International Journal of Pharmacy and Biological Sciences
ISSN: 2321-3272 (Print), ISSN: 2230-7605 (Online)
IJPBS | Volume 8 | Issue 1 | JAN-MAR| 2018 | 767-771
Research Article | Pharmaceutical Sciences | Open Access | MCI Approved| ज्ञान-विज्ञान विमुक्तये
|UGC Approved Journal |

Predicting Antibiotic Resistance Using Genomic Data and Machine Learning Algorithms

Shiji Thomas[1] and O. Jamsheela[2]

[1]Department of Microbiology, EMEA College of Arts and Science, Kondotty, Kerala, India.
[2]Department of Computer Science, EMEA College of Arts and Science, Kondotty, Kerala, India.

*Corresponding Author Email: shijiboby1@gmail.com

ABSTRACT

A major public health concern of the twenty-first century is antibiotic resistance (AR). In addition to being time-consuming, traditional diagnostic techniques for identifying antimicrobial resistance (AMR) sometimes often fail to keep pace with emerging resistant pathogens. More recently, whole-genome sequencing (WGS) has enabled scientists to create extensive microbial genomic datasets. Together with machine learning (ML) methods, these data offer strong tools for quick and precise forecasting of resistance phenotypes. In this work, the current status of research on combining machine learning algorithms and genomic data to predict antibiotic resistance is reviewed. The application of ML algorithms to the problem of AMR has gained tremendous interest in the past few years due to the growth of experimental and clinical data, heavy investment in computational capacity, advances in algorithm performance, and growing urgency for innovative approaches to tackle the problem of drug resistance. Here, we review the current applications of machine learning in improving the diagnosis, treatment and prevention of bacterial AMR.

KEY WORDS

Machine learning, genomic data, antibiotic resistance, diagnosis.

1. INTRODUCTION

Antimicrobial resistance (AMR) is a growing challenge to public health as a silent pandemic. Each year in the United States, almost 2 million people are infected with multidrug-resistant (MDR) bacteria, with several cases leading to death (1). Multidrug-resistant and pan-drug-resistant microorganisms are commonly found in a hospital environment, especially in the critical care setting (2). Lack of available antibiotics to combat drug-resistant strains is a critical issue for the healthcare sector. Moreover, antibiotic development is not considered an economically feasible investment in the pharmaceutical industry, because microbes develop resistance within a short period of time(3). Therefore, during the past 15 years, there has been a remarkable gap in the development of new antibiotics to address emerging resistance situations (4). The need for new therapies to cure MDR infections remains unfulfilled, with a paucity of new antibiotics in development. Antibiotic resistance is a severe threat to public health, food security, and economic stability. Early and precise understanding of resistance phenotypes is essential for infection control and proper therapeutic decision-making.

Microbial diagnostics have been revolutionized by whole-genome sequencing (WGS), which enables the rapid acquisition of comprehensive genomic data. The amount of microbiological data generated has increased due to the advancement of high-throughput sequencing technology. Since traditional techniques using biological cultures and microscopes are costly and time-consuming, machine-learning techniques have been progressively included into microbial research. However, WGS alone does not interpret the functional consequences of genomic variation. Machine learning,

a branch of artificial intelligence, possesses tools to identify patterns from genomic data and infer phenotypic traits, such as antibiotic resistance.

Our understanding of infectious diseases may be significantly impacted by the advent of big data, where the Internet and use of electronic health records (EHR) provide access to datasets that were unimaginable 20 years ago. Machine learning has been suggested as a solution to this problem, but its role in tackling AMR is yet to be defined. This review will provide a brief overview of machine learning and then examine recent developments in improving the diagnosis, treatment and prevention of bacterial AMR.

2. BACKGROUND

2.1 Antimicrobial Resistance (AMR)

Antimicrobial resistance (AMR) arises when bacteria evolve by developing mechanisms that enable them to resist antibiotics, rendering treatments less effective or ineffective. The production of β-lactamases, which degrade β-lactam antibiotics, is one of the main mechanisms of resistance. Other mechanisms include the modification of antibiotic target sites, such as through ribosomal RNA mutations that decrease drug binding; decreased intracellular accumulation of antibiotics through efflux pumps and decreased membrane permeability; and horizontal gene transfer, in which resistance genes are acquired through integrons, transposons, or plasmids(5). These drug resistance mechanisms allow bacteria that possess these mechanisms to survive, or even to actively grow, in the presence of a particular antimicrobial agent

2.2 Whole-Genome Sequencing in Microbiology

Whole-genome sequencing (WGS) provides complete data about an organism's genetic content, which can be used for the identification of recognized antimicrobial resistance genes, such as *bla_{CTX-M}* and *mecA*. It also enables the detection of new mutations that may be linked with resistance and provides valuable insights into bacterial phylogeny and transmission dynamics. The reduced cost of WGS has led to the use of this data in infectious diseases research and public health (6) When combined with machine learning (ML) models, WGS data permits comprehensive analyses that do not rely on traditional culture-based methods, thereby accelerating the speed and accuracy of resistance prediction and epidemiological investigations.

