Title: “Machine learning approach to predict antimicrobial resistance of *Pseudomonas aeruginosa* using whole genome sequencing data”

Combating antimicrobial resistance (AMR) requires fast and reliable antimicrobial susceptibility testing (AST) methods. This is hard to achieve with traditional approaches such as broth microdilution because the turnaround time can take days. Existing research have attempted to use whole genome sequencing (WGS) data to predict AMR as proof of principle for genome-based diagnostic. In this study, we evaluated various kinds of machine learning models to predict resistant phenotype of 1245 *P.aeruginosa* isolates with respect to 10 antipseudomonal agents (Figure 1). The pathogen is known for complex AMR mechanisms and has not been well studied in AMR prediction research. Data was collected from both public datasets (1121) and our in-house database (124). In terms of feature engineering, the presence or absence of a gene from pan-genome analysis was used. After excluding genes being present in either all isolates or less than 5 isolates, we obtained a feature set of 17485 genes. The machine learning models are all binary classifiers, including logistic regression (LR), support vector machine (SVM), random forest (RF), light gradient boosted machine (LGBM), wide and deep neural network (WDNN), and forest WDNN (fWDNN). The last one is a modified version of WDNN which feeds initial input into an array of decision trees to generate a preprocessed input layer for downstream neural network layers. This implementation aims to address the “fat” data issue, indicating that feature size significantly outnumbers sample size. From the results (Figure 2), in terms of area under the receiver operating curve (AUC), highest performance was observed in aminoglycosides. Inter-model variations were clearly spotted in combinations containing beta-lactamase inhibitor while curves tend to overlap in other medications. We further generated the list of 20 most important features from the lightGBM model (Figure 3). However, it is challenging to identify similar important features even with drugs in the same class. Therefore, despite huge potentials, genome-based method using machine learning to predict AMR needs to be fine-tuned before it can be applied in clinical setting. We also suggest developing different pathogen-specific models, rather than a one-size-fit-all one.

Chart, bar chart

Description automatically generated

Figure 1. AST results to 10 antipseudomonal agents. Phenotypic labels were converted from minimum inhibitory concentration values using The European Committee on Antimicrobial Susceptibility Testing (EUCAST) version 11 guidelines.

Diagram, engineering drawing

Description automatically generated

Figure 2. Performance of machine learning models on AMR prediction with different evaluation metrics.

Chart, bar chart, treemap chart, sunburst chart

Description automatically generated

Figure 3. Top 20 most importance feature from the lightGBM model with respect to each antimicrobial agent.