CellPress

Trends in
Pharmacological Sciences

Review

Metabolome-guided genome mining of RiPP natural products

Mitja M. Zdouc [1,*] Justin J.J. van der Hooft [1,2,*] and Marnix H. Medema [1,*]

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a chemically diverse class of metabolites. Many RiPPs show potent biological activities that make them attractive starting points for drug development. A promising approach for the discovery of new classes of RiPPs is genome mining. However, the accuracy of genome mining is hampered by the lack of signature genes shared across different RiPP classes. One way to reduce false-positive predictions is by complementing genomic information with metabolomics data. In recent years, several new approaches addressing such integrative genomics and metabolomics analyses have been developed. In this review, we provide a detailed discussion of RiPP-compatible software tools that integrate paired genomics and metabolomics data. We highlight current challenges in data integration and identify opportunities for further developments targeting new classes of bioactive RiPPs.

Highlights

Ribosomally synthesized and post-translationally modified peptides (RiPPs) from microorganisms show high chemical diversity and exhibit potent biological properties.

The computational detection of novel classes of RiPPs is hampered by their short length and the lack of universally conserved genes.

The high false-positive rate of class-independent computational detection approaches can be addressed by validation via mass spectrometry-based metabolomics.

RiPPs – a pharmacologically promising class of natural products

Microorganisms produce a vast array of low-molecular-weight metabolites known as **natural products** (NPs; see Glossary), also called 'secondary metabolites' or 'specialized metabolites'. These molecules are not immediately involved in cell survival but often display potent biological activities, a property used for the development of numerous drugs [1]. A recent large-scale survey estimated that only 3% of NP biosynthetic pathways encoded in bacterial genomes have been experimentally characterized [2]. Therefore, microorganisms represent still largely untapped sources for NP drug discovery.

Among the different classes of microbial NPs, **ribosomally synthesized and post-translationally modified peptides** (RiPPs) have received special attention due to their exceptionally large biosynthetic diversity [3]. RiPPs are known for many interesting biological properties, including antibiotic, antiviral, and antineoplastic activities [4]. For example, the recently described RiPP darobactin A (Figure 1A, structure 1) selectively kills Gram-negative bacteria by inhibition of the outer membrane protein BamA. This novel antibiotic mode of action, the first one since the 1960s, represents a promising avenue toward the development of new antibiotics [5–8]. Growing interest in the scientific and commercial potential of RiPPs has led to the discovery of no fewer than 17 new classes of RiPPs between 2011 and 2020 [9,10]. It is generally believed that the currently known 40+ distinct classes of RiPPs [10] are only the most widely distributed ones and that there is large 'hidden' RiPP biosynthetic potential left to discover.

The overwhelming majority of RiPP classes was discovered serendipitously: promising biological activity or an interesting signal in a **metabolomics** experiment was investigated, and the responsible molecules were isolated. Only after structural elucidation of the NP, followed by the genome sequencing of the producing organism, could the biosynthetic origin be elucidated [10–12]. Such 'isolation-first' strategies, also known as 'grind and find', carry the risk of rediscovery of known metabolites, are resource-intense, and are of limited compatibility with modern high-throughput

[1]Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, the Netherlands
[2]Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa

*Correspondence:
mitja.zdouc@wur.nl (M.M. Zdouc),
justin.vanderhooft@wur.nl
(J.J.J. van der Hooft), and
marnix.medema@wur.nl (M.H. Medema).

