BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Navigating and expanding the roadmap of natural product genome mining tools

Friederike Biermann‡[1,2,§], Sebastian L. Wenski‡[1,2,§] and Eric J. N. Helfrich*[1,2]

Perspective

Open Access

Address:
[1]Institute for Molecular Bio Science, Goethe University Frankfurt, Max-von-Laue Str. 9, 60438 Frankfurt am Main, Germany and
[2]LOEWE Center for Translational Biodiversity Genomics (TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany

Email:
Eric J. N. Helfrich* - eric.helfrich@bio.uni-frankfurt.de

* Corresponding author ‡ Equal contributors
§ Authors are listed in alphabetical order.

Keywords:
genome mining; natural product biosynthesis; non-canonical pathways; PKS; NRPS; RiPP

Beilstein J. Org. Chem. **2022,** *18,* 1656–1671.
https://doi.org/10.3762/bjoc.18.178

Received: 16 August 2022
Accepted: 02 November 2022
Published: 06 December 2022

This article is part of the thematic issue "Enzymes in biosynthesis".

Associate Editor: J. S. Dickschat

Abstract

Natural products are structurally highly diverse and exhibit a wide array of biological activities. As a result, they serve as an important source of new drug leads. Traditionally, natural products have been discovered by bioactivity-guided fractionation. The advent of genome sequencing technology has resulted in the introduction of an alternative approach towards novel natural product scaffolds: Genome mining. Genome mining is an in-silico natural product discovery strategy in which sequenced genomes are analyzed for the potential of the associated organism to produce natural products. Seemingly universal biosynthetic principles have been deciphered for most natural product classes that are used to detect natural product biosynthetic gene clusters using pathway-encoded conserved key enzymes, domains, or motifs as bait. Several generations of highly sophisticated tools have been developed for the biosynthetic rule-based identification of natural product gene clusters. Apart from these hard-coded algorithms, multiple tools that use machine learning-based approaches have been designed to complement the existing genome mining tool set and focus on natural product gene clusters that lack genes with conserved signature sequences. In this perspective, we take a closer look at state-of-the-art genome mining tools that are based on either hard-coded rules or machine learning algorithms, with an emphasis on the confidence of their predictions and potential to identify non-canonical natural product biosynthetic gene clusters. We highlight the genome mining pipelines' current strengths and limitations by contrasting their advantages and disadvantages. Moreover, we introduce two indirect biosynthetic gene cluster identification strategies that complement current workflows. The combination of all genome mining approaches will pave the way towards a more comprehensive understanding of the full biosynthetic repertoire encoded in microbial genome sequences.

Introduction

In 2002, the genome sequences of the model actinomycete *Streptomyces coelicolor* A3(2) [1] and the producer of the antiparasitic drug avermectin, *Streptomyces avermitilis* [2], were published. These index cases marked the transition from the pre- to the post-genomic era in microbial natural product (NP) research [3]. The introduction of next-generation sequencing technologies [4] has led to a constant decrease in sequencing costs [5]. As a result, the number of publicly available genome sequences has rapidly increased and paved the way for a completely new avenue: genome mining. Genome mining describes the targeted bioinformatic analysis of (meta-)genomes to identify gene clusters involved in the biosynthesis of NPs [3]. NPs have been shown to act as signaling metabolites (e.g., acylhomoserine lactones (**1**) [6]), siderophores (e.g., pyoverdines (**2**) [7]), virulence factors (e.g., malleicyprol (**3**) [8-10]), toxins (e.g., bongkrekic acid (**4**) [11]), antibacterial (e.g., vancomycin (**5**) [12]) or antifungal compounds (e.g., amphotericin B (**6**) [13]) (Figure 1). The identification of almost all clinically relevant antibiotics using bioactivity-guided fractionation approaches long before the beginning of the post-genomic era initiated the field of microbial NP research. In the "golden age" of antibiotic discovery from the 1940s to 1970s, microbes and especially bacteria have been identified as an almost untapped treasure trove for the discovery of bioactive NPs. For the longest time, researchers focused on a few talented NP producers, that have mainly been isolated from soil samples [14]. Since the low hanging fruits have been picked using traditional bioactivity-based workflows, this approach frequently results in the rediscovery of known metabolites. The introduction of genome mining revolutionized NP research and helped overcome the rediscovery problem frequently encountered using traditional approaches. Contrary to earlier estimations that

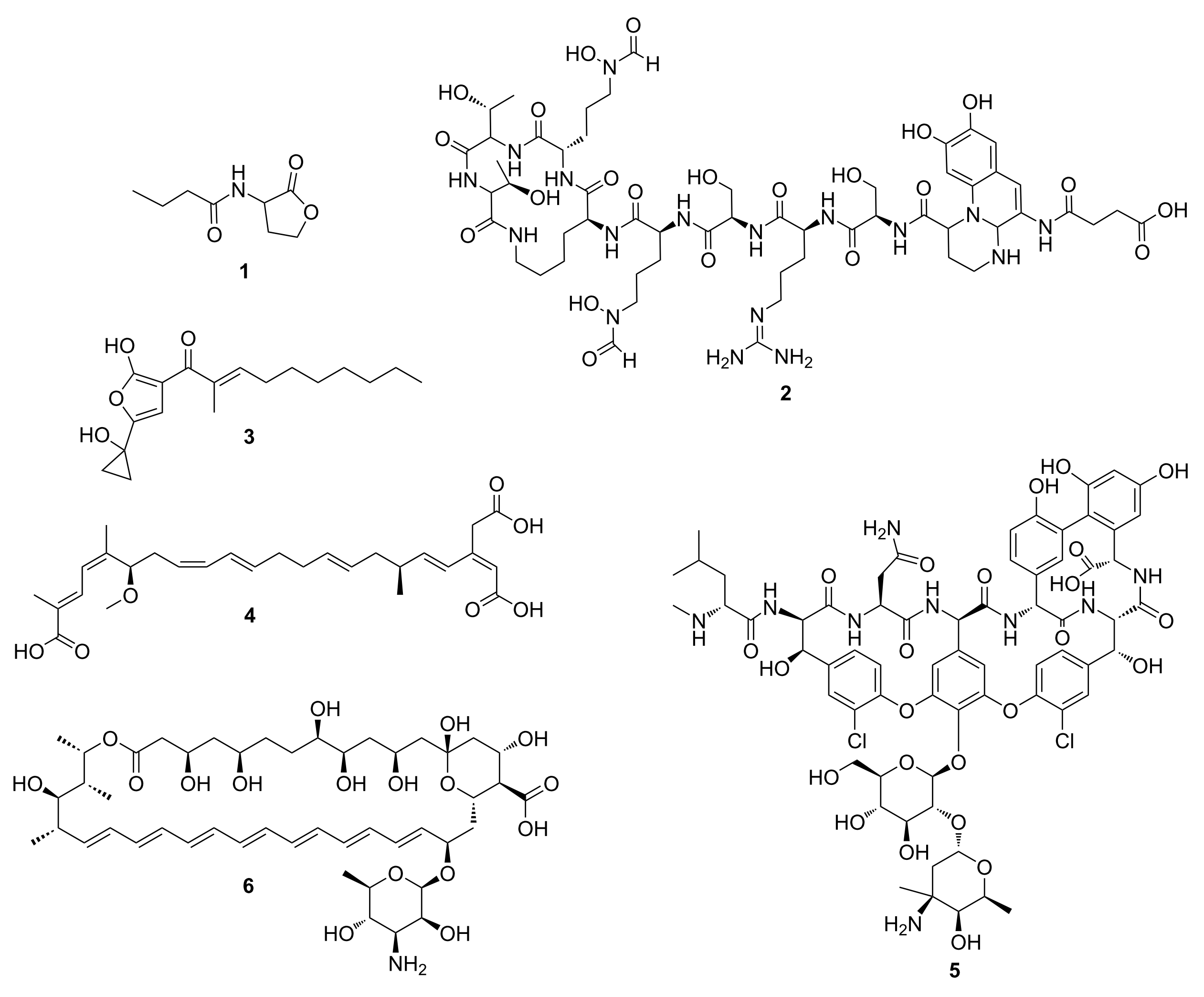


Figure 1: Examples of prominent natural products: *N*-butyrylhomoserine lactone (**1**), pyoverdin (**2**), malleicyprol (**3**), bongkrekic acid (**4**), vancomycin (**5**), and amphotericin B (**6**).

were based on bioactivity-guided discovery strategies, mining microbial genomes revealed a much higher biosynthetic potential than initially anticipated [14]. *Streptomyces hygroscopicus* sp. XM201, for instance, harbors more than 50 putative biosynthetic gene clusters (BGCs), many of which are cryptic, i.e., BGCs for which the corresponding NPs have yet to be identified [15]. A problem when it comes to the characterization of the full biosynthetic potential of an organism is the fact that many BGCs are silent. Silent BGC are not expressed under standard laboratory cultivation conditions as they might lack a specific ecological clue for their expression. As a result, two types of approaches have been developed to unleash this hidden biosynthetic potential. Several pleiotropic (non-targeted, e.g., modifying culturing conditions) and pathway-specific (e.g., heterologous expression or in situ pathway activation) approaches have been developed to awaken silent biosynthetic pathways [16]. Most importantly, however, genome mining can prevent the time-consuming re-discovery of already known metabolites [14]. In-silico dereplication can be performed on two levels: First, BGCs identified by genome mining can be compared to characterized BGCs [17]. Second, in many cases NP core structures can be predicted from genome sequence information and the predicted structures can then be used to search in NP databases for identical or related compounds [18,19]. While the BGC-centric approach might be more accurate, it is limited by the number of characterized BGCs in publicly available databases. Since significantly more NPs than NP BGCs are characterized, the search space of known metabolites is significantly larger than that of experimentally verified BGCs [20]. The accuracy of the predicted core structures on the other hand might restrict the approach.

In this perspective, we will take a closer look at the most commonly used state-of-the-art genome mining tools, ranging from algorithms based on hard-coded rules to machine learning (ML)-based approaches with regard to the natural product biosynthetic principles they are most suited for. We focus on how the different genome mining tools identify BGCs and highlight their advantages and limitations. Moreover, we will showcase two potential strategies for the targeted identification of non-canonical pathways to chart the full biosynthetic potential encoded in bacterial genomes.

