Received: 21 July 2022 | Accepted: 27 July 2022

DOI: 10.1002/ctm2.1011

CLINICAL AND TRANSLATIONAL MEDICINE Open Access WILEY

COMMENTARY

Artificial intelligence accelerates the mining of bioactive small molecules from human microbiome

Yuwei Zhang[1,2] | **Pengwei Li**[1] | **Yue Ma**[2,3] | **Jun Wang**[2,3] | **Yihua Chen**[1,2]

[1]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

[3]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

Correspondence

Yihua Chen, State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China.
Email: chenyihua@im.ac.cn

Funding information

National Natural Science Foundation of China, Grant/Award Number: 32025002

With the threat of increasing antibiotic-resistant pathogens and invasive infection and mortality worldwide, the discovery of new and potent pharmaceuticals is of great urgency. Human microbiota provides a tremendous arsenal to discover metabolites with promising antibiotic properties. By applying innovative culture methods for newly isolated microbes or adopting activity-oriented experimental discovery processes, scores of natural products with antibacterial activity have been identified from the human microbiota, for example ribosomally synthesized and post-translationally modified peptides including salivaricin,[1] and non-ribosomal peptide lugdunin[2] (Figure 1). Nonetheless, this approach faces challenges in obtaining natural products produced by unculturable microbes or undetectable under experimental conditions.

With the rapid development of DNA sequencing technology in recent decades, a vast amount of (meta)genomic data from human microbiota has become increasingly available, providing enticing opportunities to unveil the 'dark matters' hidden in their genomes. Biosynthetic genes responsible for the synthesis of bioactive natural products are usually clustered in microbial genome and have different characteristic sequences, making it possible to predict the biosynthetic gene clusters (BGCs) through *in silico* approach relying on sequence signatures. In the past few years, increasing bioinformatics tools, including antiSMASH, ClusterFinder and BAGEL4, which mainly utilize BLAST (basic local alignment search tool) searches or pHMMs (profile hidden Markov models), were well developed for the analysis of BGCs.[3] With the assistance of these tools and various genetic manipulations as well as heterologous expression systems, a number of natural products have been obtained and characterized to give more insights into their ecological roles and interaction with hosts (Figure 1). Take mutanocyclin, which is secreted by *Streptococcus mutans* strains that can cause tooth caries, as an example, this molecule was discovered by the expression of its BGC in a model strain *S. mutans* UA159.[4] Further investigation revealed that mutanocyclin has immunomodulatory activity and can suppress the filamentous growth of *Candida albicans*, indicating its role in helping the *S. mutans* host withstand the pressures from human immune system and the other microbes. Lactocillin[5] with antibacterial activity and dipeptide aldehydes[6] with cathepsin inhibitory activity also possess similar discovery paths. In addition, bioinformatic-inspired structure prediction combined with chemical synthesis, namely syn-BNP (synthetic-bioinformatic natural products), was established to dig out novel peptides derived from human-associated bacteria, such as humimycin,[7] syn-parascrofin and other syn-peptides[8] (Figure 1). This pipeline circumvented the requirements of microbial culture and BGC

Yuwei Zhang and Pengwei Li contributed equally.

