Whole Genome Sequence of *Mycobacterium tuberculosis*

Subjects: Others

Contributor: Ricardo Perea-Jacobo, Guillermo René Paredes-Gutiérrez, Miguel Ángel Guerrero-Chevannier, Dora-Luz Flores, Raquel Muñiz-Salazar

Tuberculosis (TB) remains one of the most significant global health problems, posing a significant challenge to public health systems worldwide. However, diagnosing drug-resistant tuberculosis (DR-TB) has become increasingly challenging due to the rising number of multidrug-resistant (MDR-TB) cases, despite the development of new TB diagnostic tools. Even the World Health Organization-recommended methods such as Xpert MTB/XDR or Truenat are unable to detect all the *Mycobacterium tuberculosis* genome mutations associated with drug resistance. While Whole Genome Sequencing offers a more precise DR profile, the lack of user-friendly bioinformatics analysis applications hinders its widespread use.

Keywords: NGS ; MDR-TB ; SNPs ; Mycobacterium tuberculosis ; ML ; IA

1. Introduction

Tuberculosis (TB) is a treatable and preventable infectious disease caused by *Mycobacterium tuberculosis*. Despite its curability, this ancient illness remains a major global health concern due to its high incidence and mortality rates worldwide. As of 2021, an estimated 10.6 million people had developed TB, and 1.6 million had lost their lives to the disease. The regime of drug-susceptible and drug-resistant *M. tuberculosis* isolates demands a minimum of three to four antibiotics (rifampicin, isoniazid, ethambutol, and pyrazinamide) in combination, leading to complex patterns of drug susceptibility and resistance. The World Health Organization (WHO) estimates that globally in 2020, 71% of people diagnosed with bacteriologically confirmed pulmonary TB were tested for rifampicin (RIF) resistance, up from 61% in 2019 and 50% in 2018 [1]. In 2019, about 0.5 million DR-TB cases were reported worldwide, of which 78% were MDR-TB. MDR-TB is defined as resistant to at least RIF and isoniazid (INH), the most effective first-line antituberculosis drugs. It is estimated that one in four deaths caused by antimicrobial resistance is due to rifampicin-resistant TB. Treating drug-resistant TB is more complex than treating drug-susceptible TB.

A Drug Susceptibility Test (DST) is essential for proper antituberculosis treatment, avoiding complications, and significantly reducing the treatment period. Microbiological culture is the gold standard to evaluate the DR; however, it requires two to six weeks to obtain results and must be performed at a Biosafety Level 3. Consequently, most patients start antituberculosis treatment without DST information. To effectively treat DR-TB, a rapid and specific drug sensitivity test (DST) is necessary for selecting the appropriate TB treatment. This test helps identify the most effective treatment for the patient. [2]. Today, the World Health Organization (WHO) recommended molecular tests, such as Xpert® MTB/RIF (Cepheid, Sunnyvale, CA, USA), Truenat® MTB, and the MTB Plus system (Molbio Diagnostics, Goa, India), to use as initial tests for the diagnosis of TB and rifampicin-resistant TB. However, molecular DSTs cannot detect resistance profiles when mutations occur outside the target genetic region. On the other hand, Whole Genome Sequencing (WGS) is a technique that can compensate for this weakness.

WGS allows the identification of the DR-TB profile-identified known mutations and can be used to propose new mutations that confer resistance when compared with a diverse amount of DST; it is accurate and provides a rich set of additional information for further analysis of new TB antibiotic development. Therefore, it is essential to explore the value of modern statistical approaches, such as Machine Learning (ML), which can analyze vast amounts of characteristics in large databases such as genomics and perform high-precision resistance classification. ML models can be utilized to analyze the whole genome sequencing of *M. tuberculosis* strains, helping predict resistance profiles and reducing the time delay in starting appropriate treatment. The increased availability of new artificial intelligence technologies, in particular ML and Deep Learning (DL), allows an approach to complex clinical databases, radiological images, and whole genomes to perform rapid detection and classification of the disease, support clinical decision-making, and contribute to quick and timely diagnosis. However, it is still being determined what model is recommended for these biological data or if the