Perspective

Natural product biosynthetic principles

NPs are structurally highly diverse and can be divided into several classes depending on their biosynthetic concepts. NP biosynthesis follows two fundamentally different principles: NPs can either be produced in an assembly line-like fashion (Figure 2A) or by discrete, multi-enzymatic assemblies (Figure 2B). Discrete, multi-enzymatic assemblies utilize monofunctional enzymes for the consecutive build-up and decoration of a NP scaffold. In comparison to biosynthetic assembly lines, intermediates are not permanently covalently bound to carrier proteins in discrete, multi-enzymatic assemblies. In both biosynthetic principles, the NP backbone is first assembled by core enzymes and then further modified by tailoring enzymes that decorate the NP scaffold.

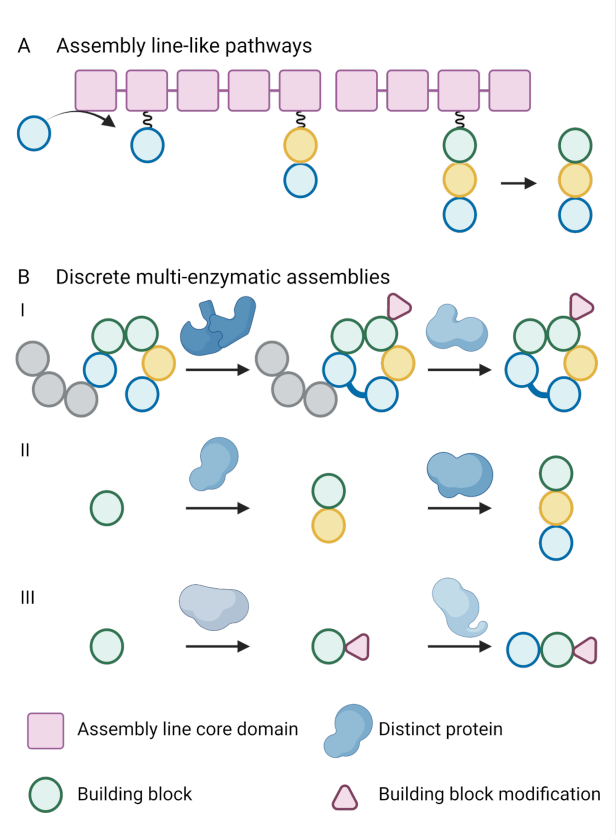

A
Assembly line-like pathways
B
Discrete multi-enzymatic assemblies
I
II
III
Assembly line core domain
Distinct protein
Building block
Building block modification

Figure 2: Biosynthetic principles of (A) assembly line-like pathways and (B) discrete multi-enzymatic assemblies. Assembly line-like pathways use large mega enzymes and generate NPs via the successive addition and/or modification of building blocks (e.g., non-ribosomal peptide biosynthesis) using conserved core domains. In discrete multi-enzymatic assemblies, distinct and mostly monofunctional enzymes catalyze the built-up of the NP scaffold and its decoration (e.g., in (I) ribosomally synthesized and post-translationally modified peptide, (II) terpene, or (III) alkaloid biosynthesis).

Assembly line-like pathways are characterized by mega enzymes, which can be subdivided into modules. Each module is responsible for the incorporation (and/or processing) of one building block into the nascent product. A “textbook” extension module minimally harbors three core domains, responsible for the activation and loading, tethering, and condensation of building blocks and intermediates. The biosynthesis is directional and starts at the N-terminal module with the activation and loading of the first building block onto the assembly line (Figure 2A) [21]. The specificity of the activating domain determines the type of building block incorporated. The growing intermediate stays permanently bound to the assembly line until the final product is released at the C-terminal module. However, modules can also be skipped, used for the modification of the nascent NP rather than its chain extension, or utilized iteratively [22,23]. In textbook assembly line-like pathways, the architecture of the mega enzyme complex correlates with the product structure, a principle that is referred to as the colinearity rule [24]. Examples of these assembly line-like pathways are canonical type I *cis*-acyltransferase polyketide synthases (PKSs) and type A non-ribosomal peptide synthetases (NRPSs) (Figure 2A) [25,26]. The substrate specificity of the specificity conferring domains in each module can be predicted from the sequences of adenylation (A) (for NRPS [26]), acyltransferase (AT) (for *cis*-AT PKS [15]), or ketosynthase (KS) domains (in *trans*-acyltransferase PKS systems [19,27]). Moreover, in the large majority of cases, the gene order within a BGC reflects the order of the corresponding enzymes during the biosynthesis of the associated NP [19]. *trans*-AT PKSs are much more complex than *cis*-AT PKS systems as they harbor non-elongating modules, cryptic domains and seemingly superfluous domains. Moreover, they frequently employ a number of *trans*-acting modifying enzymes, are characterized by modules that are split between proteins and they often harbor non-canonical module architectures and cryptic domains [19,22]. As a result, the colinearity rule cannot be applied to predict *trans*-AT PKS-derived polyketide core structures [19]. Instead, it has been observed that the amino acid sequences of the ketosynthase domains in *trans*-AT PKSs correlate with their substrate specificity [27]. This correlation can be used for the prediction of *trans*-AT PKS-derived polyketide core structures and is referred to as the correlation rule [19]. All commonly occurring domains in assembly line-like NP biosynthetic pathways as well as their non-modular homologs (e.g., type II and III PKSs) show a high degree of sequence homology. For that reason, their sequence can be used by genome mining tools as universal signature sequences to identify the genes encoding the respective domains and the remaining genes of the BGC (Figure 3A and B (e.g., bialaphos (**7**) [11])) [28].

In contrast, discrete multi-enzymatic assemblies utilize distinct, monofunctional enzymes. Examples are terpene (e.g., cyclooctatin (**8**) [29]), ribosomally synthesized and post-translationally modified peptide (RiPP), or NRPS-independent alkaloid pathways. In the case of terpene biosynthesis, terpene cyclases generate the oftentimes multicyclic, hydrocarbon scaffold via a carbocation-mediated cascade reaction [30]. Terpene cyclases are obligatory components of canonical terpene pathways and are used to identify terpene BGCs (Figure 3B) [30,31]. RiPPs, on the other hand, lack genes that are conserved across all 40 plus RiPP families [32]. However, each RiPP BGC family features genes encoding characteristic tailoring enzymes, or precursor peptides, that show a high degree of sequence conservation within the family. These conserved genes can be utilized for the targeted, family-specific identification of RiPP BGCs (Figure 3C (e.g., tryptorubin (**9**) [33])) [21]. In addition, multiple RiPP tailoring enzymes harbor a precursor peptide-binding domain, the so-called RiPP recognition element (RRE) (Figure 3D (e.g., pyrroloquinoline quinone (PQQ, **10**) [34])). RRE-derived signature motifs (i.e., short sequences that are conserved across different types of enzymes and that have a specific function) are used to identify RiPP BGCs beyond family borders as they are present in the BGCs of approximately 50% of all RiPP families [35]. BGCs without conserved signature sequences are almost impossible to identify using current bioinformatic approaches (Figure 3E (e.g., kojic acid (**11**) [36])). Therefore, the prediction of these BGCs is mainly based on the co-localization of adjacent genes encoding tailoring or additional core enzymes.

The current BGC prediction approach has its limitations, as genes involved in the biosynthesis of a NP might be dispersed (i.e., not clustered) throughout the genome and hence cannot be recognized by genome mining algorithms due to the missing proximity of the biosynthetic gene sets (BGSs) (Figure 3F (e.g., pyonitrin (**12**) [37])). NPs whose biosynthesis significantly deviates from the well-established biosynthetic principles (e.g., through the lack of signature sequences) (Figure 3E) [38] are frequently overlooked by state-of-the-art genome mining pipelines. Most genome mining algorithms rely on the identification of signature sequences (Figure 3A–D). As a result, BGCs of the most commonly studied NP classes (e.g., PKS and NRPS BGCs) can be identified with high confidence based on the sequence homology of the commonly occurring biosynthetic domains. Since chemical novelty in assembly line-like pathways is typically obtained through novel arrangements of a limited set of module architectures, a limited diversity of sequential module arrangements, and varying substrate specificities, the probability of identifying truly novel biosynthetic principles and biochemical transformations in these systems is restricted when using hard-coded biosynthetic principles that are based on the detection of the frequently encountered biosynthetic domains [21]. As a result, a lot of effort is currently being put into the development of complementing workflows to chart the “biosynthetic dark matter” (i.e., overlooked biosynthetic pathways) that we currently cannot access bioinformatically [39]. State-of-the-art genome mining tools are ideally suited for the detection of assembly line-like pathways. The focus on these pathways led to a strong bias in training sets: In the MIBiG database of characterized BGCs nearly 80% of all deposited NP BGCs are PKS, NRPS, or terpene BGCs (April 2022) [20]. As the largest database of characterized BGCs,

