Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

eMIC-AntiKP: Estimating minimum inhibitory concentrations of antibiotics towards *Klebsiella pneumoniae* using deep learning

Quang H. Nguyen [a,1], Hoang H. Ngo [a,1], Thanh-Hoang Nguyen-Vo [b], Trang T.T. Do [c], Susanto Rahardja [d,e,*], Binh P. Nguyen [b,**]

[a] *School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 100000, Viet Nam*
[b] *School of Mathematics and Statistics, Victoria University of Wellington, Wellington 6140, New Zealand*
[c] *School of Innovation, Design and Technology, Wellington Institute of Technology, Lower Hutt 5012, New Zealand*
[d] *School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China*
[e] *Infocomm Technology Cluster, Singapore Institute of Technology, Singapore 138683, Singapore*

ARTICLE INFO

Article history:
Received 6 April 2022
Received in revised form 22 December 2022
Accepted 23 December 2022
Available online 26 December 2022

Keywords:
Minimum inhibitory concentration
Antimicrobial resistance
Antibiotic
Klebsiella pneumoniae
Convolutional neural networks
k-mer counting

ABSTRACT

Nowadays, antibiotic resistance has become one of the most concerning problems that directly affects the recovery process of patients. For years, numerous efforts have been made to efficiently use antimicrobial drugs with appropriate doses not only to exterminate microbes but also stringently constrain any chances for bacterial evolution. However, choosing proper antibiotics is not a straightforward and time-effective process because well-defined drugs can only be given to patients after determining microbic taxonomy and evaluating minimum inhibitory concentrations (MICs). Besides conventional methods, numerous computer-aided frameworks have been recently developed using computational advances and public data sources of clinical antimicrobial resistance. In this study, we introduce eMIC-AntiKP, a computational framework specifically designed to predict the MIC values of 20 antibiotics towards *Klebsiella pneumoniae*. Our prediction models were constructed using convolutional neural networks and *k*-mer counting-based features. The model for cefepime has the most limited performance with a test 1-tier accuracy of 0.49, while the model for ampicillin has the highest performance with a test 1-tier accuracy of 1.00. Most models have satisfactory performance, with test accuracies ranging from about 0.70–0.90. The significance of eMIC-AntiKP is the effective utilization of computing resources to make it a compact and portable tool for most moderately configured computers. We provide users with two options, including an online web server for basic analysis and an offline package for deeper analysis and technical modification.

1. Introduction

A demi-decade has witnessed the vast evolution of various pathogenic bacteria towards antimicrobial drug resistance by two major steps: dissemination and emergence [1]. While bacterial dissemination in multiple locations weakens the human body with an increased amount of toxins, bacterial emergence accidentally occurs to resist drugs and immediately transfer these resistance genes by both vertical and horizontal transmission [2,3]. The bacterial emergence partially reveals one of the distinct mechanisms in bacterial evolution [4]. The expansion of resistance has been demonstrated to be closely associated with prudent practices of antimicrobial drugs [5,6]. Therefore, great efforts have been made to effectively use antimicrobial drugs in suitable doses with the aim of not only eradicating germs but also strictly controlling any possibilities for bacterial evolution [4]. Patients diagnosed with infections must be promptly treated with properly-dosed antimicrobial medicines to achieve fast recovery, limit unexpected complications, and reduce the use of non-specific antibiotics [7–9]. Postponed medications may reduce the survival chances of infected patients [9]. However, selecting suitable antibiotics is not a time-effective process. For a conventional approach, practitioners can only give precise drugs after they correctly determine the microbic taxonomy and evaluate

[*] Corresponding author at: School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China.
[**] Corresponding author.
E-mail addresses: susantorahardja@ieee.org (S. Rahardja), binh.p.nguyen@vuw.ac.nz (B.P. Nguyen).
[1] Equally contributed.

https://doi.org/10.1016/j.csbj.2022.12.041

minimum inhibitory concentrations (MICs) [10]. In a more complicated situation, patients may not be infected with one strain only, and performing numerous inoculations is unavoidable [11,12]. In recent years, the cost and timing of sequencing technologies have decreased in parallel with the advancement in accuracy, which has facilitated diverse 'omics' research and created an enormous source of biological data [13]. Additionally, the advent of modern molecular assays has sped up diagnostic processes with reduced repetitive steps in bacterial culture. These biological assays have now become commonly used protocols in laboratories worldwide. To identify antimicrobial resistance (AMR) phenotypes, several gene-based methods, including polymerase chain reaction, whole-genome sequencing (WGS), and microarrays, are frequently employed [14–16]. Among these approaches, VITEK 2, a fully automated system, is currently a cheaper and more available one that enables the identification of pathogens and their antibiotic susceptibility in one day [17]. Despite important advantages in processing time and accuracy, sequence-based assays still have limitations that need to be addressed. These assays depend on well-curated databases of AMR genes, and most of the entries are of well-defined AMR genes [18–20]. Besides, distinguishing AMR genes and their nonfunctional paralogs using a similarity-based matching method remains challenging. Additionally, failure in mutation detection in specific regions, such as regulatory and promoter locations, may incidentally occur to cause false-negative susceptibility prediction [21]. Also, since most of the PCR-based methods compulsorily require well-designed primers for chain reactions, annealing processes may be inert if complementary regions are found with mutations. Pre-existing knowledge of AMR regions is indispensable to obtain reliable outcomes when using these assays [10]. To predict uncommon resistances whose molecular mechanisms have not been fully explored or multifactorial, contemporary assays need to be assisted with computational advances to strengthen predictive power.