metrics are reported similarly and consistently. Furthermore, the several ways of grouping the resistance analyses by drug or treatment regimen make it challenging to compare them. There is no standardized method for analyzing sequence data to ensure a good result for resistance prediction. These methods can include, for example, analyzing mutations already known to confer drug resistance, analyzing the entire genome (considering only mutations or the whole genome compared against a reference), and analyzing specific genes. To better understand the research conducted in this area and identify any gaps in knowledge, a scoping review was conducted.

2. AI Models for Predicting Drug-Resistant Tuberculosis

After analyzing various models, researchers found that most of them use clustering for training and Random Forest for feature selection [3][4][5][6][7][8][9][10][11][12][13][14][15][16][17][18][19][20][21][22]. However, recent research shows that an increasing number of models are using neural networks [3][5][6][9][17][19], which have proven to be more effective. In fact, some studies have even implemented deep learning convolutional models [7][8][23][24], which show promising results. It's worth noting that as the models become more complex, they become harder to standardize. This requires more abstract representations of biological features used for training. Clustering models use point mutations like SNPs, whereas deep learning models use representations based on scores [24] or feature ordering [8]. This may result in difficulty correlating relevant features during the classification process.

It is revealed that there is a lack of focus on developing effective metrics to evaluate studies related to healthcare technology implementation. Researchers also found a limited number of studies on the practical application of these technologies in clinical settings. Additionally, there is insufficient consideration given to the inclusion and empowerment of healthcare professionals in the education and use of these technologies. Researchers did not come across any studies that address stakeholder relationships or the use of evaluative and iterative strategies to introduce and promote machine learning (ML) technologies in clinical practice. It was observed that the studies researchers reviewed invested more effort in improving analysis matrices than in performing standardized preprocessing that would enable comparison.

The type of characteristics used to train the machine learning models varied for each study, from binary representation for the presence and absence of mutations in resistance genes or genes complete to physicochemical characteristics of the amino acids resulting from the base arrangement in each isolate [10]. Few studies compared the performance of the models for each characteristic used; this tells researchers about the need to generate a methodology that allows for a systematic comparison of the different types of characteristics used in training machine learning models. Systematically comparing each feature's strengths and limitations and identifying which are most effective for specific machine-learning tasks will be easy for researchers. Additionally, it would allow researchers to better optimize the performance of their models by selecting the most appropriate characteristics.

Aytan-Aktug et al. (2020) [3] compared different feature representations, including binaries, scores, and combinations. They found that these representations slightly improved the prediction results, but the number of features differed significantly. For example, the amino acid representation had over 260,000 features, while the binary representation had only 6736. This stark contrast in the number of features raises the possibility of performing dimensionality reduction on the amino acid representation, which could further improve prediction results. Reducing the number of features could eliminate redundancies and noise, allowing for a more efficient and accurate representation of the data.

There is a wide range of approaches to analyzing mutations in these studies. Some focus on known resistance genes previously identified in the literature [3][7], while others explore the entire genome or specific types of single nucleotide polymorphisms (SNPs), deletions, or insertions [3][5][6][8][10][11][12][13][14][15][16][17][18][25]. Some studies look at mutations in a genome-wide context to discover new genes that may be related to resistance [19][20]. However, despite the various approaches taken, there needs to be more research that specifically examines the role of epistasis (the interaction between genes) in developing drug resistance, suggesting that there is a need for more studies that focus on understanding the complex genetic mechanisms that can contribute to the emergence of DR [9][21][24][26].

