Nerpa 2: linking biosynthetic gene clusters to nonribosomal peptide structures

Ilia Olkhovskii,[1,2,3] **Aleksandra Kushnareva**,[1,2] **Azat Tagirdzhanov**[1,2] **and Alexey Gurevich**[1,2,*]

[1] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbrücken 66123, Germany, [2]Center for Bioinformatics Saar and Saarland University, Saarland Informatics Campus, Saarbrücken, 66123, Germany and [3]Saarbrücken Graduate School of Computer Science, Saarland University, Saarbrücken 66123, Germany

*Corresponding author. Email: alexey.gurevich@helmholtz-hips.de

Abstract

Summary: Here, we introduce Nerpa 2, a tool for linking computationally-predicted biosynthetic gene clusters of nonribosomal peptides (NRPs) to their putative products. Nerpa 2 is based on our previous tool, Nerpa, but outperforms it in accuracy and sensitivity thanks to the software and algorithmic updates, including a wholly revised scoring function and extension of the training dataset. Nerpa 2 handles a variety of NRP biosynthesis use cases not supported by the original tool, is purely cross-platform, and provides output results in convenient HTML reports.

Availability and implementation: Nerpa 2 is available at https://github.com/gurevichlab/nerpa.

1. Introduction

Nonribosomal peptides (NRPs) are a promising source of bioactive compounds with high pharmaceutical potential (Agrawal *et al.*, 2017). Their biosynthesis is governed by specialized enzymes encoded in biosynthetic gene clusters (BGCs). Linking BGCs to their products helps researchers predict and manipulate the production of these compounds, potentially leading to the discovery of new therapeutic agents (Medema *et al.*, 2021). The modern genome mining methods enabled the collection of hundreds of thousands of computationally predicted BGCs encoding unknown NRPs (Bağcı *et al.*, 2024). At the same time, specialized databases like Norine (Flissi *et al.*, 2020) contain thousands of NRP chemical structures. However, peculiarities in the NRP biosynthesis make the linking between these two types of data extremely difficult. As a result, only a small fraction of currently elucidated NRPs are connected to their respective BGCs (Dejong *et al.*, 2016).

BGCs encoding NRPs have a modular structure that acts as an assembly line, where each module incorporates a single amino acid to the growing peptide chain (Ackerley, 2016). Modules, in turn, consist of domains, each having a specific function. The key domain defining the module specificity is the Adenylation (A) domain, while some others can introduce downstream modifications to the selected amino acid, such as methylation or epimerization. Some A domains are promiscuous and can attract various building blocks, enabling a single BGC to encode several closely related NRPs. Further, the gene activation order during the NRP biosynthesis is not always in line with the gene order in the corresponding BGC, leading to so-called *non-collinear* assembly lines. They might involve gene rearrangement, skipping, stuttering, and iterative reuse (Süssmuth and Mainz, 2017; Juguet *et al.*, 2009). Finally, a linear peptide sequence produced by an assembly line often undergoes macrocyclization or other tailoring reactions, resulting in a complex structure of the mature NRP product (Sieber and Marahiel, 2003).

Earlier, we developed Nerpa, a tool for high-throughput linking of BGCs to their tentative products from a database of NRP chemical structures (Kunyavskaya *et al.*, 2021). The method relied on the leading genome mining and NRP retrobiosynthesis software at that time (Blin *et al.*, 2019; Ricart *et al.*, 2019) and the knowledge-based alignment procedure. We trained the tool parameters using known BGC-NRP pairs from MIBiG, the largest database of experimentally verified BGCs (Kautsar *et al.*, 2020). Altogether, this allowed Nerpa to outperform the only competing method, GARLIC (Dejong *et al.*, 2016), in efficiency and accuracy. However, Nerpa still struggled to handle many non-collinear assembly lines and other irregularities of the NRP biosynthesis. BioCAT, introduced later, demonstrated performance comparable with Nerpa and could not solve its critical limitations (Konanov *et al.*, 2022).