In recent years, public sources of WGS data, particularly clinical AMR data, have motivated the development of computational frameworks to predict AMR phenotypes without requiring prior knowledge of AMR genes and mutations [22,23,10]. In 2016, Davis et al. [22] built a learning model using *k*-mer information and the AdaBoost algorithm [24] to predict carbapenem resistance in *Acinetobacter baumannii*, beta-lactam and co-trimoxazole resistance in *Streptococcus pneumoniae*, and methicillin resistance in *Staphylococcus aureus*, beside several types of resistance in *Mycobacterium tuberculosis*. In the same year, Drouin et al. [23] developed four prediction models using *k*-mer features and set covering machines [25] towards *Clostridium difficile*, *Pseudomonas aeruginosa*, *Streptococcus pneumoniae*, and *Mycobacterium tuberculosis*. One year later, another prediction model for the MICs of *Neisseria gonorrhoeae* was introduced using single nucleotide polymorphisms (SNPs) from various essential AMR genes [26]. In 2018, Nguyen et al. [10] proposed a computational framework to predict the MICs for *Klebsiella pneumoniae* using *k*-mer features and extreme gradient boosting. Also, in other studies, SNPs, AMR genes, and WGS data were selected as features to construct prediction models using multiple machine learning algorithms [27–30,31]. Although these methods are relatively effective for predicting the MICs of diverse antibiotics against bacterial strains, there is still room for improvement.

In this study, we introduce a compact computational framework for MIC estimation using convolutional neural networks and *k*-mer counting-based features. Our proposed method was shown to have a low computational demand, requiring fewer computing resources and a significantly shorter training time. The computational cost-effectiveness benefits our model by easily catching up with additional data. In a scenario of ceaseless expansion of "omics" data, good models are expected not only to correctly predict outcomes at a contemporary time but also to be simply updatable. Although the power of classical machine learning methods has been demonstrated in some particular problems, simply updating the model is not their strength. While deep learning models need weight tuning at several penultimate layers only, classical machine learning models demand more computationally intensive tasks (feature selection, hyper-parameter tuning, etc.). Deep learning has now become one of the most robust computational approaches which is used to address diverse problems in molecular biology [32–35–38] and biochemistry [39–42]. Our prediction framework was designed to predict the MICs of multiple antibiotics against a typical strain of infectious bacteria only to demonstrate the computational and timing efficiency of the proposed method. The dataset used to develop the framework was collected from a similar study by Nguyen et al. [10]. The framework allows users to obtain predicted MICs of 20 antibiotics against *Klebsiella pneumoniae (K. pneumoniae)*, a leading cause of global fatalities, in a very short time with competitive performances compared to those of Nguyen et al. [10], which is a state-of-the-art method. We also deploy our framework as an online web server to support the research community. Besides, an offline package was provided with all the necessary modules for deeper analysis and technical modification.

2. Materials and methods

2.1. Benchmark dataset

We used the dataset in a similar study by Nguyen et al. [10]. Between September 2011 and March 2017, the Houston Methodist Hospital System cultured *Klebsiella pneumoniae* isolates from patient samples. Using the BD-Phoenix system (BD Diagnostics, Sparks, MD, USA), strains were tested for minimum inhibitory concentrations to 20 antibiotics. Genomes were assembled using the PATRIC (Pathosystems Resource Integration Center) assembly service [20] to make up that dataset. These genomes have labels which are MIC values of 20 tested antibiotics. There are 32,312 samples for pairs of genome-antibiotic and corresponding MIC values. Table 1 provides information about the data used for model training and evaluation. Additionally, most of the sample files have an average size of 5.5 MB and contain numerous contigs of about 300 nucleotides. There are a small number of files having from 400 to 5000 contigs (76/1667 files). The MIC values were rounded up to second decimals (Table 2). For valid calculation, all mathematics notations including " $>$ ", "$\geq$ ", " $<$ ", and "$\leq$ " were removed from the MIC values. The MIC values of " $>x$" and " $<x$" were replaced with "$2x$" and "$x/2$", respectively, while the MIC values of "$\geq x$" and "$\leq x$" were all replaced with "x". The MIC values of paired antibiotics (e.g., Piperacillin/Tazobactam) were replaced with the MIC values of the first antibiotic only because those of the second ones were constant or dependent on the first ones' doses.

Table 1
Data for model development and evaluation.

Antibiotics	No. samples	Antibiotics	No. samples
Trimethoprim/ Sulfamethoxazole	1667	Cefuroxime sodium	1575
Tobramycin	1666	Ceftriaxone	1667
Tetracycline	1667	Ceftazidime	1667
Piperacillin/Tazobactam	1662	Cefoxitin	1645
Nitrofurantoin	895	Cefepime	1571
Meropenem	1660	Cefazolin	1667
Levofloxacin	1666	Aztreonam	1644
Imipenem	1666	Ampicillin/ Sulbactam	1664
Gentamicin	1667	Ampicillin	1666
Ciprofloxacin	1664	Amikacin	1667

Table 2
Minimum inhibitory concentrations of 20 antibiotics for all genome samples.

MIC	No. samples	MIC	No. samples	MIC	No. samples
0.02	1	4.00	4459	19.00	12
0.03	1	6.00	2	20.00	15
0.09	1	8.00	3737	21.00	4
0.13	282	11.00	2	23.00	1
0.25	120	12.00	2	32.00	10802
0.38	1	14.00	2	50.00	5
0.50	784	16.00	3896	64.00	1834
1.00	3118	17.00	4	128.00	1833
2.00	1775	18.00	11		