Most studies focus on model accuracy as their main performance metric. However, a predictor with a specificity and sensitivity of at least 95% [27] is generally required for clinical applications. However, this represents a significant challenge since most clinical data sets must be balanced between sensitivity and drug-resistant observations. Nevertheless, there is an imbalance, particularly noticeable for first-line TB drugs like isoniazid and rifampicin compared to other first- and second-line drugs. The model needs more examples in unbalanced sets to identify resistant cases, resulting in low sensitivity. While specificity can be high due to the model being presented with many sensitive cases, high accuracy in predicting sensitive cases does not necessarily imply good performance. This situation highlights the need for

more robust modeling strategies to improve the specificity and sensitivity of predictive models for clinical implementation. Reporting these metrics in more standardized ways in reported models is also crucial.

The research indicates the necessity for greater emphasis on creating measurable standards to assess studies and the limited attention given to implementing these studies in clinical settings. Additionally, technology inclusion and education for healthcare professionals are not being considered. Researchers could not find any studies that address the importance of building relationships and using evaluative and iterative strategies while introducing and promoting machine learning technologies in clinical practice. Furthermore, the studies in the research tend to prioritize improving analysis matrices rather than standardized preprocessing for comparison purposes.

TB continues to pose a significant global health burden, particularly with the emergence of drug-resistant strains. The COVID-19 pandemic has further underscored the urgent need to intensify efforts toward achieving the End TB strategy. However, diagnosing drug-resistant tuberculosis (DR-TB) has become increasingly challenging, despite the development of new diagnostic tools. The rising number of multidrug-resistant (MDR-TB) cases necessitates innovative approaches for accurate and timely detection of DR-TB.

In recent years, the integration of AI models has shown great promise in predicting DR-TB profiles with enhanced precision. These models leverage advanced algorithms to analyze complex genomic data, enabling the identification of key genetic mutations associated with drug resistance in *M. tuberculosis*. Although traditional diagnostic methods like Xpert MTB/XDR or Truenat, recommended by the WHO, have improved diagnostics, they may not detect all genome mutations associated with drug resistance.

It is crucial to cover all aspects of the field to comprehensively understand the current landscape of AI models for predicting DR-TB profiles. Which includes exploring the diverse range of AI techniques employed, the datasets utilized, and the performance metrics used to evaluate their effectiveness. By examining experiments or studies of impact in sufficient depth, reserachers aim to analyze the strengths and limitations of existing approaches comprehensively. By delving into the experimental setups, data sources, and evaluation methodologies employed in these studies, valuable insights can be gained, allowing for a deeper understanding of the advancements and challenges in the field. Additionally, it suggests new avenues for future research to address the current limitations and drive further progress in DR-TB prediction. By identifying gaps in knowledge and proposing novel research directions, researchers can pave the way for innovative solutions. These new avenues include the following:

(a)Integrating multiple AI techniques or combining AI with other diagnostic modalities, such as imaging or transcriptomics, to improve prediction accuracy;

(b)Developing more accurate and robust AI models that can handle complex and noisy data from different sources and settings;

(c)Exploring the use of AI for predicting resistance to other drugs besides rifampicin, isoniazid, pyrazinamide, and fluoroquinolones;

(d)Integrating AI with other technologies such as molecular diagnostics, biosensors, or nanotechnology for rapid and point-of-care detection of DR-TB;

(e)Evaluating the cost-effectiveness, feasibility, and ethical implications of implementing AI for DR-TB diagnosis in low- and middle-income countries.

Moreover, efforts should be made to expand the diversity of training datasets, encompassing various geographic regions and genetic variants of *M. tuberculosis*, to enhance the generalizability of AI models. Furthermore, developing user-friendly bioinformatics analysis applications can simplify the interpretation of WGS data and facilitate widespread adoption of this technology.

References

1. World Health Organization Global Tuberculosis Report 2022; World Health Organization: Geneva, Switzerland, 2022.
2. Jang, J.G.; Chung, J.H. Diagnosis and Treatment of Multidrug-Resistant Tuberculosis. J. Yeungnam Med. Sci. 2020, 37, 277–285.

