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

Kloosterman, A.M.

6

General discussion and conclusion

Alexander M. Kloosterman
Marnix H. Medema
Gilles P. van Wezel

Introduction

Ribosomally synthesized and post-translationally modified peptides (RiPPs), like many other natural products, comprise a dazzling array of chemical diversity [42, 48]. The simple biosynthetic logic – a precursor gene is translated, and the product is modified and cleaved – results in many different possible structures, depending on the precursor peptide and modifications applied to it. They are divided in more than 40 different subclasses, and the list of subclasses keeps steadily growing. Their functions are equally diverse, and include quorum sensing, acting as enzyme co-factors, roles in cellular development, mediating host-microbe interactions, but also the much sought-after antibacterial and antifungal properties that would make them interesting for clinical applications [261].

Next-generation sequencing efforts surprisingly revealed that the capacity of bacteria to produce natural products had been grossly underestimated. This has led to a revolution in drug discovery based on the efficient mining of the rapidly growing genome sequence data [26, 262]. Numerous tools and databases have been developed to explore, compare and catalogue biosynthetic gene clusters (BGCs) and their chemical products [29, 185, 228, 263, 264]. For most of the biosynthetic gene clusters (BGCs) discovered, the chemical compounds they produce are not yet known. These so-called cryptic BGCs represent a vast potential for new natural products with potentially interesting bioactivities. Even so, the BGCs that are currently easily detected are in some ways similar to characterized ones, since their detection is based on the identification of conserved protein domains [26, 39, 40, 228].

RiPPs present an interesting case when it comes to genome mining, as there is no single genetic marker that identifies them all, other than the prerequisite of an ORF that encodes a small peptide. Although some RiPP subclasses overlap on a genetic level, most require bioinformatic rules aimed at subclass-specific genetic markers. There is still plenty of room for innovative genome mining strategies aimed at identifying novel RiPP subclasses, each of which further charts undiscovered genetic space [52, 85, 88, 89, 114, 265, 266]. In this thesis, we have explored such strategies which prioritize novelty at the

cost of fidelity, with the primary aim being the identification of novel RiPP families. The main pipeline that was developed, decRiPPter, can function as a platform for explorative RiPP genome mining. In contrast to most tools developed for high-confidence RiPP genome mining, decRiPPter relies more heavily on user settings, and present several options for trade-offs between confidence and novelty. Using this tool, the pristinin BGC was discovered, which encodes a novel class V lanthipeptide.

Machine learning paves the way for class-independent precursor identification

Machine-learning-based and neural-network-based classifiers have risen in popularity over the last decades as tools to process and classify massive datasets with large numbers of features. The large databases of genome sequences now available provide many opportunities for these classifiers to exploit their high precision for the benefit of genome mining. Specifically for RiPPs, the identification of the precursor gene presents an interesting challenge for machine-learning classifiers. Precursor genes are not easily recognized by similarity-based methods, and are frequently missed by automatic gene annotation algorithms due to their small size. Several classifiers have been developed for RODEO that supplement classical RiPP genome mining by identifying precursors of known classes [45, 55, 72-74, 86], and several more tools have been reported for standalone precursor identification [88, 89, 114].

Detection of precursor peptides forms the core of decRiPPter (Chapter 3), and determines which genomic regions will be further investigated. As such, decRiPPter is the first reported genome mining tool that uses the detection of precursors, rather than of enzymatic domains as the basis for the identification of novel RiPP subclasses. Analysis of 1,295 *Streptomyces* genomes resulted in the discovery of 42 candidate RiPP families after manual curation. All of these families are specified by BGCs that are characterized by a promising combination of precursor, transporter, biosynthetic, regulator and peptidase genes, typically organized in a single operon-like genomic structure. While some of the BGCs contain genes previously reported in known RiPP BGCs, most of the biosynthetic genes encode enzymes not previously associated with RiPP biosynthesis,

suggesting that many more RiPP modifying enzymes exist than currently known. Characterization of these enzymes could then be translated to new RiPP genome mining rules for tools like BAGEL [62] and antiSMASH [39], standardizing their detection. Experimental investigations into one of the families showed that it did indeed encode a novel RiPP, namely a lanthipeptide, pristinin A3, that is modified by a newly discovered set of modifying enzymes. How many more of the 42 families actually specify RiPPs needs to be validated experimentally. However, if even half of these candidates encode actual RiPPs, it would represent a sizable contribution to expanding the RiPP chemical space.