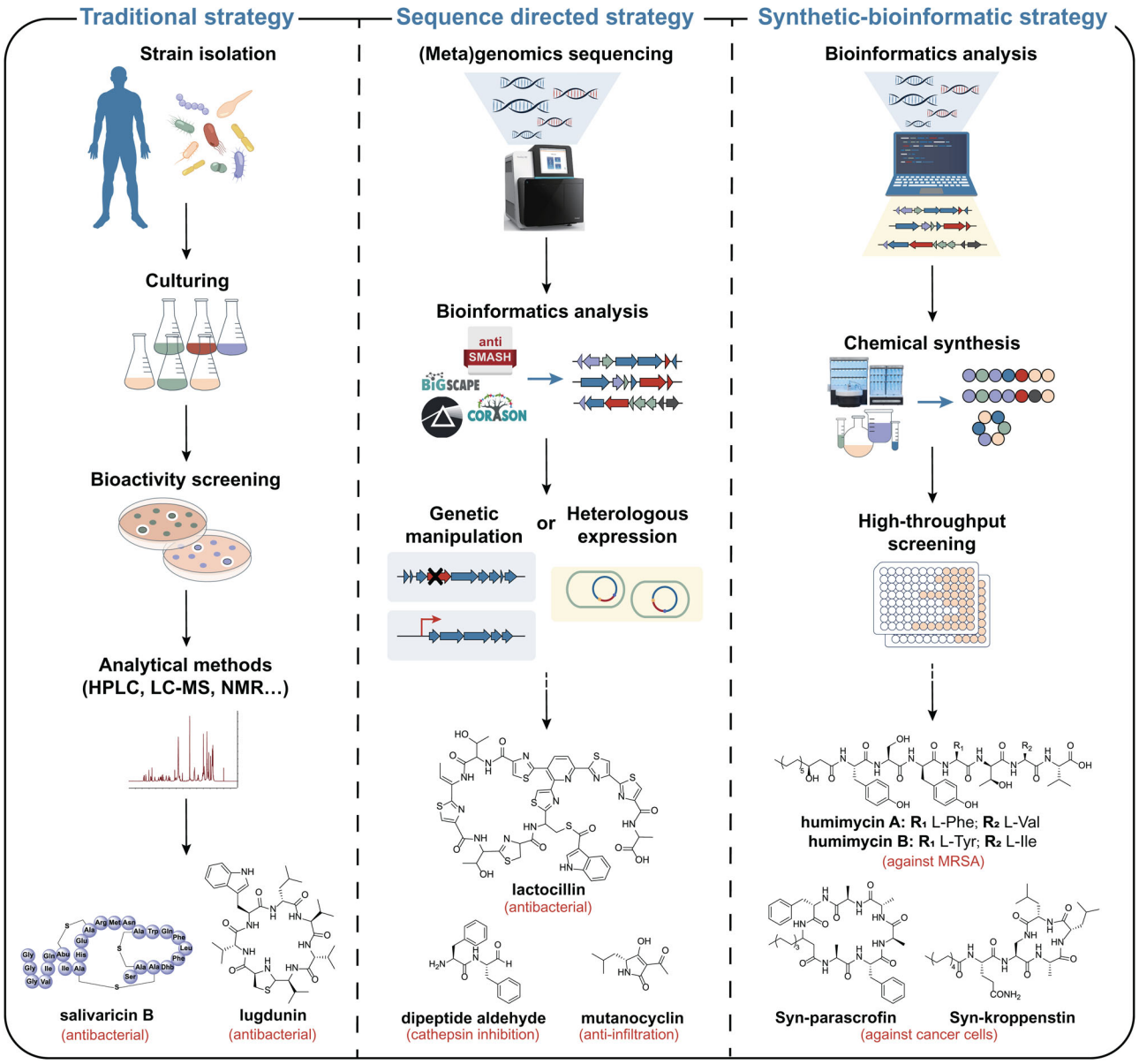

FIGURE 1 Flowcharts of mining bioactive small molecules from human microbiota by different strategies. Left, the traditional strategy relies on the cultivation of microbes, bioactive molecules were discovered through activity-oriented isolation process. Middle, the sequence-directed strategy predicts biosynthetic gene clusters (BGCs) through bioinformatics analysis and obtains the products with the help of genetic manipulation or heterologous expression. Right, the synthetic-bioinformatic strategy generates the small molecules, which are designed by analysis on certain types of BGCs, by chemical synthesis.

expression, thus greatly improving the efficiency of mining bioactive natural products or their analogues. However, methods based on BGCs prediction have certain limitations in the accuracy of the predicted structures, more effective and targeted screening methods are also needed to match the growth rate of (meta)genomic data. Therefore, more promising pipelines and strategies are of great need to expedite the exploitation of novel bioactive natural products from the immense untapped human microbiome.

