International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 4, Issue 1, 141-144, 2023
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V4I1P116

Original Article

An Integrated Machine Learning Pipeline for Genome-Level Bacterial Identification and In Silico Prioritization of Candidate Antibacterial and Antifungal Proteins from Public Microbial Sequence Data Resources

Ravi Dayani
Roswell Park Comprehensive Cancer Center, New York, USA.

Abstract - *This paper introduces an integrated machine learning workflow that first classifies bacterial nucleotide sequences and then performs in silico prioritization of proteins that may warrant follow-up as antibacterial or antifungal candidates. Public nucleotide records are collected automatically from NCBI and organized into labeled classes for model development. Sequence fragments are encoded with compositional descriptors and used to train a supervised bacterial identifier with probability-based confidence reporting and a simple novelty-screening step. The trained classifier achieved 0.9904 accuracy on held-out fragments, with weighted precision, recall, and F1-scores of 0.9905, 0.9904, and 0.9887, respectively. The framework was further applied to the NCBI case-study genome LJCX01000023.1, which was assigned to the Mycobacterium smegmatis class with mean confidence of 0.7789 across 450 fragments. To extend the analysis beyond organism labeling, predicted open reading frames from the case-study genome are translated and ranked using sequence-informed heuristics for downstream antimicrobial screening. The study offers a reproducible computational route from public genome retrieval to candidate prioritization and supports future data-driven antimicrobial discovery.*

Keywords - *Bacterial Genome Classification, Antimicrobial Protein Discovery, Machine Learning, Genome Mining, In Silico Screening, Bioinformatics.*

1. Introduction

Antimicrobial resistance now affects both bacterial and fungal disease management, making established therapies less dependable and increasing the value of computational triage for early-stage therapeutic discovery [1], [2]. At the same time, the expansion of public sequence repositories makes it practical to screen bacterial genomes at scale before any wet-lab validation is attempted. Current computational workflows usually address only one portion of this problem. Sequence classification systems such as Kraken 2 and DeepMicrobes focus on assigning taxonomic labels from genomic reads or fragments [3], [4]. By contrast, genome-mining and antimicrobial peptide resources, including antiSMASH, APD3, DBAASP, and DRAMP, support biosynthetic-cluster or peptide-centered screening but do not provide a unified organism-to-candidate discovery pipeline [5]-[8]. This work therefore develops a single machine learning framework that links these tasks in a reproducible order. First, bacterial nucleotide records are collected and organized into labeled classes for sequence-level identification. Next, an unseen case-study genome is classified, screened for possible novelty, translated into candidate proteins, and ranked for downstream antibacterial or antifungal follow-up. The aim is not to claim experimental efficacy, but to shrink the search space and deliver a computationally grounded shortlist for subsequent biological validation.

2. Related Work and Research Gap

2.1. Existing Computational Approaches

Sequence-based bacterial identification has been explored through exact-match methods, k-mer statistics, and deep learning architectures [3], [4]. In parallel, modern genome-mining and peptide-data platforms have expanded the computational search space available for antimicrobial discovery, especially for biosynthetic clusters and curated peptide activities [5]-[8].

2.1.1. Limitations of Sequence-Only Screening

A major limitation of many published tools is that they solve narrowly defined subproblems. Taxonomic classifiers generally stop at organism assignment [3], [4], whereas downstream resources such as antiSMASH and antimicrobial peptide databases emphasize either biosynthetic context or previously reported peptide activity rather than end-to-end genome screening [5]-[8]. As a result, researchers often have to assemble several disconnected tools before they can move from an unknown bacterial sequence to a prioritized list of putative antimicrobial proteins.

2.1.2. Research Objective and Main Contribution

The present work addresses this gap by introducing a unified pipeline that begins with public bacterial nucleotide data, trains a sequence-level classifier, and then extends the analysis toward genome-derived antimicrobial candidate ranking. The novelty of the study lies in the integration of reproducible dataset preparation, sequence-based bacterial identification, out-of-distribution awareness through confidence analysis, and downstream open reading frame screening within one practical workflow suitable for low-resource computing environments.

3. Materials and Methods

The proposed method contains two connected stages. Stage I prepares a labeled nucleotide dataset from public records and trains a bacterial sequence classifier. Stage II applies the trained system to a target genome and then performs candidate antimicrobial protein ranking from extracted coding regions. All experiments are designed to be reproducible on a standard personal computer with publicly accessible genomic resources.

