Project title: Applying machine learning methods to understand the distribution of rifampicin resistance mutations

across species

Course code: BIOX7011

Your name: Yu Sun

Student number: 45105764

Supervisor: Jan Engelstadter

1. Background

Antimicrobial Resistance (AMR) has become one of the most pressing threats to global public health in the 21st century. The World Health Organization estimates that by 2050, more than 10 million people will die annually from AMR-related infections [1].AMR not only makes it more difficult to treat infections, but also significantly raises healthcare costs and length of hospitalization, placing greater pressure on public health systems in low-income countries [2]. Among the many drug-resistant pathogens, drug-resistant tuberculosis (DR-TB) is of particular concern. approximately 450,000 people worldwide will have rifampicin-resistant tuberculosis (RR-TB) in 2022, with the majority of cases also showing resistance to isoniazid, thus constituting multidrug-resistant tuberculosis (MDR-TB) [3]. .

This project focuses on rifampicin, the current first-line antibiotic for the treatment of TB, whose mechanism is to inhibit mRNA synthesis by binding to the β-subunit of bacterial RNA polymerase (encoded by the rpoB gene), making it possible to block bacterial protein expression and proliferation [4] [5]. However, the widespread use of the drug has also led to the creation of a large number of resistance mutations. The most typical resistance mechanism is due to point mutations in the Rifampicin Resistance Determining Region (RRDR) of the rpoB gene, which encodes the β-subunit [6].

According to the literature ‘‘Molecular basis of rifampin resistance in Mycobacterium tuberculosis’’, the RIF resistance demonstrated in Mycobacterium tuberculosis is characterized by the following mutations: S531L, S531L, S531L, S531L, S531L, S531L, S531L and S531L. Among the mutations, S531L, H526Y, and D516V are the most common, and generally possess a highly resistant phenotype [7].

These mutation sites are not only of molecular diagnostic significance, but have also been incorporated into commercial diagnostic tools (e.g., GeneXpert MTB/RIF) as the primary basis for detecting rifampicin resistance [8]. However, with the wide application of sequencing technology and drug resistance monitoring, more and more studies have found that some clinical strains exhibit rifampicin resistance without typical RRDR mutations, suggesting the possibility of atypical or ‘‘controversial’’ mutations [9].

In addition, rifampicin resistance is not limited to Mycobacterium tuberculosis. In recent years, a large number of studies have focused on the diversity of rpoB mutations in non-tuberculous mycobacteria (NTM), such as Mycobacterium abscessus, M. avium, and M. kansasii [10]. More alarmingly, rifampicin resistance due to rpoB mutations has also been reported in some common bacteria such as Staphylococcus aureus and Escherichia coli [11] [12]. Even some viruses such as poliovirus have demonstrated selective pressure for resistance associated with rifampicin exposure in in vitro experiments [13].

Current analyses have focused on M. tuberculosis strains, and resistance mutations in other species are often missed or regarded as rare events. In addition, existing literature databases and mutation monitoring systems (e.g., TB-Profiler, WHO catalog) mostly focus on high-frequency RRDR mutations, and there is a lack of systematic collation of cross-species common mutations, rare site variants, and mechanisms of resistance beyond rpoB [14] [15]. As atypical mutations are often not recognized by current mainstream molecular testing techniques, they may lead to ‘‘false-negative’’ results, thus delaying the treatment of drug-resistant TB.

2. Aims and Objectives

Aim

This study aims to use machine learning and structural bioinformatics methods to systematically integrate and analyze the distribution characteristics and functional mechanisms of rifampicin resistance mutations in multiple bacterial species, focusing on identifying atypical mutations that have been neglected by existing studies and their potential similar cross-species resistance patterns, thereby improving the prediction ability of resistance mutations in species that have lacked research.

Objectives

Objective 1: Build a high-quality cross-species mutation literature and annotation database

Build a literature dataset with "rifampicin resistance mutation" as the keyword based on Web of Science;

Use the active learning platform ASReview to assist in completing multiple rounds of manual annotation and establish an accurate and reliable training set;

Use different language learning models to classify unannotated literature, and manually review and annotate different "contradictory" literature;

Extract mutation-related information and build a mutation annotation table containing nucleotide and amino acid sites, mutation types, species sources, literature evidence and other dimensions.