More recently, artificial intelligence (AI) has been applied to the discovery of antibiotics, some molecules with antibiotic activity were successfully screened from the existing compound library or directly designed,[9] which provides a guarantee for the feasibility of mining in human microbiome using AI-assisted approach. Antimicrobial peptides (AMPs) are considered promising antimicrobial agents due to their wide bioactivity spectra, low tendency to induce resistance and availability through chemical synthesis. However, its relatively short sequence length and high diversity hinder the application of current mining methods. In our work,[10] multiple neural network models (NNMs), including attention, long short-term memory

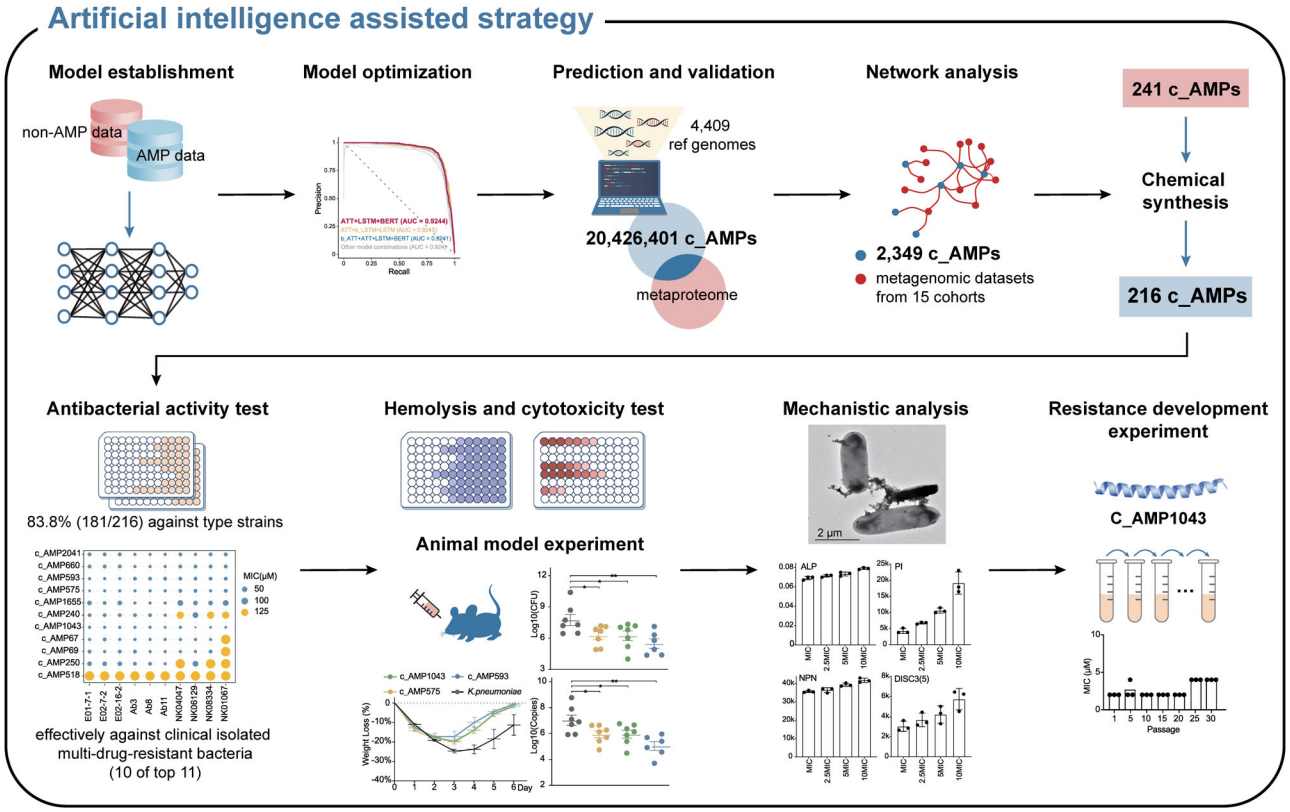

FIGURE 2 Workflow of the artificial intelligence (AI)-assisted strategy for the discovery of antimicrobial peptides (AMPs) from human microbiome

