MLMG 2022: workshop on machine learning for microbial genomics

Conference ECML/PKDD 2022

Predicting antimicrobial resistance genes from phenotypic resistance profiles: a proof-of-concept study

Gabriel Carvalho[1] (ORCID 0000-0001-6864-5815), Katy Jeannot[2], Patrick Plésiat[2], Richard Bonnet[3], Laurent Dortet[4], François Vandenesch[5], Jean-Philippe Rasigade[1,5]

[1]PHE3ID, Centre International de Recherche en Infectiologie, Institut National de la Santé et de la Recherche Médicale U1111, CNRS Unité Mixte de Recherche 5308, École Nationale Supérieure de Lyon, Université Claude Bernard Lyon 1, Lyon, France; [2]CNR *Pseudomonas*, CHU de Besançon, [3]CNR Entérobactéries, CHU de Clermont-Ferrand, [4]CNR Carbapénémases, Hôpital Bicêtre, Assistance Publique-Hôpitaux de Paris, Centre National de Référence de la Résistance aux Antibiotiques; [5]Institut des Agents Infectieux, Hospices Civils de Lyon, France

Contact: gabriel.carvalho@chu-lyon.fr or jean-philippe.rasigade@chu-lyon.fr

Abstract

Epidemics of antimicrobial resistance are increasingly linked with horizontally-transferred antibiotic resistance genes (ARGs), prompting the need for monitoring ARGs rather than specific bacterial strains for epidemic surveillance. Whole genome sequencing (WGS) has become popular but its current cost prevents its systematic use for ARG surveillance. We explore the feasibility of predicting ARGs from readily-available antimicrobial susceptibility profiles of bacteria generated by diagnostic laboratories. ARG prediction models based on random forests, support vector machines and generalized linear models were trained on an extensive collection of clinically relevant bacteria with diverse antibiotic susceptibility profiles. Model performance evaluation using leave-one-out cross validation suggests that support vector machine outperforms other methods for this task. The best-performing prediction models were then applied to predict ARG presence in all bacteria diagnosed at a large hospital group over 5years. The potential benefits and limits of this novel approach for antimicrobial resistance monitoring are discussed.

1 Introduction

Antimicrobial resistance (AMR) is one of the major threats faced by modern medicine (Mancuso et al. 2021). Beyond *de novo* AMR acquisition by mutation, bacterial pathogens frequently exchange antibiotic resistance genes (ARGs) through horizontal gene transfer (Baquero et al. 2019). This prompts the need for monitoring ARGs rather than specific bacterial strains for epidemic surveillance. Monitoring ARGs, in turn, requires efficient and low-cost methods for ARG detection in bacteria.

Whole-genome sequencing (WGS) has become increasingly popular. Combined with available ARG databases, WGS allows the annotation of ARGs from genome sequence data (Alcock et al. 2019; Florensa et al. 2022; Hendriksen et al. 2019). Understandably, most medical applications of WGS-based ARG detection focus on the prediction of the antimicrobial resistance profile from genome sequences (Kim et al. 2020; Ren et al. 2022; Van

Camp, Haslam, et Porollo 2020), rather than on the prediction of ARG presence from the resistance profile. Yet, this latter approach can have practical benefits for ARG monitoring because the current costs of WGS prevents its widespread adoption for surveillance, even in high-income settings, while antimicrobial susceptibility testing (AST) is now commonplace in virtually all hospitals and settings including middle-income countries. Hence, predicting ARG presence from resistance profiles would enable large-scale ARG monitoring at a reduced cost, considering that the predictor data are generated by routine care.

Here we investigate the feasibility of this approach by predicting the presence of relevant ARGs in pathogenic bacteria based on their resistance profile. Thus, we use the reverse approach of previous studies and we aim to predict the genotype from the phenotype using machine learning. Generalized linear model, random forest and support vector machine models were trained and compared on the critical priority pathogens defined by the World Health Organization, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and Enterobacterales. The best performing models were then applied to predict ARG presence in all bacteria of the aforementioned species, diagnosed at a large hospital group from 2017 to 2021.