2.2. *K-mer counting-based features*

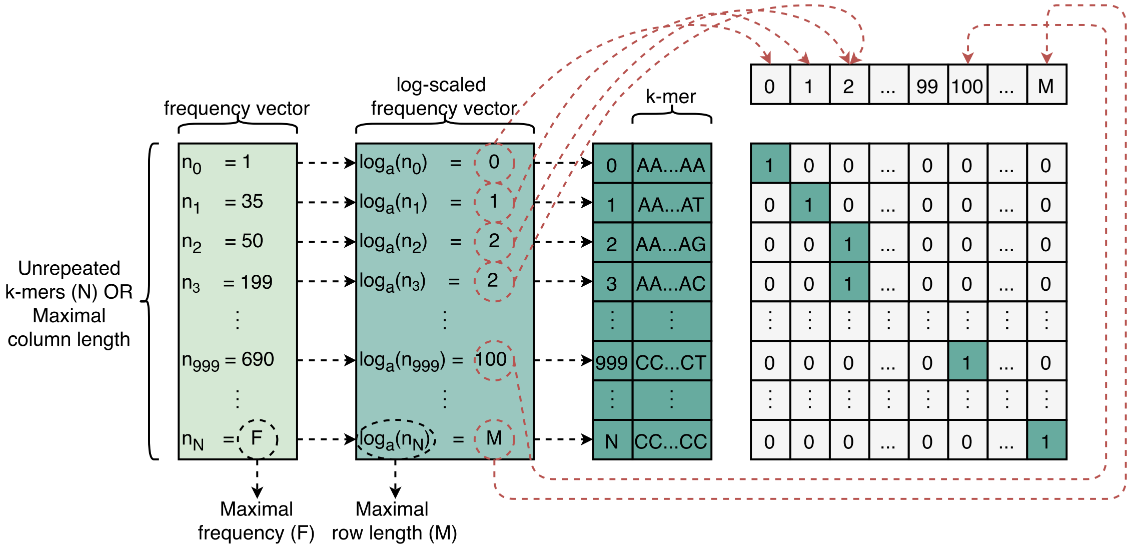
The K-mer counting (KMC) algorithm was used to compute the frequency of each *k*-mer in a single file. Initially, all frequencies of appearing *k*-mers were first computed file by file. After obtaining counts of all *k*-mers in each separate file, the counts of all *k*-mers appearing in the entire benchmark dataset were then computed. Among all possible existing *k*-mers, the highest count of any certain one is considered maximal frequency (*F*). Several values of *k* were selected to estimate the number of unrepeated *k*-mer appearing in the entire benchmark dataset. The number of unrepeated *k*-mers (*N*) for $k = 10$ is 520,000 with a maximal frequency of 4281, followed by those of $k = 8$, 6, and 5, with *k*-mer counts of 32,896, 2080, and 512, and the maximal frequencies of 6690, 30,614, and 85,308, respectively. For $k = 5$ or 6, great variations between maximal frequencies and minimal frequencies are observed. In our experiments, although $k = 8$ was selected, the variation between the maximal frequency and the minimal frequency was still large and needed to be solved. The frequencies of *k*-mers, therefore, cannot be directly used as input features because large differences will cause biases in informatics importance. To address this issue, the logarithmic scale, a non-linear mapping scale, was employed to narrow the range of values. To select the effective size of the input matrix, the base *b* of the logarithmic scale can be adjusted. The process for the log-scaled transformation of KMC-based features is visualized in Fig. 1. After all existing *k*-mers in the benchmark dataset were defined, genome files were converted into their corresponding frequency vectors, in which each component is the frequency of a particular *k*-mer. A genome file may not contain all existing *k*-mers, and the values of the unfound *k*-mers were, therefore, filled with '0'. All the created frequency vectors were same-length vectors. The log-scaled transformation was then applied to convert these frequency vectors into KMC-based log-scaled matrices. The base *b* of the logarithm function was selected as:

$$b = \sqrt[M]{F}, \tag{1}$$

where *M* and *F* are the number of desired rows and the maximal frequency, respectively. For any particular i^{th}*k*-mer, $\log_b$ value (using the integer part only) of its count is equal to j^{th} column to be filled. The value $v_{(i,\ j)}$, therefore, is assigned '1' while other values of the same row (belonging to the same *k*-mer) are assigned '0'. Each row vector (corresponding to a specific independent *k*-mer) of the created matrix now becomes a one-hot vector. The other *k*-mers are successively processed in the same manner. A frequency vector $\vec{f} = [1, 35, 50, 199, ..., 690, ..., F]$, for example, has its $\log_b(n_0 = 1)$, $\log_b(n_1 = 35)$, $\log_b(n_2 = 50)$, $\log_b(n_3 = 199)$, $\log_b(n_{999} = 690)$, and $\log_b(n_N = F)$ computed as 0, $1._{12...}$, $2._{23...}$, ..., $99._{87...}$, $100._{05...}$, ..., *M*, respectively. n_0, n_1, n_2, n_3, n_{999}, and n_N are the numbers of counts for *k*-mers 'AA…AA', 'AA…AT', 'AA…AG', 'AA…AC', 'CC…CT', and 'CC…CC', respectively. This frequency vector $\vec{f}$ is then converted into a log-scaled frequency vector $\vec{f_{log}} = [0, 1, 2, ..., 99, 100, ..., M]$. The log-scaled frequency vector $\vec{f_{log}}$ is eventually transformed into a KMC-based log-scaled matrix of size $N \times M$. Based on computed results, the values $v_{(0,\ 0)}$, $v_{(1,\ 1)}$, $v_{(2,\ 2)}$, $v_{(3,\ 2)}$, $v_{(999,\ 100)}$, and $v_{(N,\ M)}$ are assigned 1 and the other values $v_{(i,\ j)}$ are assigned 0 (Fig. 1). In our implementation, the input matrix was designed with a size of 32,896 (*N*) × 14 (*M*), which was equal to the counts of *k*-mers corresponding to $k = 8$ and the base of 2 for the logarithmic scale.