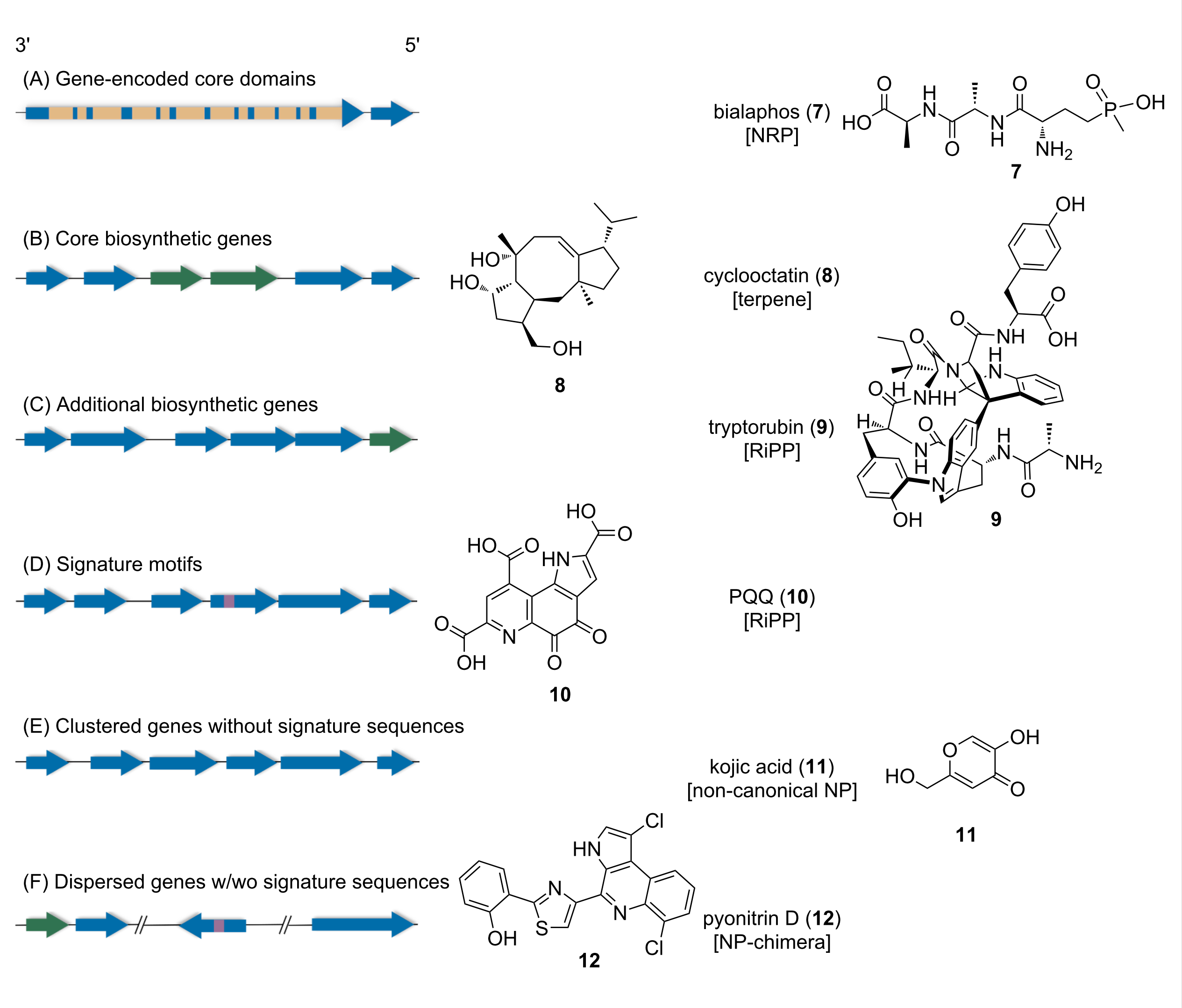

3'
5'
(A) Gene-encoded core domains
bialaphos (7)
[NRP]
7
(B) Core biosynthetic genes
8
cyclooctatin (8)
[terpene]
(C) Additional biosynthetic genes
tryptorubin (9)
[RiPP]
9
(D) Signature motifs
10
PQQ (10)
[RiPP]
(E) Clustered genes without signature sequences
kojic acid (11)
[non-canonical NP]
11
(F) Dispersed genes w/wo signature sequences
12
pyonitrin D (12)
[NP-chimera]

Figure 3: Universal (A, B, F) NP class and NP family-specific (C, D, F) signature sequences in NP BGCs and selected examples of each scenario. Genome mining tools utilize signature sequences to identify NP BGCs derived from (A) gene-encoded conserved core domains of assembly line-like pathways (e.g., modular NRPSs/PKSs), (B) distinct core enzymes (e.g., terpene or type II PKS biosynthesis), (C) tailoring enzymes (e.g., characterized families of RiPP biosynthetic pathways), (D) as well as signature motifs (e.g., RiPP biosynthetic pathways that utilize tailoring enzymes containing RREs). (E) BGCs without signature sequences (e.g., NRPS-independent alkaloid biosynthesis) or (F) genomically dispersed (i.e., not clustered) genes (here also referred to as biosynthetic gene sets) are difficult to identify. The enzymes encoded in core biosynthetic genes are responsible for assembling the NP backbone; additional biosynthetic genes encode tailoring enzymes and other components of a pathway (transporters, regulators, immunity enzymes). Conserved domains are depicted in yellow, genes in green and motifs in violet.

MIBiG is frequently used as a training data set for the development of genome mining algorithms. The imbalanced representation of NP BGCs in the database, however, might introduce a bias when it comes to the training of novel algorithms. Another obstacle to overcome is the efficient mining of the vast quantity of genomic data generated via next-generation sequencing, as a lot of genome mining algorithms are not capable of handling big data [40].

Genome mining principles and tools

Many genome mining tools are based on gene homology and rely on alignments of annotated open reading frames (ORFs). Yet, their purpose, functions, and additional features such as comparative analyses of BGCs, dereplication concepts, or NP structure prediction differ significantly. In addition, some tools implement alternative BGC identification methods like phylogenetic analyses or ML approaches. In many cases, these ML approaches are based on well-established strategies adopted from other disciplines (e.g., natural language processing or comparative genomics) that were adapted by the NP community [41].

In the following section, we will look at representative genome mining tools and discuss their underlying BGC detection principles, along with advantages and limitations of the BGC identification process.

Genome mining algorithms based on hard-coded biosynthetic principles

An early approach to identify NP BGCs in (meta-)genomic data sets were sequence alignments with known genes and domains using algorithms like BLAST (Figure 4) [42]. BLAST detects similar sequences to a given query sequence [42]. The first version of the tool BAGEL utilized BLAST analysis, among others, to identify putative BGCs of bacteriocins (= antimicrobial peptides and proteins) [43-46]. The advantage of such reference alignment methods that are based on sequence homology is their high confidence. The performance of these tools can be rapidly improved via the addition of new reference databases, which was contributing to their success at the beginning of the genome mining era. However, using BLAST-based approaches, the identification of real structural or biosynthetic novelty remains relatively sparse, as the BLAST algorithm is most suitable to detect close homologs of the query sequence. Up to this day, tools like BAGEL are predestined for the rapid and computationally cost-effective characterization of genomic data [43].

Hidden Markov Models (HMMs) are statistical models that are used by the NP community as a more flexible approach to identify BGCs (Figure 4). These models consist of a sequence of “states” (e.g., the occurrences of specific amino acids or nucleotides at a certain position of a protein or DNA sequence, respectively) with pre-determined transition probabilities from one state to the next (e.g., the transition probability in a sequence between one base at a given position to another base at the next position). A sequence of probabilities is calculated from given sequence alignments, for instance, of members of a given gene or protein family. By adding up all possibilities, the likelihood of the complete sequence being a member of the gene family can be calculated [47]. Derivatives of HMMs, so-called profile Hidden Markov Models (pHMMs), are additionally taking gaps and incomplete sequences into consideration. In addition to whole genes or proteins, sequences of conserved key domains of assembly line-like pathways like PKSs (e.g., acyl-carrier-proteins, AT or KS domains) [25] or NRPSs (e.g., peptidyl-carrier-proteins, A domains, condensation (C) domains) [26] are utilized for the generation of pHMMs. The resulting pHMMs recognize signature sequences of such conserved domains in genomic query sequences. pHMMs cannot only be employed to detect and annotate BGCs but also to predict substrate specificities that are essential for NP structure predictions [19,39,48]. After the identification of the core biosynthetic genes, co-localized genes are analyzed and the locus and borders of the BGC are predicted via hard-coded rules based on textbook biosynthetic knowledge, e.g., the minimum amount of domains in a typical NRPS. Due to their seemingly universal biosynthetic principles and modular composition, canonical PKS and NRPS BGCs are predestined for the high confidence detection of their encoded biosynthetic core domains using pHMMs. Structural novelty in these systems that predominantly comprise the same set of conserved domains arises from the novel arrangement of the limited set of different module architectures (e.g., around a dozen in *cis*-AT PKSs vs >150 in *trans*-AT PKSs [19]) along with varying substrate selectivities of specificity-conferring domains (e.g., A domains in NRPSs, AT domains in *cis*-AT PKSs, and KS domains in *trans*-AT PKSs). Moreover, since these assembly line-like pathways follow the same biosynthetic principle, they often form hybrids with other biosynthetic assembly line-like pathways [21].

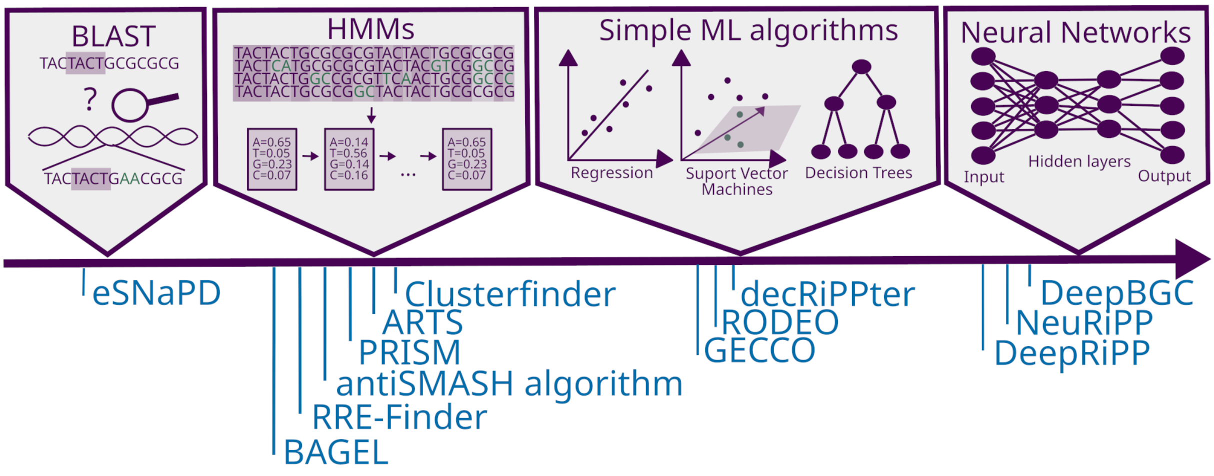


Figure 4: Concepts of algorithms in order of complexity and examples of genome mining tools that employ the respective concept.

Prominent examples of the usage of pHMMs are the original algorithm of the antibiotics & Secondary Metabolite Analysis Shell (antiSMASH) [17,29,49-52] as well as PRediction Informatics for Secondary Metabolomes (PRISM) [18,53-55]. In addition to PKSs and NRPSs, both tools identify a high number of NP classes and families using pHMMs (antiSMASH 6: 876 pHMMs, PRISM 4: 1772 pHMMS). Apart from BGC detection by pHMMs, several stand-alone tools have been implemented into antiSMASH to improve BGC identification, annotation, and substrate predictions (Table 1) (described in detail below). Therefore, we distinguish the original antiSMASH algorithm from the antiSMASH platform (Table 1). Although the BGC

Table 1: Purpose, principles, advantages, and disadvantages of selected genome mining tools. The upper part of the table contains hard-coded tools and the lower part ML-based tools. Novelty refers to the ability of a genome mining tool to chart non-canonical BGCs. Confidence refers to the ability of a genome mining tool to correctly identify a NP BGC.

