Computational and Structural Biotechnology Journal 20 (2022) 1218–1226

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

BioCAT: Search for biosynthetic gene clusters producing nonribosomal peptides with known structure

Dmitry N. Konanov [a,*,1], Danil V. Krivonos [a,1], Elena N. Ilina [a], Vladislav V. Babenko [a]

[a] Federal Research and Clinical Centre of Physical and Chemical Medicine, Federal Medical and Biological Agency of Russia, ul. Malaya Pirogovskaya., 1s3, Moscow 119435, Russian Federation

ARTICLE INFO

Article history:
Received 29 October 2021
Received in revised form 14 February 2022
Accepted 14 February 2022
Available online 04 March 2022

Keywords:
Nonribosomal peptides
Biosynthetic gene clusters
Software

ABSTRACT

Nonribosomal peptides are a class of secondary metabolites synthesized by multimodular enzymes named nonribosomal peptide synthetases and mainly produced by bacteria and fungi. NMR, LC-MS/MS and other analytical methods allow to determine a peptide structure precisely, but it is often not a trivial task to find natural producers of them. There are cases when potential producers should be found among hundreds of strains, for instance, when analyzing metagenomic data. We have developed BioCAT, a tool designed for finding biosynthetic gene clusters which may produce a given nonribosomal peptide when the structure of an interesting nonribosomal peptide has already been found. BioCAT unites the antiSMASH software and the rBAN retrosynthesis tool but some improvements were added to both gene cluster and peptide structure analysis. The main feature of the method is an implementation of a position-specific score matrix to store specificities of nonribosomal peptide synthetase modules, which has increased the alignment sensitivity in comparison with more strict approaches developed earlier. We tested the method on a manually curated nonribosomal peptide producers database and compared it with competing tools GARLIC and Nerpa. Finally, we showed the method's applicability on several external examples.

1. Introduction

Nonribosomal peptides (NRPs) are secondary metabolites produced by a wide range of taxa, such as bacteria, fungi [26], plants, and even animals [31]. Biosynthesis of NRPs in cells is provided by multidomain enzymes named nonribosomal peptide synthetases (NRPS) in an iterative way (Fig. 1). Each functional module of NRPS generally consists of three domains: the adenylation domain (A-domain) providing the activation of the substrate using ATP, the peptidyl-carrier domain (PCP-domain) binding the substrate via the 4'-phosphopantethine group, and the condensation domain (C-domain) catalyzing the amide bond or, in some cases, the ester bond [3] formation between substrates from the current and previous modules (Fig. 1). Additionally, the last module of NRPS should contain the thioesterase domain (TE-domain) which hydrolyzes the thioester bond realizing the biosynthesis product. It was shown that modules' substrate specificities were mostly provided by A-domains [35] but some reports describing C-domain specificity also were published [4,12].

The unique biosynthesis scheme has led to the tremendous diversity in the molecular structure of NRPs in comparison with ribosome-synthesized peptides, firstly, due to the possibility to combine both proteinogenic and non-proteinogenic substrates as well as to modify monomers by hydroxylation, halogenation, epimerization, and other ways simultaneously with the biosynthesis process. Moreover, joint work of different enzymes allows to build more complex structures such as NRP-polyketide hybrids [37], NRPs containing β-lactam ring [9], cyclic depsipeptides [33] and others. Generally, NMR and LC-MS/MS methods are used for the determination of NRPs chemical structure, but a number of additional technologies to analyze the structure of NRP exist. Thus, to adjust the stereochemistry of the monomers Marfey's method is often used [20]. In addition, in some cases, chirality can be determined using computational approaches [36].

An accurate prediction of potential producers of a given NRP is not a trivial task even when the NRP structure is known because of two main reasons. On the one hand, A-domains specificity prediction is based on a slightly small NRP producers dataset available these days. Thus, existing tools such as SVM-based NRPSPredictor

* Corresponding author.
E-mail address: konanovdmitriy@gmail.com (D.N. Konanov).
[1] Both authors contributed equally to this manuscript.

https://doi.org/10.1016/j.csbj.2022.02.013

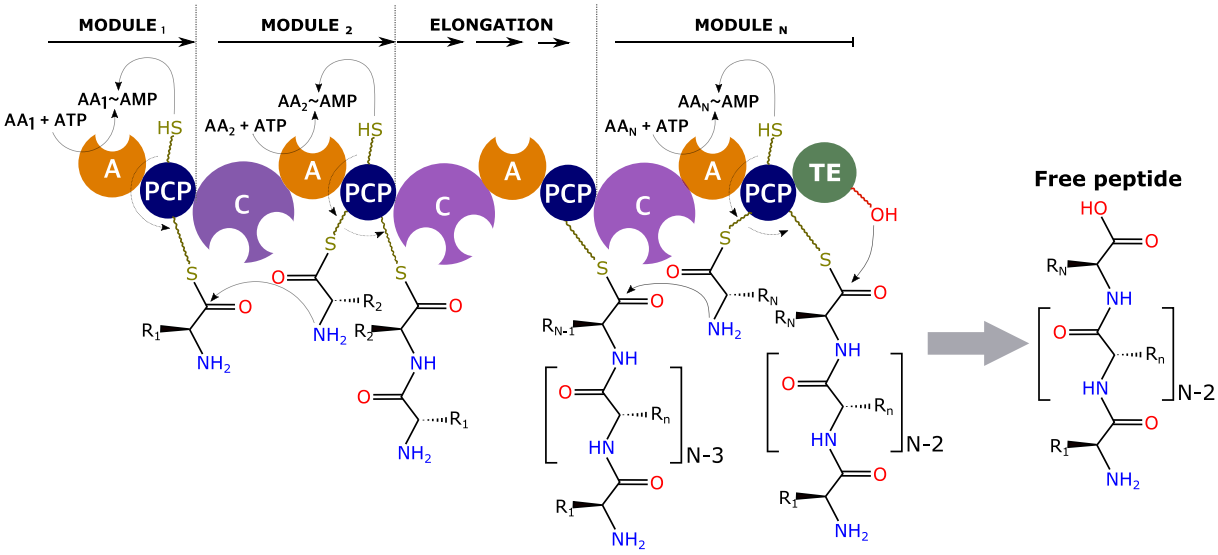

Fig. 1. Typical NRP biosynthesis scheme. The adenylation domain (A-domain) from Module 1 activates the first amino acid using ATP. Next, the neighboring peptidyl-carrier protein domain (PCP-domain) forms the thioester bond with the activated amino acid. Simultaneously, the same process occurs in Module 2. After amino acids have been connected to corresponding PCPs, the condensation domain (C-domain) from Module 2 catalyzes peptide bond formation between them. The process is iteratively repeated until the thioesterase domain (TE-domain) from the last module realizes the biosynthesis product.