2.3. *Model architecture*

Multiple empirical experiments were conducted to finally come up with the best-performing architecture for model development. Our proposed models were designed with two convolutional blocks and one fully-connected block (Fig. 2). Two convolutional blocks were similarly built of one 2-dimensional convolutional (Conv2D) layer, one max-pooling layer, and one batch normalization (BatchNorm) [43] layer while the fully-connected block comprises of three fully-connected layers (FCs). The rectified linear unit (ReLU) was used as the activation function for all blocks. The kernel size of 3 × 3 and the pooling window of 2 × 2 were applied in both the convolutional blocks. The matrices of sizes 32,896 × 14 were initially passed through the first convolutional block to turn into the smaller



0	1	2	...	99	100	...	M
1	0	0	...	0	0	...	0
0	1	0	...	0	0	...	0
0	0	1	...	0	0	...	0
0	0	1	...	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	...	0	1	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	...	0	0	...	1

Fig. 1. The processing steps in converting KMC vectors to a KMC-based log-scaled matrices.

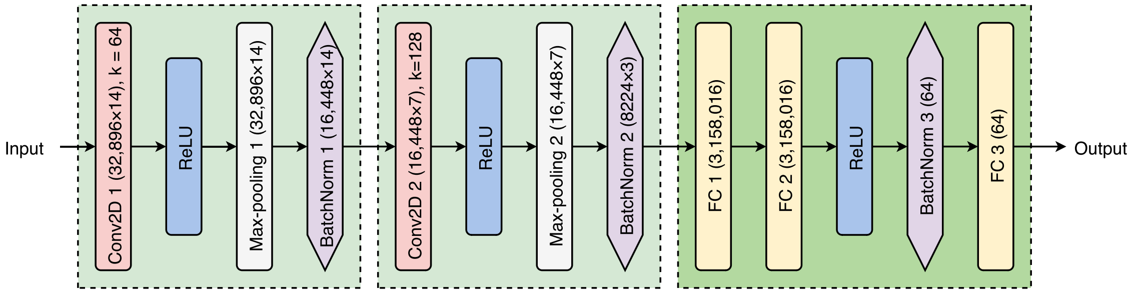


Fig. 2. Model architecture.

matrices of size 16,448 × 7 using 64 feature maps. The second convolutional block then converted these matrices into ones of size 8224 × 3 using 128 feature maps before being flattened as the first layer of 8224 × 3 × 128 = 3158,016 nodes in the fully-connected block. The second fully-connected layer rendered outputs with a size of 64, followed by a batch normalization layer before being ReLU-activated, dropped out with a rate of 0.5, and finally returning the outcomes. The model was optimized with the Adam optimizer [44] to minimize the mean square error as the model loss function.

In addition to MIC estimation, we also predict if a sample is resistant to an antibiotic in the dataset. The architecture of the models for this problem are similar to that in Fig. 2, except for the binary cross-entropy loss was used when training the models.

2.4. Evaluation Metrics

For the MIC estimation problem, the accuracy within ± 1 two-fold dilution factor (or 1-tier accuracy) of the actual MIC was used to evaluate the model performance. This evaluation metric was approved by the U.S. Food and Drug Administration and is in agreement with current standard criteria at laboratories worldwide. Fig. 3 illustrates the 1-tier accuracy.

For the prediction of resistance problem, the models were evaluated based on multiple metrics, including the accuracy, sensitivity, specificity, F1-score, area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PR), the very major error rate (VAE), defined as the percentage of resistant samples that are incorrectly predicted to be susceptible, and the major error rate (ME), defined as the percentage of susceptible samples that are incorrectly predicted to be resistant by the model. In this problem, VAE = 1 - sensitivity, and ME = 1 - specificity.

3. Results and discussion

3.1. Model development and evaluation

The processed benchmark dataset was randomly split into a training set, a validation set, and a test set, which accounts for 80 %, 10 %, and 10 % of total samples, respectively. All models for MIC prediction of 20 antibiotics were preliminarily trained over 200 epochs. In our experiments, all the deep learning models were implemented using Tensorflow version 2.0 and trained on an i5 8400 CPU, 32 GB RAM, and one NVIDIA 1080Ti GPU. It took about 30–35 s to train one epoch and about 10 s to complete testing. For predicting MIC for each antibiotic, the model at the epoch where the validation loss was minimum was selected as the optimal model. Most of the models converged around epoch 150. For predicting resistance, each model was trained with a learning rate of $1e - 3$ and the early stopping technique. After the best model was selected based on the minimum value of the validation loss, the model was trained in two more epochs with a learning rate of $1e - 4$ and the combination of the training set and the validation set.

Table 3 summarizes the model performances of the MIC prediction models for 20 antibiotics. There are 7 models with 1-tier accuracy of over 0.90 and 9 models with 1-tier accuracy of 0.74–0.90. Models for MIC prediction of Cefepime, Meropenem, Piperacillin/Tazobactam, and Tetracycline have 1-tier accuracy of 0.49, 0.59, 0.63, and 0.68, respectively.

Table 4 presents the model performances of the resistance prediction models for 19 antibiotics. Ampicillin was excluded from our experiment as there were only 4 susceptible samples corresponding to that antibiotic in the dataset. Table 4 also includes the average values of VME and ME from 10-fold cross-validation in Nguyen et al.

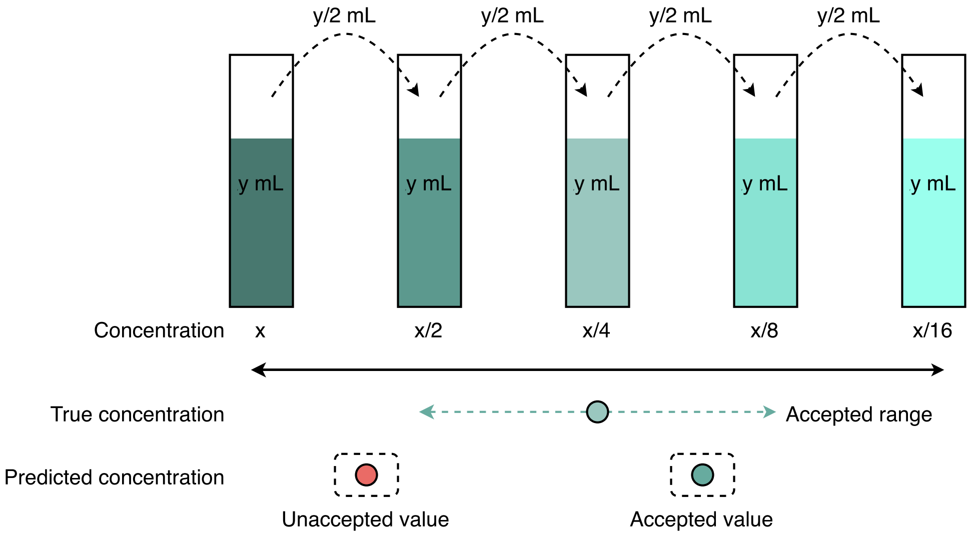


Fig. 3. The accuracy within ± 1 two-fold dilution factor.