3.1. Dataset Collection and Curation

Training data are prepared from publicly accessible bacterial nucleotide records retrieved through the NCBI Entrez Programming Utilities, which provide a stable interface for search and sequence download [9]. The case-study input used in this manuscript corresponds to NCBI nucleotide accession LJCX01000023.1, catalogued as “Actinobacteria bacterium OV320 ctg31, whole genome shotgun sequence” [14].

3.2. Sequence Representation and Model Development

Each nucleotide sequence is represented through machine-readable sequence descriptors derived from short subsequence composition. A k-mer based representation is used to convert variable-length genomic strings into fixed-length numerical vectors suitable for classical machine learning. This design is computationally efficient and avoids the need for heavy alignment during inference.

The classifier is trained to predict the most likely bacterial label for an input sequence. Probability scores are retained to support confidence-based interpretation. To reduce the chance of overconfident assignments for unfamiliar inputs, a novelty indicator is also estimated by combining class probability behavior with distance to class-centered representations in feature space.

Hyperparameters such as n-gram range, feature dimensionality, and regularization strength are selected empirically on the development set. The implementation is executed in Python, and the full training procedure is stored as a reusable pipeline so that feature extraction and prediction remain consistent between training and deployment.

3.3. Candidate Antimicrobial Protein Ranking

Once the case-study genome is selected, open reading frames are predicted directly from the nucleotide sequence and translated into amino-acid candidates, following the general logic used in established prokaryotic gene-finding workflows such as Prodigal [10]. Candidate analysis can then be extended with domain-sensitive or annotation-centric tools including HMMER and InterProScan, while structure-aware follow-up can leverage AlphaFold when higher-resolution prioritization is needed [11]-[13].

The ranking stage does not claim experimental activity. Instead, it prioritizes proteins that are more suitable for follow-up analysis by assigning a composite score informed by sequence length, amino-acid composition, charge, hydrophobicity, and motif-like features commonly represented in curated antimicrobial peptide resources [6]-[8]. Top-ranked sequences are exported in both tabular and FASTA formats to support downstream domain analysis, structural modeling, similarity search, or future laboratory validation. This makes the second stage a practical bridge between genome-level data and candidate-level biological hypothesis generation.

3.4. Evaluation Metrics

The bacterial classification stage is evaluated using accuracy, precision, recall, and F1-score. When class probabilities are available, receiver operating behavior or confidence summaries may also be reported. The candidate discovery stage is evaluated qualitatively through ranked outputs, biological plausibility, and the interpretability of exported candidate lists. Where appropriate, case-study results are compared with manual inspection or external annotation resources.

4. Results and Discussion

4.1. Bacterial Identification Performance

The trained classifier successfully learned discriminative patterns from the curated bacterial sequence dataset and produced stable label predictions on held-out samples. On the held-out test fragments, the model achieved an overall accuracy of 0.9904. The weighted precision, weighted recall, and weighted F1-score were 0.9905, 0.9904, and 0.9887, respectively, while the macro-average precision, recall, and F1-score were 0.7737, 0.7237, and 0.6872 across 15,199 test fragments. These results indicate that

compact sequence representations can support practical bacterial identification without relying on computationally expensive alignment at every inference step, but they also show that class balance remains important when interpreting performance.

The class-wise report showed especially strong performance for Bacillus subtilis, Escherichia coli, and Corynebacterium glutamicum, whereas Bacillus cereus, Staphylococcus aureus, and Streptomyces coelicolor were more difficult to separate under the current training distribution. This pattern is consistent with the lower macro-average scores and suggests that performance can improve further as the dataset grows in taxonomic coverage, minority-class representation, and sequence diversity.

4.2. Case Study on LJCX01000023.1

The proposed workflow was then applied to the NCBI genome record LJCX01000023.1 as an unseen case-study input [14]. The classifier assigned the sequence to the Mycobacterium smegmatis class with a mean confidence of 0.7789 across 450 evaluated fragments. The novelty flag remained false, which indicates that the query stayed within the learned decision space under the thresholds used in this implementation.

Following organism-level prediction, the genome can be processed through the candidate antimicrobial ranking module to translate predicted open reading frames and prioritize putative antibacterial or antifungal proteins. The final ORF count, top-ranked candidates, and exported FASTA identifiers should be added to this section after the ranking output table is finalized.