Remarkably, the wide variety of precursor sequences of many different classes can be adequately covered by the SVM-based classifier of decRiPPter. Apparently, there are certain combinations of features that are typical of RiPP precursors regardless of class. These include the enrichment of certain amino acids, like cysteine, serine and threonine, which are often modified in known RiPPs, but also frequently found in the candidate RiPP BGCs. In addition, arginine residues are particularly rare across RiPP precursors. An evolutionary link between different RiPP classes could explain these conserved features, but is made unlikely by the large disparity in precursors and modifying enzymes. Cysteine, serine and threonine residues do have oxygen- and thiol-groups, respectively, making them easier to modify. This chemical property could drive the evolutionary process towards precursor peptides containing certain residues, even if they have evolved independently. If the latter is the case, it would explain why feature-dependent classifiers that focus on amino acid frequencies are so effective at detecting precursor peptides of many different classes, and it would suggest that many more RiPP classes can be detected by them.

A difficult challenge when applying these classifiers to a large genomic space is the number of false positives. The sheer number of candidates (71 million) as opposed to the number of expected precursor genes (~6500 if each genome encodes five RiPP precursors) makes it so that even a false discovery rate of 1% would result in many more false positives than true positives. Comparisons with other machine-learning-based classifiers revealed similar numbers of hits for those, meaning that this would be a general issue. As the

number of characterized precursors increases, and therefore the training set improves, the accuracy of newer classifiers may improve as well. Alternatively, restricting the set of precursors to those for which at least two different classifiers reach a consensus would reduce the number of hits substantially, as the overlap between the three studied methods is relatively low. However, it is questionable whether precursor identification itself can become reliable enough for precursor-based RiPP genome mining without considering their genetic context. The false discovery rate would have to drop substantially while still covering the wide variety of precursor sequences. Until then, using the genetic context as shown in Chapter 3 is a viable alternative to filter the identified precursors down to a more manageable set.

Another way to filter the predicted precursors without considering genetic context is to prioritize precursors with multiple core regions. Having multiple copies of the same core region allows for the efficient production of several RiPP variants, while only needing a single leader peptide. A similar pattern was also identified in the RiPP candidate discussed in Chapter 5. These repeats are found often in eukaryotic RiPPs [51, 239, 267], and could provide a handhold for their identification without prior knowledge of their primary sequence. If these patterns occur as exact copies, their identification would be algorithmically straightforward, by taking subsets of the sequence and finding exact matches of that sequence elsewhere in the peptide. If more variation of the pattern occurs, such as in thiovarsolins, identification of these patterns would have to be based on a local alignment algorithm, such as BLAST, or a motif discovery tool such as MEME [116, 268]. Alternatively, the presence of a repeated pattern can be used as a feature in a future iteration of the classifier, so that it is taken into account during precursor prediction itself. Flagging precursors in which these patterns can be found can be used to remove many false positives, albeit at the cost of removing RiPP families which do not contain these patterns. Their presence could therefore be used as an imperfect bioinformatic handle to fine-tune precursor-based RiPP genome mining.

Prioritizing novel RiPP BGCs from the genetic context

decRiPPter uses the genetic context of predicted precursor genes to prioritize candidate BGCs. The filtering process exemplifies the trade-off between confidence and novelty, and can be set up according to user preferences. At loose conditions (e.g., mild filtering), most known RiPP BGCs are left unfiltered, but the number of false positives is estimated upwards of 84.4%, making the dataset too large to manually process. It is likely that there are still many RiPP BGCs among this dataset, which is also highlighted by the promising candidate discussed in Chapter 5, but without additional filters, selecting a suitable candidate can become difficult. In order to simplify this, the HTML-based output allows a user to browse the results. In addition, the entire set can be filtered with additional criteria of interest, such as specific biosynthetic domains, or a specific number of transporters, proteases or regulators in or nearby the precursor gene. The resulting set can then be manually investigated and a BGC of interest can be selected. Expanding the output filtering options with additional parameters, such as specific motifs within precursors, would help users browsing this large dataset and find the exact sort of BGC they are looking for.

The strict filter applied is a middle ground between confidence and novelty. On the one hand, it is permissive in the sense that many different domains are considered as possible RiPP associated enzymes and proteins. On the other hand, it is restrictive in the sense that genes for a peptidase, regulator and transporter are all required. In theory, these encode peptidases for precursor cleavage, a dedicated transporter module, and a cluster-specific regulator. Many known RiPP BGCs do not contain all of these genes, and instead their encoded pathway and products are regulated, transported and cleaved by proteins encoded elsewhere in the genome. As a result, the remaining candidate BGCs are promising, and the false positive rate was lower than with the mild filter (estimated between 46.7 and 73.0%), although many known RiPP BGCs are filtered.