Table 3
Model performances of the MIC prediction models.

Antibiotic	Validation 1-tier accuracy	Test 1-tier accuracy
Amikacin	0.88	0.91
Ampicillin	0.99	1.00
Ampicillin/Sulbactam	0.91	0.92
Aztreonam	0.80	0.79
Cefazolin	0.95	0.94
Cefepime	0.46	0.49
Cefoxitin	0.79	0.74
Ceftazidime	0.89	0.90
Ceftriaxone	0.87	0.89
Cefuroxime sodium	0.96	0.97
Ciprofloxacin	0.92	0.89
Gentamicin	0.72	0.80
Imipenem	0.70	0.75
Levofloxacin	0.86	0.89
Meropenem	0.59	0.59
Nitrofurantoin	0.93	0.93
Piperacillin/Tazobactam	0.69	0.63
Tetracycline	0.70	0.68
Tobramycin	0.80	0.80
Trimethoprim/Sulfamethoxazole	0.84	0.83

[10]. Although this is just a cursory comparison, it can be seen that the performances of our models were not much different from those of Nguyen et al. [10]. Some VME values (for Cefazolin and Trimethoprim/Sulfamethoxazole) and ME values (for Aztreonam, Ceftriaxone, and Levofloxacin) of our models are still better than the corresponding metrics in Nguyen et al. [10].

3.2. Strengths and limitations

In this study, our prediction framework was significantly optimized in terms of training time and computing resources compared to the state-of-the-art method proposed by Nguyen et al. [10]. To complete the training, only about one hour is needed. The size of the framework is only 2.4 GB, while Nguyen et al.'s framework requires up to 148 GB of disk space. During the training process, a GPU of 11 GB was used with a small requirement of random access memory (RAM), while Nguyen et al.'s framework needed a RAM of 1.5TB. Besides, their framework was trained using the XGB algorithm, which usually demands a longer time for both hyper-parameter tuning and model training. The effective use of computing resources makes our framework compact and portable for most computers with moderate configurations. Moreover, since genome data is being accumulated to become big data in genetics and biology, Nguyen et al.'s framework is bulky for updating or retraining. On the contrary, updating our model with additional future data is a simple task supported by exceptionally strong deep learning platforms.

According to Nguyen et al.'s report [10] on the model performance, their framework does not work significantly better than ours because the to-be-given antibiotic's dose must be properly adjusted based on patients' physiological conditions (e.g., age, sex). Therefore, the slight differences in performance between the two frameworks are very unlikely to result in any better or worse predicted MIC values. In particular, both methods only aim to recommend suitable doses to physicians to finally decide given doses based on each particular condition. Physicians can use these predicted MIC values as references and then combine them with other factors to give precise doses to patients. Although the feature extractor of eMIC-AntiKP is highly effective and fast, it is undeniable that our feature extraction method may partially affect the prediction accuracy. In the future, we will look into how to make a better feature extractor so that we can not only get important sample information but also speed up all the training processes.

We developed computational models to predict the minimum inhibitory concentrations of 20 antibiotics towards *Klebsiella pneumoniae* because of its prevalence in causing bacterial infection in humans. *Klebsiella pneumoniae* is an opportunistic bacterial strain normally living in human intestines and feces. They can cause diverse infection types with increased tendencies to become antibiotic-resistant. For patients with background diseases, the infection may permanently impair an organ's functions (e.g., the lung, the kidney, etc.). On the other hand, in our study, the size of the data on antibiotics treating *Klebsiella pneumoniae* was sufficiently large to develop a computation model using deep learning. The proposed method can be applied to other species of bacteria with minor adjustments, as long as there is sufficient data available.