2 Methods

2.1 Training data

Bacterial strains used for model training were selected by the French *Centres Nationaux de Référence de la Résistance aux Antibiotiques* (CNR) from their reference and clinical strain collections. The selection procedure aimed to maximize the diversity of ARGs found in healthcare settings and to include susceptible strains without acquired drug resistance. Thus, the strain collection does not represent the actual distribution of resistance genes in hospitals but it is biased to maximize ARG diversity. The selected bacterial species were those with maximal clinical impact and horizontal transfer of ARGs, namely, *Acinetobacter baumannii* (n=181), *Pseudomonas aeruginosa* (n=304) and *Enterobacterales* (n=800) including *Klebsiella pneumoniae* (n=145) and *Escherichia coli* (n=240). All bacterial strains were whole-genome sequenced using Illumina technology and ARGs were annotated using an in-house bioinformatics pipeline. Antimicrobial susceptibility testing was performed using Vitek2 (bioMérieux) devices.

Antimicrobial susceptibility testing estimates the minimal inhibitory concentration (MIC) of each drug over a limited concentration range, leading to left- and right- censoring (e.g. <=1 mg/L, >4 mg/L). To generate point estimates, censored values at the lower (upper) boundary were divided (multiplied) by 2 then all values were log2-transformed. The log2-MIC were then used as predictors of ARG presence or absence, coded as a boolean. ARGs present or absent in less/more than 3 strains were excluded.

2.2 Model training and performance evaluation

Separate models were built for *P. aeruginosa*, *A. baumannii* and enterobacteria. Enterobacteria species were pooled in a single model because virtually all species can exchange ARGs through horizontal transfer. In this pooled model, the bacterial species was included as an additional categorical predictor with one-hot encoding.

Models tested were generalized linear model (GLM), random forest (RF) and support vector machine (SVM) in the R programming environment with additional packages *randomForest* and *e1071* (Liam

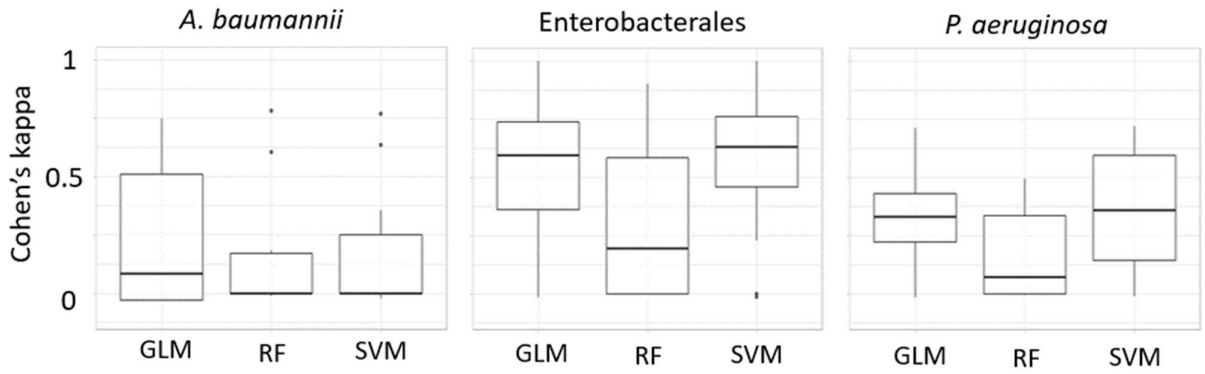

Figure 1. Cohen's kappa distribution with the different models: generalized linear model (GLM), random forest (RF) and support vector machine (SVM).

et Wiener 2002; Meyer et al. 2021). Performance evaluation metrics were obtained using leave-one-out cross validation where model training uses all data samples except the one for which a prediction is made, iterating over each sample. The confusion matrix for each drug was visualized as bar plots of true positives and negatives, and false positives and negatives. Prediction performance was summarized using the Cohen's kappa indicator, with a value of 1 for perfect prediction and a value of 0 for random (useless) prediction.

