Briefings in Bioinformatics, 22(6), 2021, 1–8

https://doi.org/10.1093/bib/bbab179
Problem Solving Protocol

A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains

Janak Sunuwar and Rajeev K. Azad

Corresponding author: Rajeev K. Azad, Department of Biological Sciences and BioDiscovery Institute; Department of Mathematics, University of North Texas, Denton, Texas 76203, USA. Tel.: +1-940-369-5078; Fax: +1-940-369-8656; E-mail: Rajeev.Azad@unt.edu

Abstract

Recently, the frequency of observing bacterial strains without known genetic components underlying phenotypic resistance to antibiotics has increased. There are several strains of bacteria lacking known resistance genes; however, they demonstrate resistance phenotype to drugs of that family. Although such strains are fewer compared to the overall population, they pose grave emerging threats to an already heavily challenged area of antimicrobial resistance (AMR), where death tolls have reached ∼700 000 per year and a grim projection of ∼10 million deaths per year by 2050 looms. Considering the fact that development of novel antibiotics is not keeping pace with the emergence and dissemination of resistance, there is a pressing need to decipher yet unknown genetic mechanisms of resistance, which will enable developing strategies for the best use of available interventions and show the way for the development of new drugs. In this study, we present a machine learning framework to predict novel AMR factors that are potentially responsible for resistance to specific antimicrobial drugs. The machine learning framework utilizes whole-genome sequencing AMR genetic data and antimicrobial susceptibility testing phenotypic data to predict resistance phenotypes and rank AMR genes by their importance in discriminating the resistance from the susceptible phenotypes. In summary, we present here a bioinformatics framework for training machine learning models, evaluating their performances, selecting the best performing model(s) and finally predicting the most important AMR loci for the resistance involved.

Key words: antimicrobial resistance (AMR); AMR gene prediction; machine learning

Introduction

Penicillin discovery in 1928 was heralded as a pivotal moment in the fight against infectious diseases; however, within the first 5 years of its use, 50% of *Staphylococcus aureus* attained resistance [1]. The golden era of antibiotic discovery reigned from 1940 through late 1960s [2], and thereafter until the 2000s, the discovery of new antibiotics kept declining [3]. During this innovation gap, the use and misuse of antibiotics continued, resulting in the evolution of a multitude of pathogens into superbugs, e.g. the emergence of new resistant strains of *Escherichia coli* and *Klebsiella pneumoniae* during this period, which were resistant to all antibiotics, including carbapenems, that were among the most effective drugs to treat various infections [4]. The festering antimicrobial resistance (AMR) problem has taken for a worse with many instances of resistance to all antibiotics emerging on a frequent basis in recent years. In November 2015, the last resort drug for multidrug-resistant infection treatment, colistin,

Janak Sunuwar is a PhD candidate at the University of North Texas. His research interests include developing and applying machine learning-based protocols and pipelines to decipher the virulence and antibiotic resistance of bacterial pathogens.
Rajeev K. Azad is an associate professor at the University of North Texas. His research interests are in the area of bioinformatics and computational biology, particularly, the development and application of mathematical and computational methods to understand how organisms, specifically microbes, innovate to adapt to changes in the environment, study of large omics datasets to determine how organisms respond to stress at the molecular and physiological levels and development of novel approaches to decipher the structural and functional features in genomes and elucidate their relationships in the context of evolution.
Submitted: 9 December 2020; **Received (in revised form):** 28 March 2021

was rendered ineffective by certain pathogens with the *mcr-1* gene, as first reported in China, and 6 months later, colistin-resistant pathogens were detected in the urine sample of a patient from Pennsylvania [5]. In 1899, C. W. Jones' foot had to be amputated due to an infection after a cut by a blade of grass, and many decades later, in July 2016, the similar procedure had to be undertaken in a hospital in Cape Town due to a multiple drug-resistant bacterial infection of a fractured ankle [6]. Similarly, a patient in Nevada died of an incurable *K. pneumoniae* infection, which was resistant to all 26 antibiotics available in the USA, including the last-resort antibiotics. This bacterium has been rightly nicknamed the 'nightmare superbug' [7]. The global public health consequences of the antibiotic misuse could be of epic proportion, leading to a huge health-care overburden, with a cost exceeding 35 billion a year, 2.8 million antibiotic-resistance infections and mortality of 35 000 plus per year in the USA alone with an enormous socioeconomic consequence [8, 9]. At this rate, the projected deaths due to AMR by 2050 would be ~10 million per year, perhaps an emerging heath crisis that calls for urgent action [10, 11].