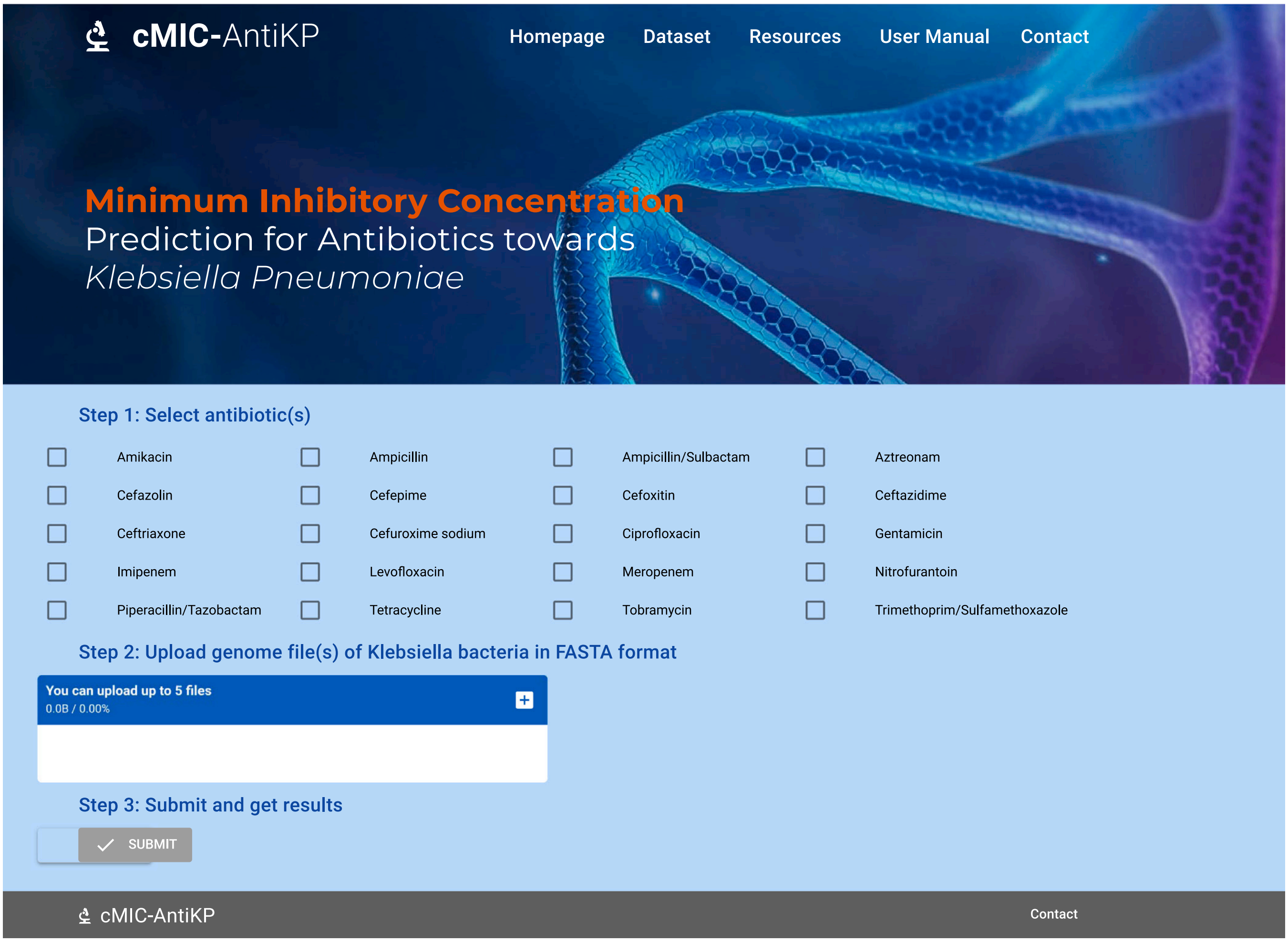
There are several models for predicting resistance to *Klebsiella pneumoniae*, for example, the AMR-Diag model [45]. However, their datasets are much smaller than the one used in our study, in terms of both the number of samples and the number of antibiotics. In the future, we will look for larger datasets for improving our models and for cross-checking to find out how the models would perform on the isolates from a different dataset.

3.3. Software availability

Our online web server was designed with a user-friendly interface to support not only expert users but also novice users. Users can freely download benchmark datasets from the website since the

Table 4
Model performances of the resistance prediction models.

Antibiotics	Accuracy	AUC-ROC	AUC-PR	Sensitivity	Specificity	F1-score	VME	ME	VME [10]	ME [10]
Amikacin	0.958	0.932	0.763	0.700	0.977	0.700	0.300	0.023	0.298	0.000
Ampicillin/Sulbactam	0.984	0.946	0.992	0.986	0.945	0.991	0.014	0.056	0.003	0.032
Aztreonam	0.871	0.850	0.956	0.901	0.682	0.924	0.099	0.318	0.001	0.398
Cefazolin	0.958	0.975	0.998	0.968	0.800	0.978	0.032	0.200	0.060	0.018
Cefepime	0.826	0.862	0.929	0.825	0.829	0.870	0.175	0.171	0.007	0.137
Cefoxitin	0.880	0.925	0.954	0.843	0.925	0.886	0.157	0.075	0.077	0.009
Ceftazidime	0.920	0.922	0.989	0.926	0.857	0.940	0.074	0.143	0.005	0.123
Ceftriaxone	0.988	1.000	1.000	0.987	1.000	0.993	0.013	0.000	0.000	0.188
Cefuroxime sodium	0.936	0.981	0.999	0.939	0.889	0.965	0.061	0.111	0.002	0.010
Ciprofloxacin	0.963	0.991	0.999	0.972	0.900	0.979	0.028	0.100	0.005	0.025
Gentamicin	0.907	0.946	0.945	0.853	0.946	0.886	0.147	0.054	0.072	0.009
Imipenem	0.951	0.983	0.952	0.958	0.948	0.920	0.042	0.052	0.040	0.032
Levofloxacin	0.970	0.994	0.999	0.961	1.000	0.980	0.039	0.000	0.016	0.020
Meropenem	0.932	0.963	0.945	0.917	0.939	0.889	0.083	0.061	0.048	0.027
Nitrofurantoin	0.910	0.803	0.962	0.931	0.667	0.950	0.069	0.333	0.018	0.227
Piperacillin/Tazobactam	0.865	0.886	0.923	0.857	0.884	0.900	0.143	0.116	0.025	0.012
Tetracycline	0.854	0.897	0.903	0.805	0.905	0.849	0.195	0.095	0.114	0.008
Tobramycin	0.894	0.926	0.916	0.849	0.949	0.899	0.151	0.051	0.040	0.012
Trimethoprim/Sulfamethoxazole	0.898	0.957	0.983	0.905	0.878	0.931	0.095	0.122	0.119	0.108



cMIC-AntiKP

Homepage Dataset Resources User Manual Contact

Minimum Inhibitory Concentration
Prediction for Antibiotics towards
Klebsiella Pneumoniae

Step 1: Select antibiotic(s)