Objective 2: Mutation feature mining and pattern recognition under unsupervised learning

Encode the feature vector of the sorted mutation table (including AA substitution type, position, species frequency, etc.);

Use principal component analysis (PCA) for dimensionality compression;

Use the nonlinear embedding method UMAP to enhance the visualization ability of mutation space;

Use the HDBSCAN clustering method to identify potential mutation clustering units and explore whether they are related to attributes such as species, habitat, and transmission ability.

Objective 3: Build a mutation classification model to improve the ability to predict drug-resistant mutations

Establish labels based on whether existing mutations are defined as "drug-resistant mutations" in the literature;

Use a variety of supervised learning methods (such as logistic regression, random forest, SVM, small neural network) to build classifiers;

Adjust model thresholds to avoid false negatives and identify high-risk mutations that may be overlooked;

Possibly try to introduce protein structure or sequence features (such as RRDR regions, conserved sites) to enhance the model's explanatory power.

Objective 4: Focus on warning of drug-resistant mutations that may exist in species that lack research

Focus on non-model species that are scarce in literature research (such as M. abscessus, M. fortuitum, etc.);

Use transfer learning methods to generalize known species models to target species and predict their potential drug-resistant mutations;

Incorporate timeline features such as the time of first report of mutations and frequency change trends to identify "warning mutations" that may be in the early evolutionary stage or at the forefront of transmission.

3. Literature Review

Summary of previous study on RIF resistance

By reading the previous literature, the authors could identify Mycobacterium tuberculosis (M. tuberculosis) as the model species where the mechanisms of rifampicin resistance have been most intensively studied. Classical studies have shown that resistance mainly stems from mutations in the β-subunit of RNA polymerase (encoded by the rpoB gene), especially in the so-called Rifampicin Resistance-Determining Region, where, for example, S531L and H526Y are considered to be the most common resistance hotspot mutations [15]. Specifically, the RNA polymerase β subunit encoded by the rpoB gene is a target of rifampicin, and its three-dimensional structure has been resolved by several research teams. Resistance mutations are mostly concentrated in the binding cleft region of the β subunit and rifampicin, which affects the drug binding ability but preserves the enzyme activity as much as possible. mutations such as S531L and H526Y are thought to reduce the binding affinity of rifampicin by altering the local amino acid charge state or spatial conformation without directly disrupting the RNA synthesis function [4]. Atypical mutations, on the other hand, may be located at structural edges or regulatory domains (e.g., linker domains), affecting enzyme function or drug

rejection effects through indirect mechanisms. And while mutations enhance drug resistance, they are often accompanied by a certain cost of adaptation [16].

Although typical rpoB mutations in the RRDR region, such as S531L, H526Y, and D516V, are widely used for molecular diagnosis and drug resistance prediction, recent studies have shown that a portion of clinical isolates exhibit significant rifampicin resistance in the absence of these mutations. These ‘‘atypical’’ or ‘‘controversial’’ mutations are located at the edge or outside of the RRDR, which may lead to the problem of ‘‘false-negative’’ resistance detection [9] [15]. For example, Shea et al. (2021) reported a class of rpoB non-RRDR mutations that result in low levels of resistance that are not easily recognized by existing tools, but may have a significant impact on treatment strategies. As such mutations are more common in non-tuberculous mycobacteria and some gram-positive bacteria, ignoring their presence may underestimate the overall risk of resistance transmission.

In addition, other non-tuberculous mycobacteria (NTM) and Gram-positive bacteria such as Staphylococcus aureus have progressively shown resistance to rifampicin, but their mutational profiles, evolutionary patterns, and physiological costs are often different from those of M. tuberculosis. For example, it was found that in vitro-induced production of S. aureus RIF resistance mutations (e.g., H481Y, S464P) may be accompanied by features of biofilm enhancement, reduced virulence, and decreased growth rate, suggesting that rpoB mutations are species-specific in their effects on bacterial phenotypes [17].