The culture-based antimicrobial susceptibility testing (AST) has been a gold standard to determine the treatment regimen for bacterial infections. The challenging aspect of this routine is that it takes a longer turnaround time, and moreover, the result variability and lack of a standardized protocol results in non-convergent outcomes of AST by different laboratories [12]. Recently, the whole genome-based machine learning approach has shown promising results. These methods have immense potential to be transformative in bacterial diagnostics and therapeutics and may be deployed as a mainline tool in clinical practice in the near future [13]. Artificial intelligence or machine learning utilizes the prevailing knowledgebase, e.g. of genetic markers, to learn complex relations or interactions in an unsupervised or supervised setting and uses the learnt patterns to find similar or related patterns in yet unseen data [14]. The effectiveness of this approach has been demonstrated by recent studies. A proof-of-concept study was done on the basis of the presence of implicated genes and molecular typing markers to distinguish between vancomycin-susceptible and vancomycin-intermediate *S. aureus* [15]. Similarly, another study employed a machine learning model to predict the phenotypes of nontyphoidal *Salmonella* strains using the genotypic AMR and phenotypic AST data [16]. Despite these advances, the dependence on known genes resistant to different classes of antibiotics limits the scope of the whole genome-based approach, particularly in predicting the phenotypic behavior in the absence of known genetic factors. In this study, we focused on important pathogen models, *K. pneumoniae*, *Salmonella enterica*, *Campylobacter jejuni*, *E. coli* and *Shigella* and *Pseudomonas aeruginosa*, and present an approach to predict the genes conferring resistance to antibiotics in strains that lack genes known to be involved in resistance to these drugs.

Our study was motivated by the investigation of *K. pneumoniae* AR_0107 strain that does not have carbapenemase and efflux pump encoding genes yet resistant to the carbapenemase family of beta-lactam antibiotics. The mechanism of resistance has not yet been elucidated. In contrast, the resistant strain AR_362 harbors, as expected, the carbapenemase and efflux pump encoding genes. Carbapenem-susceptible strain AR_376 lacks carbapenemase but harbors efflux pump encoding genes. Multidrug efflux pump systems are often thought to be involved in conferring resistance to multiple drugs, including carbapenems [17–19]; however, the genome of strain AR_107 does not contain efflux pump genes as revealed by the NCBI pathogen viewer for this genome. Furthermore, the carbapenem-susceptible strain AR_376 does have efflux pump encoding genes yet is still susceptible to carbapenem. Thus, the efflux pumps may not always confer resistance to carbapenems. This illustrates the challenges abound in characterizing strains as antibiotic-resistant or susceptible based on the prevailing knowledge of resistance genes. Furthermore, even if the strains could be characterized based on the phenotypic assays, characterization of the mechanisms of resistance remains an outstanding problem and in the absence of this, adequate therapeutic interventions cannot be attained. Clearly, there is a pressing need to investigate and understand how strains such as AR_107 attain resistance, which may provide new insights into how bacteria successfully evade antibiotics, specifically shining a new light on genetic factors that underlie the evolving traits of antibiotic resistance.

Although this study was motivated by our preliminary investigation of *K. pneumoniae* strain AR_0107, we later found several other strains of *K. pneumoniae* that lack known carbapenem-resistance genes but are phenotypically resistant to carbapenems. Basic Local Alignment Search Tool (BLAST)-based methods and databases are prone to misclassify these strains as carbapenem-sensitive as their genomes lack the already implicated carbapenem-resistance genes. We further broadened our study to other bacterial species and found similar dissonance between the presence/absence of resistance genes and expected resistance/susceptible phenotypes. Clearly, new methods and tools that can interrogate genomes and predict resistance or susceptibility independent of prior knowledge about genes involved in resistance to certain classes of drugs are sorely needed. Here, we present such a framework based on machine learning that was applied to several bacterial strains to predict their resistance or susceptibility to antibiotics and infer yet unknown genetic factors responsible for resistance to different antibiotics.

Materials and methods

Bacterial strains, AMR genotypes and AST phenotypes

All data were retrieved from the Isolates Browser at the NCBI Pathogen Detection website: https://www.ncbi.nlm.nih.gov/pathogens/. Bacterial genus and species were selected in the 'organism group', with filters checked for 'has AMR genotypes' and 'has AST phenotypes'. Based on the AMR genotypes and AST phenotypes, binary matrices of genotypes (0 for absence and 1 for presence of an AMR gene) and relevant antibiotics' phenotypes (0 for susceptibility and 1 for resistance to an antibiotic) were created (matrices are provided as supplementary excel files at https://github.com/Janaksunuwar/AMR_prediction). It is to be noted that not all bacterial strains have complete genotypic and phenotypic data available. The numbers of bacterial strains susceptible and resistant to different antibiotics and the number of AMR genes for each species group considered here are provided in Supplementary Table S1 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

Machine learning algorithms: training and validation to establish optimal model

Machine learning in Python with Scikit-learn (https://scikit-learn.org/stable/) was used to evaluate the performance of 12 machine learning algorithms, namely, Logistic Regression (logR), Gaussian Naive Bayes (gNB), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Multinominal Naive