2 [30] and an ensemble method called SANDPUMA [5] have been trained on less than one hundred manually annotated A-domains which seems insufficient for accurate specificity prediction, mainly because of the high monomers variety. The second problem is related to the complexity of accurate NRP retrosynthesis. In addition, a number of NRP structures are synthesized in non-iterative schemes including dimerization of peptide fragments (Type B biosynthesis pathway [32]) or use of one NRPS module more than one time during the biosynthesis (Type C biosynthesis pathway [32]).

Here, we present BioCAT (Biosynthesis Cluster Analysis Tool), a new tool that allows finding producers of a given NRP, using as the input a SMILES-formatted chemical structure and the genome of the potential producer in the FASTA format. Formally, the method unites antiSMASH [2] biosynthetic gene cluster (BGC) predictions and the rBAN [29] retrosynthesis tool, but there are some improvements added to both gene cluster and chemical structure analyses. Firstly, we developed a position-specific score matrix (PSSM) based approach to align NRP and BGC. Secondly, we implemented the retrosynthesis model which generates not just monomers but probable pathways of synthesis which we named core peptide chains. It should be noted, that the tool is designed to analyze only prokaryotic genomes because of the insufficient size of fungal NRP producers data.

To validate our model, we checked the quality of the full pipeline on the manually curated dataset of all known NRP/producer pairs using shuffle-split cross-validation. In addition, we showed the applicability of BioCAT on several external data, including complete genomes as well as draft ones. Finally, we compared the BioCAT pipeline with the GARLIC tool [6] and Nerpa [18] which have a similar functionality.

2. Materials and methods

2.1. Database collection

BGC annotations and corresponding chemical structures for 426 known NRPs were collected from the MIBiG database [15]. To ensure consistency of annotations all BGCs were re-annotated using antiSMASH 6 [2]. 1675 A-domain sequences with known specificity were extracted (full list of used sequences is available in Supplementary Data, A-domains table). To check genome-level applicability and train the model, 164 prokaryotic genomes containing known BGCs were collected from NCBI and corresponding structures in SMILES format were downloaded from MIBiG (the full list of collected NRP-genome pairs is available in Supplementary Data, AllProducers table). All NRPs and their producers used for external validation were collected such that they were not recorded in common NRP databases and corresponding BGCs were checked to be valid manually.

2.2. NRP retrosynthesis

In BioCAT, the NRP structure processing consists of two main parts: the monomers identification by rBAN [29] and the extracting of peptide fragments which we named core peptide chains. Additionally, there are a number of features improving the parsing of NRP chemical structures such as cycle solving and searching of inner fragments which are probably synthesized in non-classic ways (e.g. Type B or Type C biosynthesis pathways). The entire retrosynthesis stage and generation of linear peptide fragments used in the next analysis are automated. During the analysis, the molecular graph is stored as an RDKit [19] object. All processing manipulations such as bond hydrolysis and decyclization are implemented using Python 3.

2.2.1. Core peptide chain(s) prediction

Firstly, the SMILES-formatted NRP structure is processed by the rBAN software [29] (discovery mode is enabled). Next, each bond in the resulting monomeric graph (Fig. 2, A) is checked to be the peptide bond by comparing it with the peptide bond template implemented using RDKit. If some bond is not peptide, it will be removed from the graph. Thereafter, we have a number of distinct peptide fragments which are supposed to be synthesized in a linear way during the biosynthesis process. Next, each monomer is checked to be an α-amino acid strictly by comparing with α-amino acid template, and all non-amino acid monomers are removed from the fragments. If some fragments remain to be cyclic, the algorithm will hydrolyze these fragments in all possible ways to get all possible linear monomeric sequences. Simultane-

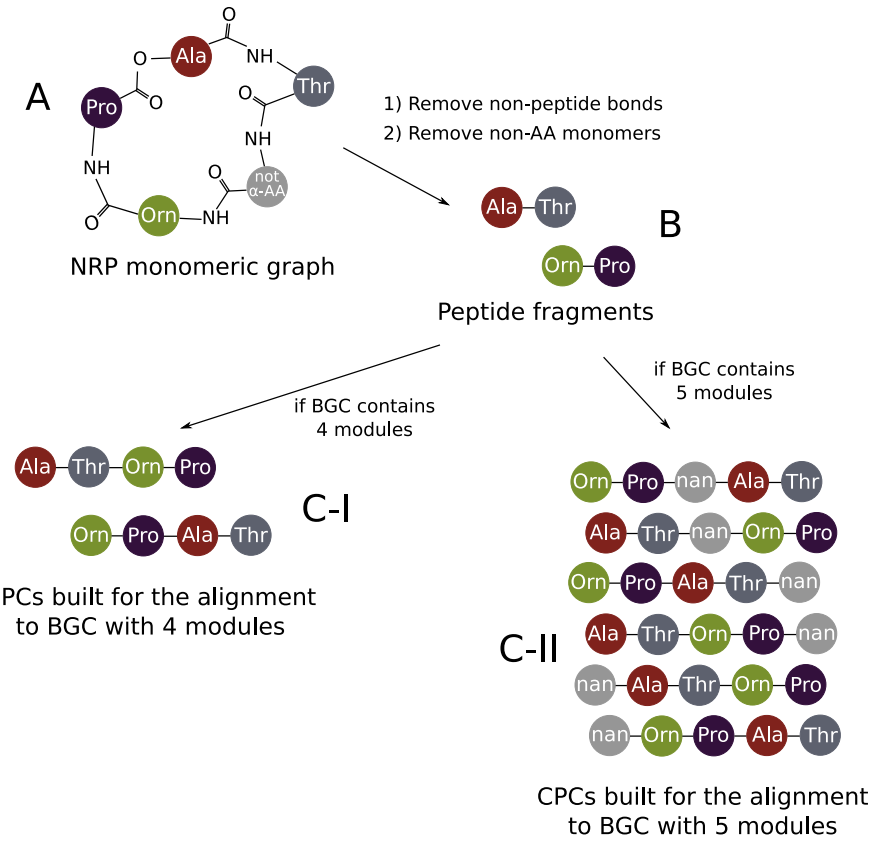

Fig. 2. Combinatorial approach to align NRP structure against BGC. Firstly, the monomeric graph generated by rBAN (A) is cut along all bonds which were recognized as non-peptide. Simultaneously, all monomers which were not recognized as α-amino acids are removed from the graph. The resulting peptide fragments (B) can be combined in different ways which depend on the size of BGC with which the current alignment is performed. If the number of modules in the BGC is the same as the sum number of monomers in the peptide fragments, these fragments will be just rearranged in all possible ways to generate core peptide chains (CPCs) (C-I). If the number of modules in the BGC is more than the sum number of monomers in the peptide fragments, gaps assigned as *nan* will be added to core peptide chains to all possible positions (C-II).

ously, the algorithm checks if the considered product can be synthesized in the Type B or Type C ways (if this option is enabled) and generates additional monomeric sequences modified according to these biosynthesis types. Thus, at this stage, we have a number of linear peptide fragments consisting only of α-amino acids connected only by peptide bonds (Fig. 2, B). To generate the product sequences which will be aligned against a given PSSM, these fragments are concatenated in all possible ways. If the length of concatenates is less than the size of the PSSM, a required number of gaps will be added between concatenated fragments. Gaps in the concatenates are assigned as *nan* (Fig. 2, C-I and C-II). In the further sections, we will call these concatenates core peptide chains (CPCs).