In Enterococcus faecium, it has also been demonstrated that rpoB mutations can lead to enhanced drug resistance, but the cost of their adaptation to survival may vary depending on the background genotype [18]. In addition, the role of epistasis (phenotypic interactions) in modulating the effects of mutations in different species backgrounds has received increasing attention, suggesting that mutational effects do not occur in isolation but need to be considered in a genome-wide context [15]

On the other hand, the mechanisms of resistance to rifampicin are not identical in different bacterial species. For example, in M. tuberculosis, the S531L mutation occurs with a frequency of more than 50% in resistant strains, whereas in NTM such as M. avium or M. kansasii, the rpoB mutation is more dispersed and has no fixed hotspot. vogwill et al. (2016) pointed out that certain mutations (e.g., D516G) exhibit different adaptive costs, suggesting different selection pressures in different host

environments. Meanwhile, Wang et al. (2019) showed that in Staphylococcus aureus, rpoB mutations (e.g., H481Y) not only confer resistance, but may also be associated with biofilm enhancement and virulence modulation, highlighting the dual effects of inter-species differences on mutation frequency and function.

Moreover, even within the same species, the mutation frequency of isolates from different regions may be jointly influenced by factors such as geographically prevalent strains, history of antibiotic use and clinical transmission patterns [19]. Therefore, comparison of resistance mutations across species and geographic regions can help to identify potential ‘‘common resistance mechanisms’’ and discover novel variants associated with transmission or evolution.

In summary, current studies on rifampicin resistance have the following limitations:

Severe species bias: most studies have focused on M. tuberculosis, and there is a scarcity of studies on other pathogens such as NTM, staphylococci, enterococci, and E. coli;

Incomplete data: non-classical mutations are poorly reported, especially low-frequency mutations with rare phenotypes are often overlooked, and data on this class of mutations are generally clinical isolates and lack a confirmed relationship with drug-resistant phenotypes;

Lack of systematic comparisons: there is a lack of a unified analytical framework for the similarities and differences, evolutionary pathways, and phenotypic impact of rpoB mutations among different species;

Lack of integration and open access to literature datasets for reproducible analysis.

4. Machine Learning and other Method Description

This study employs a variety of natural language processing (NLP) and supervised learning methods to identify and analyze high-quality literature related to rifampicin resistance mutations. These methods are based on deep learning language model embedding combined with classification and clustering algorithms to automate the screening of literature and modeling of mutation data. The following is a brief overview of the main machine learning methods involved in this study:

1. Sentence-BERT (all-MiniLM-L6-v2)

Sentence-BERT (SBERT) is an optimized BERT architecture for efficiently generating semantic embeddings at the sentence level [20]. The authors use the all-MiniLM-L6-v2 model, which is characterized by high computational efficiency, low resource consumption, and suitability for large-scale text processing. The model is capable of converting titles and abstracts into dense semantic vectors (embedding) and is used to build traditional classifiers to distinguish between relevant and irrelevant documents.

2. SciBERT (allenai/scibert_scivocab_uncased)

SciBERT is a pre-trained language model published by the Allen Institute for AI, specifically trained on biomedical and scientific literature for analyzing complex terms and structures [21]. The authors fine-tuned (fine-tuning) the model for direct use in a binary categorization task to predict whether or not literature is associated with rifampicin resistance mutations. The model was found to show better generalization ability in identifying low-frequency mutations, non-pattern species literature.

3. PCA with UMAP + HDBSCAN

To further understand the distributional structure of mutation data, the authors plan to use the following unsupervised learning methods:

Principal Component Analysis (PCA): used to reduce the dimensionality of mutation features and enhance visualization and subsequent clustering;

UMAP (Uniform Manifold Approximation and Projection): to preserve the local topology, suitable for revealing the embedding space of complex data;

HDBSCAN (Hierarchical Density-Based Spatial Clustering): a density-aware clustering algorithm that recognizes irregular mutation clusters and outliers.

4. Classification Models and Mutation Prediction

The authors will build supervised classifiers (e.g., logistic regression, random forests, lightweight neural networks, etc.) to predict whether a mutation is a ‘‘drug-resistant mutation’’. Input features will include information such as mutation location, amino acid substitutions, species, sequence context, and so on.

In addition, if a mutation co-occurrence network can be constructed, the study will attempt to apply graphical neural networks (GNNs) to model mutation interactions and propagation patterns.

5. Research innovation and differences

Although a large number of studies have elucidated the drug resistance mechanism of rpoB mutations in Mycobacterium tuberculosis, a systematic cross-species comparative perspective is lacking.Most studies have used targeted clinical strain validation or experimental induction approaches, focusing on the correspondence between known mutations and phenotypes, which makes it difficult to cover rare mutations or non-mainstream species that have been reported sporadically in the extensive literature. Meanwhile, mainstream databases (e.g., TB-Profiler, WHO mutation catalogue) also suffer from species bias, neglecting the systematic collation of rpoB mutations in other important pathogens such as E. coli and staphylococci.