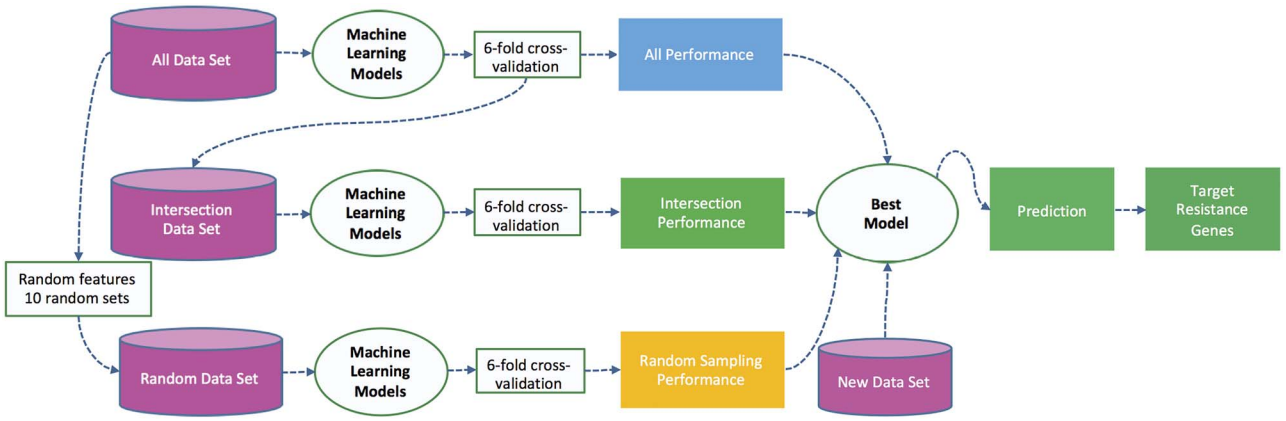

Figure 1. Schematic representation of a general workflow for establishing an optimal model to predict drug resistance and identify features of high importance (top-ranked AMR genes) that aid discrimination between the susceptible and resistance phenotypes. The performance was assessed as follows: (i) All Performance: performance metrics estimated based on the entire AMR dataset and AMR genes of high importance cataloged. (ii) Intersection Performance: performance metrics estimated based on genes that consistently appeared among top-ranked AMR genes in each round of the 6-fold validation. 'Consistent' genes were selected from top 30 high importance genes from each fold test, which were among the most important discriminating features learnt by a machine learning algorithm for predicting the resistance phenotype.(iii) Random Sampling Performance: performance metrics estimated based on randomly sampled AMR genes ('random features'). The overall random performance was obtained as the average over such 10 random sets sampled from the entire dataset (see Materials and Methods section for details).

Bayes (mNB), AdaBoost Classifier (ABC), Gradient Boosting Classifier (GBC), ExtraTrees Classifier (ETC) and Bagging Classifier (BC). We used the bacterial strains' AMR genotype and AST phenotype data for training and testing the machine learning algorithms. The performance metrics include precision, recall, F1-score, the area under the receiver operating characteristic (AU ROC), the area under the precision recall curve (AUPR) and the classification accuracy for nested 10-fold cross-validation, which were computed in an n-fold ($n = 5$, 6) cross-validation setting. Additionally, classification accuracy for leave-one-out (LOO) cross-validation was also obtained. Based on the values of the performance metrics within the 6-fold cross-validation setting, the optimal model was selected. This best model was then trained on the whole dataset, and an entirely new test dataset (AMR genotypes and AST phenotypes) of strains lacking known resistance loci yet resistant to specific antibiotics was used to predict the phenotypes.

Target gene identification and performance reassessment

AMR genes ranked high for their importance in predicting the resistance phenotype by the optimal model were enlisted for each of the n rounds of cross-validation performed. We performed 6-fold cross-validation and identified the AMR genes that were consistently ranked high in each of the six rounds of cross-validation. This dataset of AMR genes, termed intersection dataset, was then used to train the machine algorithms and their performance was reassessed. In parallel, 10 'random' datasets were generated by randomly sampling as many genes from the complete set of AMR genes for each random dataset. Performance was assessed for each of these datasets used in training and was then averaged to obtain the 'random' performance against which the performance of the model with consistently high-ranked AMR genes used in training was reassessed. This workflow is illustrated in Figure 1.

RESULTS

Assessment based on 6-fold cross-validation assessment of different machine learning algorithms on the entire dataset of bacterial strains with known genotypes and phenotypes showed that there is no single optimal machine learning model for predicting resistance phenotype across all bacterial species considered in this study. In Supplementary Table S2 (see Supplementary Data available online at http://bib.oxfordjournals.org/), we provide the information on top three classifiers that yielded F1-score values that were higher than other classifiers for each species–drug combination. This should guide the users in selecting the optimal model for the species–drug combination of their interest. Users may also decide whether to use all AMR gene set or intersection set based on the F1-score values of these models (Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/), although these differences are small despite the intersection set using much fewer genes compared to the all set. We also observed that ETC could predict the strains that lack the resistance gene(s) but have a resistance phenotype to the respective drugs as resistant better than the other models (~86% correctly predicted, Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/); note that the genotypic and phenotypic data for all these strains were not included in training the model but were included in held-out sets for testing. However, the model performance varied with species–drug combination (Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). For users interested particularly in selecting the model(s) with best performance on resistant strains that lack genes known to be involved in resistance to the specific antibiotics, we recommend using optimal models established by our pipeline for species–drug combinations, as provided in Supplementary Table S4 (see Supplementary Data available online at http://bib.oxfordjournals.org/). Note that if there are multiple best models for a species–drug combination, then the model that yields that best accuracy (F1-score) in the cross-validation test should be selected (Supplementary Table S4, see Supplementary Data available online at http://bib.oxfordjournals.org/). As ETC could predict with high accuracy the resistant phenotypes of bacterial strains that lack AMR genes implicated in those phenotypes, we examined the AMR genes that were deemed discriminative, i.e. contributing to the discrimination between susceptible and resistance strains, by using the tree-based classifiers. We considered the top 30 AMR genes, which were ranked based on their importance in discrimination, from each round of the cross-validation. We further compiled a

set of the top-ranked AMR genes that were consistently deemed features of importance in each round of cross-validation. We investigated the potential functions of these genes, specifically in the context of their potential roles in resistance to antibiotics in bacterial strains, as discussed below.