2.3. BGC analysis

2.3.1. Profile hidden markov models construction

The most common substrates were chosen such that for each of them there were at least 10 A-domain sequences in the database. Only these sequences were used in the further analysis. Profile HMMs construction was carried out as follows. Suppose that we have a number of A-domain sequences for i-th substrate. Firstly, we use these sequences as the base to build a profile HMM for i-th substrate using HMMER3 [25]. Next, we take all A-domains sequences which are known not to have the specificity to i-th substrate and align them against this profile HMM for i-th substrate. Therefore, we have a number of alignment scores which we have named the negative background for i-th substrate (NB_i). In this pipeline, E-values given by HMMER3 were used as the alignment scores. see Fig. 3.

2.3.2. PSSM construction

Suppose that $X = (x_1, x_2, \ldots, x_N)$ is an NRPS modules sequence with unknown specificity and there are N distinct A-domains already annotated. Consider i-th A-domain sequence x_i and profile HMM for j-th substrate. Raw specificity score for this pair is E-value of the sequence to HMM alignment. Let's assume that it equals *starget*. We defined the relative specificity score g_{ji} as:

$$g_{ji} = \frac{|s \subseteq NB_j : s > starget|}{|NB_j|}, g_{ji} \subseteq [0,1] \quad (1)$$

where NB_i is the negative background for j-th substrate. In simple words, the closer this number is to one, the greater the chance that i-th A-domain has a specificity to j-th substrate.

After this procedure is carried out for all N A-domains and S substrates, we get the following matrix:

$$G = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{S1} & g_{S2} & \cdots & g_{SN} \end{pmatrix} \quad (2)$$

where S is the number of possible substrates, N is the number of modules in the BGC, g_{ji} is the chance that i-th A-domain has a specificity to j-th substrate.

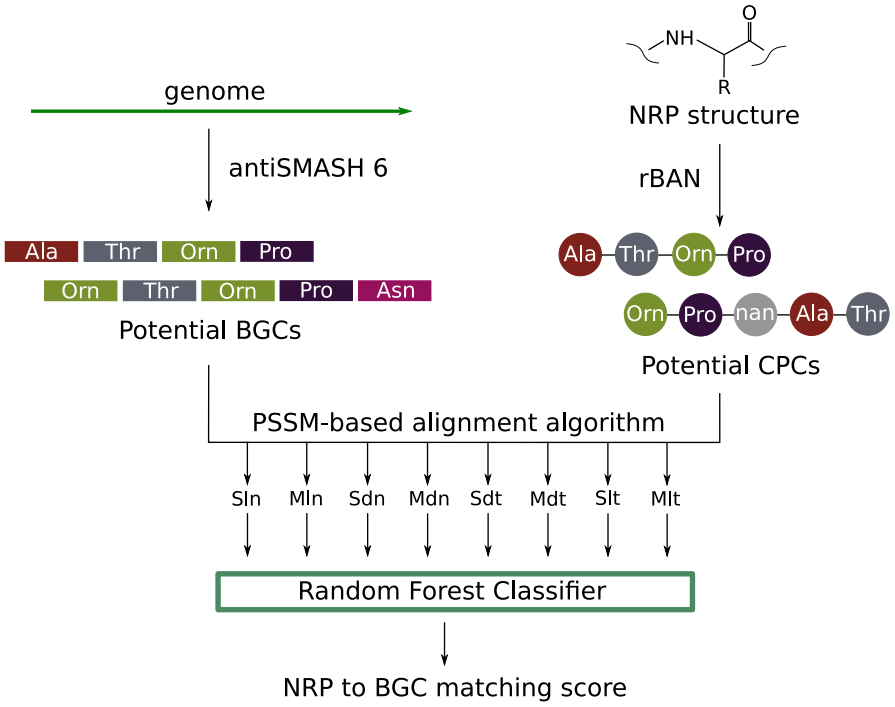

Fig. 3. Principal scheme of the BioCAT pipeline. First, the input genome is processed by antiSMASH and the NRP structure is processed by rBAN. Next, all potential biosynthesis gene clusters (BGCs) are aligned against all possible core peptide chains (CPCs) built from the monomeric graph generated by rBAN. Alignment is performed using eight different variants of the alignment score definition (Sln, Mln, Sdn, etc). Finally, for each successful matching, these scores are processed by the Random Forest Classifier, which generates the final matching score distributed from 0 to 1 and the binary matching score.

In bioinformatics, G is a classic example of a position-specific score matrix (PSSM) that can be used as an alignment template.

2.4. Alignment process

As the input for the alignment we should take one core peptide chain vector $P = (p_1, p_2, \ldots, p_N)$ and one PSSM matrix G with dimension (S, N).

2.4.1. Alignment score definition

An ensemble model has been developed for an efficient NRP to BGC alignment. First, we implemented two ways to compute raw alignment scores:

$$RawScore(P, G) = \sum_{i}^{N} g_{ji} : S_j = p_i \tag{3}$$

$$LogRawScore(P, G) = \sum_{i}^{N} \log g_{ji} : S_j = p_i \tag{4}$$

In both cases, if p_i equals *nan*, zero score will be added for i-th module.

The linear sum (Eq. 3) was the most intuitive and had a satisfactory prediction quality (F1-score = 0.596) but returned a lot of false positive matches. The logarithmic sum (Eq. 4) turned out to be more specific due to the higher influence of too small values in the PSSM on the final score. However, it had less general quality (F1-score = 0.574) compared with the linear approach.

Secondly, we build the monomers sequence $MaxSeq = [ms_1, ms_2, \ldots, ms_N]$ by the following way:

$$ms_i = S_j : g_{ij} = max(g_i), \tag{5}$$

where g_{ij} are elements of the aligned PSSM. There also were two options, how *MaxSeq* is built. The first is insertion of *nan* to the positions which contain *nan* in the core peptide chain sequence (replaced *MaxSeq*). Such operation significantly increases the method sensitivity but, again, leads to the increase in the false positive rate. If *nan* are not inserted to the *MaxSeq* (native *MaxSeq*) an absolute value of the raw alignment score tends to be much lower.