This study proposes the following innovations based on existing work:

1. a large and systematic integration of multi-species mutation information from a literature perspective using machine learning software and models, and generalization of commonalities in the target literature by extracting keywords and specific linguistic contexts;

2. Introducing pre-trained language models (comparing SciBERT and SentenceBert) to improve the semantic resolution of literature screening, and then manually reviewing and focusing on ‘‘confusing’’ literature under the divergence of the two machine models;

3. exploring automatic learning methods for mutation co-occurrence relationships, functional modules and evolutionary trends, such as graphical neural networks, time series clustering and other machine learning tools;

4. emphasizing the importance of rare mutations and cross-species co-occurrence mechanisms, which will likely provide new perspectives for future molecular diagnostics, database improvement and mechanistic studies.

Project Plan

This research plan, based mainly on the author's own and SUPERVISOR's guidance, can be roughly divided into 4 main steps, of which the first two are largely completed and the last two are still in the planning stage, thus demonstrating an idealized arrangement:

Step 1 Use Wos and ASReview to find papers reporting RIF resistance mutations and organize them into datasets

In this study, a large literature dataset of 3473 records was first constructed from Web of Science search results with the search terms "((rifampicin OR rifampin) AND (resistance OR resistant) AND (mutation OR polymorphism OR variant))". To efficiently identify key literature associated with rifampicin resistance mutations, the active learning tool ASReview was used to support efficient annotation, using manually annotated literature on ASReview as the training set, followed by two language models Sentence-BERT (all-MiniLM-L6-v2) and SciBERT (allenai/ scibert_scivocab_uncased) to compare the remaining results of the predictions, and then re-establish a new dataset for the contradictory results, and then repeat the use of ASReview to annotate the first one hundred or so of them, and then repeat the other steps mentioned above, and finally obtain 113 documents with contradictory predictions of the two models, which can be regarded as the screened-out These 113 documents can be regarded as the "highly confusing" data, and the authors believe that this data set can significantly strengthen the language model if it is used as a new training object, so they manually labeled all of them as relevant and irrelevant.

After that, the 113 "highly confusing" documents and all the manually labeled ones were merged into a new dataset, which was retrained and retested with Stratified K-Fold CV, and the parameters, especially the break-in values, were adjusted to achieve higher F1 values to build the final training model. model (using SciBERT, which performs relatively better on the author task) and then make a final prediction on all the 3139 documents other than the manually labeled ones.

Step 2 Compare mentor datasets and add new unincluded Rif resistance mutation data

Through the above process, a training set of 220 high-confidence papers was constructed, and the authors then used these 220 papers to compare with the literature dataset summarized in other ways provided by the mentor, and found up to 170 "new papers" that were not included in the mentor's dataset. The authors then extracted and standardized the information about rifampicin resistance

mutations in these 170 papers published within the last 10 years, i.e., 2015 and later, and initially developed a multidimensional mutation annotation table of more than 80 rows.

Step 3 Feature dimensionality reduction and visualization analysis and unsupervised clustering modeling

Use principal component analysis (PCA) to reduce the dimension of mutation features ;Then apply UMAP (Uniform Manifold Approximation and Projection) to further capture the nonlinear feature embedding space to enhance clustering sensitivity and visualization interpretation.

Apply HDBSCAN (Hierarchical Density-Based Spatial Clustering) to identify mutation clusters with biological consistency.Try to analyze the association between clustering results and species, environmental background and transmission ability, and explore potential "cross-species common mutations".

Step 4 Predict Rif resistance for target species that lack research through supervised classification machine learning

Build classification models (such as logistic regression, random forest, shallow neural network) to predict whether mutations have known drug resistance.Then author can focus on identifying potential "false negative" mutations that have not been confirmed in the current database.

Explore the incorporation of structural features into the classifier to improve the model's explanatory power and generalization performance.

Reference

[1] J. O'Neill, *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*, Review on Antimicrobial Resistance, 2016.

[2] World Health Organization, *Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report*, Geneva: WHO, 2023.

