Predicting Anti-microbial Resistance using Large Language Models

Hyunwoo Yoo, Bahrad Sokhansanj, James R. Brown, Gail Rosen
Drexel University
{hty23, bas44, jb4633, glr26}@drexel.edu

Abstract

During times of increasing antibiotic resistance and the spread of infectious diseases like COVID-19, it is important to classify genes related to antibiotic resistance. As natural language processing has advanced with transformer-based language models, many language models that learn characteristics of nucleotide sequences have also emerged. These models show good performance in classifying various features of nucleotide sequences. When classifying nucleotide sequences, not only the sequence itself, but also various background knowledge is utilized. In this study, we use not only a nucleotide sequence-based language model but also a text language model based on PubMed articles to reflect more biological background knowledge in the model. We propose a method to fine-tune the nucleotide sequence language model and the text language model based on various databases of antibiotic resistance genes. We also propose an LLM-based augmentation technique to supplement the data and an ensemble method to effectively combine the two models. We also propose a benchmark for evaluating the model. Our method achieved better performance than the nucleotide sequence language model in the drug resistance class prediction.

1 Introduction

The genes for antibiotic resistance have increased rapidly over the past 10 years and have become a threat to human health (Zhang et al., 2022). Moreover, dangerous infectious diseases like COVID-19 can also spread. In such times, it is important to classify the DNA sequences of antibiotic resistance genes. In bioinformatics, the main method for classifying DNA sequences has been to find similar sequences by aligning two DNA sequences using text alignment (Bonin et al., 2023). Recently, there have been methods that use language models created from the nucleotide or
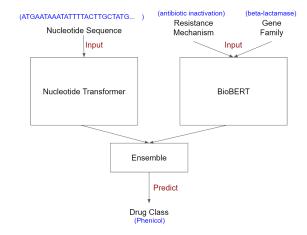

Figure 1: Overview of our approach

protein sequences of various species and fine-tune them to create classifiers(Brandes et al., 2022; Ji et al., 2021; Zhou et al., 2023). These methods have the advantage of being able to identify which parts of the nucleotide sequence are important. To fine-tune, databases containing information on antibiotic resistance genes must be used. The main databases are CARD (Jia et al., 2017) and MEGARes (Doster et al., 2020). Existing methods use the labels associated with antibiotic resistance genes, such as the class to which the resistance gene belongs, for example, the label of the antibiotic to which resistance is present. It is a prediction of a single label from a single gene sequence (Kang et al., 2022). However, if we look at the CARD or MEGARes databases, there are several attributes that describe a particular gene. There are Gene Family and Resistance Mechanism. If we use this information when predicting the antibiotic to which resistance is present, it could be helpful for prediction. Here, we get an idea and propose a model that uses human-readable information to predict antibiotic resistance genes. We also provide a method to merge the different classification systems of

Output	Input Example	BioBERT
Base	Gene Family: Beta-lactamases, Resistance Mechanism: Antibiotic incativation	78.20
Entity marker (punct)	[Gene Family]: Beta-lactamases, [Resistance Mechanism]: Antibiotic incativation	77.41
Typed entity marker	*Beta-lactamases*, #Resistance Mechanism#	77.70
Typed entity marker (punct)	*[Gene Family]: Beta-lactamases*, #[Resistance Mechanism]: Antibiotic incativation#	78.46

Table 1: Test macro F1 score of different entity representation techniques in Antibiotic Resistance Drug Class Prediction with BioBERT.

CARD and MEGARes. We will also explain the LLM-based data augmentation technique for rare classes with few samples.

2 Approaches

Our approaches include fine-tuning a pre-trained language model with various species' gene nucleotide sequence data to predict antibiotic resistance genes and their classes. We also fine-tune a pre-trained language model trained on a corpus containing diverse papers from the fields of biology and medicine to predict the names of antibiotic resistance gene properties. We provide an effective ensemble model (Kumari et al., 2021) using the above two models in a weighted soft voting method. To integrate the classes, we combine the DNA sequences and the concepts that describe them from CARD and MEGARes into one. We use the EBI ARO ontology (Cook et al., 2016) to combine CARD tagging and MEGARes tagging into one class system. For rare classes with few samples, we use BioGPT (Luo et al., 2022) prompting to perform data augmentation.