The absolute value of *RawScore* or *LogRawScore* depends on the core peptide chain length, the count of *nan* in the sequence and the nature of monomers included in the core peptide chain. To estimate the quality of an alignment, I randomly shuffled PSSM matrices are generated. The shuffling is performed in two different ways: by rows (intermodular shuffling) or by columns (intersubstrate shuffling). I was chosen to be 500 by default.

Next, after two *MaxSeq*-s and two types of shuffled matrices were formed, both native and replaced *MaxSeq* are aligned against each shuffled PSSM using both linear and logarithmic raw score calculation ways. Combining all possible computing options, we have eight arrays $F_k (k = 1, 2, .., 8)$ each of which contains I shuffled raw scores. Suppose that the observed core peptide chain was aligned to the non-shuffled PSSM with raw score equals *target*. We defined the relative alignment score for k-th method as:

$$RelScore_k = \frac{|s \subseteq F_k : s < target|}{|F_k|}, RelScore_k \subseteq [0, 1] \tag{6}$$

In other words, the relative alignment score shows the fraction of shuffled scores which are less than the non-shuffled score. Distributions of relative score for all individual models obtained on the positive dataset and negative control are shown on Supplementary Fig. 1.

Finally, these eight relative scores are processed by the Random Forest model which generates the final score also distributed between 0 to 1. Values close to 1 can be considered as successful matches.

2.4.2. Best match logic

As it was mentioned in the previous sections, a core peptide chain cannot be unambiguously determined in most cases. Due to this, the combinatorial approach was implemented to generate all possible peptide chains which can be matched to a current PSSM. The Random Forest model score is computed for each peptide chain variant independently and the highest score is chosen as the alignment result.

2.5. Output explanation

Despite only the highest alignment score influences the final alignment report, all combinatorial chain alignments are saved into the resulting file. Generally, the resulting file is a table consisting of the following columns:

- Chromosome name (column 1)
- Coordinates of BGC (column 2)
- Strand (column 3)
- Substance name (column 4)
- Cluster ID (column 5)
- Core peptide chain (column 6)
- Supposed biosynthesis type (e.g. Type A, B or C) (column 7)
- Sln, Mln, Sdn, Mdn, Sdt, Mdt, Slt, Mlt scores (columns 8–15)
- Probability of successful match for current alignment (column 16)
- Random Forest binary prediction (column 17)

Columns 8-15 of the resulting file contain scores returned by eight different individual models. Names of individual models describe their parameters as following:

- First letter means PSSM shuffling type ('S' is intersubstrate, 'M' is intermodular)
- Second letter means raw score calculation type ('l' is logarithmic, 'd' is linear)
- Third letter means *MaxSeq* processing option ('n' is with *nan* insertion, 't' is without insertion)

2.6. Method validation

First, we generated 820 incorrect genome/NRP pairs to estimate the false-positive rate of the considered methods. The number of incorrect pairs was chosen to be 5 times more than the number of correct pairs to check the specificity and selectivity of the method more accurately. To validate the Random Forest classification model implemented in BioCAT, the database of 984 genome/NRP pairs (164 correct + 820 incorrect) was divided in 80:20 ratio on train and test sample respectively. The accuracy of matching was estimated using precision, recall, F1-score, and MCC metrics. Additionally, a receiver operating curve (ROC) and a precision-recall (PR) curves were built using scores returned by the model. To estimate the stability of the model, the train/test splitting was performed randomly in 1000 iterations. Parameters used for the Random Forest model construction are available in the Supplementary Data, RFParameters table. OOB error curves and feature weights are shown in Supplementary Figs. 2 and 3.

The method was compared with the GARLIC pipeline [6] which has a similar functionality. The latest versions of GRAPE (1.0.2) and PRISM (2.1.5) tools for which command-line versions were available were used. 164 genome sequences containing BGCs with a known product were analyzed by PRISM to locate BGCs. Retrosynthesis of chemical structures was performed by GRAPE. The same list of 984 correct and incorrect genome/NRP pairs was processed by GARLIC. Because GARLIC does not return a binary matching value, the relative scores returned by GARLIC were additionally processed by a linear classifier to provide the best threshold value which was found to be 0.49. The same classification accuracy metrics as for BioCAT were calculated.

Additionally, the method was compared with the Nerpa tool published recently [18]. The same dataset of 984 genome/BGC pairs was analyzed with Nerpa. The score threshold of 6.0 recommended by the authors was used.

In the analysis, if any method did not return any possible alignments for a pair, the alignment score was assumed to be zero.

2.7. Used software and tools

MUSCLE 3.8.1551 [7] was used for multiple alignment. HMMER3 3.1b2 [25] was used to build profile HMMs. rBAN 1.0 [29] and RDKit 2021.03.4 [19] were used in the NRP retrosynthesis stage. To locate biosynthesis gene clusters antiSMASH 6.0.0 [2] was used, with Prodigal 2.6.3 [11] as a gene finding tool. The random forest model was implemented using the Scikit-learn python library (v0.24.2) [27].

3. Results

3.1. Software description and availability

We have developed a tool that estimates the likelihood that a given non-ribosomal peptide synthetase, or more generally a given organism, is capable of producing a given NRP. The tool is available as a command-line program named BioCAT (Biosynthesis Cluster Analysis Tool) on GitHub (https://github.com/DanilKrivonos/BioCAT) or can be installed via pip. The required input files necessary for the analysis are a FASTA-formatted genome and SMILES-formatted NRP-structure. In the BioCAT pipeline, the genome sequence is analyzed by antiSMASH and the structure is characterized by rBAN, so, these programs are required to be installed. Additionally, it is possible to use pre-calculated antiSMASH or rBAN results in JSON format.

Biosynthesis of NRPs can be carried out not only in the strict iterative way shown in Fig. 1. We will use the NRP biosynthesis type notation proposed in [32], where the most common canonical iterative pathway is called Type A, and two additional variants of the NRP building called Type B and Type C are defined. The Type B pathway includes a formation of two or more identical NRP fragments catalyzed by the same NRPS or the same part of NRPS which will be condensed in further biosynthesis stages. NRPs such as actinomycin D are shown to be synthesized in the Type B pathway [28]. The Type C biosynthesis variant shown for such NRPs as lugdunin [39] includes a sequential binding of two or more identical monomers activated by a single adenylation domain. In BioCAT, we implemented the support of both non-linear biosynthesis types.

We have found that the best producers' prediction quality can be reached using ensemble approaches. We have implemented the random forest classifier model which computes the final alignment score using eight pre-scores generated by slightly different algorithms, which are described in detail in the Materials and Methods section.

The result of BioCAT analysis is information about all possible NRP to BGC alignments generated in a combinatorial way. For each alignment, the final alignment score is computed independently. Final scores returned by BioCAT are distributed from 0 to 1, where values close to one show that the given BGC is likely to code the NRP synthetase providing the biosynthesis of the given NRP and vice versa.