2.3 Machine Learning Overview

Machine learning (ML) models are computer programs that use input data to classify or predict outcomes based on input data. The field of machine learning encompasses a wide range of fields, such as statistics, probability theory, approximation theory, convex analysis, and algorithm complexity theory (7). When ML models are used to predict antimicrobial resistance (AMR), they are usually trained on labeled datasets that include genome sequences and the resistance characteristics that correlate to them. Machine learning follows a different approach: instead of using 'hard coding' rules as in expert systems, the dataset is analysed to infer relationships (8).

Three major categories can be used to classify machine learning techniques. Using labeled data to carry out regression or classification tasks (such as resistant vs. susceptible) is known as supervised learning. Unsupervised learning methods, such as dimensionality reduction and clustering, are used to analyze genomic data exploratorily without labels. Deep learning, a type of ML, utilizes neural network architectures—especially convolutional and recurrent neural networks— to identify intricate patterns in high-dimensional data, such as whole-genome sequences. In microbiological research, supervised learning—particularly the support vector machine (SVM)—is always utilized.

3. Methodological Framework

3.1 Data Acquisition

Antimicrobial resistance (AMR) prediction using machine learning (ML) models requires large, high-quality datasets that consists both phenotypic and genetic data. Genotypic machine learning algorithms usually need sizable databases of high-quality sequencing data from diverse isolates that are labeled with phenotypic AST data.(9). The sequencing data input for genotypic ML models is generally shotgun DNA sequences and shotgun metagenomic DNA sequences from isolates. As an alternative, transcriptional responses induced by antibiotic treatment can be used, which has the benefit of combining genotypic and phenotypic information for AMR prediction. Such information can be found in a number of publically accessible sources, such as the Comprehensive Antibiotic Resistance Database (CARD), the European Nucleotide Archive (ENA), the Pathosystems Resource Integration Center (PATRIC), and NCBI's Bio Project and

BioSample databases. Together, these sources provide access to a variety of well-annotated datasets, which facilitates the creation and training of reliable machine learning models.

3.2 Preprocessing and Feature Engineering

Preparing genomic data for machine learning (ML) models in the prediction of antimicrobial resistance (AMR) involves several crucial phases. The majority of ML models are trained using binary or continuous vectors with a fixed width. The process of converting intricate input data into this format is called "feature extraction." Segmenting the sequencing reads or contigs into k-length subsequences, or k-mers, is a popular technique for extracting characteristics from genomic shotgun sequencing data. The presence/absence or frequency of each k-mer, of which there are 4k possibilities, can then be marked in order to count and convert the k-mers inside a given sample into a vector. K-mers typically range in length from 13 to 31 nucleotides; longer k-mers may be more specific, but they also require more training data to adequately sample the greater feature space and are more likely to experience sequencing errors. From the genomic data, feature extraction aids in the derivation of informative representations, including the presence or absence of genes, single-nucleotide polymorphisms (SNPs), k-mer frequencies, amino acid substitutions, and, more recently, sequence embeddings produced by natural language processing (NLP) based models. Dimensionality reduction approaches, like principal component analysis (PCA) or statistical feature selection, are frequently used to increase computing efficiency and lower the risk of overfitting during model training because of the high dimensionality of genomic data.

4. Machine Learning Models for AMR Prediction

4.1 Traditional Machine Learning Algorithms

Data analysis in many scientific fields still heavily relies on traditional machine learning algorithms. The ensemble-based architecture of Random Forests (RF), which mixes numerous decision trees to improve accuracy and prevent overfitting, makes them especially suitable for handling sparse and high-dimensional datasets. Because Support Vector Machines (SVM) effectively find the best hyperplanes for class separation in high-dimensional spaces, they are ideally suited for binary classification problems, particularly when working with small sample sets. Gradient Boosting Machines, such as XGBoost, exhibit high predictive accuracy and the ability to provide interpretable feature importance scores. Because of their resilience, adaptability, and excellent performance in a range of problem scenarios, these algorithms are still often utilized.

4.2 Deep Learning Approaches

The analysis of genetic data for the prediction of antimicrobial resistance (AMR) has made deep learning architectures more and more significant. The ability of Convolutional Neural Networks (CNNs) to capture local spatial patterns makes them popular for use with encoded genomic sequences or k-mer-based image representations(10). Gene or nucleotide sequences are examples of sequential data that are highly suited for Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) variations, and Transformer models(11). This allows for the capture of contextual information and long-range dependencies. Autoencoders are unsupervised neural networks that develop compressed representations of high-dimensional genomic inputs in order to extract features and reduce dimensionality(12). These models' topologies improve the model's capacity to identify intricate, non-linear patterns linked to resistance traits.