Tool [first/latest version]	Purpose	BGC identification principles	(Dis-)advantages	Novelty	Confidence
antiSMASH algorithm [17] 2011/2021	identification of a broad range of NP BGC classes and families	pHMMs, hard-coded rules	comprehensive NP class detection	low	high
antiSMASH platform [17] 2011/2021 https://antismash.secondarymetabolites.org	identification of a broad range of NP BGC classes and families, functional and comparative analyses, structure prediction	ClusterFinder[a]: pHMM RODEO: BLAST, pHMM, SVMs RRE-Finder: pHMMs/HHpred database	comprehensive analysis covering many NP classes, dereplication via comparative analysis, usage of NP BGC databases	medium	high
ARTS [59] 2017/2020 http://arts.ziemertlab.com/index	target directed genome mining for antibiotics in bacteria via resistance genes	pHMMs for BGC prediction (antiSMASH), TIGRFAM for detection of housekeeping genes, phylogenetic analysis for identification of horizontal gene transfer	targeted approach for bioactivity	low	high
BAGEL [49] 2006/2018 http://bagel4.molgenrug.nl/	identification of bacterial bacteriocins and RiPPs in (meta-) genomic sequences	BLAST analysis, HMMs, hard-coded rules	restricted to RiPP and bacteriocin BGCs	low	high
CASSIS and SMIPS [60] 2016 https://sbi.hki-jena.de/cassis/	BGC detection in fungi	CASSIS: Density of transcription factor binding sites, SMIPS: Signature sequences	precise cluster borders	low	high
ClusterFinder [61] 2014 Implemented in antiSMASH	BGC detection without functional assignment of NP class	HMM for whole cluster	comprehensive NP class detection	high	low
eSNaPD [62] 2014 http://esnapd2.rockefeller.edu/	BGC detection in non-assembled bacterial metagenomic sequences	BLAST analysis against BGC database	comprehensive NP class detection of smaller BGCs that are similar to known BGCs	low	high
EvoMining [63] 2016/2019 https://github.com/nselem/evomining	identification of BGCs integrating evolutionary principles	phylogenomic analysis in combination with antiSMASH analysis	independent of commonly used signature sequences	medium	medium

Table 1: Purpose, principles, advantages, and disadvantages of selected genome mining tools. The upper part of the table contains hard-coded tools and the lower part ML-based tools. Novelty refers to the ability of a genome mining tool to chart non-canonical BGCs. Confidence refers to the ability of a genome mining tool to correctly identify a NP BGC. (continued)

PRISM [55] 2015/2020 https://prism.adapsyn.com/	identification of a broad range of NP BGCs, structure prediction	HMMs for BGC detection, BLAST analysis, protein motifs and HMMs for domain specificity prediction, support vector machines for activity prediction	comprehensive analysis covering many NP classes, several structure suggestions, dereplication via structural comparisons	low	high
SMURF [64] 2010 http://smurf.jcvi.org/run_smurf.php	identification of fungal BGCs	HMMs, hard-coded rules	comprehensive NP class detection	low	high
transATor [19] 2019	annotation of *trans*-AT PKSs and accurate structure predictions of *trans*-AT PKS-derived polyketides	pHMMs, hard-coded rules	restricted to *trans*-AT PKSs	low	high
decRiPPter [65] 2020 https://github.com/Alexamk/decRiPPter	identification of RiPP BGCs	SVMs, pan-genomic analyses	restricted to RiPP BGCs	medium	medium
DeepBGC [41] 2019 https://github.com/Merck/deepbgc	identification of bacterial and fungal BGCs	neural network with vector- represented Pfam domains (ML)	comprehensive NP class detection	high	medium
DeepRiPP [66] 2019 http://deepripp.magarveylab.ca	identification of RiPP BGCs, structure prediction	natural language processing (deep learning)	restricted to RiPP BGCs	medium	medium
GECCO [67] 2021 https://github.com/zellerlab/GECCO	identification of bacterial and fungal BGCs	conditional random fields	comprehensive NP class detection	high	medium
NeuRiPP [68] 2019 https://github.com/emzodls/neuripp	identification of RiPP precursors	neural networks	restricted to RiPP precursors	medium	medium
RODEO [69] 2017 https://rodeo.scs.illinois.edu/	identification of RiPP BGCs	BLAST analysis of tailoring enzymes, pHMMs, SVMs for precursor detection	restricted to RiPP BGCs	low	medium

[a]ClusterFinder is not available on the antiSMASH web server any longer but is incorporated into the standalone antiSMASH command line tool.

identification approach of antiSMASH and PRISM is quite similar, both tools differ in the downstream processing of the identified BGCs. While antiSMASH focuses on functional and comparative analyses of the biosynthetic genes and BGCs, the focus of PRISM lies on a comprehensive chemical structure prediction of the associated NP [56-58]. In silico dereplication to eliminate BGCs associated with known NPs is one of the major functions of genome mining to avoid the time-consuming and costly re-isolation of known NPs. For instance, the antiSMASH platform compares putative BGCs with reference databases to detect BGCs that are similar to previously characterized BGCs [15,20,58]. However, as many NPs were isolated during the pre-genomic era, they have not been linked to their corresponding BGC. As a result, BGC databases are incomplete which is a drawback when it comes to the dereplication on a gene level. PRISM aims at overcoming this obstacle via retro-biosynthetic building block predictions of known NPs from multiple databases in combination with several BGC-derived NP structure suggestions [58].

To identify RiPP BGCs, the antiSMASH algorithm and PRISM utilize pHMMs based on RiPP-family-specific signature sequences derived from tailoring enzymes or precursor peptides (Figure 3C and D). These family-specific pHMMs are likewise used in tools like BAGEL or RODEO and enable the identification of novel members of known RiPP families [46,69]. RRE-Finder, which is integrated into the antiSMASH platform and RODEO, utilizes the presence of RREs, predicted via pHMMs, to detect RiPP BGCs (Figure 3D). Since the RRE motif is only present in approximately 50% of all RiPP families, it restricts the predictable biosynthetic space. Yet, RRE-Finder is one of the few RiPP genome mining tools which is capable of identifying RiPP BGCs in a family-independent manner [70].

Since the potential of identifying truly novel BGCs via signature sequences is limited, the tool ClusterFinder was developed and implemented into the command line version of antiSMASH [61]. ClusterFinder annotates BGCs via pHMMs from a string of contiguous Pfam domains (protein domains annotated in the protein family database) instead of individual genes. pHMMs are calculated using training sets of known BGCs and non-BGC sequences. Here, two states “BGC” and “non-BGC” are distinguished depending on the Pfam domain frequency in the training data set and the identities of adjacent domains. Consequently, the ClusterFinder algorithm is designed to detect BGCs that are overlooked by other biosynthetic pipelines. As in many other algorithms for the detection of true biosynthetic novelty, high false positive rates have to be taken into consideration, which makes the output of low-confidence/high novelty algorithms more difficult to interpret [61].

An alternative to the above mentioned classical genome mining approaches is the utilization of evolutionary information for the detection of NP BGCs. The EvoMining concept is based on the assumption that secondary metabolite biosynthetic enzymes are distant paralogs of enzymes involved in primary metabolism [63,71]. These NP biosynthetic enzymes are hypothesized to have undergone significant sequence and selectivity changes while still operating based on the same reaction mechanism (e.g., fatty acid biosynthesis → polyketide biosynthesis). As such, NP biosynthetic pathways utilize members of existing enzyme families that have evolved to perform new metabolic functions. Consequently, NP BGCs “borrow” genes encoding paralogs of enzymes that have their origin in primary metabolism and that have diverged into catalyzing alternative metabolic functions. That way, the EvoMining approach identifies members of biosynthetic enzyme families that have likely been repurposed and thus, their corresponding genes are prime targets for a closer inspection of the genomic context to identify new types of BGCs. Although EvoMining is a signature sequence independent concept and instead uses phylogenetic analysis of primary metabolite biosynthetic enzymes, it remains a “hard-coded” sequence similarity-based approach that uses phylogenetic analysis instead of pHMMs for BGC detection [63,71].

Machine learning-based genome mining tools

Some NP BGCs contain solely family-specific features, and lack universal class-specific signature sequences. In these cases, only members of the same subfamily can be identified via pHMMs. An example of the latter are RiPPs that are the most rapidly expanding NP subclass. Eighteen new RiPP families have been characterized over the span of just 8 years, suggesting that many more RiPP families have yet to be discovered [32]. To exploit these currently overlooked biosynthetic treasures, multiple recently developed genome mining tools make use of ML algorithms that have been adapted from other research fields like image recognition [65-68]. Most ML-based tools utilize “supervised learning,” a strategy that employs a dataset with known classifications to train the algorithm [72]. Traditional ML algorithms include regression, decision tree-based classifiers, and support vector machines (SVM), which construct a hyperplane that splits the n-dimensional data-space (i.e., different features/categories serve as dimensions of this space) into different areas that correspond to the different classes (Figure 4) [72]. These algorithms usually lead to robust and interpretable predictions but are limited when it comes to solving complex problems [72].

An example of an advanced combination of different approaches and methods for the identification of RiPPs is the Data-driven Exploratory Class-independent RiPP TrackER (decRiPPter) [65]. decRiPPter uses a support vector machine algorithm trained on a set of known precursor genes to detect RiPP precursor genes semi-independently of their subclass. Subsequently, a pan-genome analysis is performed to identify the corresponding BGCs with the putative RiPP precursor genes as seeds. Putative NP BGCs are identified that are organized in operon-like structures and prioritized based on the taxonomic distribution of the cluster. decRiPPter was successfully used for the identification of a new lanthipeptide subfamily, providing experimental validation of the algorithm [65].

A more advanced form of supervised learning is deep learning (Figure 4). An example of a deep learning architecture is the artificial neural network inspired by the human brain architecture. It consists of artificial neurons processing information organized in different layers and connected by synapses [73]. These advanced algorithms often provide higher accuracy in their prediction but are no longer interpretable as a result of their high level of abstraction [73]. NeuRiPP, for instance, utilizes a parallel convolutional neural network to predict novel RiPP precursor genes independent of their RiPP family. The neural network is trained on a RiPP precursor training set that is based on experimentally verified precursors and precursors predicted by other tools [68]. Both RiPP-specific tools, NeuRiPP and decRiPPter, allow a more flexible BGC identification than hard-coded algorithms but are biased in that the precursor identification depends on training sets consisting of precursors from known RiPP families.