We compared BioCAT with two competing tools called GARLIC and Nerpa which have a similar functionality. It should be men-

tioned here that all metrics for BioCAT presented in the Table 1 were calculated on the test samples when both Nerpa and GARLIC were used as is, i.e. the Nerpa and the GARLIC models were trained on full NRP datasets collected by the authors. We estimated the accuracy of the methods using recall, precision, F1-score, and MCC metrics (Table 1). After 1000 iterations of random 80:20 train/test splitting, BioCAT had a higher mean recall but GARLIC and Nerpa were shown to be more precise (Table 1). According to F1-score and MCC metrics, Nerpa and BioCAT were close to each other, and both overperformed GARLIC. ROC-AUCs and PR-AUCs additionally showed high specificity but a high false-negative rate on the Nerpa predictions, and high sensitivity and a high false-positive rate on the BioCAT results (Fig. 4).

3.2. Method benchmarking

During the analysis, the time consumption (including BGC detection and retrosynthesis stages) of considered methods was measured. We found that in all three methods used BGC detection was the limiting stage. The total time taken by the BioCAT pipeline was 338 s per NRP to genome alignment, which was faster than the full GARLIC pipeline, which averaged 527 s to run. Nerpa turned out to be faster and averaged 292 s to a full run and was shown to be significantly faster in the NRP to BGC matching stage.

Additionally, we have tested how the alignment score depends on the number of shuffling iterations. Satisfactory convergence of the results was achieved at 500 iterations (Pearson's correlation coefficient = 0.945), so this value was chosen by default (Supplementary Fig. 4).

3.3. Method application

3.3.1. Search for potential producers of a given NRP

After the model was trained and validated, a few external genome/NRP pairs were processed to show the applicability of the method. Laterocidin [21], thanamycin [13,16] and mutanobactin [14] which were not included in the curated dataset were aligned against their producers and a number of related genomes from the same genera. Laterocidin (Fig. 5A) was successfully aligned against its producer *Brevibacillus laterosporus* LMG 15441 with a relative score of 0.81. At the same time, all alignments against other *Brevibacillus* strains returned relative alignment scores less than 0.5 (Supplementary data, Laterocidine_test). Interestingly, thanamycin (Fig. 5B) had the successful matching score not only with its own producer *Pseudomonas fluorescens* DSM 11579, but with three other strains of *Pseudomonas* (Supplementary data, Thanamycin_test). One of them, *Pseudomonas* sp. 11K1, has been shown to produce brasmicin, an NRP related to thanamicin [38] (the monomeric structures of tanamycin and brasmycin are shown in Supplementary Fig. 8.). Thus, others can also be considered as potential producers of NRPs with a similar monomeric structure.

At the same time, laterocidine and thanamycin were aligned against potential producers with Nerpa and GARLIC. Nerpa has successfully predicted both laterocidine and thanamycin natural producers with Nerpa score greater than 6.0 while GARLIC has successfully detected only laterocidine producer with a score of 0.764. GARLIC score for thanamycin aligned against the natural producer was 0.185. In this analysis, for Nepra, we used a threshold of 6.0 recommended by the authors. For GARLIC we used a threshold optimized on the training dataset because the raw GARLIC score is not normalized.

3.3.2. Exploratory analysis of potential mutanobactin producers

Mutanobactin is one of the NRPs produced by *Streptococcus mutans*. In a recent work [22], the authors described in detail biosynthetic gene clusters in 17 strains of *Streptococcus mutans* isolated from dental plaque. BGC responsible for the biosynthesis of mutanobactin was found in three strains (SA41, T4, 21). Using BioCAT, we found the same three strains as potential producers of mutanobactin with relative scores higher than 0.97, when all other strains did not have any successful matches (Supplementary Fig. 5). Additionally, we collected 21 *Streptococcus mutans* complete genomes available in the RefSeq database and aligned them against the mutanobactin structure. *Streptococcus mutans* UA159 strain which had been described earlier as a producer of mutanobactin [14] had a relative score of 0.97. Moreover, 8 additional strains were shown to have similar biosynthesis clusters. Manual observation of these genomes with antiSMASH showed the presence of similar biosynthetic gene clusters (Supplementary Fig. 6; Supplementary data, Mutanobactin_test).17 genomes of *Streptococcus mutans* published by Li et al. [22] were assembled at the contig level, further showing that BioCAT is useful for performing NRP and BGC matching regardless of the assembly level when the average contig size is sufficient to successfully detect BGCs.

3.3.3. NRP producers community analysis

For a more comprehensive validation, we collected 10 NRPs of a different chemical structure and their producers described in recent works not included in common databases, e.g. MIBiG or Norine [8] (the full list of NRPs used is available in Supplementary data, Interspecies_assay_report, chemical structures are available in NRP.docx Supplementary file). Corresponding genomes downloaded from the NCBI database were used for the construction of synthetic NRP producers community and each NRP structure was aligned against the total community using three considered methods. In accordance with the results obtained on the MIBiG dataset, Nerpa showed an outstanding prediction specificity (0 false-positive matches) but producers of 5 out of 10 NRPs were not found (Table 2). BioCAT successfully discovered 9 out of 10 producers. The zelkovamycin producer [34] which was the only false-negative match obtained with BioCAT was not successfully aligned due to the occasional ambiguity of the antiSMASH BGC prediction results.

Being high-sensitive, BioCAT returned 12 false-positive results. We suppose that the presence of false matches might be partially caused by the structural homology of other NRPs produced by the analyzed strains and such cases should be considered by the user manually in further data analysis. All BioCAT results are available in Supplementary_data2.

GARLIC showed the worst result predicting only one natural producer with a score greater than the optimized threshold. Unlike BioCAT, GARLIC is based on a number of individual match bonuses and mismatch penalties optimized on the NRP dataset collected by the authors, so, such low accuracy on external data compared with the results obtained on the NRPs collected from the MIBiG database (Table 1) might be a consequence of some overfitting on the MIBiG dataset.

4. Discussion

We have developed a new high-sensitive PSSM-based approach to align an NRP structure to a biosynthesis gene cluster and implemented it as a command line tool called BioCAT (Biosynthesis Cluster Analysis Tool). In general, this tool is designed for the search of potential producers of a given non-ribosomal peptide among a number of genomes, but also can be applied for solving the reversed task when a user is interested in searching for the most likely products which can be synthesized by a given organism. In the BioCAT pipeline, antiSMASH [2] and rBAN [29] functionalities were united. To our knowledge, these tools are most commonly

Table 1
BioCAT performance compared with competing tools.