3. Aytan-Aktug, D.; Clausen, P.T.L.C.; Bortolaia, V.; Aarestrup, F.M.; Lund, O. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. mSystems 2020, 5, e00774-19.
4. Aytan-Aktug, D.; Nguyen, M.; Clausen, P.T.L.C.; Stevens, R.L.; Aarestrup, F.M.; Lund, O.; Davis, J.J. Predicting Antimicrobial Resistance Using Partial Genome Alignments. mSystems 2021, 6, e0018521.
5. Chen, M.L.; Doddi, A.; Royer, J.; Freschi, L.; Schito, M.; Ezewudo, M.; Kohane, I.S.; Beam, A.; Farhat, M. Beyond Multidrug Resistance: Leveraging Rare Variants with Machine and Statistical Learning Models in Mycobacterium tuberculosis Resistance Prediction. EBioMedicine 2019, 43, 356–369.
6. Gröschel, M.I.; Owens, M.; Freschi, L.; Vargas, R.; Marin, M.G.; Phelan, J.; Iqbal, Z.; Dixit, A.; Farhat, M.R. GenTB: A User-Friendly Genome-Based Predictor for Tuberculosis Resistance Powered by Machine Learning. Genome Med. 2021, 13, 138.
7. Kuang, X.; Wang, F.; Hernandez, K.M.; Zhang, Z.; Grossman, R.L. Accurate and Rapid Prediction of Tuberculosis Drug Resistance from Genome Sequence Data Using Traditional Machine Learning Algorithms and CNN. Sci. Rep. 2022, 12, 2427.
8. Zhang, A.; Teng, L.; Alterovitz, G. An Explainable Machine Learning Platform for Pyrazinamide Resistance Prediction and Genetic Feature Identification of Mycobacterium tuberculosis. J. Am. Med. Inf. Assoc. 2021, 28, 533–540.
9. Jamal, S.; Khubaib, M.; Gangwar, R.; Grover, S.; Grover, A.; Hasnain, S.E. Artificial Intelligence and Machine Learning Based Prediction of Resistant and Susceptible Mutations in Mycobacterium tuberculosis. Sci. Rep. 2020, 10, 5487.
10. Chowdhury, A.S.; Khaledian, E.; Broschat, S.L. Capreomycin Resistance Prediction in Two Species of Mycobacterium Using a Stacked Ensemble Method. J. Appl. Microbiol. 2019, 127, 1656–1664.
11. Deelder, W.; Napier, G.; Campino, S.; Palla, L.; Phelan, J.; Clark, T.G. A Modified Decision Tree Approach to Improve the Prediction and Mutation Discovery for Drug Resistance in Mycobacterium tuberculosis. BMC Genom. 2022, 23, 46.
12. Viveiros, M.; Coll, F.; Deelder, W.; Kouchaki, S.; Yang, Y.; Lachapelle, A.; Walker, T.M.; Walker, A.S.; Consortium, C.; Peto, T.E.A.; et al. Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking. Front. Microbiol. 2020, 11, 667.
13. Deelder, W.; Christakoudi, S.; Phelan, J.; Benavente, E.D.; Campino, S.; McNerney, R.; Palla, L.; Clark, T.G. Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance from Whole Genome Sequencing Data. Front. Genet. 2019, 10, 922.
14. Libiseller-Egger, J.; Phelan, J.; Campino, S.; Mohareb, F.; Clark, T.G. Robust Detection of Point Mutations Involved in Multidrug-Resistant Mycobacterium tuberculosis in the Presence of Co-Occurrent Resistance Markers. PLoS Comput. Biol. 2020, 16, e1008518.
15. Nguyen, M.; Olson, R.; Shukla, M.; VanOeffelen, M.; Davis, J.J. Predicting Antimicrobial Resistance Using Conserved Genes. PLoS Comput. Biol. 2020, 16, e1008319.