Training and test performance of machine learning algorithms

We assessed the performance of all 12 machine learning algorithms on *K. pneumoniae*, *E. coli* and *Shigella*, *P. aeruginosa*, *C. jejuni* and *S. enterica* genotypic and phenotypic data on several antibiotic susceptibility/resistance. For each species group, the dataset was partitioned into six equal parts and each part was used as the test set in turn with the remaining serving as the training set (6-fold cross-validation). After training, we applied the algorithms to both training and test data. Comparative performance of different algorithms on test data, in terms of the overall accuracy metric F1-score, is shown in Figure 2**A**–**C** for *K. pneumoniae*, *E. coli* and *Shigella* and *P. aeruginosa* on Doripenem, respectively; in Figure 3**A**–**C** for *K. pneumoniae*, *E. coli and Shigella*, on Ertapenem, Imipenem and Meropenem, respectively; in Figure 4**A** for *C. jejuni* on Clindamycin and in Figure 4**B** and **C** for *S. enterica* on Streptomycin and Kanamycin, respectively. Full set of performance data, including precision, recall and F1-score on training and test datasets and classification accuracy for nested 10-fold cross validation, AU ROC and AUPR for 6-fold cross-validation are provided in the supplementary files; additionally, the classification accuracy for LOO cross-validation is also provided (Supplementary Figure S1 for *K. pneumoniae*, Supplementary Figure S2 for *E. coli and Shigella*, Supplementary Figure S3 for *P. aeruginosa*, Supplementary Figure S4 for *S. enterica* and Supplementary Figure S5 for *C. jejuni*, see Supplementary Data available online at http://bib.oxfordjournals.org/). All programs using all datasets had comparable performance, in terms of F1-score and average accuracy, with these using the intersection sets for model training. Further, the intersection sets yielded better performance in many instances. Taken together, this indicates that the AMR genes deemed important by the tree-based classifiers are key players in conferring resistance/susceptible phenotypes to the bacterial strains. Consistent trends were observed with 5-fold cross-validation (Supplementary Figures S6–S13, see Supplementary Data available online at http://bib.oxfordjournals.org/); in most instances, the performance with the 5-fold cross-validation declined compared to that with the 6-fold cross-validation (Supplementary Tables S2 and S5, see Supplementary Data available online at http://bib.oxfordjournals.org/), which may be attributed to the less training data for the former compared to the latter (four-fifth versus five-sixth of the whole dataset); however, note that in both cases, the F1-score values for the top three performing methods were in the 0.7–1 range (Supplementary Tables S2 and S5, see Supplementary Data available online at http://bib.oxfordjournals.org/).

The ETC algorithm correctly predicted the resistance phenotype for most strains lacking genes known to be involved in resistance to specific antibiotics (73 of 85 strains (~86%); Supplementary Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). Overall, F1-scores from the intersection datasets were either comparable or better than those from the all datasets and were higher than those from the random sets (Figures 2 and 3).

The list of *K. pneumoniae* genes in the intersection sets for different carbapenems are provided in Table 1, and such gene lists for *E. coli* and *Shigella*, *P. aeruginosa*, *S. enterica* and *C. jejuni* are provided in Supplementary Tables S6–S9, respectively (see Supplementary Data available online at http://bib.oxfordjournals.org/). We hypothesize that since the genes in the intersection sets are being used by the algorithms to predict the resistance phenotype of bacterial strains with high accuracy, they must have some functional roles in the resistance. We discuss below potential mechanisms of resistance of some of these genes not yet known to be involved in resistance to different carbapenems for *K. pneumoniae*.

Potential resistance functions of novel genes

Enzymes that modify or neutralize antimicrobial drugs display a wide range of activities, such as hydrolysis (hydrolases include beta-lactamases, such as penicillinases, cephalosporinases, carbapenemases, esterases and epoxide hydrolases), transfer (transferases, such as acetyltransferases, phosphotransferases, nucleotidyltransferases, glycosyltransferases, ADP-ribosyltransferases and S-transferases) and lysis (redox enzymes, including monooxygenases and lyases) [20, 21]. Similarly, the target modification of DNA gyrase and evolution of efflux pumps lead to AMR [22]. Of novel genes identified in our analysis (Table 1), *ampC*, *blaCTX-M-15*, *blaSHV-11* and *blaTEM-1* encode proteins that are known to be involved in hydrolase activity. Likewise, *aac(6′)-Ib*, *aph(3′)-Ia*, *aph(4)-Ia* and *sul1* encode enzymes with transfer functions, akin to transferase activities [23]. Considering the fact that these genes encode proteins with similar functions as carbapenemase and other antibiotic-modifying enzymes, we posit that they might be the genetic factors responsible for carbapenem resistance in *K. pneumoniae* strains that lack known carbapenem-resistance genes.