2.3 Model application

To illustrate our approach with field data, the best-performing models were used to predict ARG presence in all *P. aeruginosa*, *A. baumannii* and enterobacteria diagnosed from 2017 to 2021 at a large, 5500-bed hospital group of the Hospices Civils de Lyon (HCL), France. Anonymized data were obtained for the microbiology laboratory information system. Antimicrobials with >20% missing MIC values in the dataset were excluded, then samples with >70% missing antimicrobials were excluded. The remaining missing values were estimated using random forest-based multiple imputation with R package *missForest*.

3 Results

3.1 GLM and SVM perform better than random forest

Based on the distribution of Cohen's kappa across ARG prediction models (Figure 1), GLM performed best in *A. baumannii* (11 ARGs, although with poor performance) and SVM performed best in enterobacteria (19 ARGs) and *P. aeruginosa* (11 ARGs). Random forests performed worse in all species groups.

3.2 Prediction performance varies widely across ARGs

An ARG prediction model is expected to perform best if the ARG presence is balanced across the dataset, the ARG is not too strongly correlated with other ARGs and the ARG resistance spectrum does not overlap significantly with those of other ARGs. As these conditions vary widely across ARGs (with strong correlation and overlapping resistance spectra being commonplace), the performances of prediction models varied accordingly and only a fraction of ARGs could be predicted with relevant accuracy (Figures 2BDF). Strikingly, however, those ARGs with better prediction performance were among those with maximal clinical impact including resistance to critical drugs such as carbapenems (a WHO priority), aminoglycosides or cephalosporins. In *A. baumannii*

(Figure 2AB), promising predictive power was obtained for the resistance genes *arm*A, NDM and OXA-23. The aminoglycoside resistance methyltransferase *armA* confers resistance to most aminoglycosides using in sepsis treatment. The New Delhi metallo-β-lactamase NDM and the OXA-23 β-lactamase provide resistance to virtually all currently available beta-lactams. In enterobacteria (Figure 2CD), the strongest Cohen's kappa was obtained with ACT/MIR and EC/ESC beta-lactamases that confer resistance to cephalosporins. EC/ESC is a resident gene in *E. coli*, hence its detection was provided by the species rather than the resistance profile. In addition, EC/ESC resistance depends on the expression level of the gene rather than on its presence. In *P. aeruginosa* (Figure 2EF), beta-lactamases GES, OXA-10, PER and VIM had the highest prediction accuracies but the absolute performance remained moderate, possibly due to the importance of adaptive AMR processes (such as membrane impermeability) not captured by genome sequencing in this species.