In contrast to NeuRiPP and decRiPPter, DeepBGC is not restricted to a single NP class. Comparable to some hard-coded algorithms, DeepBGC is based on HMM-generated Pfam annotations. However, instead of utilizing Pfam-domains as features, it converts the arrays of Pfam annotations into numeric vectors using a shallow two-layer neural network, an approach adopted from natural language processing [41,72]. These high-dimensional vectors are then used as input for a second two-layer neural network trained on a set of BGC and non-BGC sequences to predict NP BGCs. In the last step, the NP class is predicted using a random forest classifier (Figure 4) [41,74]. DeepBGC outperforms ClusterFinder in its accuracy and false-positive rates due to its ML approach. Like with many other tools, a major disadvantage of DeepBGC is that BGCs lacking canonical biosynthetic domains and small BGCs are filtered out in a pruning stage. Consequently, small BGCs (e.g., biarylitides **15** [75] and tryptorubins **9** [33]) or those that feature solely atypical biosynthetic genes are not recognized, which reduces the likelihood of identifying true biosynthetic novelty [41].

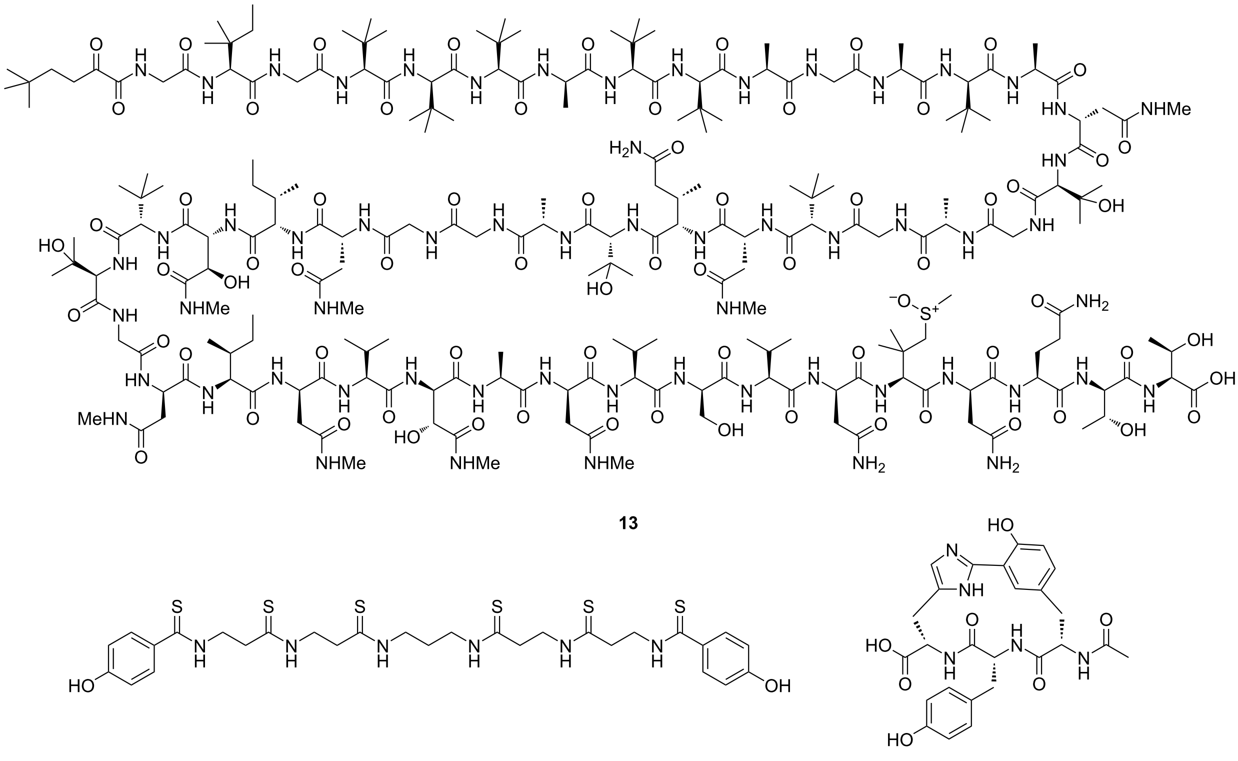
A similar approach is utilized by GECCO, that uses conditional random fields on arrays of Pfam annotations [67]. Conditional random fields belong to the statistical methods and can be classified between HMMs and simpler machine learning algorithms. An advantage of conditional random fields is their interpretability [67]. GECCO outperforms rule-based models in terms of novelty and DeepBGC in terms of accuracy while being less computationally expensive than both [67]. Like DeepBGC, GECCO currently lacks functional proof for the identification of a novel natural product guided by the tool [67].

Challenges and potential solutions to identify currently overlooked BGCs

Genome mining was pivotal for the expansion of NP chemical space in the past two decades. Despite the development of more and more sophisticated genome mining platforms, in many cases where truly novel NP scaffolds were described, the NP was isolated first and only then linked to its corresponding BGC [38]. Notable examples include the (thio-)peptides polytheonamide A (**13**) [76], closthioamide (**15**) [77] (Figure 5), and tryptorubin A (**9**) [33]. It was not until the structure of each of these peptides was determined, that manual retrobiosynthetic analysis resulted in the proposal of biosynthetic models that were subsequently experimentally verified. Once the biosynthesis of a NP is determined using this approach, the NP family can be expanded by developing genome mining algorithms to identify BGCs that follow similar biosynthetic principles [56].

One crucial challenge in the development of novel genome mining tools is balancing novelty and confidence, as one tends to fall short as the other is optimized [39]. On the one hand, genome mining tools that are focused on detecting non-canonical BGCs (high-novelty) are usually characterized by the identification of many putative BGCs that might not be involved in NP biosynthesis (high false-positive rate). These false-positive BGCs are automatically pruned and the resulting putative BGCs need to undergo a second round of manual verification and prioritization prior to functional characterization [39]. On the other hand, hard-coded algorithms detect BGC with high confidence but are restricted when it comes to the identification of BGCs that deviate significantly from what the algorithm's pHMMs have been trained to identify (true biosynthetic novelty) [39]. As most algorithms are at least to some extent signature sequence or sequence homology based, they heavily rely on the sequence space of known BGCs. The bias of hard-coded algorithms is embedded in the biosynthetic rules used for BGC detection and the dataset used to create pHMMs. The bias of ML-based algorithms results from their training sets that usually consist of characterized, canonical BGCs that are then used for the targeted identification of non-canonical BGCs [39]. Utilizing fewer gene family-based features, like the occurrence of Pfam domains or the sequence itself, for predictions can help to avoid overfitting, i.e., the problem of getting the algorithm to perform very well on the training data but underperform on unseen data [39].

Most genome mining algorithms rely on functionally annotated ORFs for the prediction of BGCs. State-of-the-art genome annotation algorithms are not yet able to recognize all ORFs



13

14

15

Figure 5: Examples of peptide NPs, the corresponding BGCs of which were determined through retrobiosynthetic analysis and then experimentally verified: polytheonamide A (**13**) [76], closthioamide (**14**) [77], and biarylitide YYH (**15**) [33].

correctly, especially very short ORFs like RiPP precursor genes [78]. Combined with many false ORF annotations, missing annotations impair BGC predictions downstream of the annotation process. Moreover, the BGCs of certain NP families are inherently easier to identify than others. For example, domains of canonical NRPSs and PKSs can be identified by signature sequence-based pHMMs with high confidence (Figure 3). Furthermore, pHMMs of conserved domains can be subdivided into dozens of individual pHMMs used to determine the substrate specificity of a conserved domain [17,19]. However, BGCs lacking known signature sequences are inherently more difficult to identify. In addition, the size of the BGC of interest impacts the predictive power of the algorithms: Extremely small BGCs, harboring only a few genes, are frequently overlooked as they usually do not pass hard-coded thresholds. For instance, the 1.2 kb gene cluster linked to tryptorubin (**9**) biosynthesis only encodes a 26 amino acid precursor peptide and a single cytochrome P450 monooxygenase [33,79], and hence it was overlooked by genome mining algorithms. On the other hand, large PKS or NRPS BGCs can be split across multiple contigs. This mosaic-like distribution of a single BGC makes the identification of the entire BGC a challenging endeavor especially if multiple assembly line-like BGCs are present in a genome. Moreover, the quality of assembled genomes obtained from short reads decreases with highly repetitive sequences present in many large PKS or NRPS genes [39].

Although the traditional hard-coded rule- and ML-based approaches differ fundamentally when it comes to the implementation of the respective NP BGC identification, they are both based on the same principle: The direct identification of NP BGCs. Both approaches heavily rely on training sets to generate pHMMs or to train the respective ML algorithm. As a consequence, they are both hypothesis-driven approaches resulting in an inherent bias “to identify what the algorithm was trained to identify” rather than to chart the entire biosynthetic space. This bias is largely based on the fact that both approaches use the characterized NP biosynthetic space as a training set for its expansion. Even though there might be no truly unbiased approach towards the expansion of NP biosynthetic space, indirect NP BGC detection methods might be capable of complementing the current strategies. These indirect approaches are exclusively based on the assumption that NP biosynthetic genes are clustered in microbial genomes (even though this might not be true for all NP biosynthetic pathways) and do not require prior knowledge about characterized biosynthetic pathways as training data sets. Below, we are showcasing two putative solutions to complement existing approaches to expand NP biosynthetic space and to chart biosynthetic dark matter.

Genome-wide characterization of all clustered genes as an approach to identify non-canonical pathways

One concept that is based on the above outlined indirect approach is the genome-wide characterization of all clustered genes (gcBGC). In comparison to state-of-the-art genome mining tools, gcBGC inverts the current BGC identification process. Instead of identifying NP BGCs, all clustered genes involved in primary and secondary metabolite biosynthesis are identified. To specifically target non-canonical BGCs, BGCs that can be unambiguously assigned to primary metabolism and those BGCs that are detected by state-of-the-art genome mining pipelines are filtered out. Based on the initial hypothesis underlying the gcBGC approach, the remaining BGCs are likely involved in non-canonical NP biosynthesis (“biosynthetic dark matter” in Figure 6).

The gcBGC concept is based on the assumption that secondary metabolite BGCs evolve from primary metabolite biosynthetic pathways, and that the transition between both is fluid [71]. First, gcBGC identifies all clustered genes in a signature sequence-independent manner via analysis of operon-like structures (e.g., promoters or transcription start sites) as shown in fungi by the tool CASSIS/SMIPS [60]. This concept contrasts the commonly used principles that rely on the direct detection of genes via (p)HMMs- or ML-based approaches, both of which typically require a training data set.

As this approach leads to the identification of a large number of primary and secondary metabolite BGCs that are likewise detected by state-of-the-art genome mining pipelines, a filtering step is required to prioritize the putative non-canonical BGCs that are currently overlooked by existing genome mining tools [17,41,55,80] (Figure 6). Moreover, additional information on taxonomic relationships, pan-genome analyses, or whole-genome comparisons of all members of the pan-genome can be used for further prioritization (Figure 6) [81]. gcBGC is restricted to well-studied organisms where primary metabolite gene cassettes can be confidentially identified. However, the inverted BGC identification concept combined with the focus on as-of-yet unidentified BGCs suggests gcBGC-like approaches to be promising alternatives for the detection of non-canonical pathways.

