ResPred: Biomarker Discovery and Antibiotic Resistance Prediction Through Machine Learning-based Genomic Analysis of Neisseria gonorrhoeae

Goyal, Rushank (School: Rajeev Gandhi Higher Secondary School)

Antibiotic resistance is a global problem projected to kill 10 million each year by 2050. The CDC lists Neisseria gonorrhoeae among the most urgent threats in this area due to its widespread resistant mutations. In this study, eight models were trained on three datasets of azithromycin, ciprofloxacin and cefixime, drugs used against N. gonorrhoeae. Each dataset had 3000+ samples and their corresponding resistance values, and each sample consisted of a unique pattern of certain consensus regions of the genome. XGBoost was the highest-performing model on the testing dataset, with a sensitivity of 95.4%, specificity of 97.8%, accuracy of 97.5%, and AUC of 0.966 on azithromycin resistance prediction, with similar results on other drugs. Support Vector Machine and Logistic Regression were not far behind, with similar accuracies and slightly lower sensitivities and specificities. The models were used to determine potential consensus regions correlated with resistance, which were then analyzed using chi-square tests to determine significance. Out of 584,362 regions, 135, 2612 and 6 regions were identified as significant for azithromycin, ciprofloxacin and cefixime respectively. The most predictive contigs of each antibiotic could accurately predict resistance with accuracies of 95%, 96% and 86% respectively; these contigs were identified as denoting mutations in certain proteins of the cell. Overall, this study led to the creation of a highly accurate machine learning model and identified resistance biomarkers for three drugs in N. gonorrhoeae. The model can be used for genotype-based resistance diagnosis and biomarkers can be further researched for point-of-care test development.

Awards Won:
American Statistical Association: Certificate of Honorable Mention
Embark China: First Award