Here, we introduce Nerpa 2, a major update of our previous BGC-NRP linking tool. This version expands Nerpa's applicability to complex non-collinear assembly lines and promiscuous A domains. We demonstrate how these changes allowed Nerpa 2 to outperform its predecessor in accuracy and sensitivity. Furthermore, Nerpa 2 complements the output with

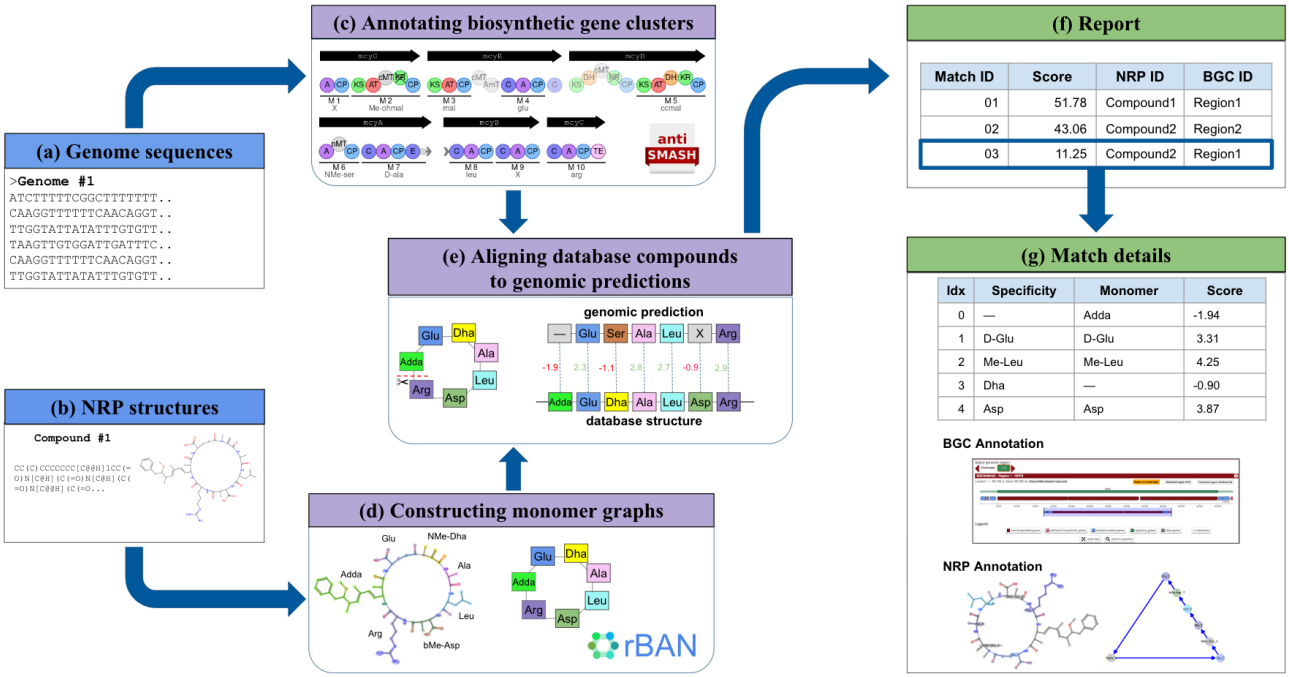

Fig. 1: **Nerpa 2 pipeline**. Input genome sequences (**a**) and nonribosomal peptide (NRP) structures (**b**) are preprocessed with third-party tools antiSMASH (**c**) and rBAN (**d**) to extract putative biosynthetic gene clusters (BGCs) and NRP monomer graphs, respectively. Nerpa 2 converts them to linear sequences of monomers, aligns them against each other, and scores the alignments (**e**). The best scoring hits per NRP and BGC are reported (**f**) with an option to view the detailed information about each match (**g**).

visual and interactive HTML reports, making interpreting and analyzing the results much more convenient.