A comparative genomics approach to identify non-canonical BGCs

Another concept for the expansion of NP biosynthetic space is based on a Comparative Genomics Approach (CGA). This approach relies on the fact that many BGCs are introduced into microbial genomes via horizontal gene transfer (HGT). A genome can be subdivided into groups of genes called syntenic blocks [82]. Among related strains, the order of these syntenic blocks, as well as their gene composition, is highly similar

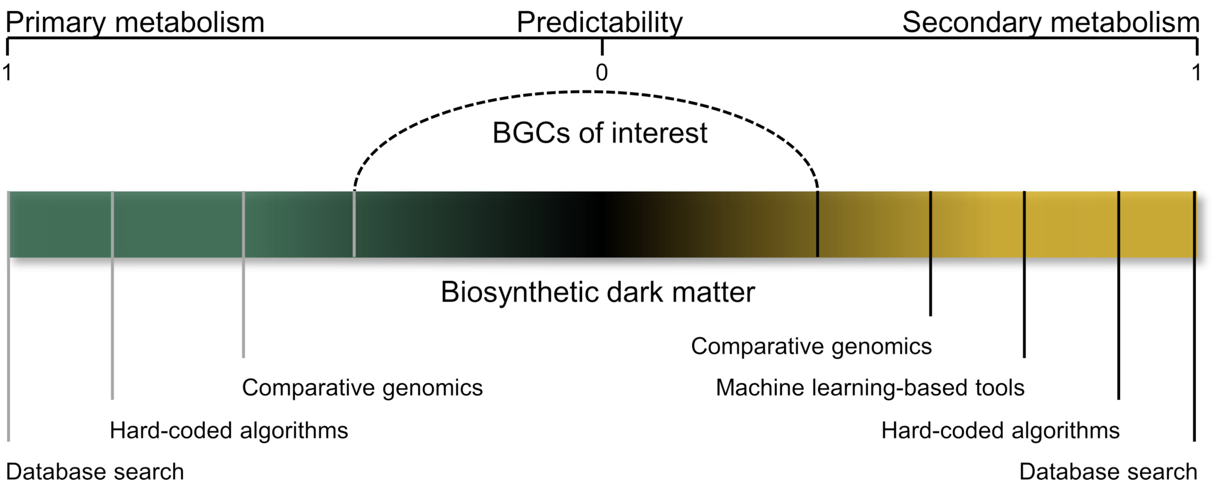


Figure 6: gcBGC prioritization process. The bar represents all identified gene clusters in one organism and is subdivided into primary (green) on the left and secondary metabolism (yellow) on the right. Gene clusters associated with primary or secondary metabolism (both ends of the spectrum) are identified using the strategies listed below the bar and subsequently excluded from further analysis. gcBGC aims at identifying as-of-yet overlooked BGCs that cannot be detected by state-of-the-art bioinformatic platforms, here referred to as “biosynthetic dark matter.”

(Figure 7). Evolutionary young HGT events in single/few strains can disrupt this order, leading to the insertion of non-syntenic blocks (Figure 7) [82]. These insertions are detectable by comparing multiple closely related strains utilizing whole genome alignments, a technique adopted from the field of comparative genomics [83]. In a recent study, 10 *Aspergillus* genomes were compared to identify BGCs in non-syntenic blocks, leading to the confirmation of all previously known BGCs using the CGA concept [84]. As a proof of concept, the previously characterized kojic acid (**11**) BGC, which escaped detection by state-of-the-art genome mining algorithms, was identified [84]. The kojic acid (**11**) BGC lacks the classical biosynthetic signature sequences typically used for BGC identification, thus showing the potential of the approach (Figure 3) [84].

CGA aims at scaling this approach and comparing all sequenced strains of one genus (e.g., *Streptomyces*) to find non-syntenic blocks that might code for NP BGCs. Comparable to the genome-wide characterization of all clustered genes concept, CGA focuses on BGC detection independently of signature sequences and known NP families to expand the known NP chemical space via the identification of non-canonical pathways.

The first step of CGA consists of the homogenous functional annotation of all genes of the selected genomes to reduce false positive rates of non-syntenic blocks due to different annotations of genes using different annotation algorithms. Subsequently, all annotated genes are clustered based on sequence similarity to improve functional annotations [85]. The obtained sequential arrangements of gene annotations representing the different genomes are aligned to compare the genomes not on a sequence level, but instead on the gene-function level [86]. Whole genome alignments are performed to detect single diverging gene loci that are subsequently expanded by their genomic neighborhood to detect genomic islands. Therefore, the genomic neighborhoods of the identified genes are analyzed for differences in their synteny to detect HGT regions composed of multiple genes. In addition, these regions are analyzed for genetic characteristics like promoters or transposase genes to identify operon-like structures. Single gene duplication events are filtered out and all known BGCs are excluded in a similar fashion as in the gcBGC approach. The presence of prototypical tailoring enzymes ubiquitously distributed in secondary metabolism might serve as an additional line of evidence for a functional NP BGC. This approach is computationally expensive yet feasible with the availability of high-performance computer clusters but requires excellent quality of the analyzed genomes. This method has the advantage over simpler approaches to detect HGT events, like for example comparing GC contents of different regions, that it can be used to detect HGT events from closely related strains.

Conclusion

The development of next-generation sequencing technologies [4] and the resulting availability of a seemingly exponentially increasing number of genome sequences enabled or revolutionized several biological fields including comparative genomics [81], functional genomics [87], and NP-genome mining [88]. From simple BLAST analyses through pHMM-based algorithms to ML-based approaches, genome mining is a continuously evolving field that has benefited from other disciplines, such as mathematics, image processing, or linguistics. State-of-the-art sequence homology- and ML-based genome mining tools identify BGCs that share even low levels of similarity with known BGCs with high confidence. Traditional pHMMs-based approaches are ideally suited to chart the biosynthetic space of

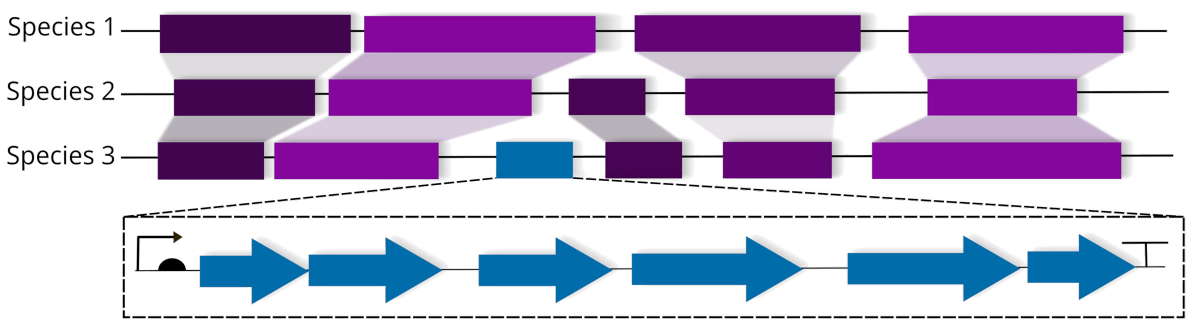


Figure 7: Genome alignments of related organisms revealing the presence of syntenic (purple) as well as non-syntenic blocks (blue). The order of syntenic blocks can be disrupted by putative HGT events, leading to the integration of non-syntenic blocks that are subsequently screened for operon-like structures containing multiple continuous gene arrangements (blue) without large gaps.

assembly line-like pathways that are typically composed of novel arrangements of recurring module architectures with varying specifications of the substrate specificity-conferring domains. ML-based approaches on the other hand are more frequently employed to target non-homogeneous NP classes such as RiPPs whose BGCs do not share sequence homologies across all 40 plus RiPP-families and to identify NP BGCs that are currently overlooked by state-of-the-art sequence homology-based tools. Even though the scope and implementation of both approaches differs significantly, the underlying concept is the same: The direct, hypothesis-driven identification of clustered NP biosynthetic genes based on a training data set that requires a database of characterized BGCs. This training data set might comprise individual domains from characterized pathways to generate pHMMs to complex features that are extracted from characterized BGCs. The bias introduced through the dependence on these reference datasets is likely to result in an inherent limitation when it comes to the identification of truly non-canonical pathways that share low to no similarity to characterized pathways. To address these limitations, indirect approaches that do not rely on training data sets of characterized BGCs might be capable of complementing the current suite of highly sophisticated genome mining tools as they might be ideally suited to identify non-canonical pathways that are overlooked by direct identification approaches. We showcased two such hypothetical indirect approaches that we named “genome-wide characterization of all clustered genes” and “comparative genomics-based identification of non-canonical BGCs”. These indirect BGC detection concepts are solely based on the assumption that biosynthetic genes are clustered in bacterial genomes. Both approaches are based on the sequence similarity-independent identification of non-canonical BGCs via recognition of operon-like structures or usage of comparative genomics to detect horizontally transferred gene clusters. In a subsequent prioritization step, clustered genes that are involved in primary metabolite biosynthesis or that can be likewise detected by state-of-the-art genome mining pipelines can be excluded to target uncharted biosynthetic space also referred to as biosynthetic dark matter. These indirect concepts might serve as an inspiration for further innovative tools for the targeted discovery of hidden biosynthetic treasures.

Acknowledgements

Figure 1 and the graphical abstract were created in BioRender.

Funding

EJNH gratefully acknowledges funding by the LOEWE Center for Translational Biodiversity Genomics (LOEWE TBG) and the German Chemical Industry Association. FB acknowledges funding from a Kekulé Fellowship of the German Chemical Industry Association.