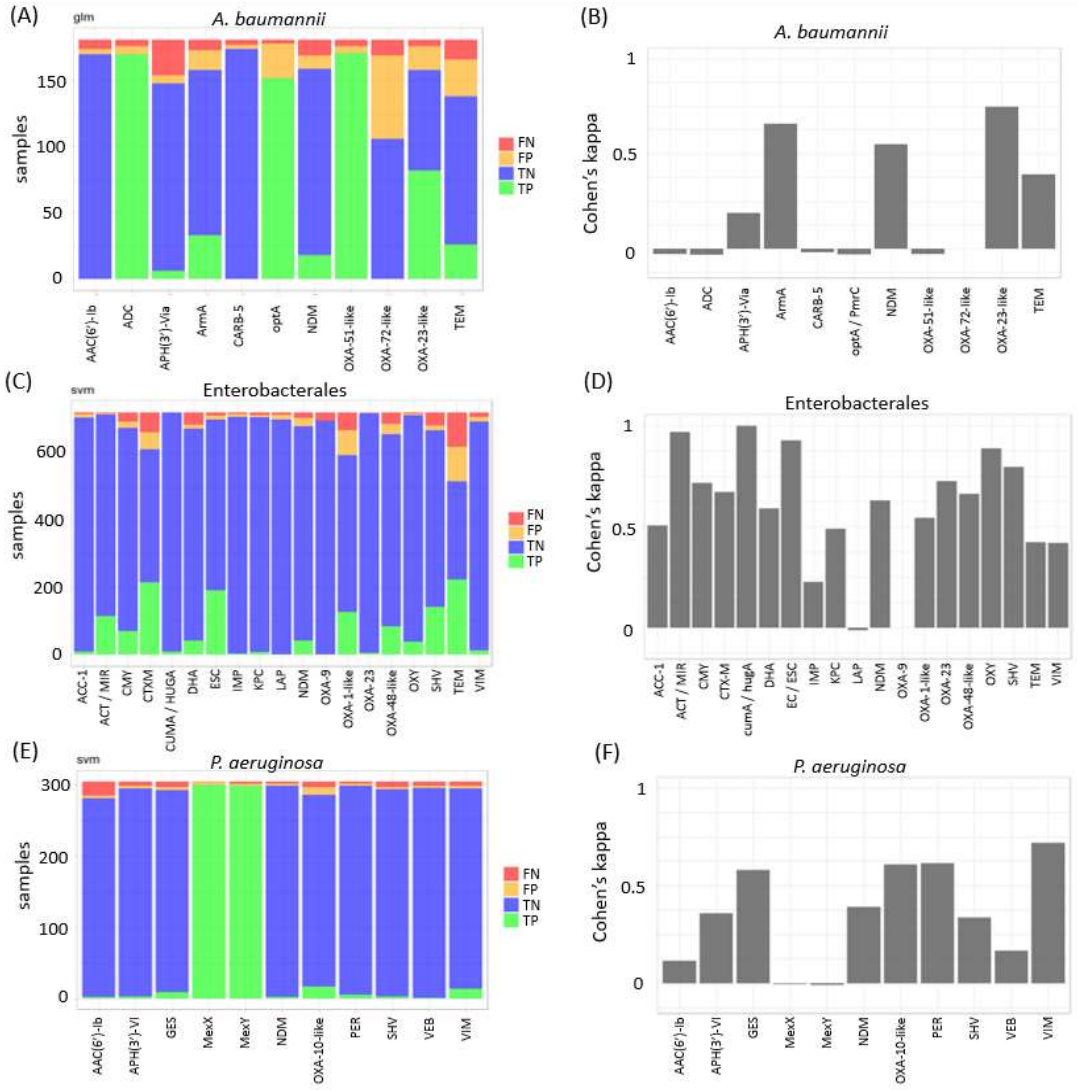

Figure 2 Models' performance. Results from the best model is displayed for each species group. (A,C,E) True positive TP, true negative TN, false positive and false negative from leave one out cross-validation. (B,D,F) Cohen's kappa for the prediction of the different ARGs.