2. Materials and methods

2.1. Pipeline overview

The Nerpa 2 pipeline resembles the predecessor's backbone (Kunyavskaya *et al.*, 2021). Briefly, the tool preprocesses the input genomes and NRP structures (Figure 1ab) into linear sequences of NRP building blocks (Figure 1cd), aligns and scores them against each other (Figure 1e), summarizes the results and reports the best matches (Figure 1fg). However, all pipeline steps were substantially updated or completely revisited in the new version of the tool.

2.2. Updated input and preprocessing

2.2.1. Supported data and third-party tools

We adapted Nerpa 2 to the latest versions of the state-of-the-art preprocessing methods: antiSMASH v.7 (Blin *et al.*, 2023) for the genome mining of BGCs and rBAN v.1.0 (Ricart *et al.*, 2019) for the retrobiosynthesis of NRP structures. All standard input data formats, such as FASTA and GenBank for genome sequences, SMILES and InChI for chemical structures, are supported. Alternatively, Nerpa 2 supports genomic input in the form of locally stored antiSMASH outputs or job IDs in the web version of the tool. The latter are automatically downloaded from the antiSMASH webserver. We also extended the supplied preprocessed database of known putative NRP structures (pNRPdb). The updated pNRPdb v.2 contains 9,506 unique compounds, including all NRPs added to Norine (accessed 05.11.2024) and MIBiG v.4.0rc1 after the pNRPdb v.1 release.

2.2.2. Nerpa monomers

Nerpa 2 operates with NRP building blocks (further referred to as "monomers"), which include amino acids, lipid tails, and other types of residues present in NRP structures. Each monomer is characterized by a core residue and additional features such as its methylation and stereochemistry (D-/L-configuration), when applicable. We manually annotated 249 monomers supported by antiSMASH and 397 monomers from the Norine database used by rBAN to classify them into core residues and additional features. The final Nerpa 2 monomer dictionary enables streamlined conversion of both preprocessing tools' outputs into the unified set of 42 common core residues and a special "unknown" label.

2.2.3. Genome and NRP preprocessing

Given a genome, antiSMASH detects BGCs related to NRP synthesis and annotates them into modules and domains (Figure 1c). In parallel, rBAN parses a provided NRP structure into a monomer graph (Figure 1d). However, these raw outputs cannot be directly used for the BGC-NRP matching and thus undergo downstream manipulations in Nerpa 2.

Our tool converts A domain annotations reported by antiSMASH into the probability distribution over the supported core residues. This procedure involves a newly trained random forest classifier and a recalibration step that considers possible A domain promiscuity and weakly characterized A domains (Supplementary Note S1).

Further, Nerpa 2 accounts for non-collinear and iterative NRP assembly lines by composing possible rearrangements of genes in BGCs and linearizing rBAN monomer graphs.

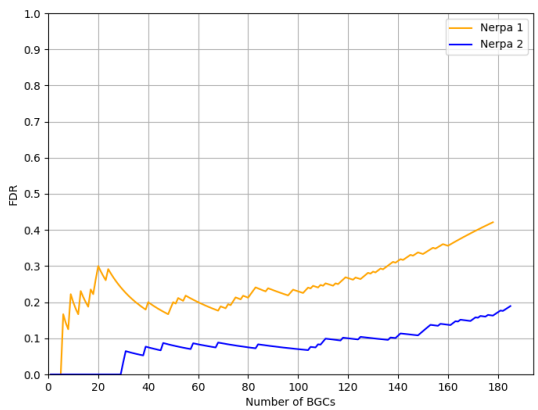

Fig. 2: **Nerpa 1 and 2 accuracy in the MIBiG v.2 benchmark experiment**. The y-axis shows the tools' false discovery rate (FDR) for x top-scoring BGC identifications.