Discussion

In this study, we demonstrated the usefulness of a machine learning framework in identifying the genetic components in bacterial strains responsible for resistance to antibiotics, particularly where the strains lack the known resistance factors. Furthermore, we developed an approach to determine the optimal models for resistance phenotype prediction and the underlying genetic factor identification. We showed here the effectiveness of a machine learning-based approach in predicting the resistance phenotype of the bacterial strains, with the exception of *Enterobacter*, with high accuracy. The ETC algorithm correctly predicted the resistance trait of over 85% bacterial isolates that lacked genes known to be responsible for resistance to specific antibiotics, a primary motivation behind developing such a pipeline. The ETC classifier is similar to the RF classifier in that the ensemble or Classification and Regression Trees (CART) creates random bootstraps that constructs decision nodes based on the random subset of features [24]. However, they differ in the insertion of randomness during the algorithm training where ETC introduces higher grade randomness that results in more independent trees; this decreases the variance because of which ETC tends to yield better results than RF [25], which might be indicative in our case as well.

Literature survey provided support to several genes predicted made by our pipeline. Eventually, the predicted resistance genes will need to be verified using experimental assays to the extent possible. One of the ways to perform the experimental validation to support the prediction made by our pipeline is to conduct tests in strains, such as *E. coli* and *Shigella*, that are resistant to all four carbapenems, i.e. Doripenem, Ertapenem, Imipenem

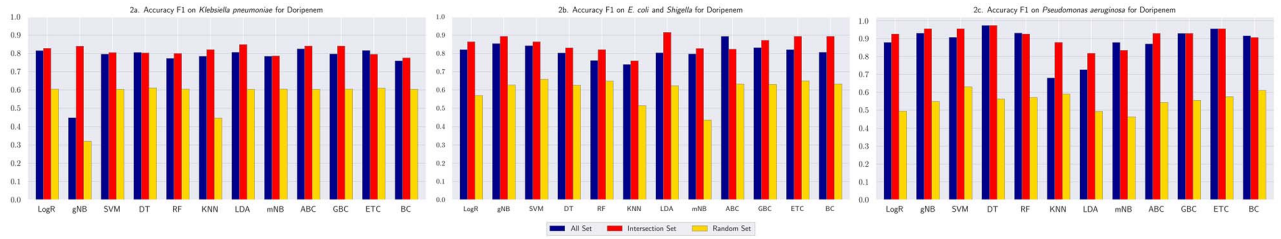

Figure 2. Assessment of the performance of machine learning algorithms in predicting resistance to Doripenem by (**A**) *K. pneumoniae*, (**B**) *E. coli* and *Shigella* and (**C**) *P. aeruginosa*. 'All' denotes all AMR genes for training (as in the cross-validation partitioning), 'Intersection' refers to training using AMR genes that consistently ranked high across all six rounds of cross-validation and 'Random' refers to training using randomly sampled AMR genes.