Several other methods for prioritizing gene clusters of interest can be envisioned, which would each represent a different trade-off in confidence and

novelty. Integrating these into decRiPPter would further expand the possibilities for more fine-tuned search strategies in which several criteria can be combined. The tool for one of these, RRE-Finder, was discussed in Chapter 2. RiPP Recognition Elements (RREs) are involved in the precursor recognition of many different RiPP classes, and could function as a class-independent bioinformatics handle for RiPP discovery. With RRE-Finder, RREs can be detected at a faster rate than with HHPred, allowing for the analysis of large amounts of queries. Exploratory mode of RRE-Finder, which is based on HHPred, detected several novel RRE-enzyme fusions in the UniProt database, which could lead to the discovery of novel RiPP modifying enzymes. Unfortunately, the false discovery rate of exploratory mode is higher than for precision mode, which makes it questionable which of the newly discovered RRE-enzyme fusions would be worth investigating. This disadvantage can be mitigated by imposing other mild criteria of decRiPPter, i.e. a predicted precursor gene nearby, one or two biosynthetic domains in an operon-like gene organization, and not being part of the core genome. Integration of RRE-Finder therefore would be a valuable addition to the decRiPPter pipeline, and help increase the confidence for both tools.

RRE-Finder itself could be further improved by using a machine-learning classifier for the detection of RREs. Like RiPP precursor peptides, RREs are generally no longer than 120 amino acids long. A candidate sequence of this length can be used completely as an input vector in a neural network, as is done in NeuRiPP, without having to select specific features. This approach would allow for detection of discrete RREs by using part of the sequence, e.g. the N- or C-terminal regions, as raw input for the network. These classifiers might be able to better distinguish between regulators and RREs, as they can recognize more complicated patterns than only secondary structure. A possible discriminatory feature are the sequence residues that are known to interact with the precursor peptide. Several of these residues have been shown to co-evolve with the precursor peptide, and likely stand out from a sequence-based point of view when compared to similar domains found in regulators. Further research is required to determine if machine-learning classifiers are indeed suitable for the detection of RREs.

Insights into RiPP evolution guide discovery of novel RiPPs

Understanding how different RiPPs have evolved can provide useful insights for the prioritization of RiPP BGCs, especially if these principles are class-independent. For secondary metabolism in general, it has been hypothesized that their enzymes have evolved from primary metabolism enzymes. An example of this can be seen for polyketide synthetases (PKSs), which descend from fatty acid synthetases, but have diverged to take in different substrates, and apply extra tailoring [269]. This property has been used earlier to mine for BGCs in EvoMining [160, 161]. By searching for enzymes that have evolved from primary metabolism enzymes, many BGCs of known classes like NRPS and PKS, but also of novel classes, can be identified.

Interestingly, the RiPP candidates prioritized by decRiPPter included several BGCs that encode proteins previously identified in a different context. HypD, HypE, MauD and MauE are thought to be involved in protein maturation, by creating crosslinks or modifying specific amino acids [182, 183]. These proteins could have easily evolved towards modifying small peptides rather than proteins, and could thus have become RiPP-modifying enzymes. A similar example was recognized earlier: QhpD, an enzyme that catalyzes the synthesis of a thioether bond in a protein, and radical SAM enzymes involved in thioether crosslink formation in sactipeptides and ranthipeptides, show moderate similarity [55, 270]. Protein modification is a widely occurring phenomenon in all branches of life, and it is possible that more RiPP modifying enzymes evolved from them. An approach similar to EvoMining, using protein-modifying enzymes as a query, could aid in the identification of more of these RiPP subclasses.

Another sizable contributor to RiPP BGC biodiversity is the occurrence of gene swaps. The genes for YcaOs [95], rSAMs [96], lanDs [194], for example, are encoded by BGCs of several RiPP subclasses. The newly reported lanthipeptide class V further contributes to this list, as its BGC contains elements from both linaridins [271] and thioamitides [52, 94, 272], further suggesting that gene swaps contribute significantly for RiPP diversity. An automated procedure might be able to prioritize genes present in many RiPP-like clusters, even if they were not previously functionally associated with RiPPs before. If, from a

candidate RiPP BGC, a gene or set of genes can be detected in BGCs of other candidate RiPP families as detected by decRiPPter, this would make it more likely that the gene product is involved in RiPP maturation. This can be seen to some extent with the *mauE* and *mauD* genes, which are present in three different RiPP families, and also with the core enzymes of the novel lanthipeptide subclass, described in Chapter 4. In a simple form, this procedure can be automated by searching for biosynthetic domains that are seen among several different RiPP families. A more sophisticated pipeline could involve the usage of CORASON to identify gene islands widespread across many different RiPP-like contexts. Successful identification of these islands would help prioritize RiPP modifying enzymes, and by extension, RiPP families.