16. Sergeev, R.S.; Kavaliou, I.S.; Sataneuski, U.V.; Gabrielian, A.; Rosenthal, A.; Tartakovsky, M.; Tuzikov, A.V. Genome-Wide Analysis of MDR and XDR Tuberculosis from Belarus: Machine-Learning Approach. IEEE/ACM Trans. Comput. Biol. Bioinform. 2019, 16, 1398–1408.
17. Yang, Y.; Walker, T.M.; Walker, A.S.; Wilson, D.J.; Peto, T.E.A.; Crook, D.W.; Shamout, F.; Zhu, T.; Clifton, D.A.; Arandjelovic, I.; et al. DeepAMR for Predicting Co-Occurrent Resistance of Mycobacterium tuberculosis. Bioinformatics 2019, 35, 3240–3249.
18. Kouchaki, S.; Yang, Y.Y.; Walker, T.M.; Walker, A.S.; Wilson, D.J.; Peto, T.E.A.; Crook, D.W.; Clifton, D.A.; Hoosdally, S.J.; Gibertoni Cruz, A.L.; et al. Application of Machine Learning Techniques to Tuberculosis Drug Resistance Analysis. Bioinformatics 2019, 35, 2276.
19. Müller, S.J.; Meraba, R.L.; Dlamini, G.S.; Mapiye, D.S. First-Line Drug Resistance Profiling of Mycobacterium tuberculosis: A Machine Learning Approach. AMIA Annu. Symp. Proc. 2021, 2021, 891–899.
20. Kavvas, E.S.; Catoiu, E.; Mih, N.; Yurkovich, J.T.; Seif, Y.; Dillon, N.; Heckmann, D.; Anand, A.; Yang, L.; Nizet, V.; et al. Machine Learning and Structural Analysis of Mycobacterium tuberculosis Pan-Genome Identifies Genetic Signatures of Antibiotic Resistance. Nat. Commun. 2018, 9, 4306.
21. Li, X.; Lin, J.; Hu, Y.; Zhou, J. PARMAP: A Pan-Genome-Based Computational Framework for Predicting Antimicrobial Resistance. Front. Microbiol. 2020, 11, 578795.
22. Zabeti, H.; Dexter, N.; Safari, A.H.; Sedaghat, N.; Libbrecht, M.; Chindelevitch, L. INGOT-DR: An interpretable classifier for predicting drug resistance in M. tuberculosis. Algorithms Mol. Biol. 2021, 16, 17.
23. Green, A.G.; Yoon, C.H.; Chen, M.L.; Freschi, L.; Gröschel, M.I.; Kohane, I.; Beam, A.; Farhat, M. A Convolutional Neural Network Highlights Mutations Relevant to Antimicrobial Resistance in Mycobacterium tuberculosis. Nat.

Commun. 2022, 13, 3817.

24. Safari, A.H.; Sedaghat, N.; Zabeti, H.; Forna, A.; Chindelevitch, L.; Libbrecht, M. Predicting Drug Resistance in M. tuberculosis Using a Long-Term Recurrent Convolutional Network. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2021, Gainesville, FL, USA, 1–4 August; 2021; Volume 1.
25. Jiang, Z.; Lu, Y.; Liu, Z.; Wu, W.; Xu, X.; Dinnyés, A.; Yu, Z.; Chen, L.; Sun, Q. Drug Resistance Prediction and Resistance Genes Identification in Mycobacterium tuberculosis Based on a Hierarchical Attentive Neural Network Utilizing Genome-Wide Variants. Brief. Bioinform. 2022, 23, bbac041.
26. Kavvas, E.S.; Yang, L.; Monk, J.M.; Heckmann, D.; Palsson, B.O. A Biochemically-Interpretable Machine Learning Classifier for Microbial GWAS. Nat. Commun. 2020, 11, 2580.
27. Su, M.; Satola, S.W.; Read, T.D. Genome-Based Prediction of Bacterial Antibiotic Resistance. J. Clin. Microbiol. 2019, 57, e01405-18.