genome will likely contain only 1 to 5 RiPP BGCs [67]. Assuming that on average, a small ORF can be found between each pair of genes, there will be 8,000 small ORFs to analyze. A RiPP classification method with a low false positive discovery rate of 1% and a perfect true positive discovery rate of 100% will detect all positive hits, but also roughly 80 negative hits, outweighing the positive hits 40 to 1. Therefore, these models should either have a very low false positive rate or be integrated into larger pipelines in order to be useful for novel RiPP discoveries. This proves a difficult challenge, especially considering that there are relatively few positive examples of RiPP precursors compared to the number in the negative training set. Besides RODEO, three other tools use different machine-learning models to detect and classify precursor peptides: RiPPMINER, NeuRiPP and DeepRiPP.

Like RODEO, RiPPMINER uses a trained SVM to distinguish precursor peptides from other peptides, although unlike RODEO, a single SVM is used for all RiPP classes [87]. The predicted precursor peptides are then classified to their respective RiPP class using a multi-class SVM. The tool can identify well-characterized RiPP families such as lanthipeptides, lasso peptides and linaridins, by training on a manually curated training set of more than 500 RiPP precursors. The precision and sensitivity for identification were 0.93 and 0.90, respectively. While these values are fairly high, the tool might still report a fair number of false positives when used as a starting point for genome mining for the reasons stated above. Nevertheless, RiPPMINER was used in a pipeline with the ClusterFinder algorithm [41] and transcriptome data analysis to detect novel candidate RiPP BGCs in the fungus *Trichoderma* spp [114] showing that tools with relatively low sensitivity can find novel results when used in conjunction with other datasets.

NeuRiPP [88] uses a neural network rather than an SVM for precursor peptide classification. The model trained here takes the raw sequence as input, instead of calculated features like amino acid frequency or hydrophobicity. Several network architectures were tested, of which the parallel convoluted neural network (CNN) performed best. The tool separated precursor peptides from non-precursor peptides with an accuracy of 99.84% on the entire training set. In addition, it was capable of detecting bottromycin precursor peptides, despite the fact that these were lacking from the training data. This suggests that the network was capable of identifying precursor-specific features from the raw sequences across different RiPP families. The author suggested integration of the tool into RiPPer [52] as an effective means to filter precursor peptides in a class-independent manner. As a proof of concept, the precursor peptide networks associated with the *tfua* gene, also identified by RiPPer (see above) were analyzed. Of the 12 peptide networks prioritized by the authors, 8 were identified as precursor peptides by NeuRiPP, despite the fact to no precursor peptides of these RiPP classes were present in the training data.

DeepRiPP is an assembly of three modules and the first tool described that fully integrates precursor-centric mining with comparative genomics and metabolomics tools [89]. The first module, NLPPrecursor, uses a Universal

Language Model Fine-Tuning (ULMFiT) neural network to detect encoded precursor peptides. This is a neural network architecture used for language processing that has shown to be highly effective in building models from training sets with low amount of data. It further classifies detected precursor peptides into specific RiPP families. BARLEY, the second part of the pipeline, then tries to estimate all possible RiPPs that may be produced from a detected precursor peptide, using known modifying enzymes in the surrounding genomic context, and a machine-learning model to estimate the cleavage site. All possible final products are compared to each other and to a database of known RiPPs and the distance between all products is calculated, either using the genomic predictions of the final products of the known products. In this way, known products can be easily dereplicated, and BGCs can be identified whose predicted products are distant from known RiPP BGCs, increasing the odds that new RiPP variants will be discovered. In the final step, extracts of all the strains analyzed made under various growth conditions are analyzed by LCMS. CLAMS, the third module of the pipeline, then tries to find correlations between the absence and/or presence of predicted RiPP products with detected LCMS peaks. The fragmentation data of these peaks are also matched to possible peptide fragments calculated from the precursors. By mining 65,421 bacterial genomes, 19,498 new possible RiPP products were identified. The authors then extract 463 of these strains under various conditions, creating a metabolomics database of 10,498 extracts. In these extracts, three new RiPPs could be identified with CLAMS, belonging to the lanthipeptide, lasso peptide and thiopeptide RiPP families, respectively.

In summary, machine-learning methods provide an excellent way to accurately predict precursor peptides. These methods can supplement RiPP genome mining of known classes, as is done in RODEO and antiSMASH. More importantly, precursor-centric genome mining can lead to the discovery of novel RiPP BGCs, without needing to first identify characteristic modifying enzymes. Since precursor-centric genome mining carries an inherent risk for a high amount of false positives, the results of these classifiers need to be carefully analysed. Indeed, these methods proved most successful when combined with more extensive data analysis, such as analysis of genomic context, comparative genomics, transcriptomics and metabolomics.

1 Integration with omics in larger pipelines

As illustrated by DeepRiPP described above, the integration of metabolomics data could accelerate the identification of RiPPs. In contrast to normal proteins, however, RiPPs contain modified amino acids and are rarely linear. For known RiPP classes, the modifications can be predicted based on genomic information. Predicted peptide fragments containing these modifications can be matched to the spectra with tools like DEREPLICATOR [90] (recently updated with NPS [115]) and CLAMS (available within the DeepRiPP pipeline [89]). DeepRiPP is perhaps the most integrative pipeline for RiPP discovery. Besides structure prediction based on the identification of known modifications, it also combines comparative genomics with comparative metabolomics, to prioritize peaks whose presence/absence matches that of the BGCs of interest. However, it could even be extended further, by also considering transcriptomics and proteomics data. Elicitors should therefore be added to activate the expression of cryptic BGCs, whereby comparative metabolomics combined with transcriptomics or proteomics will allow linkage of BGC expression profiles to changes in metabolites. This will allow scientists not only to observe more metabolites than under one specific growth condition, but also to predict which metabolites are produced by which BGCs.