(LSTM) and Bidirectional Encoder Representations from Transformers (BERT), were combined to form a pipeline for AMPs mining, which achieved 91.31% precision and maintained a low false-positive rate (Figure 2). Using this pipeline, a total of 20 426 401 AMPs were predicted from 4409 qualified representative genomes from human microbiome. To ensure that the obtained AMPs are expressed in human body, we performed cross-validation with metaproteomic data and refined the range of candidate AMPs to 2349. Considering the potential negative correlations between functional AMPs and the bacteria they can inhibit, AMP correlation networks were constructed using metagenomic datasets from 15 independent cohorts, and the list of candidate AMPs was further narrowed down to 241. Subsequently, 216 successfully synthesized AMPs were examined in the initial antibacterial activity test, and results showed a positive rate up to 83.8% (181/216). In addition, most of the 181 AMPs showed less than 40% identity with previously reported antibacterial AMPs, indicating that our pipeline was able to identify novel AMPs based on internal relationships of amino acids in the sequences instead of relying on sequence similarity. The top 11 AMPs were further tested and 10 AMPs showed high antibacterial activity against clinically isolated multi-drug-resistant bacteria. Among them, three AMPs with relatively low haemolysis and cytotoxicity were tested on mouse models infected with *Klebsiella pneumoniae* and showed significant therapeutic effects. Moreover, the resistance development experiment of one AMP against *Escherichia coli* DH5α showed no observed resistance after 30-day passaging, which further demonstrates the therapeutic potential of AMPs we identified.

To sum up, the high positive rate, effectiveness and novelty of the AMPs we identified exhibit the high efficiency of AI-assisted approach in the mining of bioactive molecules from human microbiome. Besides the application of NNMs, *in silico* high-throughput and targeted screening using metaproteomic data and association network analysis significantly reduced the excessive workload and cost of traditional in vitro experiments. As shown in our results, the long-term symbiosis and competition existing in human microbiota have derived bioactive molecules like AMPs with clinical application potential. It is worth noting that with different sources of training data and different strategies in the *in silico* screening, the pipeline we developed can also be applied for the exploration of different types of natural products with other desired activities. With the acquisition of more omics data

(including clinical data) and the progress of sequencing technology at the strain-level resolution, AI-assisted mining strategy can effectively accelerate the high-throughput and goal-oriented screening process that was time-consuming in the past, which can be very useful in the discovery of druggable molecules from microbes in different habitats.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant no. 32025002).

ORCID

Yue Ma https://orcid.org/0000-0001-8989-2786
Jun Wang https://orcid.org/0000-0001-9362-512X

REFERENCES

1. Donia MS, Fischbach MA. Small molecules from the human microbiota. *Science*. 2015;349(6246):1254766. doi:10.1126/science.1254766
2. Zipperer A, Konnerth MC, Laux C, et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*. 2016;535(7613):511-516. doi:10.1038/nature18634
3. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol*. 2015;11(9):639-648. doi:10.1038/nchembio.1884
4. Hao TT, Xie ZJ, Wang M, et al. An anaerobic bacterium host system for heterologous expression of natural product biosynthetic gene clusters. *Nat Commun*. 2019;10(1):1-13. doi:10.1038/s41467-019-11673-0
5. Donia MS, Cimermancic P, Schulze CJ, et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*. 2014;158(6):1402-1414. doi:10.1016/j.cell.2014.08.032
6. Guo CJ, Chang FY, Wyche TP, et al. Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell*. 2017;168(3):517-526. doi:10.1016/j.cell.2016.12.021
7. Chu J, Vila-Farres X, Inoyama D, et al. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat Chem Biol*. 2016;12(12):1004-1006. doi:10.1038/nchembio.2207
8. Chu J, Vila-Farres X, Brady SF. Bioactive synthetic-bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome. *J Am Chem Soc*. 2019;141(40):15737-15741. doi:10.1021/jacs.9b07317
9. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*. 2020;180(4):688-702. doi:10.1016/j.cell.2020.01.021
10. Ma Y, Guo ZY, Xia BB, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol*. 2022;40(6):921-931. doi:10.1038/s41587-022-01226-0

How to cite this article: Zhang Y, Li P, Ma Y, Wang J, Chen Y. Artificial intelligence accelerates the mining of bioactive small molecules from human microbiome. *Clin Transl Med*. 2022;12:e1011. https://doi.org/10.1002/ctm2.1011