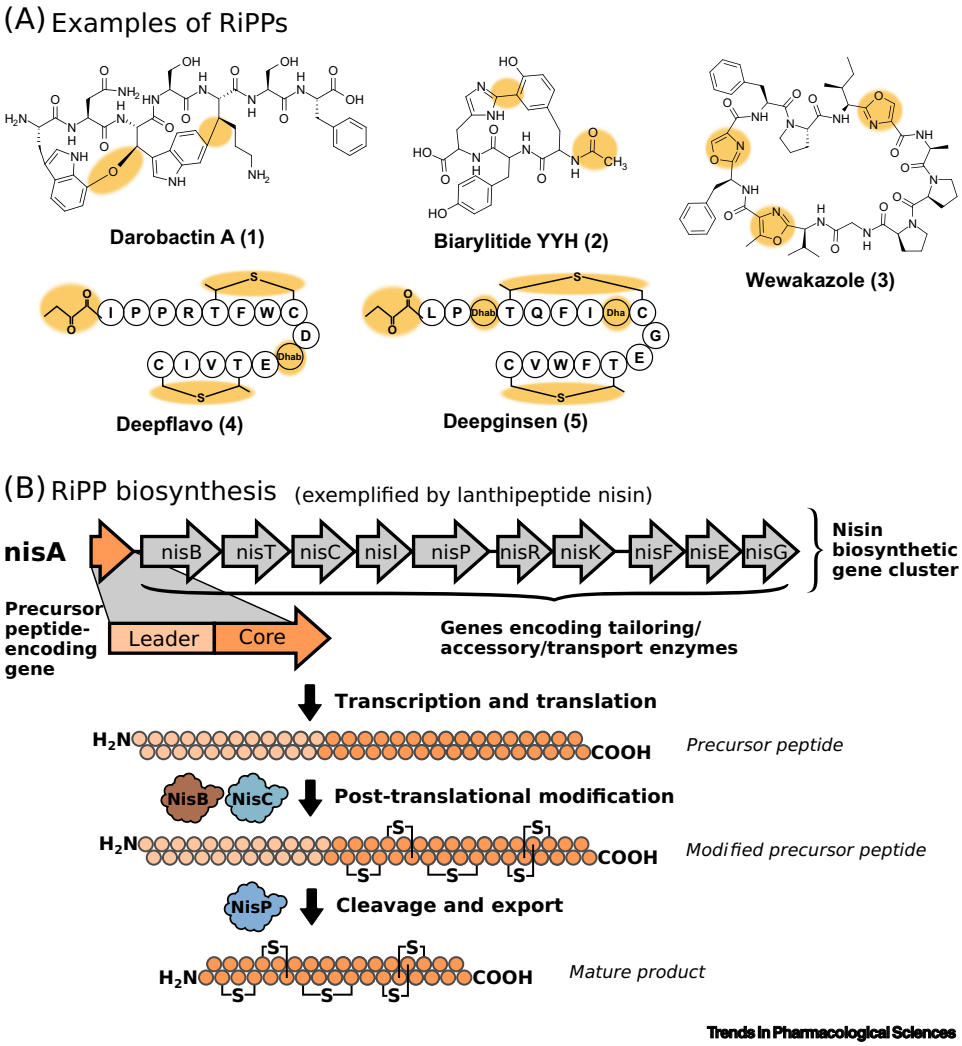

Figure 1. Overview showing examples of ribosomally synthesized and post-translationally modified peptides (RiPPs) and their canonical biosynthetic pathway. (A) RiPP chemical structures mentioned throughout the article: darobactin A (1), biarylitide YYH (2), wewakazole (3), deepflavo (4), deepginsen (5). Highlighted in yellow are different post-translational modifications introduced by a variety of class-dependent enzymatic reactions. (B) RiPP biosynthesis exemplified by lanthipeptide nisin. The structural gene *nisA* is transcribed and translated, resulting in the precursor peptide, consisting of a 'leader' and core peptide part. The core peptide is modified by tailoring enzymes that are encoded in the biosynthetic gene cluster. The 'leader' peptide part is cleaved from the modified core peptide, and the mature product is exported [76].

approaches. Therefore, computational methods for the detection and prioritization of biosynthetic pathways in **genomics** data have been developed. Predictions can be further validated by using metabolomics data, but automated data integration is not yet trivial. In this review, we first discuss the biosynthetic principles that complicate the detection of new classes of RiPPs by genome mining. We continue with an overview of recently developed generalist and RiPP-specialized software tools for automated integration of genomic and metabolomics data, then address current challenges, and finally highlight opportunities for further development.

Fundamentals of RiPP biosynthesis

Canonically, the biosynthesis of microbial NPs is governed by a set of genes colocalized in the same genomic region, known as a **biosynthetic gene cluster** (BGC). RiPPs follow this

Glossary

Biosynthetic gene cluster: genes responsible for the biosynthesis of a natural product, colocalized in a genomic region.
Feature-based approaches: paired omics approach using extracted (sub) structural or chemical compound class information for connecting biosynthetic gene clusters and molecular families.
Gene cluster family: subnetwork resulting from clustering of biosynthetic gene clusters based on pairwise similarity analysis.
Genomics: methods to study genes, their functions, and spatial distributions in a given genome.
Liquid chromatography–tandem mass spectrometry: analytical technique for the separation and spectrometric analysis of molecules and their fragmentation.
Metabolomics: methods to study the total set of small metabolites produced by an organism.
Molecular family: subnetwork resulting from MS/MS fragmentation-based spectral similarity analysis, with nodes representing individual or grouped spectra.
Natural products: small-molecule specialized or secondary metabolites not immediately involved in cell survival.
Paired omics: pairing of genomic and metabolomic data with regard to the investigation of natural products.
Precursor peptide: short peptide resulting from a structural gene, which is modified and results in the mature product.
Ribosomally synthesized and post-translationally modified peptides: natural products that are produced by the proteinogenic way, in contrast to nonribosomal peptides.
Strain-correlation-based approaches: paired omics approach using strain/sample presence/absence overlap to assess correlation between biosynthetic gene clusters/gene cluster families and MS/MS spectra/molecular families.

biosynthetic logic and consist of at least two components: first, one or more small structural genes encoding short **precursor peptides**, and second, one or more genes encoding precursor-peptide-modifying 'tailoring' enzymes. These two components alone can be sufficient to produce a mature product [11,12]. Additionally, accessory genes related to maturation, transport, autoresistance, or regulation are commonly colocalized in RiPP BGCs [3].