2.1 Nucleotide Sequence Based Antibiotic Resistance Drug Class Classification

Following the structure of (Dalla-Torre et al., 2023), we uses a large pre-training language model based on nucleotide sequences and fine-tune a classifier based on Drug Class data. The nucleotide sequence input is limited to a length of 1000, the input size of the pre-training model. The tokenizer uses a 6-mer tokenizer. A 6-mer tokenizer is a type of k-mer tokenizer. A k-mer tokenizer is a technique used in genome analysis and bioinformatics research that splits a biological sequence into substrings of length k (Mejía-Guerra and Buckler, 2019). The pre-training model uses NT, which is pre-trained on multi-species including bacteria, fungi, invertbate, protozoa, verterbate gene sequences. Unlike other nucleotide sequence-based pre-training models that mostly use human genes, this model is trained on multi-species genes, providing a better

representation. Fine-tuning is done using LoRA tuning. LoRA tuning is a method that fixes the weights of a pre-trained large-scale language model and inserts a low-rank decomposed matrix into each transformer layer, dramatically reducing the number of trainable parameters for the downstream task (Hu et al., 2021). This allows for more effective fine-tuning.

2.2 Text Information Based Antibiotic Resistance Drug Class Classification

Text information based antibiotic resistance drug class classification uses a BioBERT language model pre-trained on a large medical and biological text corpus as the pre-training model. BioBERT is a pre-trained biomedical language representation model that uses a large-scale biomedical text corpus including PubMed abstracts, PMC full-text articles, and the Genia corpus. (Lee et al., 2020) We fine-tune this model to extract antibiotic resistance drug classes, such as Drug Class or Gene Family, from text that describes antibiotic resistance genes. We aim to improve the performance of the classifier by utilizing a pre-trained biomedical text-based model. Instead of using multiple classification layers, we create a single classification layer and fine-tune it. The training data is structured as [Resistance Mechanism] followed by a description of the attribute, such as Antibiotic inactivation. To further improve performance, we create a format that encloses special characters (Zhou and Chen, 2021), such as *[Gene Family]: Beta-lactamases*, #[Resistance Mechanism]: Antibiotic inactivation#.

2.3 Weighted Soft-voting Ensemble

To combine the pre-trained nucleotide sequence-based language model and the pre-trained text-based language model mentioned earlier, we use a soft-voting ensemble model. Additionally, we find the optimal weights through validation data and apply them to create a weighted soft voting ensemble model. A more detailed explanation of the validation data will be provided in the Experiment section. This data is a third dataset separate from

Method	Accuracy	Macro F1	Precision	Recall
NT	84.15	64.04	72.78	59.28
NT with data augmentation	83.42	64.85	80.15	58.65
NT with reads	82.85	61.02	68.32	57.06
NT with reads and data augmentation	83.11	62.82	74.81	57.32

Table 2: Result of data augmentation for the class which has small samples. Data augmentation increases the F1 score.

the training and test data. This allows us to use both nucleotide sequence information and the text information that describes it. This model requires both types of input. It receives the nucleotide sequence and information about Gene Family and Resistance Mechanism in the format [Resistance Mechanism]: Antibiotic Effuls, #[Gene Family]: Bata-Lactamases#.