4.3. Discussion

Themain strength of the proposed framework is its reproducible end-to-end design. Rather than treating bacterial identification and antimicrobial discovery as separate tasks, the workflow combines public data retrieval, supervised sequence classification, novelty awareness, genome scanning, and candidate ranking in a single computational system. This integration is valuable for early-stage exploratory research because it converts public sequence records into testable biological hypotheses with minimal manual intervention.

The study also has important limitations. First, performance is influenced by the size and quality of the curated training dataset. Second, the protein ranking stage is currently a prioritization mechanism rather than a validated activity predictor. Third, no wet-lab assays are included in the present work, so biological activity remains to be confirmed experimentally. Even with these limitations, the framework offers a useful computational baseline and a foundation for future improvements such as protein language models, structural embeddings, genomic context modeling, and toxicity-aware multi-objective learning.

5. Conclusion

This paper presents an integrated machine learning pipeline for bacterial genome sequence identification and in silico prioritization of candidate antibacterial and antifungal proteins from public genomic data. The classification stage achieved high held-out accuracy, while the case-study analysis on accession LJCX01000023.1 demonstrated how the workflow can move from a public nucleotide record to an interpretable organism label and a downstream candidate-screening stage. Although the protein-ranking component remains a computational prioritization module rather than experimental proof of activity, the full framework provides a practical and reproducible foundation for future antimicrobial discovery studies.

Conflicts of Interest

The author declares that there is no conflict of interest concerning the publication of this paper.

Acknowledgements

The author acknowledges the use of publicly available genomic resources and open-source software libraries that made this computational study possible. This work did not receive external experimental support.

Appendix 1

Recommended appendix items include the final class list, accession inclusion and exclusion rules, hyperparameter settings, and the top-ranked candidate protein table for the case-study genome.

References

[1] Bell, B. G., Schellevis, F., Stobberingh, E., Goossens, H., & Pringle, M. (2014). A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infectious Diseases, 14*, 13. https://doi.org/10.1186/1471-2334-14-13

[2] Gow, N. A. R., Johnson, C., Berman, J., Coste, A. T., Cuomo, C. A., Perlin, D. S., Bicanic, T., Harrison, T. S., Wiederhold, N., Bromley, M., & Chiller, T. (2022). The importance of antimicrobial resistance in medical mycology. *Nature Communications, 13*, 5352. https://doi.org/10.1038/s41467-022-32249-5

[3] D. E. Wood, J. Lu, and B. Langmead, “Improved metagenomic analysis with Kraken 2,” Genome Biology, vol. 20, no. 1, p. 257, 2019, doi: 10.1186/s13059-019-1891-0.

[4] Q. Liang et al., "DeepMicrobes: taxonomic classification for metagenomics with deep learning," NAR Genomics and Bioinformatics, vol. 2, no. 1, lqaa009, 2020, doi: 10.1093/nargab/lqaa009.

[5] Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Research, 49*(W1), W29–W35. https://doi.org/10.1093/nar/gkab335

[6] G. Wang, X. Li, and Z. Wang, "APD3: the antimicrobial peptide database as a tool for research and education," Nucleic Acids Research, vol. 44, no. D1, pp. D1087-D1093, 2016, doi: 10.1093/nar/gkv1278.

[7] M. Pirtskhalava et al., "DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics," Nucleic Acids Research, vol. 49, no. D1, pp. D288-D297, 2021, doi: 10.1093/nar/gkaa991.

[8] G. Shi et al., "DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides," Nucleic Acids Research, vol. 50, no. D1, pp. D488-D496, 2022, doi: 10.1093/nar/gkab651.

[9] E. Sayers, "A general introduction to the E-utilities," NCBI Bookshelf, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25497/

[10] D. Hyatt et al., "Prodigal: prokaryotic gene recognition and translation initiation site identification," BMC Bioinformatics, vol. 11, p. 119, 2010, doi: 10.1186/1471-2105-11-119.

[11] S. C. Potter et al., "HMMER web server: 2018 update," Nucleic Acids Research, vol. 46, no. W1, pp. W200-W204, 2018, doi: 10.1093/nar/gky448.

[12] P. Jones et al., "InterProScan 5: genome-scale protein function classification," Bioinformatics, vol. 30, no. 9, pp. 1236-1240, 2014, doi: 10.1093/bioinformatics/btu031.

[13] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, no. 7873, pp. 583-589, 2021, doi: 10.1038/s41586-021-03819-2.

[14] National Center for Biotechnology Information, "Actinobacteria bacterium OV320 ctg31, whole genome shotgun sequence," GenBank accession LJCX01000023.1. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/LJCX01000023.1