The RiPP precursor peptides consist of a core peptide, usually flanked by an N-terminal 'leader' peptide (Figure 1B). In some cases, a C-terminal recognition sequence (the 'follower') is present, either on its own or together with the 'leader' peptide [3]. After transcription and translation, the precursor peptide is modified by tailoring enzymes, which introduce post-translational modifications (PTMs). PTMs greatly expand the chemical space of proteinogenic amino acids, including the introduction of β- or D-amino acids, alterations to the peptide conformation, and additions of heteroatoms or other functional groups [10]. RiPPs are grouped into classes (or families) based on shared structural and biosynthetic concepts. These range from 'simple' macrocyclization (e.g., the lasso-fold structure observed in lassopeptides) to complex biosynthetic cascades (e.g., thiopeptides, also known as pyritides) [3,13]. After modification by tailoring enzymes, the leader and/or recognition sequences flanking the core peptide are removed by proteolysis, resulting in the mature modified core peptide, which is eventually exported from the cell [3].

Genome mining for RiPPs: principles and challenges

The conserved architecture of colocalized genes in the BGCs of microbial NPs can be detected and annotated computationally by a strategy known as 'genome mining' [14–16]. Most commonly, BGCs are detected by using hardcoded rulesets based on conserved 'signature' genes (e.g., antiSMASH [17,18] or PRISM [19,20]). Detected BGCs can be annotated by matching against experimentally characterized BGCs, using community resources such as MIBiG [21,22]. Large databases of putatively detected BGCs are available for comparisons (e.g., antiSMASH-DB [23], IGM-ABC [24]). On the basis of the observation that similar BGCs often produce similar compounds, BGCs can be further grouped into so-called **gene cluster families** (GCFs). In GCFs, annotations of identified BGCs can be propagated to their neighbors in the network, which allows one to formulate hypotheses about their encoded products [25,26]. Furthermore, subcluster analysis can predict putative substructures of the encoded (unknown) metabolites [27,28]. Therefore, genome mining allows automated assessment of the 'theoretical' biosynthetic capacity encoded in a microbial genome (i.e., the 'biosynthetic blueprint') and to compare it with the existing body of knowledge [15].

Genome mining is also suitable for the detection of RiPP BGCs: antiSMASH can detect at least 28 different classes of RiPPs [18], whereas RiPP-PRISM can detect no fewer than 21 different classes [9]. In the antiSMASH database (version 3), the 14 most abundant classes of RiPPs amount to at least 44 000 predicted RiPP BGCs across publicly available bacterial, archaeal, and fungal reference genomes [23]. Once an RiPP class is described, the involved enzymatic machinery can easily be detected by gene homology-based approaches. However, genome mining for completely novel RiPP classes is much more challenging: because RiPP biosynthetic classes do not share universally conserved core enzymes or motif sequences, they remain 'invisible' to rule-based genome mining tools. Furthermore, RiPP structural genes encoding precursor peptides can be extremely short: the smallest reported structural gene [*bytA*, encoding the biarylitide YYH (Figure 1A, structure 2) precursor] is only 18 base pairs long, making it also the shortest known coding gene [11]. Considering all possible short open reading frames in a genome may lead to a prohibitively high number of potential candidates, including many false-positives, whereas defining a minimal gene length for structural peptides may also exclude novel classes of short RiPPs.

To address the limitations of homology-dependent BGC detection, tools using alternative concepts for BGC detection were developed: besides tools using concepts applicable to all classes of microbial BGCs, such as ClusterFinder [29], EvoMining [30], or DeepBGC [31], a few tools have been designed specifically for the detection of novel RiPP BGCs. The tool DeepRiPP uses a deep-learning approach based on natural language processing (NLPPrecursor) to identify new RiPP precursor peptides linked to known classes [32]. Similarly, neuRiPP uses a deep neural network architecture to recognize RiPP structural genes independent of their biosynthetic class [33]. Another tool, decRiPPter, uses a support vector machine and a set of rules to differentiate putative RiPP precursor peptides from small noncoding genes [34]. A drawback of such homology-independent methods is their high rate of false-positive detection due to lack of indicatory signature enzymes, requiring extensive manual follow-up validation [34].

Pruning of false-positives: pairing genomics and metabolomics data

One strategy to reduce false-positives and to improve throughput in the discovery of novel classes of RiPPs is to validate predictions from genome mining *via* detection of products using **liquid chromatography–tandem mass spectrometry** (LC-MS/MS)-based metabolomics [35]. In LC-MS/MS analysis, NPs are separated, ionized, and fragmented by collisional dissociation. In the resulting tandem mass (MS/MS) fragmentation spectra, individual fragments typically correspond to parts of the parent molecule structures (i.e., substructures). This makes MS/MS spectra useful for diagnostic purposes, such as the annotation of substructures and the identification of the chemical compound class [36–41]. MS/MS fragmentation spectra can also be considered as characteristic molecular fingerprints, with similar molecules usually showing similar MS/MS fragmentation. Modification-tolerant matching of spectra allows clustering of molecules into networks based on MS/MS spectral similarity [also known as '**molecular families**' (MFs)], thereby organizing data and propagating annotations [42–45].

Therefore, experimentally observed NPs can be annotated and 'mapped' back to BGCs to confirm initial predictions. This matching also allows one to prioritize BGCs that show expression over those that do not (many BGCs are 'silent' under laboratory conditions). Hence, genomic and metabolomic data are complementary in forming and confirming hypotheses and reducing false-positives. Such integrated metabolomics and genomics data are generally referred to as **paired omics** datasets [35]. In recent years, different tools for the processing and analysis of paired omics datasets have been developed [35,46–48]. We first survey generalist tools that are also applicable to RiPP NPs, followed by tools that are specifically designed for the analysis of RiPPs (see overview in Table 1). We limit our discussion to tools that require both genomics and metabolomics data as input.