Examples of novel RiPPs and their classification

To validate decRiPPter's capabilities to detect novel RiPP classes, we selected two BGCs of different candidate families to experimentally characterize. One of these encodes a novel lanthipeptide, pristinin A3, containing the classical thioether bridge, a C-terminal aminovinylcysteine and serine-to-alanine conversions (Chapter 4). Importantly, two candidate genes appear likely candidates for the formation of the thioether bridge. Their presence in many genetic contexts shows that this class is widespread across several taxonomic clades, and that these genes are excellent candidates to add to the rulesets of high-confidence RiPP genome mining tools. Furthermore, lanthipeptides frequently possess antimicrobial activity [273, 274], so the discovery of a novel class of these could in time lead to the discovery of novel antibiotics.

Another promising BGC (Chapter 5) has many features that suggest it specifies a RiPP. This BGC contains many genes that encode enzymes previously associated with RiPP biosynthesis, like an rSAM and an ATP-grasp ligase. Despite this, the BGC was not directly recognized by other RiPP genome mining tools, and encodes several more predicted modifying enzymes that were not recognized. The repeated, conserved patterns observed in the precursor peptides are likely multiple core regions. Several masses were detected exclusively when the gene cluster was activated, which were no longer present when the gene cluster was inactivated. These masses were within 200 Da of the

mass of the predicted core peptide. Unfortunately, none of the masses could be matched to the core peptide, and it remains unclear whether any of the masses are directly derived from it. It seems likely that the many predicted enzymes extensively modify the core peptides, meaning more sophisticated analytical chemistry is required to relate the structure to the peptide. Furthermore, heterologous expression of the BGC could help prioritize which masses are exclusively derived from the BGC, and not produced due to any secondary effects, like the activation of another BGC.

The two BGCs described in this work both contain genes that have homologs encoded by BGCs of other RiPP subclasses. Despite this, they both would still likely specify members of a novel RiPP class, due to a unique combination of modifications or novel enzymatic machinery that installs it. In general, however, the discovery of RiPP classes that are produced mostly by a combination of modifying enzymes already known makes their classification more complicated. The consensus for classification of RiPPs is based on designating modifications as core or accessory, and determining which core modifications are required for one RiPP family [42]. This methodology is becoming more and more difficult to uphold. Given that modifications can be swapped between different RiPP families, which one is considered a core modification and which one is considered an accessory one is context-dependent. If the lanthionine bridge of pristinin A3 is considered the core modification, as for other lanthipeptides, then all other modifications would be considered secondary. These include the formation of dehydrated serine residues, which are considered a core modification in linaridins.

As a result, what makes up a novel RiPP class becomes somewhat arbitrary. Lipolanthines, for example, are considered a standalone RiPP class, but they are clearly very related to other lanthipeptides [80]. By contrast, glycosylated lanthipeptides are not considered their own class. Since the definition of a RiPP class determines the rules for genome mining of that class, we should take care not to restrict ourselves too much with these definitions. Many more interesting RiPP variants can be found by alleviating the strictest of rules. Rather than focus on the identification of novel RiPP classes, which could be considered arbitrary, perhaps the priority should be the identification of

RiPP-associated reactions and their corresponding modifying enzymes. The RiPP classes can be considered examples in which specific modifications have been found combined. But any RiPP-associated enzyme could arguably lead to the discovery of new RiPP classes and variants, whether core or accessory.

Conclusion

Natural products and their BGCs come in many shapes and sizes, resulting in a rich diversity to explore. In this thesis, we have explored methods aimed at finding novel types of natural products, specifically novel RiPP subclasses. The biosynthetic logic of a RiPP can be made up of many different precursors and modifying enzymes. There are several features, however, which can be exploited for their detection. RiPP BGCs should always encode a precursor peptide, providing a handhold for identification with machine-learning classifiers. Encoded modifying enzymes in the BGC should be capable of recognizing the precursor peptide, which can be exploited through the detection of RREs or through their association with other RiPP classes. We have combined these methods to prioritize many different gene clusters, and illustrated that one of these gene clusters indeed specified a novel type of lanthipeptide (pristinin). The pipeline can be expanded further in many ways, including the integration of RRE-Finder, new precursor classifiers, or detection methods using evolutionary principles, which will help expanding the large chemical diversity of this class of natural products.