4.3 Hybrid and Ensemble Models

By improving model accuracy, robustness, and generalizability, hybrid and ensemble machine learning algorithms provide notable benefits in the context of antimicrobial resistance (AMR) prediction(13). To provide a more dependable final result, ensemble techniques like stacking, bagging, and boosting combine the predictions of several basic classifiers. For instance, stacking combines different algorithms, including neural networks, support vector machines, and decision trees, into a layered architecture in which the outputs of each model are used as inputs for a meta-classifier. This method utilizes the strengths of each model, mitigating individual weaknesses and reducing the risk of overfitting.

This idea is expanded upon by hybrid models, which integrate biological knowledge specific to a certain domain straight into the machine learning pipeline(6). Data-driven predictors like k-mer profiles or sequence embeddings may be used in conjunction with curated features of known resistance determinants (e.g., presence of bla_{CTX-M}, mecA, or efflux pump genes)(14).

Both interpretability and excellent prediction performance can be attained by hybrid models, which combine complicated patterns discovered from the data with biomarkers that have been empirically validated.

5. CONCLUSION

Using whole-genome sequencing data, machine learning has become an efficient method for predicting antibiotic resistance. It offers quick, scalable, and possibly more precise substitutes compared to the conventional techniques. ML models could play a key role in clinical microbiology and public health surveillance as availability to microbial genomes increases and algorithmic transparency improves. However, widespread implementation will require close attention to model interpretability, data quality, and integration into current healthcare systems.

These models can assist detect patterns in resistance at the population level and help clinicians choose therapies that work. However, careful consideration of a few crucial issues is necessary for effective incorporation into healthcare settings. First and foremost, model interpretability is essential—especially in high-stakes situations, healthcare providers need to be able to trust and comprehend forecasts. Second, to prevent biases and enhance generalizability across pathogens and environments, training data quality and variety must be guaranteed. Finally, real-world implementation requires smooth interface with current electronic health record systems and laboratory operations. ML-driven antibiotic resistance prediction has the potential to completely transform antimicrobial stewardship and infectious disease management when these considerations are carefully taken into account.

REFERENCES

1.Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health. 2015 Oct;109(7):309–18.
2.Basak S, Singh P, Rajurkar M. Multidrug Resistant and Extensively Drug-Resistant Bacteria: A Study. J Pathog. 2016; 2016:4065603.
3.Wraith DC. The future of immunotherapy: a 20-year perspective. Frontiers in immunology. 2017 Nov 28; 8:1668.
4.Luepke KH, Suda KJ, Boucher H, Russo RL, Bonney MW, Hunt TD, et al. Past, Present, and Future of Antibacterial Economics: Increasing Bacterial Resistance, Limited Antibiotic Pipeline, and Societal Implications. Pharmacother J Hum Pharmacol Drug Ther. 2017;37(1):71–84.
5.Kapoor G, Saigal S, Elongavan A. Action and resistance mechanisms of antibiotics: A guide for clinicians. J Anaesthesiol Clin Pharmacol. 2017 Sep;33(3):300.
6.Liu YY, Chen CC. Computational Analysis of the Molecular Mechanism of RamR Mutations Contributing to Antimicrobial Resistance in Salmonella enterica. Sci Rep. 2017 Oct 17;7(1):13418.
7.Qu K, Han K, Wu S, Wang G, Wei L. Identification of DNA-binding proteins using mixed feature representation methods. Molecules. 2017;22(10):1602.
8.Macesic N, Polubriaginof F, Tatonetti NP. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. Curr Opin Infect Dis. 2017;30(6):511–7.
9.Shakoor N, Lee S, Mockler TC. High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. Curr Opin Plant Biol. 2017 Aug 1;38:184–92.
10.Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics. 2017 Dec 1;18(13):478.
11.Jurtz VI, Johansen AR, Nielsen M, Almagro Armenteros JJ, Nielsen H, Sønderby CK, et al. An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics. 2017 Nov 15;33(22):3685–90.
12.Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017 Nov 17;18(9):845.
13.Villa AE, Masulli P, Rivero AJ, editors. Artificial neural networks and machine learning–icann 2016: 25th international conference on artificial neural networks, barcelona, spain, september 6-9, 2016, proceedings, part ii. Springer; 2016 Aug 26.
14.Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016 Sep 16;7(1):12797.

***Corresponding Author:**
Shiji Thomas
Email: shijiboby1@gmail.com