Generalist and RiPP-specific tools for omics data pairing

Generalist tools pair BGCs to MS/MS spectra by relying on information that is applicable to all biosynthetic classes [35]. A common strategy is the analysis of presence–absence patterns of BGCs and MS/MS spectra associated with microbial strains, so-called **strain-correlation-based approaches** (Figure 2). BGCs and MS/MS spectra are first organized into GCFs and MFs, respectively, using different clustering tools. Therefore, GCFs and MFs each can be traced back to sets of strains, allowing the calculation of linking scores based on strain overlap [35]. Such a generalist approach was first introduced under the name 'metabologenomics' by Doroghazi and colleagues, who matched GCFs and detected molecules using a point-based system relying on strain contribution, followed by manual verification of the putative links [26]. Similarly, Duncan and others applied 'pattern-based genome mining', which relied on a manual comparison of the presence–absence patterns of GCFs and MFs [49]. Some other generalist tools use a 'hybrid' approach by combining both correlation- and feature-based concepts in

Table 1. Recently developed paired genomics and metabolomics software addressing ribosomally synthesized and post-translationally modified peptides with several key factors to consider upon their use

Tool [latest version]	Approach	RiPP specific?	Open source	Free academic license?	Note	Refs
Ripp2Path [2016]	Feature- based	Yes	Yes	Yes	Part of Pep2Path package	[56]
RippQuest [2014]	Feature- based	Yes	No	–	Superseded by MetaMiner	[57]
MetaMiner [2019]	Feature- based	Yes	No	Yes	NPDtools package, GNPS website	[58]
DeepRiPP [2021]	Feature- based	Yes	No	Yes[a]	–	[32]
Metabolo-genomics [2023]	Correlation- based	No	No	No	No public release of program	[26,63]
NPLinker [2023]	Hybrid	No	Yes	Yes	Undergoing refactoring, see https://github.com/NPLinker/nplinker	[50]
NPOmix [2022]	Hybrid	No	Yes	Yes	Input must be similar to reference database; undergoing refactoring, https://github.com/tiagolbiotech/NPOmix_python	[52]

[a]Requires login and approval of extensive end user license agreement.

pairing. One of them, NPLinker [50], expands and refines the scoring algorithm first introduced by the 'metabologenomics' approach [26] and combines it with the feature-based IOKR score, which calculates binary molecule fingerprints from MS/MS fragmentation spectra and structures predicted from BGCs for improved pairing. Recently, NPLinker was enhanced by a new scoring function called 'NPClassScore', which uses chemical compound classes predicted from BGCs and MS/MS fragmentation patterns to eliminate a substantial number of false-positive BGC-MS/MS links [51]. Another hybrid tool is NPOmix, which uses a *k*-nearest neighbor-based classifier to compare similarity fingerprints calculated from the association of microbial strains to GCFs and MFs. NPOmix further uses information regarding predicted molecular substructures

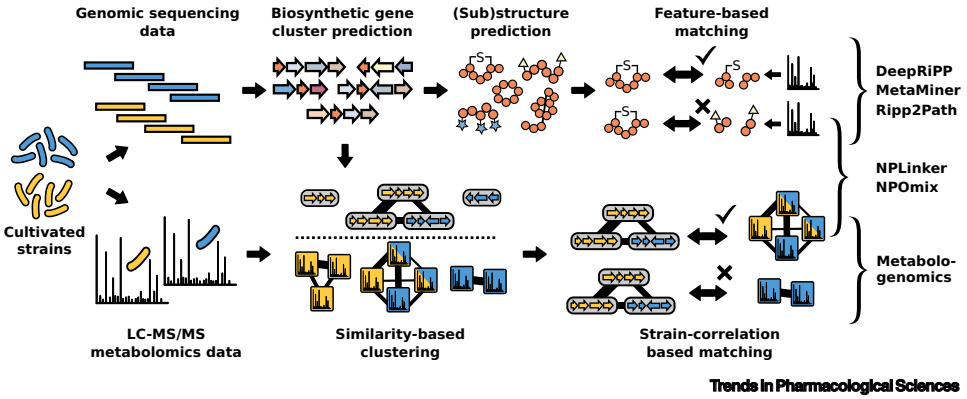

Figure 2. Schematic visualization of a generalized paired omics data workflow for the prioritization of ribosomally synthesized and post-translationally modified peptides. Starting from microbial strains, genomes are sequenced and metabolomics data are generated. After the prediction of biosynthetic gene clusters, two strategies of data analysis can be followed. One strategy focuses on the prediction of chemical (sub)structures from biosynthetic gene clusters and from metabolomics data. The presence or absence of predicted substructures is then used for feature-based matching. Another strategy organizes biosynthetic gene clusters and mass spectrometry data based on similarity, followed by strain co-occurrence (correlation)-based matching. Examples of software tools using these strategies are mentioned on the right-hand side of the figure. Abbreviation: LC-MS/MS, liquid chromatography–tandem mass spectrometry.

and biosynthetic class to supplement the classifier-based score [52]. Intuitive and generally applicable, these correlation-based concepts were used in manual or semiautomated fashion for the discovery of new RiPPs from known classes, such as the chymotrypsin inhibitor microviridin 1777 [53] or new congeners of the antibiotic siomycin [54].