Method	recall	precision	F1-score	MCC	Mean time consupmtion, s
BioCAT	0.735	0.515	0.600	0.519	338
GARLIC	0.363	0.766	0.487	0.468	527
Nerpa	0.419	0.948	0.577	0.589	292

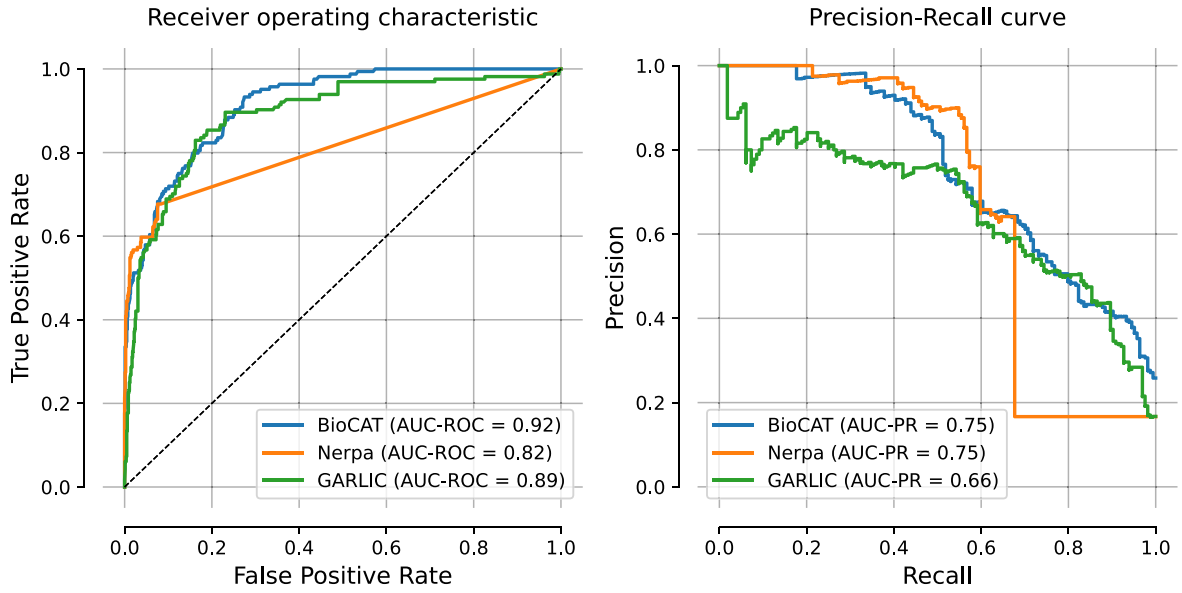

Fig. 4. Performance of BioCAT compared with GARLIC and Nerpa. The graphs show receiver operating characteristic curves (left) and recall-precision curves (right) obtained using BioCAT (blue), Nerpa (orange) and GARLIC (green) tools.

used for BGC annotation and NRP retrosynthesis respectively, which is why they were chosen to be implemented in the pipeline.

During the BioCAT development, we were trying to avoid complex machine learning methods to keep the transparency of the interpretation of the results. We tested several different approaches to calculate the alignment score without machine learning usage, but their accuracies did not seem satisfactory for us. Observing individual scores' results, we supposed that an ensemble model is capable of providing higher accuracy than individual ones. We implemented a simple Random Forest Classifier combining eight slightly different methods to calculate the relative alignment score and showed its efficiency using common classification quality metrics.

We compared our method with the GARLIC pipeline [6] based on other BGC detecting and structure retrosynthesis tools. Although GARLIC had a higher specificity, in general, NRP to BGC matching quality obtained on the BioCAT results turned out to be higher than the GARLIC quality. The Nerpa tool [18] recently published and based on the same external software has shown similar general matching quality and as well as BioCAT overperformed GARLIC. However, it should be noted here that BioCAT was designed to be more sensitive than specific, unlike Nerpa which has shown high matching specificity but moderate sensitivity. Thus, when analyzing hundreds of genomes, BioCAT can be used as an additional filtering stage to narrow down the list of potential producers with a low chance of rejecting the real producer. At the same time, the main scope of Nerpa usage is to search for the most likely producer among a huge number of candidates even if a part of native producers might be rejected. An additional advantage of BioCAT compared with the Nerpa tool is the native support of type B and type C biosynthetic pathways. Thus, we found that natural producers of NRPs such as actinomycin [28] and valinomycin [24] (the Type B biosynthesis pathway) or lugdunin [39] (the Type C biosynthesis pathway) are not predicted by Nerpa but can be detected with BioCAT. Moreover, we have shown that due to high sensitivity BioCAT is capable of detecting not only producers of the target NRP structure but producers of close chemical homologs too such as brasmycin and thanamycin.

The method we developed has a number of limitations, mainly related to the quality of NRP chemical structures retrosynthesis. The rBAN tool is able to determine a wide range of substrates but some unusual chemical modifications such as acylation of proline lead to the appearance of excessive unrecognized elements in a molecular graph. Moreover, some chemical features such as fatty acid residues or poly-ketide fragments are not used in the PSSM construction and are not taken into account during the processing. Also, in NRP biosynthesis, some condensation domains are known to be able to form not peptide bonds but ester ones [3]. In these cases, peptide chains will be restricted at the ester bond and resulting fragments will be combined more aggressively which can increase the chance of false-positive results. Generally, if peptide chains generated by the model include too many substrates assigned as *nan*, we recommend users to try simplifying the chemical structure of the interesting NRP manually before the analysis, e.g. to remove modifications from the substrates. For instance, we have found that echinomycin which is formally synthesized in the type B biosynthesis pathway but containing cysteine modified by methylation [17] can be successfully matched to its producers only after manual removal of this modification from the SMILES string.

The BGC prediction stage also has some drawbacks. The main is the lack of formal rules to define edges of biosynthetic gene clusters. For example, some gene clusters such as nunamycin/-

