Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Biomed Eng. Author manuscript; available in PMC 2024 November 05.

Published in final edited form as:

Nat Biomed Eng. 2023 June ; 7(6): 707–708. doi:10.1038/s41551-023-01027-z.

Mining for antimicrobial peptides in sequence space

Fangping Wan[1,2,3], **Cesar de la Fuente-Nunez**[1,2,3,✉]

[1]Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA

[3]Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA, USA

Abstract

A machine-learning pipeline identifies potent antimicrobial peptides by gradually narrowing down the search space of polypeptide chain sequences.

Short antimicrobial peptides (AMPs; proteins that are up to 50 amino acids in length) can be cytotoxic to bacteria, viruses and fungi, and can be used to treat bacterial infections. Unlike traditional antimicrobials, which interfere with the synthesis of bacterial nucleic acids or proteins, most AMPs inhibit bacteria by disrupting the bacterial cell membrane, a mechanism of action that makes bacteria less likely to become resistant to AMPs. Antimicrobial resistance, which is exacerbated by the overuse of antimicrobial drugs, is a major global health threat that in 2019 led to an estimated 4.95 million deaths worldwide[1]. Moreover, the large gap between the demand for antimicrobial therapies and the limited number of peptide drugs on the market[2] (about 80) requires high-throughput strategies for AMP discovery. Now reporting in *Nature Biomedical Engineering*, Jian Ji, Junbo Zhao, Peng Zhang and colleagues[3] describe a machine-learning pipeline designed to identify potent AMPs by virtually screening the entire combinatorial space of polypeptide chains of a specified length (Fig. 1).

Models for antimicrobial activity prediction[4–7], AMP-sequence generation[8–10], and haemolysis analyses[11] have shown that machine learning can accelerate the discovery and design of effective AMPs, and models for predicting antimicrobial activity have largely been formulated as a classification problem. Ji and co-authors designed an AMP-screening pipeline leveraging empirical knowledge as well as machine-learning models for performing classification, ranking and regression tasks (Fig. 1). The authors used the pipeline to mine whole libraries of peptide-sequence spaces (for a sequence of length L, there are 20^L possible canonical peptides). Specifically, the pipeline first uses empirical rules to identify peptides that have both a positive net charge and an amphipathic structure. Most known

✉ cfuente@upenn.edu .

Competing interests

The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

AMPs possess these physicochemical properties, so this filtering step quickly rules out molecules that are unlikely to be active as antimicrobial molecules. The filtered peptides are then successively passed on to a binary classifier, a ranker and a regressor. The classifier categorizes the peptides according to whether they are predicted to have antimicrobial activity. The ranker sorts the peptides according to antimicrobial potential. And the regressor predicts their antimicrobial activities (their minimum inhibitory concentration; MIC). The peptides that passed all these selection criteria were considered as AMP candidates.

The models were trained with data from 1,762 AMPs targeting *Staphylococcus aureus* and with experimental MIC values available from the database GRAMPA (for 'giant repository of AMP activities'[12]), and from 5,898 peptides without antimicrobial activity from the UniProt database. The three models used the antimicrobial-activity information in a coarse-to-fine manner. The classifier used the coarsest label information by binarizing the antimicrobial-activity data into two distinct classes (active and inactive). Although a binary classification is more robust to noise and variance in the experimental MIC values, neglecting the antimicrobial-activity values during model fitting may cause the classifier to fail to retrieve peptides with substantially low MICs (that is, with high potency). The ranker used the peptides' rank order according to their MIC values, thus providing a way to determine which of two peptides would have better antimicrobial properties, and the regressor directly predicted the experimental MIC values. Because the experimental MICs used for the training of the regression model were compiled from multiple sources, the heterogeneous experimental conditions, as well as inherent experimental variance, may have reduced the quality of the training data and compromised the reliability of the regressor. However, this effect was mitigated by the classifier and the ranker, which used coarser information from experimental MICs during training, and which therefore may be less sensitive to the noise in the training data. By jointly considering predictions from multiple models, the machine-learning pipeline can be considered as an ensemble-learning framework that uses multiple types of label, feature and model to improve its predictive performance. In addition, to further improve the robustness of the regressor, Ji and co-authors used an incremental learning approach. The authors fine-tuned the regression model on 67 datapoints from peptides randomly selected from the training dataset and whose antibacterial activity was experimentally validated using MIC assays. For model selection, the authors explored different types of model (in particular, a decision-tree-based model taking the physicochemical properties of the peptides as input, as well as a neural network with peptide sequences as input), and selected the best model on the basis of its performance in predicting AMPs from unseen peptides under a standard train/validation/test split on the training dataset.

Ji and co-authors applied the machine-learning pipeline to screen all possible hexapeptides, heptapeptides, octapeptides and nonapeptides. The authors synthesized the 10 highest-ranked hexapeptides and the top 15 heptapeptides, octapeptides and nonapeptides, and experimentally validated them for antimicrobial activity against *S. aureus*. Among the top 55 peptides, 54 of them showed antimicrobial activity (a success rate of 98.2%). Moreover, by removing specific modules in the pipeline and then experimentally validating the resulting top peptides, and on the basis of the distribution of experimental MICs, the authors confirmed that each module in the pipeline improved the identification of AMPs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Furthermore, three hexapeptides showed broad-spectrum antimicrobial activity against Gram-positive and Gram-negative bacteria in vitro (as expected, the peptides disrupted the membrane of the bacteria), as well as efficacy in infected mice, exhibiting low cell toxicity and negligible haemolytic activity.