Most RiPP-specific tools use **feature-based approaches**: molecular 'building blocks' (e.g., substructures, scaffolds, or specific functional groups such as amino acids, PTMs) are first inferred or predicted from BGCs and MS/MS fragmentation spectra. These structural features are then used to create profiles for individual BGCs and MS/MS spectra, and the overlap between these profiles is used to calculate pairwise linking scores (Figure 2) [35]. This approach was first applied to RiPPs (and nonribosomal peptides) by Kersten and others using the 'peptidogenomics' method: short amino acid sequence tags from MS/MS fragmentation spectra were identified and manually matched against predicted peptide sequences from BGCs [55]. This initial approach was later automated by Pep2Path's RiPP2Path algorithm. The latter converted a mass shift sequence into a list of candidate peptide sequences and matched them against predicted precursor peptides from the genome, regardless of PTMs [56]. The tool RiPPQuest used a similar approach but focused on the identification of lanthipeptides, and the authors validated their approach with the identification and structural prediction of a putative new lanthipeptide named 'informatipeptin' [57]. Recently introduced tools further advance automation and annotation capabilities: MetaMiner expands the scope of its predecessor RiPPQuest in terms of covered RiPP classes and allows processing of metagenomic data [58]. MetaMiner creates a combinatorial library of putative peptide sequences with a variety of PTMs. Predicted MS/MS spectra are compared against experimental ones in a modification-tolerant way, and matches are scored by statistical significance (estimating the false discovery rate). The authors demonstrated the applicability of MetaMiner by linking the known RiPP wewakazole (Figure 1A, structure 3) to its BGC [58]. In another study, MetaMiner was used to annotate putatively novel lassopeptides [59]. A similarly automated tool, DeepRiPP, uses a natural language processing approach to detect RiPP precursor peptides and predict their cleavage patterns [32]. It is further integrated with the RiPP-PRISM tool for prediction of putative tailoring reactions [9] and includes algorithms for precursor annotation by comparison against known RiPPs and identification of matching MS/MS fragmentation spectra. The applicability of DeepRiPP was demonstrated by the detection, prioritization, characterization, and isolation of novel members of known RiPP classes [e.g., the lanthipeptides deepflavo (Figure 1A, structure 4) and deepginsen (Figure 1A, structure 5)].

Challenges and opportunities in RiPP paired omics analysis

Toward class-agnostic feature-based pairing tools

RiPP-specific paired omics tools usually use feature-based pairing approaches, relying on substructure recognition for scoring of putative links. However, current substructure prediction strategies are mostly restricted to PTMs of characterized RiPP classes. For example, the MetaMiner tool uses a hardcoded ruleset of tailoring reactions of nine classes of RiPPs for the creation of a combinatorial library of putative products, relying on antiSMASH [17,18] and BOA [60] for BGC detection [58]. Similarly, the RIPP-PRISM algorithm used by DeepRiPP is limited to tailoring reactions of well-known RiPP classes to create a combinatorial library of putative MS/MS fragments for matching against experimental data [32]. In our literature survey, we could not find any account of a novel RiPP class being discovered by using feature-based paired omics tools. Therefore, a pressing issue is the improved prediction of tailoring reactions acting on putative RiPP precursor peptides, resulting from tailoring enzymes that may only be distantly related to characterized ones. This is a crucial step for the prediction especially of novel RiPP substructures and consecutively linking them to metabolite spectra. One possible approach could be the use of machine learning-based models trained on RiPP-associated, generalist tailoring enzymes

(e.g., halogenases, oxidases, dehydratases). Having such models in place would allow generalized *in silico* biosynthesis of putative substructures, extending the concepts used by tools such as DeepRiPP and MetaMiner for known RiPP classes. Alternatively, RetroRules-like generalized reaction rules [61] for enzyme classes commonly involved in RiPP biosynthesis could be extracted and used to predict biosynthesis. Both approaches could predict amino acid sequence tags modified with putative PTMs to be used for direct matching against experimental MS/MS fragmentation spectra or to predict molecular fingerprints. Although perfect structure prediction remains elusive for the foreseeable future, matches only need to be 'good enough' to allow annotation and hypothesis-driven prioritization for follow-up experimental validation.