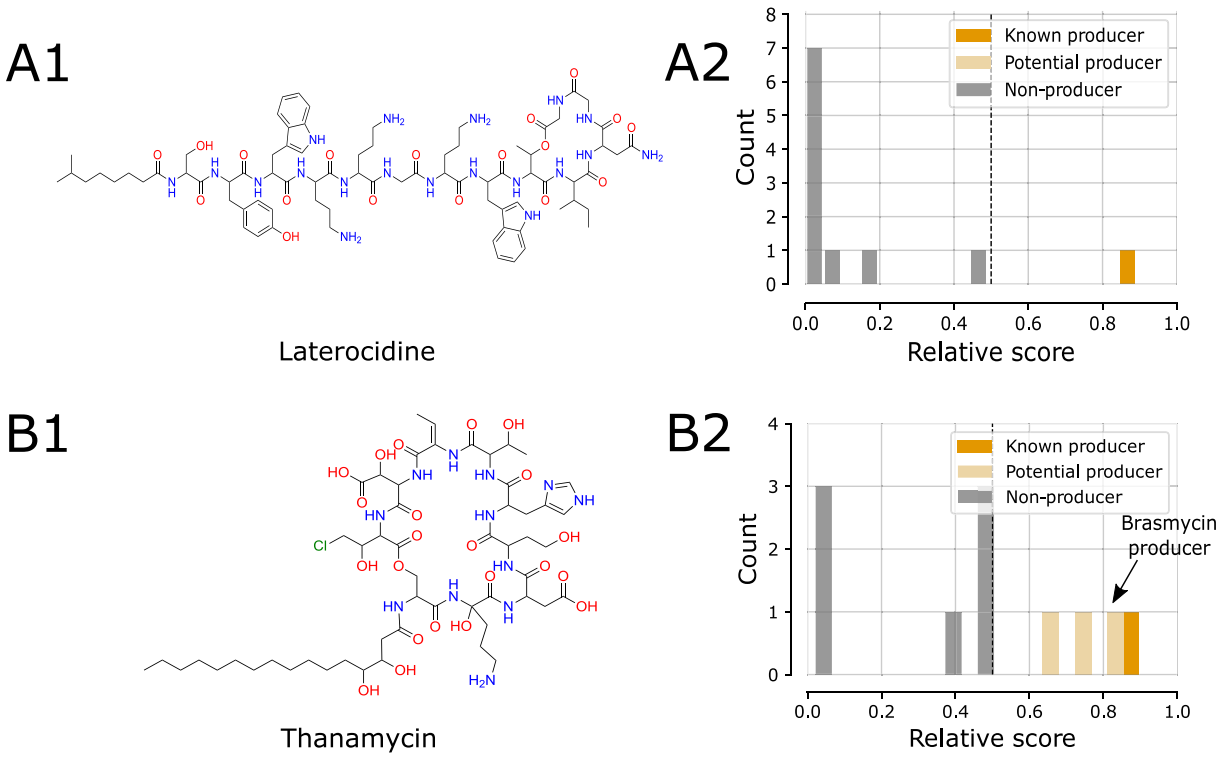

Fig. 5. Applicability of BioCAT for the identification of potential producers of a given NRP. A) Laterocidine chemical structure (A1) was aligned against its natural producer *Brevibacillus laterosporus* LMG 15441 and 10 close *Brevibacillus* strains. Only the producer (A2, orange bar) had an alignment score higher than 0.5. B) Thanamycin (B1) was aligned against its producer *Pseudomonas fluorencens* DSM 11579 (B2, orange bar) and 10 other *Pseudomonas* strains. The producer was successfully aligned with the resulting score of 0.86. In addition, there were three *Pseudomonas* strains that were assigned as potential producers of thanamycin (B2, light orange bars). *Pseudomonas* sp. 11K1 strain which had the highest alignment score was described earlier as a producer of brasmycin, NRP related to thanamycin.

Table 2
Interspecies testing of NRP to BGC matching tools. In this test, there were 100 genome/NRP pairs, with 10 a priori correct and 90 incorrect.

Method	True positives	False positives
BioCAT	9/10	12/90
GARLIC	1/10	0/90
Nerpa	5/10	0/90

nunapeptin BGC from *Pseudomanas* sp. Ln5 encode two different NRPSs located close to each other because of regulatory reasons [10]. Fortunately, these NRPSs are encoded in different DNA strands, so, in BioCAT we have implemented additional fragmentation of clusters based on the strand direction. However, there are cases, for example, himastatine biosynthesis cluster from *Streptomyces himastatinicus* ATCC 53653, when genes located in both DNA strands are responsible for the biosynthesis of only one NRP product [23].

Despite the drawbacks described above, BioCAT showed a satisfactory matching accuracy and can be useful for high-throughput exploratory analysis of genomic data to identify possible producers of an NRP of interest or structures homologous to it. Going forward, the method can be improved in several ways. First, we are planning to include to the model information about additional gene cluster domains such as halogenation and hydroxylation which may increase the specificity of the alignment algorithm. Secondly, core peptide chains generated during the BioCAT analysis often contain non-recognized substrates assigned as *nan*, so, these unrecognized peptide chain positions can be represented in the same PSSM way, using special metrics such as Tanimoto chemical similarity coefficient [1]. However, it can significantly increase the model complexity, so, we decided not to implement it in this BioCAT version due to the insufficient size of the NRP library available nowadays.

Simplification and unification of genomic data processing are becoming more important with the intensive development of sequencing technologies. Thus, the more massive genomes are sequenced, the more time is consumed to perform an accurate prediction of NRP producers among them manually using antiSMASH or Prism software. The authors do not declare that the BioCAT tool can completely replace manual BGC annotation but hope that it will help to automatize the preliminary genomic data observation to narrow down a list of possible producers of a given NRP.

5. Conclusion

We have developed a novel tool, called BioCAT, which has united the antiSMASH and the rBAN pipelines and allows to find potential producers of a given NRP. BioCAT was shown to be slightly faster and more accurate in comparison with the GARLIC tool published earlier. The second competing tool Nerpa has been shown to be more specific rather than sensitive, unlike BioCAT which was first designed as a tool useful for preliminary filtering of a huge number of potential producers with a minimal chance of rejecting a real producer. The applicability of the method was additionally shown on several external data.

6. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Dmitry N. Konanov: Conceptualization, Methodology, Writing - original draft, Visualization. **Danil V. Krivonos:** Software, Writing - original draft, Data curation, Formal analysis, Visualization. **Elena N. Ilina:** Supervision, Writing - review & editing. **Vladislav V. Babenko:** Conceptualization, Writing - review & editing, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Federal Research and Clinical Center of Physical–Chemical Medicine of Federal Medical Biological Agency for providing computational resources for this project.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2022.02.013.