[3] World Health Organization, *Global Tuberculosis Report 2022*, Geneva: WHO, 2022.

[4] E. A. Campbell, O. Muzzin, M. Chlenov, et al., "Structural mechanism for rifampicin inhibition of bacterial RNA polymerase," *Cell*, vol. 104, no. 6, pp. 901–912, 2001.

[5] G. Hartmann, J. A. Honikel, and H. Knusel, "Mode of action of rifampin on the RNA polymerase of *Escherichia coli*," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 57, no. 6, pp. 1794–1801, 1967.

[6] A. Telenti, P. Imboden, F. Marchesi, et al., "Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*," *The Lancet*, vol. 341, no. 8846, pp. 647–650, 1993.

[7] Y. Zhang, S. Heym, B. Allen, D. Young, and S. Cole, "The molecular basis of rifampin resistance in *Mycobacterium tuberculosis*: update," *Am. J. Med. Sci.*, vol. 318, no. 1, pp. 29–35, 1999.

[8] C. C. Boehme, P. Nabeta, D. Hillemann, et al., "Rapid molecular detection of tuberculosis and rifampin resistance," *New Engl. J. Med.*, vol. 363, pp. 1005–1015, 2010.

[9] J. Shea, A. H. Halse, B. Kohlerschmidt, et al., "Low-level rifampin resistance and rpoB mutations in *Mycobacterium tuberculosis*," *J. Clin. Microbiol.*, vol. 59, no. 3, e01885-20, 2021.

[10] E. Tortoli, A. Kohl, R. Brown-Elliott, et al., "Non-tuberculous mycobacteria: pathogenicity and clinical significance," *Clin. Microbiol. Infect.*, vol. 23, no. 10, pp. 683–688, 2017.

[11] Y. Wang, Y. Liu, X. Ma, et al., "Study on the relationship between rpoB mutations, biofilm formation, virulence and growth rate in *Staphylococcus aureus*," *Chinese Journal of Antibiotics (中国抗生素杂志)*, vol. 44, no. 9, pp. 991–997, 2019.

[12] B. P. Goldstein, "Resistance to rifampicin: a review," *J. Antibiot.*, vol. 67, no. 9, pp. 625–630, 2014.

[13] J. Deval, A. Navarro, A. Selmi, et al., "Molecular basis for the resistance of poliovirus RNA polymerase to inhibitors," *J. Virol.*, vol. 78, no. 18, pp. 8413–8425, 2004.

[14] P. Miotto, G. T. Zhang, R. Cirillo, and E. Cabibbe, "A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*," *Eur. Respir. J.*, vol. 50, no. 6, 1701354, 2017.

[15] R. Vogwill, T. Kojadinovic, and R. C. MacLean, "Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of *Pseudomonas*," *Proc. R. Soc. B: Biol. Sci.*, vol. 283, no. 1830, 20160151, 2016, doi: 10.1098/rspb.2016.0151.

[16] D. I. Andersson and D. Hughes, "Effect of rpoB mutations conferring rifampin resistance on fitness of *Mycobacterium tuberculosis*," *Antimicrob. Agents Chemother.*, vol. 48, no. 4, pp. 1289–1294, 2004, doi: 10.1128/AAC.48.4.1289-1294.2004.

[17] Y. Wang, Y. Liu, X. Ma, et al., "Study on the relationship between rpoB mutations, biofilm formation, virulence and growth rate in *Staphylococcus aureus*," *Chinese Journal of Antibiotics*, vol. 44, no. 9, pp. 991–997, 2019.

[18] A. K. Karki, R. Biswas, and R. S. Ghosh, "Rifampicin resistance and its fitness cost in *Enterococcus faecium*," *Indian J. Med. Microbiol.*, vol. 36, no. 3, pp. 426–429, 2018, doi: 10.4103/ijmm.IJMM_18_144.

[19] I. Comas, M. C. Coscolla, T. Luo, et al., "Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans," *Nature Genet.*, vol. 45, no. 10, pp. 1176–1182, 2013, doi: 10.1038/ng.2744.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proc. EMNLP-IJCNLP, 2019. doi: 10.48550/arXiv.1908.10084

[21] I. Beltagy, K. Lo, and A. Cohan, ''SciBERT: A pretrained language model for scientific text,'' Proc. EMNLP, 2019. doi: 10.48550/arXiv.1903.10676