2.4 Integrating Classes Based on Antibiotic Resistance Ontology

The databases provided in the literature (CARD, MEGARes) have different classification systems and hierarchical relationships. EBI ARO provides hierarchical information on antibiotic resistance genes. EBI stands for European Bioinformatics Institute. These diverse antibiotic resistance classification systems, gene groupings, and resistance mechanisms can be combined through the EBI ontology, and the model can store integrated concept representations. Each database's header is read and the EBI API is searched. The mapped items are used as new Gene Family. Rather than using very small and specific hierarchical classes, more general hierarchical classes are employed. The third level from the top in the EBI ARO hierarchy is used as the basis.
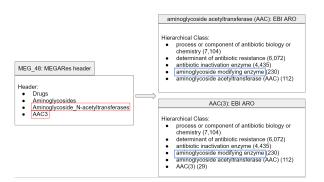

Figure 2: EBI ARO Gene Family mapping: search to find mapping information with header and ontology by using API.

2.5 Data Augmentation Using a Large Language Model

The categories were integrated based on the EBI ARO Ontology's gene group and CARD Resistance Mechanism. However, there are still cases where the number of samples corresponding to a class is small. Data augmentation was conducted for these cases. BioGPT was used for data augmentation. Similar data were created through prompting. Through this, it was possible to see that performance improved as follows: In particular, the accuracy in classes with a small number of samples increased.

3 Experiments

3.1 Datasets

The CARD and MEGARes v3 datasets are used for training and evaluation. Classes with fewer than 15 samples are removed because obtaining meaningful results from the data split is difficult. The remaining data is split into 75% for training data, 20% for test data, and 5% for validation data. EBI ARO ontology search is used to integrate the data, which is then split similarly to the above. Classes with difficult-to-obtain meaningful results are also removed. The MEGARes dataset consists of 9733 Reference Sequences, 1088 SNPs, 4 antibiotic types, 59 resistance classes, and 233 mechanisms. The CARD dataset consists of 5194 Reference Sequences and 2005 SNPs, 142 Drug Classes, 331 Gene Families, and 10 Resistance Mechanisms. The EBI ARO ontology provides hierarchical group information for genes. Using the EBI ARO Ontology, Gene Family class information can be integrated into a higher-level hierarchy. The number of Gene Family text information classes in the case of MEGARes is 589, while for CARD, it is 331. There are 300 and 166 datasets with only one sample in their respective classes for Gene Family in the case of MEGARes and CARD, respectively. Resistance Mechanism is integrated based on the 6 categories of CARD. The original

Dataset	Method	Accuracy	Macro F1	Precision	Recall
CARD	NT	87.92	63.08	66.46	61.51
CARD	BB	97.22	89.68	92.09	90.54
CARD	Ensemble	97.55	93.44	95.72	92.86
MEGARes	NT	89.61	46.42	54.92	43.94
MEGARes	BB	99.64	99.47	99.96	99.03
MEGARes	Ensemble	99.99	99.99	99.99	99.99
Integrated	NT	82.89	65.79	81.84	58.67
Integrated	BB	90.26	79.34	84.05	77.14
Integrated	Ensemble	92.11	80.95	83.52	78.94
Integrated with reads	NT	83.11	62.82	74.81	57.32
Integrated with reads	BB	90.24	79.34	84.05	77.14
Integrated with reads	Ensemble	93.40	81.85	84.34	80.25

Table 3: Result of using the CARD, MEGARes, and Integrated databases for antibiotic resistance drug class prediction using Nucleotide Transformer(NT), BioBERT(BB), and a weighted ensemble of both. The weighted ensemble with Nucleotide Transformer(NT) and BioBERT(BB) shows better performance in every datasets.

8 categories were reduced to 6, excluding cases of various class combinations and those with very few samples. Drug Class is integrated using 9 common Drug Classes found in competing models. Integration is done based on names and theories and has been verified. Macro f1 score, accuracy, balanced accuracy, and precision are used as performance metrics, and the results are listed in the table 3.

3.2 Implementation Details

Basic structure of the model and fine-tuning follow the methods proposed by BioBERT and Nucleotide Transformer. The layers and information of the model are in the Appendix.