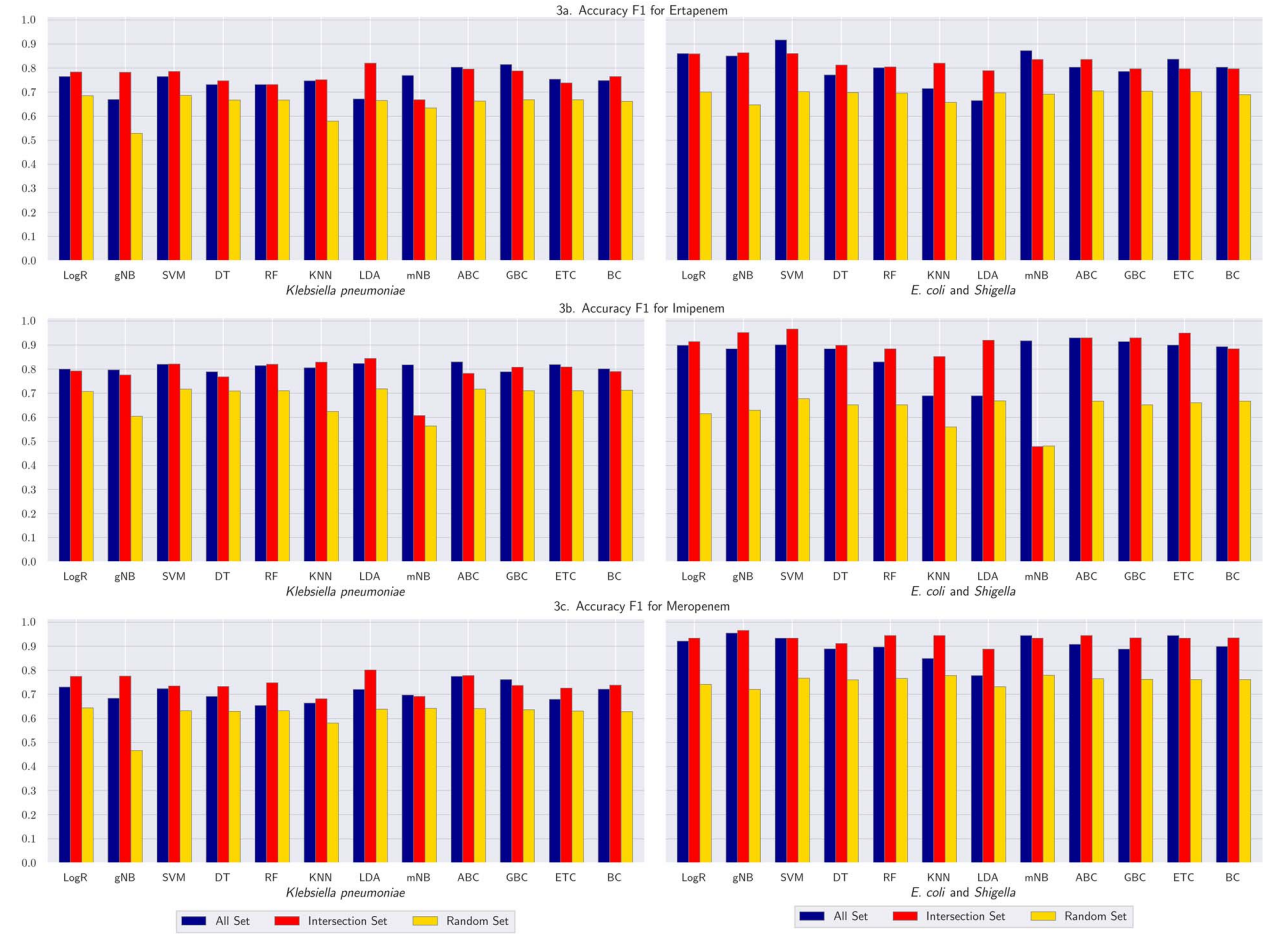

Figure 3. Assessment of the performance of machine learning algorithms in predicting resistance to (**A**) Ertapenem, (**B**) Imipenem and (**C**) Meropenem by *K. pneumoniae* and *E. coli* and *Shigella*. 'All' denotes all AMR genes for training (as in the cross-validation partitioning), 'Intersection' refers to training using AMR genes that consistently ranked high across all six rounds of cross-validation and 'Random' refers to training using randomly sampled AMR genes.

and Meropenem, however, they lack the already implicated resistance genes. Given the ETC algorithm predicts one such strain, *E. coli* AR_0006, as resistant, a knockout of gene(s) predicted for carbapenem resistance in *E. coli* AR_0006 should demonstrate susceptible trait in this strain. An RNAi probe or a CRISPR system seems to be a plausible molecular technique that can be executed to inhibit gene expression. The other way is to clone and complement the predicted gene in a susceptible strain and examine if this results in a gain of resistance trait. We expect future studies to focus on these aspects and the protocol presented here to serve as a tool to prioritize target genes for experimental assays.

Although the machine learning models trained on *Enterobacter* genotypic and phenotypic data attained high accuracy in predicting the resistance phenotype, comparable to that observed with the other organisms, none of the models could predict the resistance phenotype of the *Enterobacter* strains that lack the AMR genes known to be involved in resistance to different classes of drugs. This might be due to limited training data size available for *Enterobacter* in comparison to the other species in our study. Sparse training data may render models that may not generalize well to be a good predictor on yet unseen data. Limited training data make it difficult to obtain a model that generalizes well while balancing the bias-variance trade-off to minimize

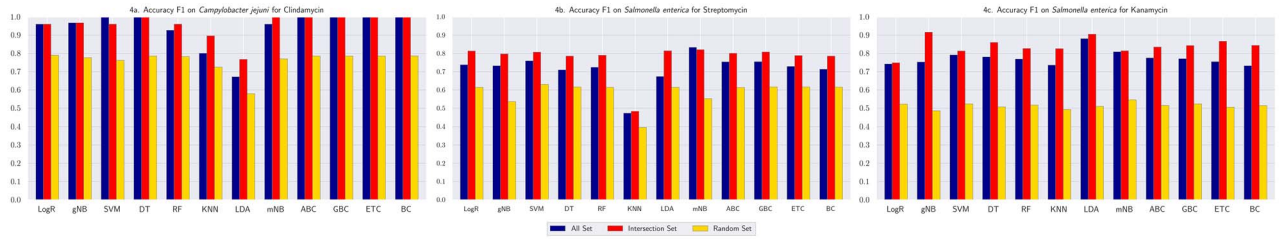

Figure 4. Assessment of the performance of machine learning algorithms in predicting resistance to (**A**) Clindamycin by *C. jejuni*, (**B**) Streptomycin by *S. enterica* and (**C**) Kanamycin by *S. enterica*. 'All' denotes all AMR genes for training (as in the cross-validation partitioning), 'Intersection' refers to training using AMR genes that consistently ranked high across all six rounds of cross-validation and 'Random' refers to training using randomly sampled AMR genes.