Limitations of correlation-based generalist tools

Contrary to feature-based approaches, strain correlation-based approaches are independent of prior or inferred biosynthetic or chemical knowledge. Therefore, they are in principle suitable for known and unknown RiPPs alike. A disadvantage of correlation-based approaches is the high number of possible pairwise links that can arise between GCFs and MFs with similar source strain contributions/sample occurrence [50]. In this case, the linking scores have low differentiating power, requiring manual sifting through the top *n* best matches to identify plausible ones. This not only is resource-intensive but also requires expert knowledge. Therefore, correlation-based approaches do not scale well to large datasets and struggle in differentiating strains with similar biosynthetic profiles. Another drawback of correlation-based approaches is their reliance on similarity-based grouping of BGCs into GCFs and metabolites into MFs, respectively. Usually, a range of cutoff values can be used to construct GCFs and MFs, with 'looser' cutoff values leading to more permissive groups with a larger number of members than 'stricter' values [35]. Currently, there is no generally accepted consensus or definition for the minimum similarity two BGCs or two MS/MS spectra need to display to be considered related [40]. Cutoff values are therefore often chosen empirically and are specific to the research question. Moreover, RiPP BGCs tend to be rather small, and their grouping can therefore be significantly affected by included flanking regions that are added to the biosynthetic regions by 'greedy' approaches such as antiSMASH. Similarity-based grouping always carries a certain amount of arbitrariness, thereby strongly affecting downstream processing. The parameter dependence of similarity-based tools is discussed in more detail elsewhere [62,63].

Opportunities in combining feature- and correlation-based approaches

Some tools, such as NPLinker [50], apply a hybrid strategy to combine correlation- with feature-based approaches to filter out false-positive connections, such as by considering the biosynthetic class of the encoded product using the NPClassScore [51]. However, to our knowledge, there is no RiPP-specific pairing tool that combines both correlation- and feature-based approaches. Ideally, such a tool would (i) selectively detect RiPP structural genes independent of known classes; (ii) accurately predict RiPP substructures and annotate members of known RiPP classes; (iii) use strain correlation-based strategies to identify possible pairwise links between BGCs and MS/MS spectra; (iv) use substructure information (inferred chemical compound classes, precursor peptide sequences, and/or PTMs) to accurately prune false-positive connections; (v) present results organized into novel versus known RiPP classes, including confidence scores; and (vi) suggest promising candidates for follow-up experimental characterization in terms of novelty, association with orthogonal data (e.g., bioactivity), and isolation feasibility. Some components of such a hypothetical tool already exist in one way or another: tools for class-independent detection of precursor peptides (e.g., DeepRiPP [32], neuRiPP [33], decRiPPter [34]) or substructure prediction (e.g., iPRESTO [28], PRISM4 [20,44], MS2LDA [41], CANOPUS [44], MSNovelist [64]) are available. Furthermore, tools exist to match MS/MS spectra or BGCs against databases (e.g., DEREPLICATOR+ [65], Nerpa [66]), which reduces the risk of reisolation of known molecules. Integration of additional sources of information, such as data on biological activity (e.g., FERMO

[67], NP Analyst [68]), or transcriptomic data (e.g., BiG-MAP [69]), promises to improve prioritization and minimize manual validation of putative matches. However, the creation of such an RiPP discovery tool as described in the preceding text is still hampered by heterogeneity in terms of input and output data formats, software architectures, and terms of software use. Furthermore, not all developers publish their software source code, hampering comprehensibility and accessibility. Here, we emphasize the importance of the open source model (https://opensource.org/osd/) for the development of scientific software: making source code freely available in a well-documented form drastically facilitates the use of scientific software in such custom pipelines.

Besides technical challenges, limitations in available training data impede the development of models for the prediction of putative PTMs. To build better software tools, more and better annotated training data need to be made available in machine-readable form. Public data repositories such as the Paired Omics Data Platform framework (PoDP [70]) allow users to register paired genomics and metabolomics data and specify validated BGC-metabolite matches, using FAIR (Findable, Accessible, Interoperable, Reusable) data principles [71]. We encourage researchers to submit their data to the PoDP and similar initiatives, such as MIBiG [22] for experimentally verified BGCs, MetaboLights [22,72], Metabolomics Workbench [73], or GNPS-MassIVE [39] for metabolomics data, and the Natural Product Atlas [74] for newly elucidated NPs. Depositing both raw and annotated data preserves the manual effort invested in studying BGC–metabolite connections, makes them easily findable and accessible for future work, and allows the development of better software tools. We recognize that deposition of curated research data is time-consuming, which further increases the workload in the publication process. A possible solution to incentivize data submission would be the acknowledgment of original data contributors (e.g., via ORCID) by the developers of machine-learning tools who use their data for training purposes. This easily implementable solution not only would increase visibility of previous work but also would make the time invested in data deposition and curation creditable.

Concluding remarks and future perspectives

The discovery of new RiPP classes by genome mining is complicated by the lack of universally conserved signature genes, leading to a high number of false-positive or false-negative annotations, depending on the approach taken. The automated integration of LC-MS/MS metabolomics data with genomic information promises to accelerate the prioritization process and to eliminate false-positives generated by exploratory algorithms. Currently available tools are either specifically designed for the annotation of already known RiPP classes or too generic to lead to a feasible number of matches when working with large-scale datasets. There is a lack of tools that specifically target novel RiPP BGCs by applying both correlation- and feature-based approaches in a complementary fashion. An important issue is the current inability to account for unknown tailoring reactions in novel RiPP classes, resulting from a general lack of well-annotated training data (see Outstanding questions). Even with the availability of better tools, the discovery of putatively novel RiPP classes will remain a balancing act between sensitivity of detection and confidence of annotation. An even grander future challenge is the correct detection of microbial RiPP BGCs where precursor-peptide-encoding genes and tailoring-enzyme-coding genes are not colocalized. Such noncanonical BGCs are a general problem in genome mining and require special consideration, as recently reviewed elsewhere [16]. This issue may be addressed by integration of further omics data types (e.g., transcriptomics, facilitating coexpression analysis). This could provide additional information about the correct detection of novel classes of RiPPs but may introduce additional challenges in terms of data integration [48]. Nevertheless, follow-up experimental validation will remain essential, and recent developments involving automation via biological foundries are a promising approach to scale up experimental work on RiPPs [75]. Despite the current challenges, the detection of novel RiPP classes is a highly promising