2.3. New alignment and scoring

The Nerpa 2 alignment procedure is a variation of the Viterbi algorithm (Forney, 1973) that considers the quality of core residue and additional features predictions and also takes into account possible skipping and stuttering of BGC modules. The scoring parameters were estimated similarly to (Kunyavskaya *et al.*, 2021) based on a manually curated set of 349 ground truth BGC-NRP alignments found in the literature (Terlouw *et al.*, 2023).

2.4. Extended output

In contrast to the plain text output of the previous version, Nerpa 2 produces machine-readable JSON files, suitable for the tool integration into large computational pipelines, as well as a user-friendly HTML report, suitable for manual analysis of the results (Figure 1fg). The HTML report offers a summary table of Nerpa 2 matches, which enables sorting and filtering them by score, BGC, or NRP (Figure 1f). The summary table is complemented with a detailed view featuring the alignment information, antiSMASH output for the corresponding BGC, and depiction of the matched NRP as an annotated molecular structure and a monomer graph (Figure 1g). The HTML report is built with JavaScript, and the molecular structure and monomer graph figures are drawn using RDKit (Landrum *et al.*, 2006) and NetworkX (Hagberg *et al.*, 2008) Python libraries.

3. Results

3.1. Benchmarking accuracy

We compared Nerpa 2 with the previous version using the benchmarking setup from the original publication (Kunyavskaya *et al.*, 2021). All NRP-related entries from the MIBiG v.2 database (Kautsar *et al.*, 2020) (194 BGCs) were searched against the corresponding structural annotations (ground truth) complemented with our database of putative NRP structures, pNRPdb v.1 (Kunyavskaya *et al.*, 2021) (8449 compounds in total). In contrast to the original experiment, we processed all BGCs with antiSMASH v.7.0.1. We adjusted Nerpa 1 to work with the latest antiSMASH version to exclude the effect of potential improvements in the third-party genome mining step and thus make the core algorithms comparison fair.

A BGC was considered *identified* by a tool if the ground truth NRP was among the tool's top 10 matches for that BGC and *false positive* otherwise. Each identified BGC was associated with the score of the correct match, while false positive BGCs were linked to the score of the best (incorrect) hit. We sorted the resulting list of BGCs in descending order of their scores and computed the false discovery rate (FDR) as

$$FDR(n) = \frac{FP(n)}{n},$$

where $FP(n)$ is the number of false positives among the first n BGCs in the sorted list.

Figure 2 demonstrates the FDR curves for both Nerpa versions. Nerpa 2 consistently outperformed its predecessor and was able to handle more BGCs in general (185 vs 178). Note that neither of the tools processed all 194 BGCs since some of them were filtered out as unsuitable for matching, e.g., due to a small size (less than two NRP building blocks).

3.2. Inspecting performance on non-standard assembly lines

Improvements in the alignment algorithm allow Nerpa 2 to handle complex non-collinear NRP assembly lines. Supplementary Figures S1 and S2 illustrate the alignments produced by Nerpa 1 and Nerpa 2 for an interactive NRP and an assembly line missing some modules and reusing one of the genes. While Nerpa 1 fails to match most NRP monomers properly in these cases, the new version produced the correct alignments in both.

4. Conclusion

We have introduced Nerpa 2, a tool for linking nonribosomal peptide (NRP) biosynthetic gene clusters (BGCs) to their putative products. By integrating a probabilistic model for Adenylation domain specificity and refining the alignment scoring scheme, Nerpa 2 effectively addresses the challenges posed by the NRP biosynthesis, including domain promiscuity and non-collinear assembly lines.

Performance evaluations indicate that Nerpa 2 surpasses its predecessor in both accuracy and sensitivity, particularly in handling complex cases that were previously unsupported.

In addition to the enhanced capabilities, Nerpa 2 offers a more user-friendly interface and provides easily interpretable output. It is also cross-platform compatible. We hope that our tool will serve as a valuable resource for the NRP community and assist researchers in the annotation of computationally-predicted BGCs.