3.3 Main Results

Tables 3 show metrics using our method with the latest techniques (SOTA) in the text-based information model for the CARD and MEGARes experiments, showing that our method surpasses previous SOTA. Additionally, the method using integrated data shows superiority over previous SOTA. Our method also demonstrates competitive results compared to other competing models and SOTA.

4 Discussion

Does text information help?

In all datasets, using a text information-based language model shows a 9.53 accuracy and 30.34 macro f1 score improvement in CARD and 10.38 accuracy and 50.57 macro f1 score improvement

in MEGARes. Adjusted ratio ensemble models show better performance compared to other cases through experiments. Existing NT and other nucleotide sequence-based models find it difficult to process natural language. Our fine-tuned text-based language model was trained using a small amount of pre-training resources (40GB A100 GPU). By constructing an ensemble model, it achieves better performance compared to competing models such as AMR-meta (Marini et al., 2022), Meta-MARC (Lakin et al., 2019), and Deep ARG (Arango-Argoty et al., 2018).

Does text information class integration help?

To compare with other models, we integrated the class system. This enables comparison with competing models. It also allows us to create models for predicting Gene Family and Resistance Mechanism. In particular, the number of samples corresponding to classes in Gene Family and Resistance Mechanism is very small in many cases. This integration helps to implement Gene Family and Resistance Mechanism prediction models. The integrated class system shows better performance compared to cases where it is not. The number of genes available for training increases.

Sequencing Read Generation

In some competing models, it is recommended to use reads instead of full genes. In the case of AMR-meta, it aims to predict paired end genes. To compare with these models, it is necessary to

generate reads. Reads generation uses ART. ART is a simulator for analyzing nucleotide sequences, and it helps with accurate modeling of biological information data as a software (Huang et al., 2012). ART has the advantage of customizable indel error rates (Milhaven and Pfeifer, 2023). The learning and experiments using these reads are presented in Table . In this experiment, the proposed model also demonstrates strong competitiveness.

5   Related Work

AMR-meta is a method for classifying antibiotic resistance in high-speed metagenomic data. This method uses a sequence alignment-free approach based on k-mers and meta-features, and it utilizes both resistant and non-resistant genes as training data. As a result, AMR-meta can more accurately identify antibiotic resistance genes and reduce false-positive rates for non-resistant genes. However, it uses a complex matrix decomposition method to generate meta-features, which can be computationally intensive. Additionally, the prediction performance of AMR-meta may vary depending on the type of antibiotic used or the diversity of the resistance genes. These characteristics make AMR-meta useful for analyzing high-speed metagenomic data, but at the same time, they suggest that it may be limited in certain situations.

AMR++ is a customized bioinformatics pipeline that uses high-throughput sequencing data to predict the diversity and abundance of antibiotic resistance genes (ARGs). This pipeline is integrated with the MEGARes database, allowing for efficient analysis of ARGs in large-scale metagenomic sequencing data. The main advantage of AMR++ is its high throughput and efficiency, enabling users to quickly and accurately analyze complex datasets. In addition, this software can distinguish between types of ARGs, including cases where resistance genes require specific mutations. However, this pipeline requires high-quality assembled and/or translated data, which may cause difficulties or limitations in generating metagenomic datasets. Furthermore, AMR++ may require advanced bioinformatics skills and resources, potentially limiting accessibility for some researchers.

Meta-MARC is a machine learning classifier developed to enhance the detection and classification of antibiotic resistance genes. This system is based on the MEGARes database and uses DNA-based hierarchical Hidden Markov Models (HMMs) to classify antibiotic resistance genes in high-throughput sequencing data. Meta-MARC is robust against various gene mutations, which is particularly useful for non-standard databases and sequences. This tool provides high sensitivity and specificity, playing a crucial role in accurate antibiotic resistance detection. However, Meta-MARC is computationally demanding, particularly when dealing with large datasets, which can result in increased processing time and memory usage. Additionally, high sensitivity settings may potentially increase false positives, so users must carefully interpret the results.