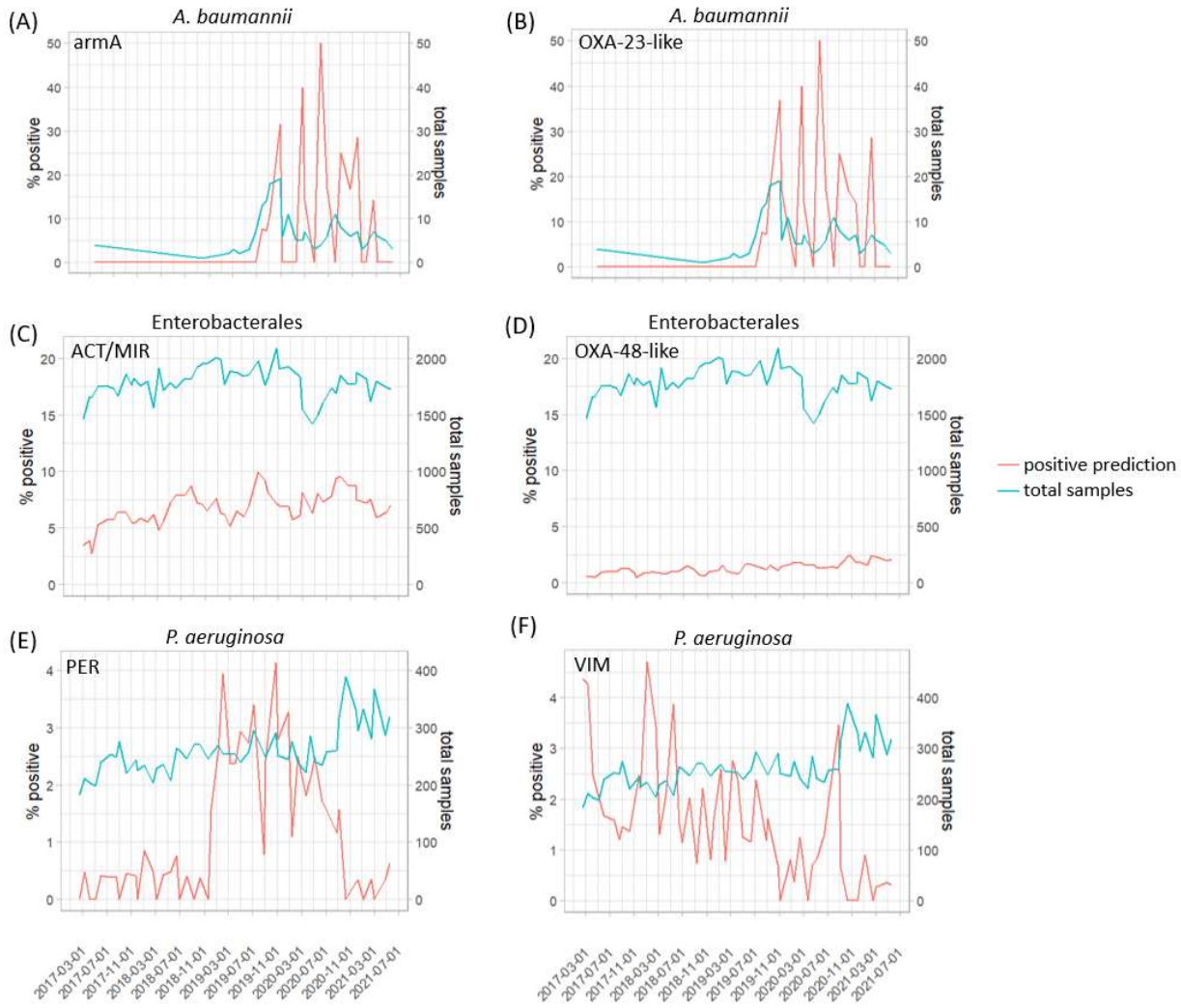

Figure 3 ARG prediction on the HCL data from 2017 to 2021 for the most predictable ARG. GLMs were used for *A. baumannii* and SVM models for enterobacteria and *P. aeruginosa*.

3.3 Retrospective ARG monitoring using ARG prediction from antibiograms

To illustrate an application of ARG prediction, we applied the best-performing models (with Cohen's kappa >0.5) on retrospective bacterial susceptibility profiles obtained from clinical isolates from the Hospices Civils de Lyon over the period 2017 – 2021 (Figure 3). *A.baumannii* was rare in the HCL ecology and *arm*A and OXA-23-like genes seemed to be present at very variable levels. Enterobacteria, in contrast, had more than a thousand samples per month, a sample size which is expected to yield more accurate tendencies of ARG fluctuations. For example, ACT/MIR was more frequent than OXA-48 in the hospital data and the predicted proportions of both ARGs steadily increased over the study period. In *P. aeruginosa*, PER seemed to emerge abruptly in November 2018 while VIM predictably decreased over the study period. However, the PER apparent increase seemed to correlate with a temporary change in the methodology of the Vitek2 (data not shown).