□ Amikacin
□ Ampicillin
□ Ampicillin/Sulbactam
□ Aztreonam
□ Cefazolin
□ Cefepime
□ Cefoxitin
□ Ceftazidime
□ Ceftriaxone
□ Cefuroxime sodium
□ Ciprofloxacin
□ Gentamicin
□ Imipenem
□ Levofloxacin
□ Meropenem
□ Nitrofurantoin
□ Piperacillin/Tazobactam
□ Tetracycline
□ Tobramycin
□ Trimethoprim/Sulfamethoxazole

Step 2: Upload genome file(s) of Klebsiella bacteria in FASTA format

You can upload up to 5 files
0.0B / 0.00%

Step 3: Submit and get results

✓ SUBMIT

cMIC-AntiKP

Contact

Fig. 4. The user interface of eMIC-AntiKP.

model performance of our proposed method is expected to be improved by users' contributions. Besides an online tool, we also created a Python module for offline analysis and a detailed guideline to assist users to appropriately and effectively perform prediction tasks. Expert users are highly recommended to download the Python module to run on their personal computers for the best experience and convenience. Input (bacterial genome) files must be uploaded to the server for further processing. The web server supports multifaceted options for 20 antibiotics. The interface of eMIC-AntiKP is presented in Fig. 4. The data and code used in our study can be downloaded from the project web page at https://github.com/ngphubinh/eMIC-AntiKP, and our web server is available at https://homepages.ecs.vuw.ac.nz/~nguyenb5/apps/emic-antikp.

4. Conclusions

Our prediction framework for MIC values of 20 antibiotics towards *Klebsiella pneumoniae* is a compact and less computationally-demanding analytic tool which is suitable for most of the current personal computers. Both the online web server and the offline package are available with clear instructions and an easy-to-use interface.

CRediT authorship contribution statement

Quang H. Nguyen: Methodology, Software, Resources, Writing - review & editing. **Hoang H. Ngo**: Methodology, Software, Investigation, Data curation, Writing - original draft. **Thanh-Hoang Nguyen-Vo**: Formal analysis, Validation, Writing - original draft, Writing - review & editing, Visualization. **Trang T. T. Do**: Visualization, Writing - review & editing. **Susanto Rahardja**: Formal analysis, Resources, Writing review & editing, Supervision. **Binh P. Nguyen**: Conceptualization, Methodology, Formal analysis, Writing - review & editing, Visualization, Supervision.

Funding

The work of Trang T. T. Do was supported in part by the Whitireia and WelTec Contestable fund.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Courvalin P. Antimicrobial drug resistance: "prediction is very difficult, especially about the future. Emerg Infect Dis 2005;11(10):1503. https://doi.org/10.3201/eid1110.051014

[2] Courvalin P, Cuot PT. Minimizing potential resistance: the molecular view. Clin Infect Dis 2001;33(Supplement_3):S138–46. https://doi.org/10.1086/321840

[3] Seppälä H, Klaukka T, Lehtonen R, Nenonen E, Huovinen P. Outpatient use of erythromycin: link to increased erythromycin resistance in group A streptococci. Clin Infect Dis 1995;21(6):1378–85. https://doi.org/10.1093/clinids/21.6.1378

[4] Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 2010;74(3):417–33. https://doi.org/10.1128/MMBR.00016-10

[5] Andersson DI. Persistence of antibiotic resistant bacteria. Curr Opin Microbiol 2003;6(5):452–6. https://doi.org/10.1016/j.mib.2003.09.001

[6] Chiew Y-F, Yeo S-F, Hall LMC, Livermore DM. Can susceptibility to an antimicrobial be restored by halting its use? the case of streptomycin versus Enterobacteriaceae. J Antimicrob Chemother 1998;41(2):247–51. https://doi.org/10.1093/jac/41.2.247

[7] Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. Ther Adv Drug Safety 2014;5(6). https://doi.org/10.1177/2042098614554919

[8] Palmer H, Palavecino E, Johnson J, Ohl C, Williamson J. Clinical and microbiological implications of time-to-positivity of blood cultures in patients with gram-negative bacilli bacteremia. Eur J Clin Microbiol Infect Dis 2013;32(7):955–9. https://doi.org/10.1007/s10096-013-1833-9

[9] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 2006;34(6):1589–96. https://doi.org/10.1097/01.CCM.0000217961.75225.E9

[10] Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an *in silico* minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci Rep 2018;8(1):1–11. https://doi.org/10.1038/s41598-017-18972-w

[11] Reller LB, Weinstein M, Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clin Infect Dis 2009;49(11):1749–55. https://doi.org/10.1086/647952

[12] Opota O, Croxatto A, Prod'hom G, Greub G. Blood culture-based diagnosis of bacteraemia: state of the art. Clin Microbiol Infect. 2015;21(4):313–22. https://doi.org/10.1016/j.cmi.2015.01.003

[13] Mardis ER. DNA sequencing technologies: 2006–2016. Nature Prot. 2017;12(2):213–8. https://doi.org/10.1038/nprot.2016.182

[14] Goldberg B, Sichtig H, Geyer C, Ledeboer N, Weinstock GM. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. MBio 2015;6(6). https://doi.org/10.1128/mBio.01888-15

[15] Leski TA, Vora GJ, Barrows BR, Pimentel G, House BL, Nicklasson M, et al. Molecular characterization of multidrug resistant hospital isolates using the antimicrobial resistance determinant microarray. PloS One 2013;8(7):e69507https://doi.org/10.1371/journal.pone.0069507

[16] Pulido MR, García-Quintanilla M, Martín-Peña R, Cisneros JM, McConnell MJ. Progress on the development of rapid methods for antimicrobial susceptibility testing. J Antimicrob Chemother 2013;68(12):2710–7. https://doi.org/10.1093/jac/dkt253