Outstanding questions

How can new RiPP classes be discovered in both a selective and sensitive manner?

What new approaches can effectively detect false-positive RiPP annotations?

How can noncanonical RiPP BGCs without coclustering of genes be addressed?

How can structural predictions of RiPPs in terms of unknown tailoring reactions be improved?

What strategies need to be implemented to guarantee a better integration of existing and future software?

How can prioritization help to make experimental validation become more cost- and time-effective?

What incentives would motivate researchers to deposit their annotated data in public repositories?

endeavor, and new computational tools integrating the full omics cascade can be expected to lead to exciting discoveries.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and suggestions. This work was funded by the European Union Horizon 2020 project MARBLES (101000392).

Declaration of interests

J.J.J.v.d.H. is a member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy. M.H.M. is a member of the scientific advisory board of Hexagon Bio and cofounder of Design Pharmaceuticals. M.M.Z. declares no competing interests.

References

1. Newman, D.J. and Cragg, G.M. (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83, 770–803
2. Gavriilidou, A. *et al.* (2022) Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat. Microbiol.* 7, 726–735
3. Arnison, P.G. *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
4. Zhong, G. *et al.* (2023) Recent advances in discovery, bioengineering, and bioactivity-evaluation of ribosomally synthesized and post-translationally modified peptides. *ACS Bio. Med. Chem. Au.* 3, 1–31
5. Lewis, K. (2013) Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* 12, 371–387
6. Imai, Y. *et al.* (2019) A new antibiotic selectively kills Gram-negative pathogens. *Nature* 576, 459–464
7. Ritzmann, N. *et al.* (2022) Monitoring the antibiotic darobactin modulating the β-barrel assembly factor BamA. *Structure* 30, 350–359.e3
8. Seyfert, C.E. *et al.* (2023) Darobactins exhibiting superior antibiotic activity by cryo-EM structure guided biosynthetic engineering. *Angew. Chem. Int. Ed. Engl.* 62, e202214094
9. Skinnider, M.A. *et al.* (2016) Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc. Natl. Acad. Sci. U. S. A.* 113, E6343–E6351
10. Montalbán-López, M. *et al.* (2021) New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* 38, 130–239
11. Zdouc, M.M. *et al.* (2021) A biaryl-linked tripeptide from *Planomonospora* reveals a widespread class of minimal RiPP gene clusters. *Cell Chem. Biol.* 28, 733–739.e4
12. Nanudorn, P. *et al.* (2022) Atropopeptides are a novel family of ribosomally synthesized and posttranslationally modified peptides with a complex molecular shape. *Angew. Chem. Int. Ed Engl.* 61, e202208361
13. Kunakom, S. *et al.* (2023) Cytochromes P450 involved in bacterial RiPP biosyntheses. *J. Ind. Microbiol. Biotechnol.* 50, kuad005
14. Ziemert, N. *et al.* (2016) The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.* 33, 988–1005
15. Medema, M.H. *et al.* (2021) Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* 22, 553–571
16. Biermann, F. *et al.* (2022) Navigating and expanding the roadmap of natural product genome mining tools. *Beilstein J. Org. Chem.* 18, 1656–1671
17. Medema, M.H. *et al.* (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 39, W339–W346
18. Blin, K. *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* Published online May 4, 2023. https://doi.org/10.1093/nar/gkad344
19. Skinnider, M.A. *et al.* (2015) Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* 43, 9645–9662
20. Skinnider, M.A. *et al.* (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* 11, 6058
21. Medema, M.H. *et al.* (2015) Minimum Information about a biosynthetic gene cluster. *Nat. Chem. Biol.* 11, 625–631
22. Terlouw, B.R. *et al.* (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* 51, D603–D610
23. Blin, K. *et al.* (2021) The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.* 49, D639–D643
24. Palaniappan, K. *et al.* (2020) IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* 48, D422–D430
25. Navarro-Muñoz, J.C. *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* 16, 60–68
26. Doroghazi, J.R. *et al.* (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* 10, 963–968
27. Del Carratore, F. *et al.* (2019) Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Commun. Biol.* 2, 83
28. Louwen, J.J.R. *et al.* (2023) iPRESTO: Automated discovery of biosynthetic sub-clusters linked to specific natural product substructures. *PLoS Comput. Biol.* 19, e1010462
29. Cimermancic, P. *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158, 412–421
30. Sélem-Mojica, N. *et al.* (2019) EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genom.* 5, e000260
31. Hannigan, G.D. *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* 47, e110
32. Merwin, N.J. *et al.* (2020) DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci. U. S. A.* 117, 371–380
33. de Los Santos, E.L.C. (2019) NeuRiPP: Neural network identification of RiPP precursor peptides. *Sci. Rep.* 9, 13406
34. Kloosterman, A.M. *et al.* (2020) Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS Biol.* 18, e3001026
35. van der Hooft, J.J.J. *et al.* (2020) Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* 49, 3297–3314
36. Niessen, W.M.A. *et al.* (2017) *Interpretation of MS-MS Mass Spectra of Drugs and Pesticides*, John Wiley & Sons
37. Beniddir, M.A. *et al.* (2021) Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. *Nat. Prod. Rep.* 38, 1967–1993
38. van der Hooft, J.J.J. *et al.* (2016) Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* 113, 13738–13743
39. de Jonge, N.F. *et al.* (2023) MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nat. Commun.* 14, 1752