4 Discussion

We propose a novel approach to monitor the prevalence of ARGs in healthcare settings using standard-of-care data on

antimicrobial resistance profiles. The models show encouraging performances for some ARGs and this method could be used as an additional tool for ARG monitoring. Due to the low-resolution of the MIC vectors used as predictors, especially in comparison with the large number of potential ARGs, the models struggle to discriminate ARGs with similar resistance profiles. We propose that performances might be improved by using the reference method, microdilution MICs in the training data (data not shown), rather than the estimated MICs obtained using automated, routine-care devices such as Vitek2. However, microdilution MICs are not expected to be readily available in most settings due to cost and workforce constraints. The inclusion of isolates from the local ecology in the training dataset might also improve performance, although at an added cost. Because the prevalence of each ARG in the training dataset is not expected to match the prevalence in the local ecology, the detection thresholds of the models might be strongly biased. Optimizing detection thresholds based on the local prevalence of resistance, possibly by applying weighs to training samples to reflect their prevalence in the local setting, might help improve performance. Finally, ARGs could be predicted as groups of genes (such as those harbored together on a plasmid) rather than as individual genes. In any case, ARG prediction results from resistance profiles should be compared to random whole genome sequencing from the local ecology to confirm observations and to provide accurate ARG monitoring.

5 Acknowledgments

This work was supported by the French National Research Agency under grant ANR-20-CE35-0012 to JPR and grant ANR-18-RHUS-0013 to FV.

6 References

Alcock, Brian P, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, et al. 2019. « CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database ». *Nucleic Acids Research*, octobre, gkz935. https://doi.org/10.1093/nar/gkz935.

Baquero, Fernando, Teresa M. Coque, José-Luis Martínez, Sonia Aracil-Gisbert, et Val F. Lanza. 2019. « Gene Transmission in the One Health Microbiosphere and the Channels of Antimicrobial Resistance ». *Frontiers in Microbiology* 10 (décembre): 2892. https://doi.org/10.3389/fmicb.2019.02892.

Florensa, Alfred Ferrer, Rolf Sommer Kaas, Philip Thomas Lanken Conradsen Clausen, Derya Aytan-Aktug, et Frank M. Aarestrup. 2022. « ResFinder – an Open Online Resource for Identification of Antimicrobial Resistance Genes in next-Generation Sequencing Data and Prediction of Phenotypes from Genotypes ». *Microbial Genomics* 8 (1). https://doi.org/10.1099/mgen.0.000748.

Hendriksen, Rene S., Valeria Bortolaia, Heather Tate, Gregory H. Tyson, Frank M. Aarestrup, et Patrick F. McDermott. 2019. « Using Genomics to Track Global Antimicrobial Resistance ». *Frontiers in Public Health* 7 (septembre): 242. https://doi.org/10.3389/fpubh.2019.00242.

Kim, Jiwoong, David E. Greenberg, Reed Pifer, Shuang Jiang, Guanghua Xiao, Samuel A. Shelburne, Andrew Koh, Yang Xie, et Xiaowei Zhan. 2020. « VAMPr: VAriant Mapping and Prediction of Antibiotic Resistance via Explainable Features and Machine Learning. » *PLoS Computational Biology* 16 (1): e1007511. https://doi.org/10.1371/journal.pcbi.1007511.

Liam, Andy, et Matthew Wiener. 2002. « Classification and Regression by randomForest ». *R news* 2 (3): 18-22.

Mancuso, Giuseppe, Angelina Midiri, Elisabetta Gerace, et Carmelo Biondo. 2021. « Bacterial Antibiotic Resistance: The Most Critical Pathogens ». *Pathogens* 10 (10): 1310. https://doi.org/10.3390/pathogens10101310.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, et Friedrich Leisch. 2021. « e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien », 2021. https://CRAN.R-project.org/package=e1071.

Ren, Yunxiao, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, et Dominik Heider. 2022. « Prediction of Antimicrobial Resistance Based on Whole-Genome Sequencing and Machine Learning ». Édité par Inanc Birol. *Bioinformatics* 38 (2): 325-34. https://doi.org/10.1093/bioinformatics/btab681.

Van Camp, Pieter-Jan, David B. Haslam, et Aleksey Porollo. 2020. « Bioinformatics Approaches to the Understanding of Molecular Mechanisms in Antimicrobial Resistance ». *International Journal of Molecular Sciences* 21 (4): 1363. https://doi.org/10.3390/ijms21041363.