[17] Ligozzi M, Bernini C, Bonora MG, Fatima MD, Zuliani J, Fontana R. Evaluation of the VITEK 2 system for identification and antimicrobial susceptibility testing of medically relevant gram-positive cocci. J Clin Microbiol 2002;40(5):1681–6. https://doi.org/10.1128/JCM.40.5.1681-1686.2002

[18] McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 2013;57(7):3348–57. https://doi.org/10.1128/AAC.00419-13

[19] Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 2015;9(1):207–16. https://doi.org/10.1038/ismej.2014.106

[20] Antonopoulos DA, Assaf R, Aziz RK, Brettin T, Bun C, Conrad N, et al. PATRIC as a unique resource for studying antimicrobial resistance. Briefings Bioinf 2019;20(4):1094–102. https://doi.org/10.1093/bib/bbx083

[21] Macintyre G, Yepes AJ, Ong CS, Verspoor K. Associating disease-related genetic variants in intergenic regions to the genes they impact. PeerJ 2014;2:e639https://doi.org/10.7717/peerj.639

[22] Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep 2016;6:27930. https://doi.org/10.1038/srep27930

[23] Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genom 2016;17(1):1–15. https://doi.org/10.1186/s12864-016-2889-6

[24] Freund Y, Schapire R. A short introduction to boosting. J Japn Soc Artif Intell 1999;14(5):771–80.

[25] Marchand M, Shawe-Taylor J. The set covering machine (Dec). J Machine Learn Res 2002;3:723–46.

[26] Eyre DW, Silva DD, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. J Antimicrob Chemother 2017;72(7):1937–47. https://doi.org/10.1093/jac/dkx067

[27] Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun 2015;6(1):1–15. https://doi.org/10.1038/ncomms10063

[28] K.E. Niehaus, T.M. Walker, D.W. Crook, T.E. Peto, D.A. Clifton, Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis, in: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2014, pp.618–621.10.1109/BHI.2014.6864440.

[29] Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C-AD, et al. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. Front Microbiol 2016;7:1887. https://doi.org/10.3389/fmicb.2016.01887

[30] Coelho JR, Carriço JA, Knight D, Martínez J-L, Morrissey I, Oggioni MR, et al. The use of machine learning methodologies to analyse antibiotic and biocide susceptibility in Staphylococcus aureus. PLoS One 2013;8(2):e55582https://doi.org/10.1371/journal.pone.0055582

[31] Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Elias CDO, et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother 2013;68(10):2234–44. https://doi.org/10.1093/jac/dkt180

[32] Nguyen BP, Nguyen QH, Doan-Ngoc G-N, Nguyen-Vo T-H, Rahardja S. iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks. BMC Bioinform 2019;20(23):1–12. https://doi.org/10.1186/s12859-019-3295-2

[33] Nguyen QH, Nguyen-Vo T-H, Le NQK, Do TT, Rahardja S, Nguyen BP. iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks. BMC Genom 2019;20(9):951. https://doi.org/10.1186/s12864-019-6336-3

[34] Le N-Q-K, Nguyen QH, Chen X, Rahardja S, Nguyen BP. Classification of adaptor proteins using recurrent neural networks and PSSM profiles. BMC Genom 2019;20(Suppl 9):1–9. https://doi.org/10.1186/s12864-019-6335-4

[35] Nguyen-Vo T-H, Trinh QH, Nguyen L, Nguyen-Hoang P-U, Rahardja S, Nguyen BP. iPromoter-Seqvec: identifying promoters using bidirectional long short-term memory and sequence-embedded features. BMC Genom 2022;23(681). https://doi.org/10.1186/s12864-022-08829-6

[36] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2

[37] Nguyen QH, Tran HV, Nguyen BP, Do TTT. Identifying transcription factors that prefer binding to methylated DNA using reduced g-gap dipeptide composition. ACS Omega 2022;7(36):32322–30. https://doi.org/10.1021/acsomega.2c03696

[38] Rahardja S, Wang M, Nguyen BP, Franti P, Rahardja S. A lightweight classification of adaptor proteins using transformer networks. BMC Bioinform 2022;23(461). https://doi.org/10.1186/s12859-022-05000-6

[39] Nguyen-Vo T-H, Nguyen L, Do N, Le PH, Nguyen T-N, Nguyen BP, et al. Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features. ACS Omega 2020;5(39):25432–9. https://doi.org/10.1021/acsomega.0c03866

[40] Nguyen-Vo T-H, Trinh QH, Nguyen L, Nguyen-Hoang P-U, Nguyen T-N, Nguyen DT, et al. iCYP-MFE: identifying human Cytochrome P450 inhibitors using multitask learning and molecular fingerprint-embedded encoding. J Chem Inform Model 2021;62(21):5059–68. https://doi.org/10.1021/acs.jcim.1c00628

[41] Nguyen-Vo T-H, Trinh QH, Nguyen L, Do TTT, Chua MCH, Nguyen BP. Predicting antimalarial activity in natural products using pretrained bidirectional encoder representations from transformers. J Chem Inform Model 2021;62(21):5050–8. https://doi.org/10.1021/acs.jcim.1c00584

[42] Nguyen L, Nguyen-Vo T-H, Trinh QH, Nguyen BH, Nguyen-Hoang P-U, Le L, et al. iANP-EC: identifying anticancer natural products using ensemble learning incorporated with evolutionary computation. J Chem Inform Model 2022;62(21):5080–9. https://doi.org/10.1021/acs.jcim.1c00920

[43] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.031670 2015.

[44] D.P. Grzybowski, Adam: A method for stochastic optimization. arXiv:1412.6980 2014.

[45] Avershina E, Sharma P, Taxt AM, Singh H, Frye SA, Paul K, et al. AMR-Diag: Neural network based genotype-to-phenotype prediction of resistance towards β-lactams in Escherichia coli and Klebsiella pneumoniae. Comput Struct Biotechnol J 2021;19:1896–906. https://doi.org/10.1016/j.csbj.2021.03.027