The machine-learning pipeline relies on empirical data of AMPs, and hence its performance may improve by training on MIC data for other clinically relevant bacterial species (also available in public databases such as GRAMPA). Still, the filtering module based on the peptides' net charge and amphipathicity inevitably biases the search space (for example, towards peptides that are membrane targeting). Hence, expanding the search space for the discovery of AMPs with other mechanisms of action may require the relaxation of the filter's 'exploration–exploitation trade-off'. Ji and co-authors note that mining the whole nonapeptide space (20^9 peptides) with the machine-learning pipeline and synthesizing and characterizing the lead candidates took about 27 days. Because the number of possible peptides grows exponentially with sequence length, discovering AMPs with much longer sequences would be computationally prohibitive; still, implementing informed searching techniques such as genetic algorithms or Bayesian optimization into the pipeline to heuristically explore the peptide-sequence space could lessen the computational burden, and the modular architecture of the pipeline could serve as a general framework for the prediction and discovery of biomolecular properties.

References

1. Murray CJL et al. Lancet 399, 629–655 (2022). [PubMed: 35065702]
2. Muttenthaler M, King GF, Adams DJ & Alewood PF Nat. Rev. Drug Discov 20, 309–325 (2021). [PubMed: 33536635]
3. Huang J. et al. Nat. Biomed. Eng 10.1038/s41551-022-00991-2 (2023).
4. Ma Y. et al. Nat. Biotechnol 40, 921–931 (2022). [PubMed: 35241840]
5. Xu J. et al. Brief. Bioinform 22, bbab083 (2021). [PubMed: 33774670]
6. García-Jacas CR, Pinacho-Castellanos SA, García-González LA & Brizuela CA Brief. Bioinform 23, bbac094 (2022). [PubMed: 35380616]
7. Torres MDT et al. Nat. Biomed. Eng 6, 67–75 (2022). [PubMed: 34737399]
8. Porto WF et al. Nat. Commun 9, 1490 (2018). [PubMed: 29662055]
9. Das P. et al. Nat. Biomed. Eng 5, 613–623 (2021). [PubMed: 33707779]
10. Wan F, Kontogiorgos-Heintz P & de la Fuente-Nunez C Digit. Discov 1, 195–208 (2022). [PubMed: 35769205]
11. Capecchi A. et al. Chem. Sci 12, 9221–9232 (2021). [PubMed: 34349895]
12. Antimicrobial Peptides (GitHub, 2019); https://github.com/zswitten/Antimicrobial-Peptides

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

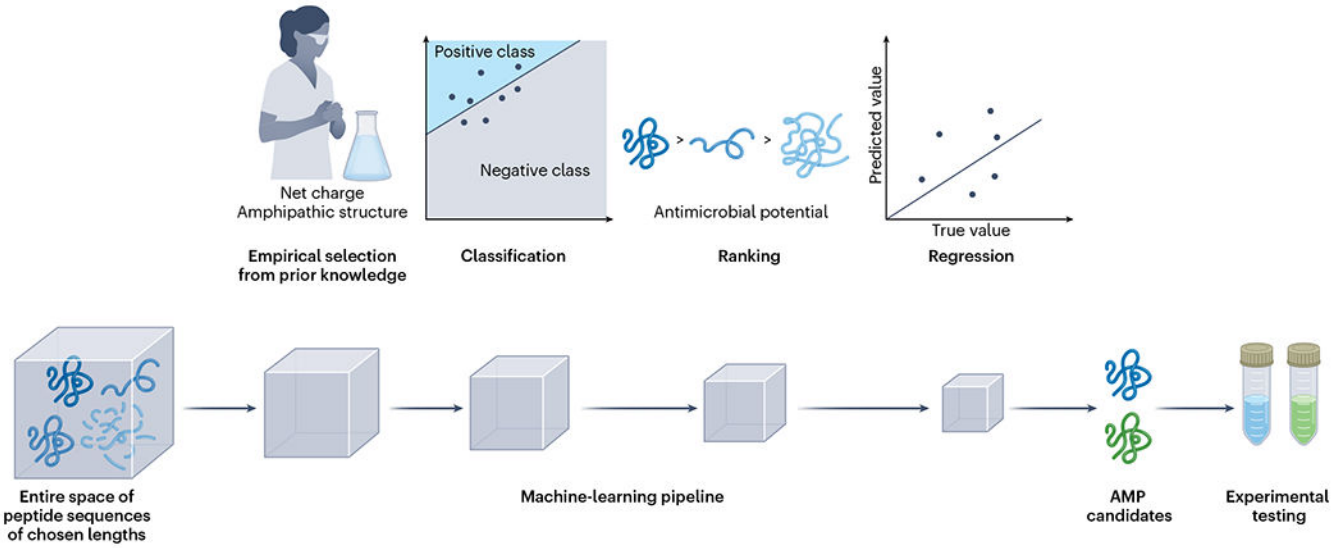

Fig. 1 |. A machine-learning pipeline for the discovery of AMPs.

The pipeline sequentially screens the sequence space of peptides of selected lengths by combining empirical knowledge of known AMPs as well as classifier, ranker and regressor machine-learning modules. The classifier filters peptides according to whether they are predicted to have antimicrobial activity (the ‘positive class’), the ranker orders the peptides according to their antimicrobial potential, and the regressor predicts MICs of the top ranked peptides. The selected peptides are then synthesized and experimentally tested.