A major challenge for automated MS/MS analysis that remains is dealing with new modifications. VarQuest [91], an extension of DEREPLICATOR, can identify peptide variants based on known peptides, even if these variants contain unknown modifications. MetaMiner [92] combines genomics and metabolomics to predict precursor modifications and find associated spectra, which can contain unknown modifications. Completely *de novo* identification of novel RiPPs with only unknown modification has yet to be explored by tools like these, but represents a sizable computational challenge. Even so, just matching a small sequence of unmodified amino acids to part of a candidate novel RiPP precursor is a valuable addition to more explorative RiPP searches. Identified, novel precursors could then be fed back to the training data of the precursor classifiers, creating an iterative process in which the classifiers will become increasingly specific and tuned toward a larger variety of RiPP classes.

Outline of the thesis: towards the detection of completely novel RiPP subclasses

The diversity in tools described above highlights the challenges in RiPP genome mining. Traditional RTE-based approaches do an excellent job at increasing the number of members of a RiPP family, as long as a well-defined set of enzymes characteristic of that class is known. Increasingly well-polished methods for the identification of RiPP precursors make an excellent supplement to these methods, providing additional information to properly identify the final product. Integration with metabolomics further streamlines RiPP identification, and can unite metabolomic and genomic information.

Most of the tools described above, however, do not focus on the discovery of completely novel RiPP subclasses. The discovery of these is a difficult challenge – after all, if no modifying enzymes can be used as queries, BGCs cannot be discovered with methods that target specific domains. Even though RiPPer gives a user more freedom in this regard, it still relies on the selection of a query domain, which biases the results. However, one feature is always present in almost all RiPP subclasses: each RiPP BGC should encode a precursor, and contain at least one modifying enzyme. These domain-independent features could be exploited to mine RiPPs in a less restricted manner, and lead to the discovery of new RiPP subclasses.

Detection of precursors in a class-independent manner could be accomplished with machine-learning-based classifiers. The high confidence of these classifiers has already led to precursor-based genome mining, but mostly of known RiPP subclasses. Interestingly, NeuRiPP was capable of predicting some RiPP precursors of RiPP classes for which it had not been trained. Apparently, this neural network is capable of identifying some property of combination of properties that distinguishes precursors of any class from other peptides. If this set of properties is shared among not-yet-discovered RiPPs, it is possible that precursor-based genome mining could lead to the identification of completely novel RiPPs, as no restrictions would be placed onto the genomic contexts in terms of known RTEs.

In **Chapter 2**, we describe a novel tool for the detection of RiPP Recognition Elements (RREs) – the domain that is shared among the most different bacterial RiPP classes. Specific profile Hidden Markov Models have been designed for each of the different types of RRE. This allows for high-confidence detection of RREs of known classes in precision mode. A second mode, called exploratory mode, is based on HHPred and can detect more distantly related RREs, at the cost of computational power and more false positives. These methods allow the detection of novel RRE-enzyme fusions, that can lead to the discovery of novel RiPP subclasses.

In **Chapter 3**, we describe an innovative tool for the identification of novel RiPPs, called decRiPPter (Data-driven Explorative Class-independent RiPP TrackER). This tool utilizes an SVM-based RiPP precursor classifier, which is independent of RiPP subclass, and can therefore be used to identify novel RiPPs. Instead of focusing on the amino acid sequence, decRiPPter examines the genomic contexts of encoded precursor peptides for possible RTEs, associated with RiPPs or otherwise. The results have been combined across many genomes to form candidate RiPP families. The work underlines the power of artificial intelligence approaches for the discovery of new candidate bioactive molecules.

Chapter 4 describes the application of decRiPPter for the identification of a novel class of lanthipeptides. BGCs of this family are widespread among Actinobacteria and Firmicutes, but so far their function was unknown. Experimental characterization of a gene cluster from *Streptomyces pristinaespiralis* revealed that it indeed specifies a novel RiPP, that we called pristinin A3. Pristinin A3 contains many modifications also found in other types of lanthipeptides. Lanthipeptides are further classified by their modifying enzymes. Since the modifying enzymes involved in the generation of this family of RiPPs are novel, we classified this RiPP as a new lanthipeptide subclass, called class V. The complex two-dimensional structure of pristinin A3 was elucidated by mass spectrometry and NMR.

In **Chapter 5,** a different type of RiPP BGC is characterized. This BGC shows distant similarity to known RiPP BGCs of different classes, as it contains genes encoding a radical SAM enzyme and an ATP-grasp ligase. Still, the presence of these genes alone do not place it clearly in any known RiPP subclass.

In addition, two well-conserved genes encoding predicted precursors show a unique motif that is repeated multiple times. A detailed bioinformatic description is given explaining the homologies of this BGC and it's relation to other known RiPP BGCs. In addition, experimental work is presented describing the activation of the BGC and the analysis of chemical extracts aimed at identifying the final product.

In **Chapter 6**, the results are summarized and reviewed in a general discussion. The explorative approach taken towards RiPP genome mining, and the use of machine learning classifiers for this purpose, are reviewed. the challenges encountered in this thesis are described and possible solutions are proposed. Further extenstions for the decRiPPter pipeline are outlined, which could futher help future efforts in class-independent RiPP genome mining. Also, the RiPP BGCs that were studied in this work are further discussed, including their possible classifications with regard to currently accepted schemes.

1