The application of machine learning models in predicting antimicrobial resistance with genomic data: the opportunities and challenges

Student: Bella, Liuyue Yang (2nd year MPhil)

Supervisor: Prof. Margaret Ip

Joint postgraduate seminar

13 Dec 2022

Outline

1. Introduction
   - Antimicrobial resistance (AMR)
   - Machine learning (ML) problem

2. Development and application
   - the GenTB predictor

3. Opportunities and challenges

I. Antimicrobial resistance (AMR)

- A global threat to human health
- Spreading in different settings (hospitals, communities, and farms)
- Causing infections that are difficult, and sometimes impossible to treat.

-> Improve the antibiotic prescription

(Centers for Disease Control and Prevention, 2021; Anahtar et al, 2021)

II. Antimicrobial susceptibility testing (AST)

– *in vitro* phenotypic AST methods

- Gold standard
- Labour-intense and time-consuming (slow-growing bacteria)
- Turnaround time: from days to weeks

III. Alternative: sequenced-based machine learning

Machine learning question: development of supervised machine learning (ML) models.

★ **1) Qualitative:** Is an isolate resistant or susceptible to one/several antibiotics?

-> Machine learning model for classification problem (binary).

2) Quantitative: Predict the exact value of minimum inhibitory concentration (MIC)

-> Machine learning model for regression problem (continuous)
e.g. linear regression, random forest, and lasso regression

(Anahtar et al, 2021)

Type 1: Machine learning models for classification

R S

1. Decision tree
2. Random forest (i.e. a collection of decision tree)
3. Logistic regression
4. Naïve Bayes
5. Support vector machines
6. Artificial neural networks

(Lv, 2021)

Type 1: Machine learning models for classification

R S

1. Decision tree
2. Random forest
3. Naïve Bayes
4. Logistic regression
5. Support vector machines
6. Artificial neural networks

How to choose model?

1. Ability handling missing values
2. Sensitivity to outlier data
3. Interpretability
4. Speed of learning

(Lv, 2021)

Step 1: Featurization – “feature extraction”

Aim: to capture variation in the genomic data

- Presence or absence of genes
- Single nucleotide variants (SNVs)

  e.g. single nucleotide polymorphism (e.g. SNPs)
- Insertion or deletion of bases (Indels)

Featurization methods:

- *k*-mers-based modelling (presence/absence or frequency of each *k*-mer)
- Alignment-based or functional orthologs-based

(Anahtar et al, 2021)

Step 2: a routine supervised machine learning model

Labels (Phenotype): R or S (“ground truth”)

Train data and test data:

DNA sequence + labels

Model validation

- Independent dataset (e.g. external)
- Performance indicator:
  - **Specificity** (true positive rate)
  - **Sensitivity** (true negative rate)
  - **ROC curve** (receiver operating characteristic curve) and **AUC** (area under the ROC curve)

Example - GenTB: a predictor for tuberculosis resistance

- **Target bacteria: *Mycobacterium tuberculosis***

  Treatment barrier: Diagnosing drug resistance

  - Culture-based method: grow slowly
  - Long turnaround time: *In vitro* antibiotic susceptibility tests (AST)
  - PCR-based method: limited drugs and poor detection accuracy
  - WGS becomes more affordable

(Gröschel et al, 2021)

Example - GenTB: a predictor for tuberculosis resistance

- **Target bacteria: *Mycobacterium tuberculosis***
- Open and web-based prediction tool
- Phenotypic prediction for 10 to 13 antibiotics, such as rifampicin and amikacin

(Gröschel et al, 2021)

Example - GenTB: a predictor for tuberculosis resistance

- User-friendly: for non-expert user
- Batch upload: up to 300 isolates
- Pre-processing of the dataset (e.g. quality check)
- Computing time (median): 35 mins
- Get email when prediction results are ready

https://gentb.hms.harvard.edu

Example - GenTB: a predictor for tuberculosis resistance

- **Two multivariate machine learning models** (random forest & wide and deep neural network)
- **Trained with rich variants dataset**
- **Empowered by two machine learning models**

Model 1: Random forest
Training dataset: 1,397 strains
Featurization: alignment
Features: 238 mutations (i.e. SNPs, deletion, and insertion)

Model 2: Wide and deep neural network
Input - training dataset: 3,601 strains
Featurization: alignment
Features: 222 mutations (i.e. SNPs, deletion, and insertion)

(Gröschel et al, 2021; Chen et al, 2019; Farhat et al, 2016)

Example - GenTB: a predictor for tuberculosis resistance

Performance validation:

- A ground truth dataset of **20,408 isolates**
- **Indicators**: sensitivity, specificity, and AUC.
- **High prediction accuracy** to first-line tuberculosis drugs (rifampicin and isoniazid)
- **Lower prediction accuracy** to second-line tuberculosis drugs (low sensitivity), such as amikacin.
- **Reasons for prediction bias**: undescribed resistance variants in training dataset, undetected genetic loci, and the reproducibility of AST results.

