Improving antimicrobial resistance surveillance and diagnostics with machine learning predictions from genomic data

1^{st} Jee In Kim
PhD Candidate
Faculty of Graduate Studies
Dalhousie University
Halifax, Canada
jeein.j.kim@gmail.com

2^{nd} Dr. Robert Beiko
Faculty of Computer Science
Dalhousie University
Halifax, Canada
beiko@cs.dal.ca

***Abstract*—Antimicrobial resistance (AMR) refers to the phenomenon where antibiotic drugs can no longer inhibit the growth of bacteria. AMR is rapidly increasing worldwide, with grave consequences for our ability to treat infectious diseases. Multidrug-resistant pathogens (or 'superbugs') are rapidly emerging in healthcare and agriculture settings due to the high volume of antibiotic use and misuse that pressures microbes to quickly develop defence mechanisms. With the development of rapid and affordable genome sequencing, we can study the complete set of genes in an organism, including genes conferring AMR. However, it is still challenging to predict an organisms' resistance behaviour based on the AMR genes alone. This investigation uses machine-learning (ML) tools to predict how a pathogen will respond to antibiotic treatment. The performance of ML models with 300+ *Enterococcus faecium* genomes as the learning data demonstrated maximum accuracy of 98% when predicting resistance against antibiotics like vancomycin. Other genomic data encoding methods and feature selection demonstrated that ML can accurately predict organisms' resistance to specific antibiotics and appropriately select highly relevant features that contribute to resistance. The newly established genetic features that have no previous connection to resistance will be investigated in the laboratory to confirm their contribution to bacteria's resistance behaviour. The potential research findings will hopefully assist in the early prevention and mitigation effort of AMR.**

***Index Terms*—antimicrobial resistance, machine learning, feature evaluation, genomics**

I. Introduction

Antimicrobial resistance (AMR) refers to the phenomenon where antibiotic drugs can no longer inhibit the growth of bacteria, posing a threat to modern medicine. The high volume of antibiotic use in various sectors, such as human health and agriculture, has fueled the acceleration of AMR emergence, which is considered a global health crisis [4]. The increasing use of antibiotics will exert pressure on pathogens to develop stronger resistance, jeopardizing the health and well-being of citizens around the world. As well, continuing care is strongly impacted by AMR as failed antibiotic treatments lead to complications and hospitalization, increased morbidity and mortality, and higher economic costs. To avoid such adverse outcomes, fostering innovative projects focusing on surveillance and rapid diagnostics of AMR should be one of the top public health priorities. AMR often spreads via the emergence of resistance genes which can disseminate across geographic or human-animal boundaries. Surveillance at the genetic and genomic level and accurate resistance behaviour prediction from humans, animals, and environments are, therefore, the most effective ways to prevent and manage AMR. Surveillance will also aid in rapid and accurate diagnostics to avoid mistreatment of patients and prevent sequelae. This project aims to develop and validate genomic machine learning approaches to predict resistance accurately, improving the surveillance and rapid diagnostics efforts of AMR.

In this investigation, 300+ *Enterococcus faecium* genomes collected from hospitals, community sources, and ruminants, through collaboration with Agriculture and Agri-Food Canada (AAFC), is used. Resistant enterococci are among the leading causes of hospital-acquired infections and are widespread in the agri-food industry of Canada yet relatively poorly characterized [2, 5, 6]. In Canada, the vancomycin-resistant enterococci bloodstream infection rate has been increasing at a rate of approximately 28% each year, which poses a severe risk [5].

II. Results

Machine learning approaches successfully predicted AMR phenotype from enterococci genomic data

The Comprehensive Antibiotic Resistance Database (CARD), manually curated by experts, contains a collection of well-characterized, peer-reviewed, experimentally-validated AMR contributing genetic factors [1]. Using CARD's Resistance Gene Identifier (RGI v5.1.0) software, we identified AMR genes from the 300+ *Enterococcus faecium* genomes. The RGI identified AMR genes were used as the input features for machine-learning (ML) methods such as random forest and logistic regression. The highest performance accuracy of 98% for resistance prediction against antimicrobials such as vancomycin was achieved. However, RGI only looks for previously defined AMR genes, limiting the opportunity to discover new genetic factors that contribute to AMR.