References

[1] Bajusz D, Rácz A, Héberger K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminformatics 2015;7:1–13.
[2] Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel G.P, Medema MH, Tilmann W. antismash 6.0. Nucleic Acids Research; 2021..
[3] Bloudoff K, Schmeing TM. Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1865; 2017, 1587–1604..
[4] Calcott MJ, Owen JG, Ackerley DF. Efficient rational modification of non-ribosomal peptides by adenylation domain substitution. Nat Commun 2020;11:1–10.
[5] Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. Sandpuma: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across actinobacteria. Bioinformatics 2017;33:3202–10.
[6] Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, Skinnider MA, Webster AL, Magarvey NA. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol 2016;12:1007–14.
[7] Edgar RC. Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 2004;5:1–19.
[8] Flissi A, Ricart E, Campart C, Chevalier M, Dufresne Y, Michalik J, Jacques P, Flahaut C, Lisacek F, Leclère V, et al. Norine: update of the nonribosomal peptide resource. Nucl Acids Res 2020;48:D465–9.
[9] Gaudelli NM, Long DH, Townsend CA. β)lactam formation by a non-ribosomal peptide synthetase during antibiotic biosynthesis. Nature 2015;520:383–7.
[10] Hennessy RC, Phippen CB, Nielsen KF, Olsson S, Stougaard P. Biosynthesis of the antimicrobial cyclic lipopeptides nunamycin and nunapeptin by pseudomonas fluorescens strain in5 is regulated by the luxr-type transcriptional regulator nunf. Microbiologyopen 2017;6:e00516.
[11] Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 2010;11:1–11.
[12] Izoré T, Ho YC, Kaczmarski JA, Gavriilidou A, Chow KH, Steer DL, Goode RJ, Schittenhelm RB, Tailhades J, Tosin M, et al. Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. Nat Commun 2021;12:1–14.
[13] Johnston CW, Skinnider MA, Wyatt MA, Li X, Ranieri MR, Yang L, Zechel DL, Ma B, Magarvey NA. An automated genomes-to-natural products platform (gnp) for the discovery of modular natural products. Nat Commun 2015;6:1–11.
[14] Joyner PM, Liu J, Zhang Z, Merritt J, Qi F, Cichewicz RH. Mutanobactin a from the human oral pathogen streptococcus mutans is a cross-kingdom regulator of the yeast-mycelium transition. Organic Biomol Chem 2010;8:5486–9.
[15] Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJ, Van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, et al. Mibig 2.0: a repository for biosynthetic gene clusters of known function. Nucl Acids Res 2020;48:D454–8.
[16] Kirchner N, Cano-Prieto C, van der Voort M, Raaijmakers JM, Gross H. Draft genome sequence of lipopeptide-producing strain pseudomonas fluorescens dsm 11579 and comparative genomics with pseudomonas sp. strain sh-c52, a closely related lipopeptide-producing strain. Microbiol Resource Announ 2020;9:e00304–20.
[17] Kong D, Park EJ, Stephen AG, Calvani M, Cardellina JH, Monks A, Fisher RJ, Shoemaker RH, Melillo G. Echinomycin, a small-molecule inhibitor of hypoxia-inducible factor-1 dna-binding activity. Cancer Res 2005;65:9047–55.
[18] Kunyavskaya O, Tagirdzhanov AM, Caraballo-Rodríguez AM, Nothias LF, Dorrestein PC, Korobeynikov A, Mohimani H, Gurevich A. Nerpa: A tool for discovering biosynthetic gene clusters of bacterial nonribosomal peptides. Metabolites 2021;11:693.
[19] Landrum G. Rdkit documentation. Release 2013;1:4.
[20] Li Y, Liu L, Zhang G, He N, Guo W, Hong B, Xie Y. Potashchelins, a suite of lipid siderophores bearing both l-threo and l-erythro beta-hydroxyaspartic acids, acquired from the potash-salt-ore-derived extremophile halomonas sp. mg34. Front Chem 2020;8:197.
[21] Li YX, Zhong Z, Zhang WP, Qian PY. Discovery of cationic nonribosomal peptides as gram-negative antibiotics through global genome mining. Nat Commun 2018;9:1–9.
[22] Li ZR, Sun J, Du Y, Pan A, Zeng L, Maboudian R, Burne RA, Qian PY, Zhang W. Mutanofactin promotes adhesion and biofilm formation of cariogenic streptococcus mutans. Nat Chem Biol 2021;17:576–84.
[23] Ma J, Wang Z, Huang H, Luo M, Zuo D, Wang B, Sun A, Cheng YQ, Zhang C, Ju J. Biosynthesis of himastatin: assembly line and characterization of three cytochrome p450 enzymes involved in the post-tailoring oxidative steps. Angew Chem Int Ed 2011;50:7797–802.
[24] Matter AM, Hoot SB, Anderson PD, Neves SS, Cheng YQ. Valinomycin biosynthetic gene cluster in streptomyces: conservation, ecology and evolution. PloS one 2009;4:e7194.
[25] Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. Nucl Acids Res 2013;41:e121.
[26] Oide S, Turgeon BG. Natural roles of nonribosomal peptide metabolites in fungi. Mycoscience 2020;61:101–10.
[27] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[28] Pfennig F, Schauwecker F, Keller U. Molecular characterization of the genes of actinomycin synthetase i and of a 4-methyl-3-hydroxyanthranilic acid carrier protein involved in the assembly of the acylpeptide chain of actinomycin in streptomyces. J Biol Chem 1999;274:12508–16.
[29] Ricart E, Leclère V, Flissi A, Mueller M, Pupin M, Lisacek F. rban: retro-biosynthetic analysis of nonribosomal peptides. J Cheminformatics 2019;11:1–14.
[30] Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. Nrpspredictor2–a web server for predicting nrps adenylation domain specificity. Nucl Acids Res 2011;39:W362–7.
[31] Shou Q, Feng L, Long Y, Han J, Nunnery JK, Powell DH, Butcher RA. A hybrid polyketide–nonribosomal peptide in nematodes that promotes larval survival. Nat Chem Biol 2016;12:770–2.
[32] Süssmuth RD, Mainz A. Nonribosomal peptide synthesis–principles and prospects. Angew Chem Int Ed 2017;56:3770–821.
[33] Taevernier L, Wynendaele E, Gevaert B, De Spiegeleer B. Chemical classification of cyclic depsipeptides. Curr Protein Pept Sci 2017;18:425–52.
[34] Tarantini FS, Brunati M, Taravella A, Carrano L, Parenti F, Hong KW, Williams P, Chan KG, Heeb S, Chan WC. Actinomadura graeca sp. nov.: A novel producer of the macrocyclic antibiotic zelkovamycin. Plos One 2021;16:e0260413.
[35] Throckmorton K, Vinnik V, Chowdhury R, Cook T, Chevrette MG, Maranas C, Pfleger B, Thomas MG. Directed evolution reveals the functional sequence space of an adenylation domain specificity code. ACS Chem Biol 2019;14:2044–54.
[36] Tyurin AP, Alferova VA, Paramonov AS, Shuvalov MV, Kudryakova GK, Rogozhin EA, Zherebker AY, Brylev VA, Chistov AA, Baranova AA, et al. Innentitelbild: Gausemycins a, b: Cyclic lipoglycopeptides from streptomyces sp. (angew. chem. 34/2021). Angew Chem 2021;133:18498.
[37] Zhang JM, Liu X, Wei Q, Ma C, Li D, Zou Y. Berberine bridge enzyme-like oxidase-catalysed double bond isomerization acts as the pathway switch in cytochalasin synthesis. Nat Commun 2022;13:1–10.
[38] Zhao H, Liu YP, Zhang LQ. In silico and genetic analyses of cyclic lipopeptide synthetic gene clusters in pseudomonas sp. 11k1. Front Microbiol 2019;10:544.
[39] Zipperer A, Konnerth MC, Laux C, Berscheid A, Janek D, Weidenmaier C, Burian M, Schilling NA, Slavetinsky C, Marschal M, et al. Human commensals producing a novel antibiotic impair pathogen colonization. Nature 2016;535:511–6.