ORCID® iDs

Friederike Biermann - https://orcid.org/0000-0002-9152-2562
Eric J. N. Helfrich - https://orcid.org/0000-0001-8751-3279

References

1. Bentley, S. D.; Chater, K. F.; Cerdeño-Tárraga, A.-M.; Challis, G. L.; Thomson, N. R.; James, K. D.; Harris, D. E.; Quail, M. A.; Kieser, H.; Harper, D.; Bateman, A.; Brown, S.; Chandra, G.; Chen, C. W.; Collins, M.; Cronin, A.; Fraser, A.; Goble, A.; Hidalgo, J.; Hornsby, T.; Howarth, S.; Huang, C.-H.; Kieser, T.; Larke, L.; Murphy, L.; Oliver, K.; O'Neil, S.; Rabbinowitsch, E.; Rajandream, M.-A.; Rutherford, K.; Rutter, S.; Seeger, K.; Saunders, D.; Sharp, S.; Squares, R.; Squares, S.; Taylor, K.; Warren, T.; Wietzorrek, A.; Woodward, J.; Barrell, B. G.; Parkhill, J.; Hopwood, D. A. *Nature* **2002,** *417,* 141–147. doi:10.1038/417141a
2. Ikeda, H.; Ishikawa, J.; Hanamoto, A.; Shinose, M.; Kikuchi, H.; Shiba, T.; Sakaki, Y.; Hattori, M.; Ōmura, S. *Nat. Biotechnol.* **2003,** *21,* 526–531. doi:10.1038/nbt820
3. Bachmann, B. O.; Van Lanen, S. G.; Baltz, R. H. *J. Ind. Microbiol. Biotechnol.* **2014,** *41,* 175–184. doi:10.1007/s10295-013-1389-9
4. Goodwin, S.; McPherson, J. D.; McCombie, W. R. *Nat. Rev. Genet.* **2016,** *17,* 333–351. doi:10.1038/nrg.2016.49
5. van Dijk, E. L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. *Trends Genet.* **2018,** *34,* 666–681. doi:10.1016/j.tig.2018.05.008
6. Waters, C. M.; Bassler, B. L. *Annu. Rev. Cell Dev. Biol.* **2005,** *21,* 319–346. doi:10.1146/annurev.cellbio.21.012704.131001
7. Wendenbaum, S.; Demange, P.; Dell, A.; Meyer, J. M.; Abdallah, M. A. *Tetrahedron Lett.* **1983,** *24,* 4877–4880. doi:10.1016/s0040-4039(00)94031-0
8. Biggins, J. B.; Ternei, M. A.; Brady, S. F. *J. Am. Chem. Soc.* **2012,** *134,* 13192–13195. doi:10.1021/ja3052156
9. Franke, J.; Ishida, K.; Hertweck, C. *Angew. Chem., Int. Ed.* **2012,** *51,* 11611–11615. doi:10.1002/anie.201205566
10. Trottmann, F.; Franke, J.; Richter, I.; Ishida, K.; Cyrulies, M.; Dahse, H.-M.; Regestein, L.; Hertweck, C. *Angew. Chem., Int. Ed.* **2019,** *58,* 14129–14133. doi:10.1002/anie.201907324
11. Moebius, N.; Ross, C.; Scherlach, K.; Rohm, B.; Roth, M.; Hertweck, C. *Chem. Biol.* **2012,** *19,* 1164–1174. doi:10.1016/j.chembiol.2012.07.022
12. Felnagle, E. A.; Jackson, E. E.; Chan, Y. A.; Podevels, A. M.; Berti, A. D.; McMahon, M. D.; Thomas, M. G. *Mol. Pharmaceutics* **2008,** *5,* 191–211. doi:10.1021/mp700137g
13. Caffrey, P.; Lynch, S.; Flood, E.; Finnan, S.; Oliynyk, M. *Chem. Biol.* **2001,** *8,* 713–723. doi:10.1016/s1074-5521(01)00046-1
14. Baltz, R. H. *J. Ind. Microbiol. Biotechnol.* **2019,** *46,* 281–299. doi:10.1007/s10295-018-2115-4
15. Kautsar, S. A.; Blin, K.; Shaw, S.; Weber, T.; Medema, M. H. *Nucleic Acids Res.* **2021,** *49,* D490–D497. doi:10.1093/nar/gkaa812
16. Liu, Z.; Zhao, Y.; Huang, C.; Luo, Y. *Front. Bioeng. Biotechnol.* **2021,** *9,* 632230. doi:10.3389/fbioe.2021.632230
17. Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; van Wezel, G. P.; Medema, M. H.; Weber, T. *Nucleic Acids Res.* **2021,** *49,* W29–W35. doi:10.1093/nar/gkab335
18. Skinnider, M. A.; Dejong, C. A.; Rees, P. N.; Johnston, C. W.; Li, H.; Webster, A. L. H.; Wyatt, M. A.; Magarvey, N. A. *Nucleic Acids Res.* **2015,** *43,* 9645–9662. doi:10.1093/nar/gkv1012

19. Helfrich, E. J. N.; Ueoka, R.; Dolev, A.; Rust, M.; Meoded, R. A.; Bhushan, A.; Califano, G.; Costa, R.; Gugger, M.; Steinbeck, C.; Moreno, P.; Piel, J. *Nat. Chem. Biol.* **2019,** *15,* 813–821. doi:10.1038/s41589-019-0313-7
20. Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Muñoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Suarez Duran, H. G.; Pascal Andreu, V.; Selem-Mojica, N.; Alanjary, M.; Robinson, S. L.; Lund, G.; Epstein, S. C.; Sisto, A. C.; Charkoudian, L. K.; Collemare, J.; Linington, R. G.; Weber, T.; Medema, M. H. *Nucleic Acids Res.* **2020,** *48,* D454–D458. doi:10.1093/nar/gkz882
21. Wenski, S. L.; Thiengmag, S.; Helfrich, E. J. N. *Synth. Syst. Biotechnol.* **2022,** *7,* 631–647. doi:10.1016/j.synbio.2022.01.007
22. Helfrich, E. J. N.; Piel, J. *Nat. Prod. Rep.* **2016,** *33,* 231–316. doi:10.1039/c5np00125k
23. He, J.; Hertweck, C. *ChemBioChem* **2005,** *6,* 908–912. doi:10.1002/cbic.200400333
24. Traitcheva, N.; Jenke-Kodama, H.; He, J.; Dittmann, E.; Hertweck, C. *ChemBioChem* **2007,** *8,* 1841–1849. doi:10.1002/cbic.200700309
25. Hertweck, C. *Angew. Chem., Int. Ed.* **2009,** *48,* 4688–4716. doi:10.1002/anie.200806121
26. Süssmuth, R. D.; Mainz, A. *Angew. Chem., Int. Ed.* **2017,** *56,* 3770–3821. doi:10.1002/anie.201609079
27. Nguyen, T.; Ishida, K.; Jenke-Kodama, H.; Dittmann, E.; Gurgui, C.; Hochmuth, T.; Taudien, S.; Platzer, M.; Hertweck, C.; Piel, J. *Nat. Biotechnol.* **2008,** *26,* 225–233. doi:10.1038/nbt1379
28. Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Res.* **2011,** *39* (Suppl. 2), W339–W346. doi:10.1093/nar/gkr466
29. Aoyama, T.; Naganawa, H.; Muraoka, Y.; Aoyagi, T.; Takeuchi, T. *J. Antibiot.* **1992,** *45,* 1703–1704. doi:10.7164/antibiotics.45.1703
30. Helfrich, E. J. N.; Lin, G.-M.; Voigt, C. A.; Clardy, J. *Beilstein J. Org. Chem.* **2019,** *15,* 2889–2906. doi:10.3762/bjoc.15.283
31. Dickschat, J. S. *Nat. Prod. Rep.* **2016,** *33,* 87–110. doi:10.1039/c5np00102a
32. Montalbán-López, M.; Scott, T. A.; Ramesh, S.; Rahman, I. R.; van Heel, A. J.; Viel, J. H.; Bandarian, V.; Dittmann, E.; Genilloud, O.; Goto, Y.; Grande Burgos, M. J.; Hill, C.; Kim, S.; Koehnke, J.; Latham, J. A.; Link, A. J.; Martínez, B.; Nair, S. K.; Nicolet, Y.; Rebuffat, S.; Sahl, H.-G.; Sareen, D.; Schmidt, E. W.; Schmitt, L.; Severinov, K.; Süssmuth, R. D.; Truman, A. W.; Wang, H.; Weng, J.-K.; van Wezel, G. P.; Zhang, Q.; Zhong, J.; Piel, J.; Mitchell, D. A.; Kuipers, O. P.; van der Donk, W. A. *Nat. Prod. Rep.* **2021,** *38,* 130–239. doi:10.1039/d0np00027b
33. Reisberg, S. H.; Gao, Y.; Walker, A. S.; Helfrich, E. J. N.; Clardy, J.; Baran, P. S. *Science* **2020,** *367,* 458–463. doi:10.1126/science.aay9981
34. Evans, R. L., III; Latham, J. A.; Xia, Y.; Klinman, J. P.; Wilmot, C. M. *Biochemistry* **2017,** *56,* 2735–2746. doi:10.1021/acs.biochem.7b00247
35. Burkhart, B. J.; Hudson, G. A.; Dunbar, K. L.; Mitchell, D. A. *Nat. Chem. Biol.* **2015,** *11,* 564–570. doi:10.1038/nchembio.1856
36. Terabayashi, Y.; Sano, M.; Yamane, N.; Marui, J.; Tamano, K.; Sagara, J.; Dohmoto, M.; Oda, K.; Ohshima, E.; Tachibana, K.; Higa, Y.; Ohashi, S.; Koike, H.; Machida, M. *Fungal Genet. Biol.* **2010,** *47,* 953–961. doi:10.1016/j.fgb.2010.08.014
37. Mevers, E.; Saurí, J.; Helfrich, E. J. N.; Henke, M.; Barns, K. J.; Bugni, T. S.; Andes, D.; Currie, C. R.; Clardy, J. *J. Am. Chem. Soc.* **2019,** *141,* 17098–17101. doi:10.1021/jacs.9b09739
38. Biermann, F.; Helfrich, E. J. N. *mSystems* **2021,** *6,* e00846-21. doi:10.1128/msystems.00846-21
39. Blin, K.; Kim, H. U.; Medema, M. H.; Weber, T. *Briefings Bioinf.* **2019,** *20,* 1103–1113. doi:10.1093/bib/bbx146
40. Baltz, R. H. *J. Ind. Microbiol. Biotechnol.* **2021,** *48,* kuab044. doi:10.1093/jimb/kuab044
41. Hannigan, G. D.; Prihoda, D.; Palicka, A.; Soukup, J.; Klempir, O.; Rampula, L.; Durcak, J.; Wurst, M.; Kotowski, J.; Chang, D.; Wang, R.; Piizzi, G.; Temesi, G.; Hazuda, D. J.; Woelk, C. H.; Bitton, D. A. *Nucleic Acids Res.* **2019,** *47,* e110. doi:10.1093/nar/gkz654
42. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990,** *215,* 403–410. doi:10.1016/s0022-2836(05)80360-2
43. van Heel, A. J.; de Jong, A.; Song, C.; Viel, J. H.; Kok, J.; Kuipers, O. P. *Nucleic Acids Res.* **2018,** *46,* W278–W281. doi:10.1093/nar/gky383
44. de Jong, A.; van Heel, A. J.; Kok, J.; Kuipers, O. P. *Nucleic Acids Res.* **2010,** *38* (Suppl. 2), W647–W651. doi:10.1093/nar/gkq365
45. van Heel, A. J.; de Jong, A.; Montalbán-López, M.; Kok, J.; Kuipers, O. P. *Nucleic Acids Res.* **2013,** *41,* W448–W453. doi:10.1093/nar/gkt391
46. de Jong, A.; van Hijum, S. A. F. T.; Bijlsma, J. J. E.; Kok, J.; Kuipers, O. P. *Nucleic Acids Res.* **2006,** *34,* W273–W279. doi:10.1093/nar/gkl237
47. Eddy, S. R. *Bioinformatics* **1998,** *14,* 755–763. doi:10.1093/bioinformatics/14.9.755
48. Röttig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. *Nucleic Acids Res.* **2011,** *39* (Suppl. 2), W362–W367. doi:10.1093/nar/gkr323
49. Blin, K.; Medema, M. H.; Kazempour, D.; Fischbach, M. A.; Breitling, R.; Takano, E.; Weber, T. *Nucleic Acids Res.* **2013,** *41,* W204–W212. doi:10.1093/nar/gkt449
50. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Müller, R.; Wohlleben, W.; Breitling, R.; Takano, E.; Medema, M. H. *Nucleic Acids Res.* **2015,** *43,* W237–W243. doi:10.1093/nar/gkv437
51. Blin, K.; Wolf, T.; Chevrette, M. G.; Lu, X.; Schwalen, C. J.; Kautsar, S. A.; Suarez Duran, H. G.; de los Santos, E. L. C.; Kim, H. U.; Nave, M.; Dickschat, J. S.; Mitchell, D. A.; Shelest, E.; Breitling, R.; Takano, E.; Lee, S. Y.; Weber, T.; Medema, M. H. *Nucleic Acids Res.* **2017,** *45,* W36–W41. doi:10.1093/nar/gkx319
52. Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S. Y.; Medema, M. H.; Weber, T. *Nucleic Acids Res.* **2019,** *47,* W81–W87. doi:10.1093/nar/gkz310
53. Skinnider, M. A.; Merwin, N. J.; Johnston, C. W.; Magarvey, N. A. *Nucleic Acids Res.* **2017,** *45,* W49–W54. doi:10.1093/nar/gkx320
54. Skinnider, M. A.; Johnston, C. W.; Edgar, R. E.; Dejong, C. A.; Merwin, N. J.; Rees, P. N.; Magarvey, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2016,** *113,* E6343–E6351. doi:10.1073/pnas.1609014113
55. Skinnider, M. A.; Johnston, C. W.; Gunabalasingam, M.; Merwin, N. J.; Kieliszek, A. M.; MacLellan, R. J.; Li, H.; Ranieri, M. R. M.; Webster, A. L. H.; Cao, M. P. T.; Pfeifle, A.; Spencer, N.; To, Q. H.; Wallace, D. P.; Dejong, C. A.; Magarvey, N. A. *Nat. Commun.* **2020,** *11,* 6058. doi:10.1038/s41467-020-19986-1
56. Medema, M. H.; de Rond, T.; Moore, B. S. *Nat. Rev. Genet.* **2021,** *22,* 553–571. doi:10.1038/s41576-021-00363-7
57. Johnston, C. W.; Skinnider, M. A.; Wyatt, M. A.; Li, X.; Ranieri, M. R. M.; Yang, L.; Zechel, D. L.; Ma, B.; Magarvey, N. A. *Nat. Commun.* **2015,** *6,* 8421. doi:10.1038/ncomms9421