AMR prediction with additional features that include genes not explicitly associated with resistance led to comparable high accuracy as the prediction based on RGI selected genes. The highest-ranked features deduced from feature selection methods coincided with AMR genes that RGI had identified. However, some features that were previously unrelated to AMR were also highly ranked. Further investigation in the laboratory is required to potentially discover new genetic factors associated with AMR.

ML prediction with other feature representations, such as encoding the 300+ genomes into k-mers, was followed. 31-mer features with a rule-based algorithm [3] helped obtain a model with the maximum sensitivity of 1.0 for predicting resistance against vancomycin antibiotic. The top 31-mer rules were mapped to the previously identified vancomycin-resistance genes.

Overall, the results demonstrate that ML models can accurately predict antibiotic resistant pathogens based on genetic information alone, and have the potential to aid in uncovering new genetic features associated with AMR.

III. Impact & Future Work

The global threat of AMR requires urgent investigation to develop better methods for early prediction of resistance emergence. This investigation's outcome will help rapidly and accurately predict AMR outbreak potentials based on genomic data. The prediction models' highly ranked genetic features not yet associated with resistance will be further assessed and validated in the laboratory setting. A comprehensive understanding of the genetic features related to resistance can potentially improve the feature selection process that will increase the prediction performance of ML models. The successfully validated predictive models can be applied for AMR surveillance, diagnostic and improve antimicrobial usage guidelines for humans and animals. The research outcome has the potential to be integrated into front-line practice for resistance prevention and mitigation that will ultimately benefit the health of citizens and save costs for governing authorities.

References

[1] Brian P Alcock, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, Sally Y Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E Werfalli, Jalees A Nasir, Martins Oloni, David J Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N Sharma, Emily Bordeleau, Andrew C Pawlowski, Haley L Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L Winsor, Robert G Beiko, Fiona S L Brinkman, William W L Hsiao, Gary V Domselaar, and Andrew G McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, pages 1–9, 10 2019.

[2] Grace A. Blackwell, Martin Hunt, Kerri M. Malone, Leandro Lima, Gal Horesh, and Nicholas R Iqbal Zamin Alako, Blaise T.F. Thomson. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *bioRxiv*, 2021.

[3] Alexandre Drouin, Gaël Letarte, Frédéric Raymond, Mario Marchand, Jacques Corbeil, and François Laviolette. Interpretable Genotype-to-Phenotype Classifiers with Performance Guarantees. *Scientific Reports*, 9(1):4071, 12 2019.

[4] J O'Neill. Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. Technical Report December, 2014.

[5] Public Health Agency of Canada. Canadian Nosocomial Infection Surveillance Program (CNISP): Summary Report of Healthcare Associated Infection (HAI), Antimicrobial Resistance (AMR) and Antimicrobial Use (AMU) Surveillance Data from January 1, 2013 to December 31, 2017. Technical report, Public Health Agency of Canada, 2019.

[6] Rahat Zaheer, Shaun R Cook, Ruth Barbieri, Noriko Goji, Andrew Cameron, Aaron Petkau, Rodrigo Ortega Polo, Lisa Tymensen, Courtney Stamm, Jiming Song, Sherry Hannon, Tineke Jones, Deirdre Church, calvin W Booker, Kingsley Amoako, Gary Van Domselaar, Ron R Read, and tim A McAllister. Surveillance of Enterococcus spp. Reveals Distinct Species and Antimicrobial Resistance Diversity Across a One-Health Continuum. *Scientific Reports*, 10(1):3937, 12 2020.