(Gröschel et al, 2021; Chen et al, 2019; Farhat et al, 2016)

Example - GenTB: a predictor for tuberculosis resistance

- **Performance validation - compared with other predictors:**
  - **Rule-based model:** Mykrobe and TB-Profiler
  - **Test data:** A ground truth dataset of 20,408 isolates

(Gröschel et al, 2021)

Example - GenTB: a predictor for tuberculosis resistance

	Sensitivity (True negative)	Specificity (True positive)
GenTB-RF	**77.6% (95% CI 76.6–78.5%)**	**96.1% (95% CI 96.0 – 96.3%)**
GenTB-WDNN	**75.4% (95% CI 74.5–76.4%)**	**96.2% (95% CI 96.0 – 96.4%)**
Mykrobe	71.9% (95% CI 70.9–72.9%)	97.6% (95% CI 97.5–97.7%)
TB-Profiler	74.4% (95% CI 73.4–75.3%)	96.9% (95% CI 96.7 to 97.0%)

- Trade-off between specificity and sensitivity
- Low sequencing depth data: lower sensitivity (the need for quality control)

(Gröschel et al, 2021)

Example - GenTB: a predictor for tuberculosis resistance

Advantages:

- Captures both common and rare mutations
- Multivariate prediction models:
  - Gene-gene interaction
- Higher sensitivity but slightly lower specificity

(Gröschel et al, 2021; Chen et al, 2019; Farhat et al, 2016)

Example - GenTB: a predictor for tuberculosis resistance

Limitations:

-Short-read sequencing lowers the sensitivity in detecting genomic variants (e.g. structural variation).

-Does not cover recently introduced or repurposed drugs.

-Does not re-test the laboratory-based drug susceptibility profiles for isolates with discordant predictions

-Diagnostic accuracy maybe vary among datasets from different countries.

(Gröschel et al, 2021; Chen et al, 2019; Farhat et al, 2016)

Challenges

1. **Data availability:**

   1) Class-imbalance

      e.g. more antibiotic-susceptible isolates than resistant isolates

   2) New antibiotic

   4) Data sharing (e.g. FDA-ARGOS database)

2. **Risk of overfitting**

3. **Understanding of AMR mechanisms**

   e.g. complex AMR mechanisms

(Anahtar et al, 2021)

Opportunities:

1. Reduced cost of WGS
2. Improvements in bioinformatic software
3. Continued growth of data and research interest
4. Providing more profound insight into the mechanisms of AMR
5. Providing organism identification, and information on virulence factors

(Anahtar et al, 2021)

Reference:

ANAHTAR, M. N., YANG, J. H. & KANJILAL, S. 2021. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *Journal of clinical microbiology,* 59**,** e01260-20.

CENTERS FOR DISEASE CONTROL AND PREVENTION. 2021. *Antimicrobial resistance* [Online]. Available: https://www.cdc.gov/drugresistance/index.html [Accessed].

CHEN, M. L., DODDI, A., ROYER, J., FRESCHI, L., SCHITO, M., EZEWUDO, M., KOHANE, I. S., BEAM, A. & FARHAT, M. 2019. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine,* 43**,** 356-369.

FARHAT, M. R., SULTANA, R., IARTCHOUK, O., BOZEMAN, S., GALAGAN, J., SISK, P., STOLTE, C., NEBENZAHL-GUIMARAES, H., JACOBSON, K. & SLOUTSKY, A. 2016. Genetic determinants of drug resistance in Mycobacterium tuberculosis and their diagnostic value. *American journal of respiratory and critical care medicine,* 194**,** 621-630.

GRÖSCHEL, M. I., OWENS, M., FRESCHI, L., VARGAS, R., MARIN, M. G., PHELAN, J., IQBAL, Z., DIXIT, A. & FARHAT, M. R. 2021. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome medicine,* 13**,** 1-14.

LI, Y. & CHEN, Z. 2018. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math,* 7**,** 212-216.

LV, J., DENG, S. & ZHANG, L. 2021. A review of artificial intelligence applications for antimicrobial resistance. *Biosafety and Health,* 3**,** 22-31.

Thank you!

Q & A