Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240381

Prediction of *A. Baumannii* Amikacin Resistance in Clinical Metagenomics

Michael Aaron SY [a], Mattia PROSPERI[a], Mohammadali SERAJIAN[b], Christina BOUCHER[b], Panayotis Takis BENOS [a] and Simone MARINI[a,1]
[a] *Department of Epidemiology, University of Florida, Gainesville, FL, USA*
[b] *Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA*

Abstract. Respiratory tract infections are a serious threat to health, especially in the presence of antimicrobial resistance (AMR). Existing AMR detection methods are limited by slow turnaround times and low accuracy due to the presence of false positives and negatives. In this study, we simulate 1,116 clinical metagenomics samples on both Illumina and Nanopore sequencing from curated, real-world sequencing of *A. baumannii* respiratory infections and build AI models to predict resistance to amikacin. The best performance is achieved by XGBoost on Illumina sequencing (area under the ROC curve = 0.7993 on 5-fold cross-validation).

Keywords. Antimicrobial resistance, AMR, Simulation, AI, metagenomics

1. Introduction

A critical complication of respiratory tract infections is the presence of antimicrobial resistance (AMR), impairing antibiotic treatment [1]. The current AMR clinical detection methods are multiplex PCR (mPCR) and antimicrobial susceptibility tests (ASTs). While mPCR is fast (24-hour turnaround), it is prone to false positives and negatives. On the other hand, AST is precise but slow (up to 5 days). The potential of AI in predicting AMR has been reviewed in recent literature [2]. Here we present a novel method combining AI and clinical metagenomics to address the need for rapid and accurate AMR detection.

2. Methods

We collected 184 *A. baumannii* genomes from human respiratory infections and related amikacin antimicrobial susceptibility test data from the BV-BRC database [3] and 88 commensal genomes from NCBI based on respiratory infections literature [4,5]. These genomes were used to create 1116 simulated (584 resistant, 532 susceptible) clinical metagenomics samples using PBSIM2 [6] for Nanopore and InScilicoSeq [7] for Illumina sequencing. Each simulated sample had 250 million bases, with 10% pathogen sequences. Nanopore had a median pathogen coverage of 28.72x (IQR: 22.62) with 11,104-base reads; Illumina had 27.74x (IQR: 22.92) with 150-base reads. We split the

[1] Corresponding Author: Simone Marinil; E-mail: simone.marini@ufl.edu.

sequences into *k*-mer (*k*=13 as suggested by literature [8]) and use them as features. We selected the top 100 *k*-mers (chi-squared, performed independently for each fold of a cross-validation) to train our models. We trained various AI models with scikit-learn (Table 1) default parameters. To avoid overfitting, we tested their performance using 5-fold cross-validation.

3. Results, Discussion and Conclusions

For Illumina and Nanopore, the best-performing approach is XGBoost, with the area under the receiver operating characteristic curve (AUROC) of 0.7993; for Nanopore, it is the Support Vector Classifier (0.61 AUROC). While longer, Nanopore reads are less accurate than Illumina, and this might explain the generally lower performance. These results are promising especially considering the low amounts of bacterial DNA used in the samples.

Table 1. A Illumina Simulation Metrics (Inscilicoseq)

Metrics 5-FoldCV	Random Forest	XGboost	SVC	Decision Trees	Logistic Regression	Naïve Bayes
Accuracy	0.7104	**0.7265**	0.6452	0.6424	0.6066	0.6227
F1 Score	0.7337	**0.7342**	0.6845	0.6589	0.6363	0.6058
AUROC	0.7766	**0.7993**	0.6948	0.6437	0.6665	0.6740

Table 1. B Nanopore Simulation Metrics (PBsim2)

Metrics 5-FoldCV	Random Forest	XGboost	SVC	Decision Trees	Logistic Regression	Naïve Bayes
Accuracy	0.5105	0.5446	0.5258	0.5330	**0.5652**	0.5429
F1 Score	0.5821	0.5789	0.5700	0.5565	**0.5915**	0.5731
AUROC	0.4789	0.5491	**0.6172**	0.5318	0.5967	0.5643

References

[1] Schuetz P, Albrich W, Christ-Crain M, Chastre J, Mueller B, Stolz D, et al. Economic evaluation of procalcitonin-guided antibiotic therapy in acute respiratory infections: a US health system perspective. Clin Chem Lab Med. 2015 Mar;53(4):583-92. doi: 10.1515/cclm-2014-1015.

[2] Sakagianni A, Varlamis I, Pratikakis P. Using Machine Learning to Predict Antimicrobial Resistance-A Literature Review. Antibiotics (Basel). 2023 Feb;12(3):452. doi: 10.3390/antibiotics12030452.

[3] Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res. 2022 Nov;51(D1). doi: 10.1093/nar/gkac1003.

[4] Pienkowska K, Wiehlmann L, Tümmler B. Airway microbial metagenomics. Microbes Infect. 2018 Oct;20(9-10):536-42. doi: 10.1016/j.micinf.2017.07.005.

[5] Kullberg RFJ, Tsonaka R, Rijkers GT, Meijers JCM, Cremer OL, Nossent EJ, et al. Lung Microbiota of Critically Ill Patients with COVID-19 Are Associated with Nonresolving Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2022 Oct;206(7):846-56. doi: 10.1164/rccm.202202-0274OC.

[6] Ono Y, Asai K, Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. Bioinformatics. 2021 Mar;37(5):589-95. doi: 10.1093/bioinformatics/btaa835.

[7] Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics. 2019 Feb;35(3):521-2. doi: 10.1093/bioinformatics/bty630.

[8] Marini S, Andersson AF, Berendonk TU, Berglund F, Fick J, Kristiansson E, et al. AMR-meta: a k-mer and metafeature approach to classify antimicrobial resistance from high-throughput short-read metagenomics data. GigaScience. 2022 Jan;11. doi: 10.1093/gigascience/giac029.