DeepARG is a deep learning-based system used for predicting antibiotic resistance genes (ARGs) in metagenomic data. It utilizes two models, DeepARG-SS and DeepARG-LS, for classifying short and full-length gene sequences. Compared to the traditional 'best hit' approach, it has the advantage of identifying a wider range of ARG diversity with lower false negative rates. However, the performance of this system heavily depends on the quality of the training database, and it has limitations when it comes to predicting new categories of ARGs. Despite these limitations, DeepARG is a useful tool for evaluating the presence and diversity of ARGs in environmental samples.

6   Conclusion

As far as we know, our work is the first to combine natural language models and biological sequence models to predict antibiotic resistance genes. We proposed a model that combines two different attribute language models into an ensemble. By using both nucleotide sequence information and its description, including Gene family and resistance mechanism information, it enables more accurate drug class predictions. We also integrated various databases using the EBI ontology and used a large language model (LLM) for data augmentation in classes with insufficient data. As a result, we achieved performance close to the state-of-the-art. We believe this fusion has significant meaning. Moreover, we tested the structure we trained using only nucleotide sequences and obtained acceptable results. This seems promising for future research.

Acknowledgements

References

Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S. Heath, Peter Vikesland, and Liqing Zhang. 2018. Deeparg: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):23.

Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S Morley, and Christina Boucher. 2023. Megares and amr++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Research*, 51(D1):D744–D752.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. 2016. The european bioinformatics institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Genomics*.

Enrique Doster, Steven M Lakin, Christopher J Dean, Cory Wolfe, Jared G Young, Christina Boucher, Keith E Belk, Noelle R Noyes, and Paul S Morley. 2020. Megares 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research*, 48(D1):D561–D569.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv*. ArXiv:2106.09685v2.

Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. Art: A next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E. Williams, Jonathan G. Frye, Tariq Elsayegh, Daim Sardar, Erin L. Westman, Andrew C. Pawlowski, Timothy A. Johnson, Fiona S.L. Brinkman, Gerard D. Wright, and Andrew G. McArthur. 2017. Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573.

Hyeunseok Kang, Sungwoo Goo, Hyunjung Lee, Jung-Woo Chae, Hwi-Yeol Yun, and Sangkeun Jung. 2022. Fine-tuning of bert model to accurately predict drug-target interactions. *Pharmaceutics*, 14(8):1710.

Saloni Kumari, Deepika Kumar, and Mamta Mittal. 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2:40–46.

Steven M. Lakin, Alan Kuhnle, Bahar Alipanahi, Noelle R. Noyes, Chris Dean, Martin Muggli, Rob Raymond, et al. 2019. Hierarchical hidden markov models enable accurate and diverse detection of antimicrobial resistance sequences. *Communications Biology*, 2(1):294.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Simone Marini, Marco Oliva, Ilya B Slizovskiy, Rishabh A Das, Noelle Robertson Noyes, Tamer Kahveci, Christina Boucher, and Mattia Prosperi. 2022. Amr-meta: A k -mer and metafeature approach to classify antimicrobial resistance from high-throughput short-read metagenomics data. *GigaScience*, 11. Giac029.

María Katherine Mejía-Guerra and Edward S. Buckler. 2019. A k-mer grammar analysis to uncover maize regulatory architecture. *BMC Plant Biology*, 19(1):103.

Mark Milhaven and Susanne P. Pfeifer. 2023. Performance evaluation of six popular short-read simulators. *Heredity*, 130(2):55–63.

Zhenyan Zhang, Qi Zhang, Tingzhang Wang, Nuohan Xu, Tao Lu, Wenjie Hong, Josep Penuelas, Michael Gillings, Meixia Wang, Wenwen Gao, and Haifeng Qian. 2022. Assessment of global health risk of antibiotic resistance genes. *Nature Communications*, 13.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv*. ArXiv:2102.01373v4.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv*. ArXiv:2306.15006v1.