40. Ernst, M. *et al.* (2019) MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* 9, 144
41. Dührkop, K. *et al.* (2021) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* 39, 462–471
42. Bandeira, N. (2007) Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications. *Biotechniques* 42, 687–695
43. Watrous, J. *et al.* (2012) Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1743–E1752
44. Wang, M. *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 34, 828–837
45. Gurevich, A. *et al.* (2018) Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* 3, 319–327
46. Soldatou, S. *et al.* (2019) Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. *FEMS Microbiol. Lett.* 366, fnz142
47. Caesar, L.K. *et al.* (2021) Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities. *Nat. Prod. Rep.* 38, 2041–2065
48. Louwen, J.J.R. and van der Hooft, J.J.J. (2021) Comprehensive large-scale integrative analysis of omics data to accelerate specialized metabolite discovery. *mSystems* 6, e0072621
49. Duncan, K.R. *et al.* (2015) Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chem. Biol.* 22, 460–471
50. Hjörleifsson Eldjárn, G. *et al.* (2021) Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput. Biol.* 17, e1008920
51. Louwen, J.J.R. *et al.* (2023) Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. *Microbiome* 11, 13
52. Leão, T.F. *et al.* (2022) NPOmix: a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. *PNAS Nexus* 1, gac257
53. Sieber, S. *et al.* (2020) Microviridin 1777: a toxic chymotrypsin inhibitor discovered by a metabologenomic approach. *J. Nat. Prod.* 83, 438–446
54. Zdouc, M.M. *et al.* (2021) *Planomonospora*: a metabolomics perspective on an underexplored actinobacteria genus. *J. Nat. Prod.* 84, 204–219
55. Kersten, R.D. *et al.* (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* 7, 794–802
56. Medema, M.H. *et al.* (2014) Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* 10, e1003822
57. Mohimani, H. *et al.* (2014) Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* 9, 1545–1551
58. Cao, L. *et al.* (2019) MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. *Cell Syst.* 9, 600–608.e4
59. Rajwani, R. *et al.* (2021) Genome-guided discovery of natural products through multiplexed low-coverage whole-genome sequencing of soil actinomycetes on Oxford Nanopore Flongle. *mSystems* 6, e0102021
60. Morton, J.T. *et al.* (2015) A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. *BMC Bioinformatics* 16, 381
61. Duigou, T. *et al.* (2019) RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.* 47, D1229–D1235
62. de Jonge, N.F. *et al.* (2022) Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* 18, 103
63 Caesar, L.K. *et al.* (2023) Correlative metabologenomics of 110 fungi reveals metabolite-gene cluster pairs. *Nat. Chem. Biol.* Published online March 6, 2023. https://doi.org/10.1038/s41589-023-01276-8
64. Stravs, M.A. *et al.* (2022) MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* 19, 865–870
65. Mohimani, H. *et al.* (2018) Dereplication of microbial metabolites through database search of mass spectra. *Nat. Commun.* 9, 4035
66. Kunyavskaya, O. *et al.* (2021) Nerpa: a tool for discovering biosynthetic gene clusters of bacterial nonribosomal peptides. *Metabolites* 11, 693
67. Zdouc, M.M. *et al.* (2022) FERMO: a dashboard for streamlined rationalized prioritization of molecular features from mass spectrometry data. *bioRxiv* Published online December 22, 2022. DOI: https://doi.org/10.1101/2022.12.21.521422
68. Lee, S. *et al.* (2022) NP Analyst: an open online platform for compound activity mapping. *ACS Cent. Sci.* 8, 223–234
69. Pascal Andreu, V. *et al.* (2021) BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. *mSystems* 6, e0093721
70. Schorn, M.A. *et al.* (2021) A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* 17, 363–368
71. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018
72. Haug, K. *et al.* (2013) MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, D781–D786
73. Sud, M. *et al.* (2016) Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44, D463–D470
74. van Santen, J.A. *et al.* (2022) The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.* 50, D1317–D1323
75. Ayikpoe, R.S. *et al.* (2022) A scalable platform to discover antimicrobials of ribosomal origin. *Nat. Commun.* 13, 6135
76. Trmčić, A. *et al.* (2011) Expression of nisin genes in cheese – a quantitative real-time polymerase chain reaction approach. *J. Dairy Sci.* 94, 77–85