Predicting Drug Resistance and Susceptibility in Foodborne Illnesses with Random Forests

Yanru Liao[1], Breanna Richards[1], Jina Yang[1]
[1]Brown University School of Public Health Department of Biostatistics, Providence, RI

School of Public Health
BROWN UNIVERSITY

Overview

We predicted antibiotic susceptibility and resistance in Salmonella, E.coli, and Campylobacter foodborne illnesses in order to create an effective drug recommendation system applicable for treating new cases.

Background

- In recent years, the excessive incorrect use of antibiotics for the treatment of infectious diseases in humans and animals has become a major area of concern for human health.
- When this behavior of misusage persists, bacteria are not killed but instead develop survival traits, or **resistance**, against these treatments, which they can pass on to even more bacteria. This often results in increased risks of severe infections, illnesses, and even death.
- The designation of *effective* treatment for these infections requires learning of organisms' **susceptibility** to various antibiotics, though the traditional clinical processes to do so are often time-consuming.
- High-throughput sequencing technology based on genomic data of bacteria has the potential to fast guide this decision-making process.
- Thus, we utilized binary **random forest classification** for antibiotic responses in consideration of two cases of data availability **1)** source and genetic predictors and **2)** source predictors only (to capture cases where genomic information is unavailable) to predict pathogens' susceptibility or resistance to prominent antibiotics.

Methodology

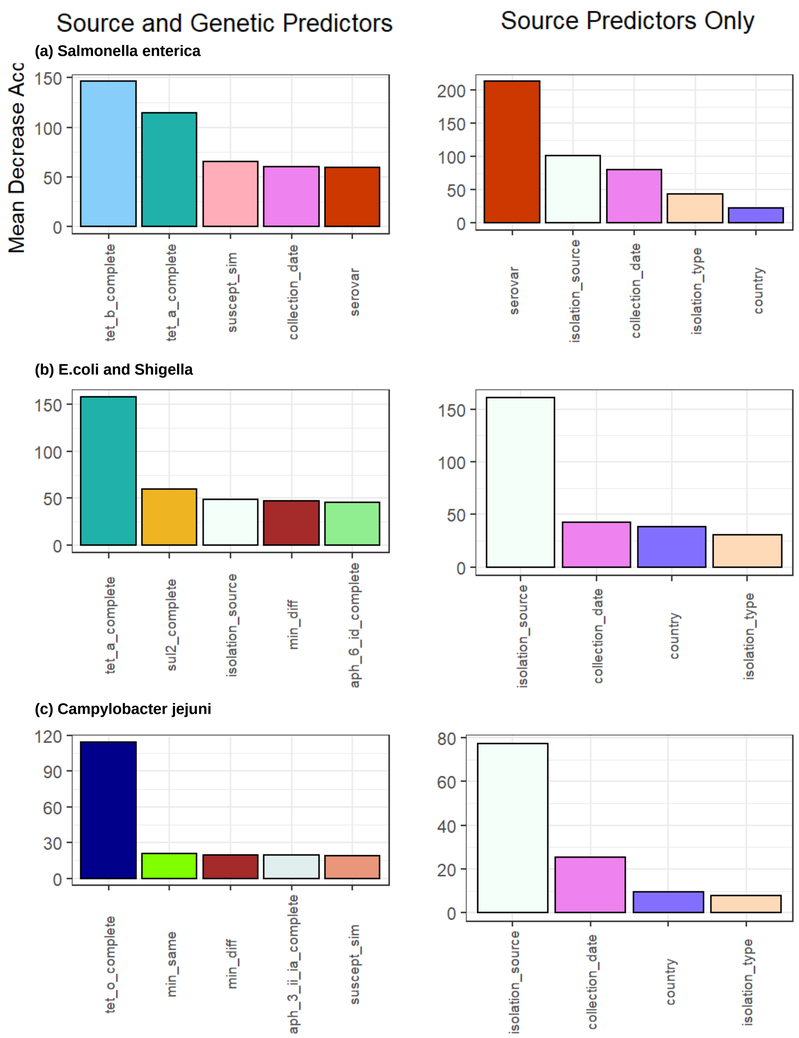
Data: Pathogen detection records collected from the Salmonella enterica, E.coli and Shigella, and Campylobacter jejuni species in the National Center for Biotechnology Information's (NCBI) Pathogen database.

- Missing values were imputed using multiple imputation (predictive mean matching and logistic regression for continuous and categorical data, respectively).
- For each antibiotic of interest, we built and optimized a random forest model using hyperparameter tuning and variable selection to predict either a susceptible or resistant response to that given drug.
- Mean decrease accuracy captured strong predictors of drug response, and the performance of these models were assessed using accuracy, sensitivity, specificity, and area under the ROC curve (AUC).
- Ultimately, we'll develop a system to receive genetic and baseline information from new cases, to then return the most probable antibiotic those cases would garner susceptible responses to.

Random Forest Results

- Salmonella enterica[a], E.coli and Shigella[b], and Campylobacter jejuni[c] isolates' antibiotic responses to tetracycline on separate models with (left column) and without (right column) the presence of information on antimicrobial resistant genes illustrates some of the strongest predictor associations observed with the outcome.
- As expected, when information on antimicrobial resistant genes are available, they tend to be most paramount to predicting antibiotic response. In their absence, however, serovar and isolation source knowledge are commonly the top drivers of this prediction process.

Top 5 Most Important Variables in Classifying Antibiotic Response to Tetracycline



Test Set Performance on Most Common Antibiotic Models (two for each species)	Salmonella enterica		E.Coli and Shigella		Campylobacter jejuni	
	Tetracycline	Streptomycin	Tetracycline	Meropenem	Tetracycline	Gentamicin
	Source and Genetic Model **(Source Only Model)**					
Accuracy	0.99 **(0.78)**	0.99 **(0.80)**	0.89 **(0.68)**	0.99 **(0.90)**	0.99 **(0.60)**	0.99 **(0.98)**
Sensitivity	0.99 **(0.77)**	0.99 **(0.88)**	0.96 **(0.61)**	0.99 **(0.89)**	0.99 **(0.53)**	1 **(0.99)**
Specificity	0.99 **(0.79)**	0.99 **(0.68)**	0.83 **(0.74)**	0.92 **(0.96)**	0.99 **(0.67)**	0.90 **(0.35)**
Area Under the ROC curve (AUC)	0.99 **(0.78)**	0.99 **(0.78)**	0.89 **(0.68)**	0.96 **(0.93)**	0.99 **(0.60)**	0.95 **(0.67)**

Conclusion

- In the absence of traditional clinical-based procedures, we can turn to methods in machine learning to help prevent foodborne illnesses and outbreaks through the appropriate prescription of antibiotics to heterogeneous cases.
- Though the use of genetic information directly is most favorable to aid in this process, accessible source information still prove to have strong predictive power in select and non-negligible instances.

References

Van Camp, P.-J., Haslam, D. B., & Porollo, A. (2020). Prediction of Antimicrobial Resistance in GramNegative Bacteria From Whole-Genome Sequencing Data. Frontiers in Microbiology, 11, 1013. https://doi. org/10.3389/fmicb.2020.01013

National Center for Biotechnology Information. (n.d.). Home - Pathogen Detection - NCBI. National Center for Biotechnology Information. Retrieved October 23, 2022, from https://www.ncbi.nlm.nih.gov/pathogens/