Table 1. A list of target genes in *K. pneumoniae* that appeared consistently among top-ranked features of importance in each round of the 6-fold cross-validation for each carbapenem (Doripenem, Ertapenem, Imipenem and Meropenem); these genes have not previously been implicated for carbapenem resistance; however, we hypothesize that these genes could have potential roles in the resistance to carbapenems as their protein products show many similar enzymatic activities as carbapenemases

Set of consistent genes	Doripenem	Ertapenem	Imipenem	Meropenem	NCBI nomenclature
aac(3)-IIa	×	✓	×	×	Aminoglycoside N-acetyltransferase AAC(3)-IIa
aac(3)-IId	×	✓	✓	×	Aminoglycoside N-acetyltransferase AAC(3)-IId
aac(3)-IV	×	✓	✓	×	Aminoglycoside N-acetyltransferase AAC(3)-IV
aac(6′)-Ib	✓	✓	✓	✓	Bifunctional aminoglycoside N-acetyltransferase AAC(3)-Ib/aminoglycoside N-acetyltransferase AAC(6′)-Ib″
aac(6′)-Ib-cr5	×	✓	✓	✓	Fluoroquinolone-acetylating aminoglycoside 6'-N-acetyltransferase AAC(6′)-Ib-cr5
aadA1	×	✓	✓	✓	ANT(3″)-Ia family aminoglycoside nucleotidyltransferase AadA1
aadA2	✓	✓	✓	✓	ANT(3″)-Ia family aminoglycoside nucleotidyltransferase AadA2
ampC	✓	×	✓	✓	Class C beta-lactamase
aph(3″)-Ib	×	✓	×	×	Aminoglycoside O-phosphotransferase APH(3″)-Ib
aph(4)-Ia	×	✓	✓	×	Aminoglycoside O-phosphotransferase APH(4)-Ia
blaCTX-M-15	✓	✓	✓	✓	Class A extended-spectrum beta-lactamase CTX-M-15
blaOXA-1	×	✓	×	×	Oxacillin-hydrolyzing class D beta-lactamase OXA-1
blaSHV-11	✓	×	×	×	Class A broad-spectrum beta-lactamase SHV-11
blaSHV-12	✓	×	×	×	Class A broad-spectrum beta-lactamase SHV-12
blaTEM-1	✓	✓	✓	✓	Class A broad-spectrum beta-lactamase TEM-1
catB3	×	✓	×	×	Type B-3 chloramphenicol O-acetyltransferase CatB3
cmlA1	×	✓	✓	×	Chloramphenicol efflux MFS transporter CmlA1
dfrA12	✓	✓	✓	✓	Trimethoprim-resistant dihydrofolate reductase DfrA12
gyrA_D87N	×	×	×	✓	DNA gyrase subunit A
gyrA_S83I	✓	×	×	✓	DNA gyrase subunit A
gyrA_S83T	✓	×	×	✓	DNA gyrase subunit A
mph(A)	×	×	×	✓	Mph(A) family macrolide 2′-phosphotransferase
oqxB	✓	×	×	×	Multidrug efflux RND transporter permease subunit OqxB
parC_S80I	✓	×	×	✓	DNA topoisomerase IV subunit A
pmrB_R256G	✓	×	×	✓	Two-component sensor histidine kinase
sul1	✓	✓	✓	✓	Sulfonamide-resistant dihydropteroate synthase Sul1
sul2	×	×	✓	×	Sulfonamide-resistant dihydropteroate synthase Sul2
sul3	×	×	✓	×	Sulfonamide-resistant dihydropteroate synthase Sul3

Note. '✓' indicates 'appeared consistently among top-ranked genes of importance in each round of the cross-validation test' and '×' indicates otherwise.

both underfitting and overfitting of the model; such models could be of limited predictive value on new data [26]. Perhaps the approach to address this problem is to sequence more strains, annotate them and create MIC data in the databases. Although the machine learning models for a species–drug design could be specific and not applicable to species that are responsive to other drugs, it is possible that a model trained on a species may predict the traits of its close relatives that respond to the same drugs; these aspects can be explored in future studies.

One of the other problems that afflict machine learning-based classification is the imbalances in data representing different classes, which is often reflected in the biases toward certain classes [24], and this was also obvious with the genetic data at the NCBI pathogen database. This database has a high overrepresentation of clinically important resistant bacterial strains vis-a-vis susceptible strains, and this can affect the predictive power of machine learning models. Although we could find substantial numbers of both resistant and susceptible strains of some species, it was not so for several other species represented in this database, including *Enterobacter*. With advances in sequencing and high throughput technologies, and perhaps a desire to enrich the databases in an unbiased way as

evidenced by a spurt in numerous genome and metagenome projects, we expect the databases to be more enriched and balanced in the near future, which will render the machine learning approaches, such as the one presented here, to be of profound value in addressing many intriguing problems, including the prevailing antibiotic resistance crisis. On the other hand, advances in methodology may mitigate the effects of database imbalances, making possible a reliable prediction even when faced with sparse or skewed data. Artificial intelligence or machine learning, in particular, holds a great promise in addressing emerging health problems, including the AMR.

Software availability

Custom codes and associated datasets are available at GitHub: https://github.com/Janaksunuwar/AMR_prediction. Machine learning in Python with Scikit-learn is available at https://scikit-learn.org/stable/.

Data availability

All genotype and phenotype data used in this study were retrieved from the Isolates Browser at the NCBI Pathogen Detection website: https://www.ncbi.nlm.nih.gov/pathogens/. All other data generated by the authors are available in the article and in its online supplementary material.