58. Dejong, C. A.; Chen, G. M.; Li, H.; Johnston, C. W.; Edwards, M. R.; Rees, P. N.; Skinnider, M. A.; Webster, A. L. H.; Magarvey, N. A. *Nat. Chem. Biol.* **2016,** *12,* 1007–1014. doi:10.1038/nchembio.2188
59. Mungan, M. D.; Alanjary, M.; Blin, K.; Weber, T.; Medema, M. H.; Ziemert, N. *Nucleic Acids Res.* **2020,** *48,* W546–W552. doi:10.1093/nar/gkaa374
60. Wolf, T.; Shelest, V.; Nath, N.; Shelest, E. *Bioinformatics* **2016,** *32,* 1138–1143. doi:10.1093/bioinformatics/btv713
61. Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Wieland Brown, L. C.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. *Cell* **2014,** *158,* 412–421. doi:10.1016/j.cell.2014.06.034
62. Reddy, B. V. B.; Milshteyn, A.; Charlop-Powers, Z.; Brady, S. F. *Chem. Biol.* **2014,** *21,* 1023–1033. doi:10.1016/j.chembiol.2014.06.007
63. Sélem-Mojica, N.; Aguilar, C.; Gutiérrez-García, K.; Martínez-Guerrero, C. E.; Barona-Gómez, F. *Microb. Genomics* **2019,** *5,* e000260. doi:10.1099/mgen.0.000260
64. Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D. *Fungal Genet. Biol.* **2010,** *47,* 736–741. doi:10.1016/j.fgb.2010.06.003
65. Kloosterman, A. M.; Cimermancic, P.; Elsayed, S. S.; Du, C.; Hadjithomas, M.; Donia, M. S.; Fischbach, M. A.; van Wezel, G. P.; Medema, M. H. *PLoS Biol.* **2020,** *18,* e3001026. doi:10.1371/journal.pbio.3001026
66. Merwin, N. J.; Mousa, W. K.; Dejong, C. A.; Skinnider, M. A.; Cannon, M. J.; Li, H.; Dial, K.; Gunabalasingam, M.; Johnston, C.; Magarvey, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2020,** *117,* 371–380. doi:10.1073/pnas.1901493116
67. Carroll, L. M.; Larralde, M.; Fleck, J. S.; Ponnudurai, R.; Milanese, A.; Cappio, E.; Zeller, G. *bioRxiv* **2021**. doi:10.1101/2021.05.03.442509
68. de los Santos, E. L. C. *Sci. Rep.* **2019,** *9,* 13406. doi:10.1038/s41598-019-49764-z
69. Tietz, J. I.; Schwalen, C. J.; Patel, P. S.; Maxson, T.; Blair, P. M.; Tai, H.-C.; Zakai, U. I.; Mitchell, D. A. *Nat. Chem. Biol.* **2017,** *13,* 470–478. doi:10.1038/nchembio.2319
70. Kloosterman, A. M.; Shelton, K. E.; van Wezel, G. P.; Medema, M. H.; Mitchell, D. A. *mSystems* **2020,** *5,* e00267–20. doi:10.1128/msystems.00267-20
71. de los Santos, E. L. C. *Sci. Rep.* **2019,** *9,* 13406. doi:10.1038/s41598-019-49764-z
72. Nasteski, V. *Horizons* **2017,** *4,* 51–62.
73. Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. *J. Big Data* **2021,** *8,* 53. doi:10.1186/s40537-021-00444-8
74. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. *arXiv* **2013,** 1301.3781. doi:10.48550/arxiv.1301.3781
75. Zdouc, M. M.; Alanjary, M. M.; Zarazúa, G. S.; Maffioli, S. I.; Crüsemann, M.; Medema, M. H.; Donadio, S.; Sosio, M. *Cell Chem. Biol.* **2021,** *28,* 733–739.e4. doi:10.1016/j.chembiol.2020.11.009
76. Oiki, S.; Muramatsu, I.; Matsunaga, S.; Fusetani, N. *Folia Pharmacol. Jpn.* **1997,** *110,* 195–198. doi:10.1254/fpj.110.supplement_195
77. Lincke, T.; Behnken, S.; Ishida, K.; Roth, M.; Hertweck, C. *Angew. Chem., Int. Ed.* **2010,** *49,* 2011–2013. doi:10.1002/anie.200906114
78. Chang, Y.-C.; Hu, Z.; Rachlin, J.; Anton, B. P.; Kasif, S.; Roberts, R. J.; Steffen, M. *Nucleic Acids Res.* **2016,** *44,* D330–D335. doi:10.1093/nar/gkv1324
79. Nanudorn, P.; Thiengmag, S.; Biermann, F.; Erkoc, P.; Dirnberger, S. D.; Phan, T. N.; Fürst, R.; Ueoka, R.; Helfrich, E. J. N. *Angew. Chem., Int. Ed.* **2022,** *61,* e202208361. doi:10.1002/anie.202208361
80. Pascal Andreu, V.; Roel-Touris, J.; Dodd, D.; Fischbach, M. A.; Medema, M. H. *Nucleic Acids Res.* **2021,** *49,* W263–W270. doi:10.1093/nar/gkab353
81. Costa, S. S.; Guimarães, L. C.; Silva, A.; Soares, S. C.; Baraúna, R. A. *Bioinf. Biol. Insights* **2020,** *14,* 1177932220938064. doi:10.1177/1177932220938064
82. Medema, M. H.; Fischbach, M. A. *Nat. Chem. Biol.* **2015,** *11,* 639–648. doi:10.1038/nchembio.1884
83. Beskrovnaya, P.; Melnyk, R. A.; Liu, Z.; Liu, Y.; Higgins, M. A.; Song, Y.; Ryan, K. S.; Haney, C. H. *mBio* **2020,** *11,* e01906-20. doi:10.1128/mbio.01906-20
84. Takeda, I.; Umemura, M.; Koike, H.; Asai, K.; Machida, M. *DNA Res.* **2014,** *21,* 447–457. doi:10.1093/dnares/dsu010
85. Emms, D. M.; Kelly, S. *Genome Biol.* **2015,** *16,* 157. doi:10.1186/s13059-015-0721-2
86. Armstrong, J.; Hickey, G.; Diekhans, M.; Fiddes, I. T.; Novak, A. M.; Deran, A.; Fang, Q.; Xie, D.; Feng, S.; Stiller, J.; Genereux, D.; Johnson, J.; Marinescu, V. D.; Alföldi, J.; Harris, R. S.; Lindblad-Toh, K.; Haussler, D.; Karlsson, E.; Jarvis, E. D.; Zhang, G.; Paten, B. *Nature* **2020,** *587,* 246–251. doi:10.1038/s41586-020-2871-y
87. Morozova, O.; Marra, M. A. *Genomics* **2008,** *92,* 255–264. doi:10.1016/j.ygeno.2008.07.001
88. Ziemert, N.; Alanjary, M.; Weber, T. *Nat. Prod. Rep.* **2016,** *33,* 988–1005. doi:10.1039/c6np00025h

License and Terms