Supplementary data

Supplementary data are available online.

Conflict of interest

None declared.

Funding

Saarland University [the NextAID project to support the position of A.T.].

References

Ackerley, D. F. (2016). Cracking the nonribosomal code. *Cell chemical biology*, **23**(5), 535–537.

Agrawal, S., Acharya, D., Adholeya, A., Barrow, C. J., and Deshmukh, S. K. (2017). Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential. *Frontiers in pharmacology*, **8**, 828.

Bağcı, C., Nuhamunada, M., Goyat, H., Ladanyi, C., Sehnal, L., Blin, K., Kautsar, S. A., Tagirdzhanov, A., Gurevich, A., Mantri, S., *et al.* (2024). Bgc atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. *Nucleic Acids Research*, page gkae953.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019). antismash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research*, **47**(W1), W81–W87.

Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J., *et al.* (2023). antismash 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic acids research*, **51**(W1), W46–W50.

Dejong, C. A., Chen, G. M., Li, H., Johnston, C. W., Edwards, M. R., Rees, P. N., Skinnider, M. A., Webster, A. L., and Magarvey, N. A. (2016). Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat Chem Biol*, **12**(12), 1007–1014.

Flissi, A., Ricart, E., Campart, C., Chevalier, M., Dufresne, Y., Michalik, J., Jacques, P., Flahaut, C., Lisacek, F., Leclère, V., *et al.* (2020). Norine: update of the nonribosomal peptide resource. *Nucleic acids research*, **48**(D1), D465–D469.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, **61**(3), 268–278.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Juguet, M., Lautru, S., Francou, F.-X., Nezbedová, Š., Leblond, P., Gondry, M., and Pernodet, J.-L. (2009). An iterative nonribosomal peptide synthetase assembles the pyrrole-amide antibiotic congocidine in streptomyces ambofaciens. *Chemistry & biology*, **16**(4), 421–431.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J., Van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., *et al.* (2020). Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research*, **48**(D1), D454–D458.

Konanov, D. N., Krivonos, D. V., Ilina, E. N., and Babenko, V. V. (2022). Biocat: Search for biosynthetic gene clusters producing nonribosomal peptides with known structure. *Computational and Structural Biotechnology Journal*, **20**, 1218–1226.

Kunyavskaya, O., Tagirdzhanov, A. M., Caraballo-Rodríguez, A. M., Nothias, L.-F., Dorrestein, P. C., Korobeynikov, A., Mohimani, H., and Gurevich, A. (2021). Nerpa: a tool for discovering biosynthetic gene clusters of bacterial nonribosomal peptides. *Metabolites*, **11**(10), 693.

Landrum, G. *et al.* (2006). Rdkit: Open-source cheminformatics.

Medema, M. H., de Rond, T., and Moore, B. S. (2021). Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics*, **22**(9), 553–571.

Ricart, E., Leclère, V., Flissi, A., Mueller, M., Pupin, M., and Lisacek, F. (2019). rban: retro-biosynthetic analysis of nonribosomal peptides. *Journal of cheminformatics*, **11**(1), 1–14.

Sieber, S. A. and Marahiel, M. A. (2003). Learning from nature's drug factories: Nonribosomal synthesis of macrocyclic peptides. *Journal of Bacteriology*, **185**(24), 7036–7043.

Süssmuth, R. D. and Mainz, A. (2017). Nonribosomal peptide synthesis—principles and prospects. *Angewandte Chemie International Edition*, **56**(14), 3770–3821.

Terlouw, B. R., Blin, K., Navarro-Munoz, J. C., Avalon, N. E., Chevrette, M. G., Egbert, S., Lee, S., Meijer, D., Recchia, M. J., Reitz, Z. L., *et al.* (2023). Mibig 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic acids research*, **51**(D1), D603–D610.