Key Points

- We present a computational framework that assesses different machine learning algorithms' ability to predict the antibiotic resistance phenotype of bacterial strains that lack known AMR genes yet demonstrate resistance to drugs of those families.
- We employ machine learning algorithms and evaluate their performance in predicting resistance trait using the whole-genome sequencing AMR genetic data and AST phenotypic data.
- We demonstrate how a machine learning-based protocol can be leveraged to decipher novel genes that have not yet been implicated in resistance to different families of antibiotics.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

References

1. Wenzel RP. The antibiotic pipeline—challenges, costs, and values. *N Engl J Med* 2004;**351**:523–6.
2. Davies J. Where have all the antibiotics gone? *Can J Infect Dis Med Microbiol* (Journal Canadien des Maladies Infectieuses et de la Microbiologie Medicale) 2006;**17**:287–90.
3. Luepke KH, Suda KJ, Boucher H, *et al*. Past, present, and future of antibacterial economics: increasing bacterial resistance, limited antibiotic pipeline, and societal implications. *Pharmacotherapy* 2017;**37**:71–84.
4. Dantas G, Sommer MOA. How to fight back against antibiotic resistance. *Am Sci* 2014;**102**:42.
5. Centers for Disease Control and Prevention. *Newly Reported Gene, mcr-1, Threatens Last-Resort Antibiotics*. 2016. https://www.cdc.gov/drugresistance/solutions-initiative/stories/gene-reported-mcr.html.
6. Laxminarayan R. *Antibiotic Resistance: Crisis Response Journal*. 2016;**12**:26–28. www.crisis.response.com.
7. Chen L, Todd R, Kiehlbauch J, *et al*. *Notes from the Field: Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing Klebsiella pneumoniae—Washoe County, Nevada , 2016*. MMWR Morb Mortal Wkly Rep. 2017;**66**:33. doi:http://dx.doi.org/10.15585/mmwr.mm6601a7externalicon
8. Kadri S. Key takeaways from the U.S. CDC's 2019 antibiotic resistance threats report for frontline providers. *Crit Care Med* 2020;**48**:939–45.
9. CDC, Antibiotic Resistance Threats in the United States, 2019. Atlanta, GA: U.S Department of Helth and Human Services. *CDC*. 2019. https://www.cdc.gov/drugresistance/pdf/threats-report/2019-ar-threats-report-508.pdf
10. Jim O'N. *Review on Antimicrobial Resistance commissioned by the UK Government and the Wellcome Trust*. 2016. https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf
11. Brogan DM, Mossialos E. A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *Glob Health* 2016;**12**:8.
12. Bradley P, Gordon NC, Walker TM, *et al*. Erratum: corrigendum: rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun* 2016;**7**:11465.
13. Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. *J Clin Microbiol* 2019;**57**: 1–15. doi: 10.1128/JCM.01405-18 Print 2019 Mar.
14. Larrañaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. *Brief Bioinform* 2006;**7**:86–112.
15. Rishishwar L, Petit RA, III, Kraft CS, *et al*. Genome sequence-based discriminator for vancomycin-intermediate *Staphylococcus aureus*. *J Bacteriol* 2014;**196**:940–8.
16. Maguire F, Rehman MA, Carrillo C, *et al*. Identification of primary antimicrobial resistance drivers in agricultural nontyphoidal *Salmonella enterica* serovars by using machine learning. *mSystems* 2019;**4**:00211–19.
17. Meletis G, Exindari M, Vavatsi N, *et al*. Mechanisms responsible for the emergence of carbapenem resistance in *Pseudomonas aeruginosa*. *Hippokratia* 2012;**16**:303–7.
18. Schweizer HP. Efflux as a mechanism of resistance to antimicrobials in *Pseudomonas aeruginosa* and related bacteria: unanswered questions. *Genet Mol Res* 2003;**2**:48–62.
19. Zheng J, Lin Z, Sun X, *et al*. Overexpression of OqxAB and MacAB efflux pumps contributes to eravacycline resistance and heteroresistance in clinical isolates of *Klebsiella pneumoniae*. *Emerg Microbes Infect* 2018;**7**:1–11.
20. Egorov AM, Ulyashova MM, Rubtsova MY. Bacterial enzymes and antibiotic resistance. *Acta Naturae* 2018;**10**:33–48.
21. Banerjee R, Humphries R. Clinical and laboratory considerations for the rapid detection of carbapenem-resistant Enterobacteriaceae. *Virulence* 2017;**8**:427–39.
22. Vila J, Ruiz J, Marco F, *et al*. Association between double mutation in gyrA gene of ciprofloxacin-resistant clinical isolates of *Escherichia coli* and MICs. *Antimicrob Agents Chemother* 1994;**38**:2477–9.
23. Alcock BP, Raphenya AR, Lau TTY, *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive

antibiotic resistance database. *Nucleic Acids Res* 2019;**48**: D517–25.
24. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front Aging Neurosci* 2017;**9**:329.
25. Götz M, Weber C, Blöcher J, *et al*. Extremely randomized trees based brain tumor segmentation. Sep 14, 2014.
26. Ding Y, Tang S, Liao SG, *et al*. Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics* 2014;**30**:3152–8.