CHENNAI
MATHEMATICAL
INSTITUTE

Industry Project Proposal

M.Sc. Data Science, Chennai Mathematical Institute

Predictive Modeling for TB Strain Identification, Drug Resistance, and Epidemiological Spread Analysis

Supervisor: Dr. Bikul Das
Director and Senior Scientist, KaviKrishna Laboratory
bdas@kavikrishnalab.org

Student: Gauranga Kumar Baishya
M.Sc. Data Science,
Chennai Mathematical Institute
Contact: +91 7086090441
gauranga.mds2023@cmi.ac.in

Student: Siddhesh Maheshwari
M.Sc. Data Science,
Chennai Mathematical Institute
Contact: +91 6350130496
siddheshm.mds2023@cmi.ac.in

August 19, 2024

1 Introduction

Tuberculosis (TB) remains one of the world's deadliest infectious diseases, claiming over a million lives annually. A staggering one-third of the global population carries latent or active tuberculosis (TB) infection, and India is unfortunately among the TB hotspot countries. Despite advancements in medicine, the rapid spread of TB, especially drug-resistant strains, poses a significant public health challenge. Failure to develop precise predictive models for TB transmission in different Epidemiological settings is one of the major concerns as per WHO's end TB strategy.

2 Problem Statement

Managing TB effectively is hindered by two major challenges: the increasing prevalence of drug-resistant strains and the intricate patterns of TB spread within different populations. Current diagnostic methods, while effective, are often slow and may not provide the level of detail required for personalized treatment. Furthermore, traditional epidemiological models lack the granularity needed to accurately predict the spread of TB, particularly in diverse regions with varying healthcare infrastructure. By incorporating drug information derived from patient metadata, we can develop a model to predict drug resistance profiles using molecular docking. This approach will pave the path for novel TB drug discovery.

3 Methodology

The project is divided into two key parts:

Part 1: Prediction TB Strain Identification and its corresponding drug resistance profile :- The first part focuses on predicting drug resistance and identifying TB strains using a dataset that includes Whole Genome Sequencing (WGS) data for TB strains, and clinical subject metadata. The initial step involves processing the RAW data obtained from the Kavikrishna Lab to ensure it is suitable for analysis. The post-processed mapped data will be used to train deep learning models, such as Recurrent Neural Networks (RNNs), to identify TB strains based on their mutation profile and associated metadata. By integrating drug information from the provided patient metadata, we can develop a model to identify drug resistance profiling using the molecular docking approach thereafter coming up with new drug discovery.

Part 2: Epidemiological Modeling of TB Spread:- Using Whole Genome Sequencing (WGS) data, we can identify strain-specific markers crucial for understanding and predicting TB spread. We will collect epidemiological data on TB incidence, population mobility, and healthcare access, along with environmental data such as climate and etiological factors, from open-source datasets available on the Internet. The model design will focus on constructing a Bayesian network to represent key variables and their dependencies, estimating conditional probabilities using historical data. This model will be trained on a subset of the data and optimized using techniques like expectation-maximization. Finally, the model will be validated on unseen data to ensure accuracy, with cross-validation employed to reduce overfitting. The refined model will be used to predict TB spread, identify high-risk areas, and provide insights into drug resistance.

These two interconnected parts of the project collectively aim to tackle tuberculosis from both a clinical and epidemiological perspective. By first identifying TB strains and predicting drug resistance using advanced deep-learning techniques, we lay the groundwork for understanding the genetic factors influencing the pathogen behavior and possible drug discovery. This knowledge is then integrated into epidemiological models, leveraging Bayesian networks to predict how these strains spread within populations under various conditions. Together these models will allow us to develop a novel TB control strategy applicable to existing TB hotspots worldwide.

4 Planning

- We can dedicate 2-3 hours daily on weekdays and over 3-4 hours on weekends to model development and data analysis, with weekly meetings with my supervisor to discuss the progress and next steps.
- We will focus on refining models one at a time, starting with the classical ML techniques, Deep Neural Networks in Part 1, & Bayesian networks, and agent-based models in Part 2.

5 Learning Outcomes

This project will provide significant insights and hands-on experience in the following areas:

- Advanced machine learning techniques, including Deep Neural networks, Bayesian networks, and agent-based modeling.
- Integration and analysis of complex datasets, combining genetic, epidemiological, and population data.
- Practical applications of AI/ML in systems biology, and public health, with a focus on infectious disease modeling.

6 References

- Getoor, L., Rhee, J. T., Koller, D., & Small, P. (2004). Understanding tuberculosis epidemiology using structured statistical models. *Artificial Intelligence in Medicine*, 30(3), 233-256. https://doi.org/10.1016/j.artmed.2003.11.003
- Sharma, A., Machado, E., Lima, K. V. B., Suffys, P. N., & Conceição, E. C. (2022). Tuberculosis drug resistance profiling based on machine learning: A literature review. *Brazilian Journal of Infectious Diseases*, 26(1), 102332. https://doi.org/10.1016/j.bjid.2022.102332
- Kostyukova, I., Pasechnik, O., & Mokrousov, I. (2023). Epidemiology and drug resistance patterns of Mycobacterium tuberculosis in a high-burden area in Western Siberia, Russia. *Microorganisms*, 11(2), 425. https://doi.org/10.3390/microorganisms11020425
- Liu, M., Xu, P., Liao, X., Li, Q., Chen, W., Gao, Q., Li, N., Luo, T., & Chen, L. (2021). Molecular epidemiology and drug-resistance of tuberculosis in Luodian revealed by whole genome sequencing. *Infection, Genetics and Evolution*, 93, 104979. https://doi.org/10.1016/j.meegid.2021.104979
- Das, B., Kashino, S. S., Pulu, I., Kalita, D., Swami, V., Yeger, H., Felsher, D. W., & Campos-Neto, A. (2013). CD271(+) bone marrow mesenchymal stem cells may provide a niche for dormant Mycobacterium tuberculosis. *Science Translational Medicine*, 5(170), 170ra13. https://doi.org/10.1126/scitranslmed.3004912

- Garhyan, J., Bhuyan, S., Pulu, I., Kalita, D., Das, B., & Bhatnagar, R. (2015). Preclinical and Clinical Evidence of Mycobacterium tuberculosis Persistence in the Hypoxic Niche of Bone Marrow Mesenchymal Stem Cells after Therapy. *The American Journal of Pathology*, 185(7), 1924-34. https://doi.org/10.1016/j.ajpath.2015.03.028

- Pathak, L., & Das, B. (2021). Initiation of Post-Primary Tuberculosis of the Lungs: Exploring the Secret Role of Bone Marrow Derived Stem Cells. *Frontiers in Immunology*, 11, 594572. https://doi.org/10.3389/fimmu.2020.594572

- Mitra, S., Saikia, P. J., Dutta, A., Das, R., Das, G., Baishya, T., Das, C., Sarma, T., Das, S., Pathak, L., & Das, B. (2023). A novel comprehensive Tuberculosis (TB) control programme methodology based on the nexus of participatory action research inspired public health and precision treatment approach. *medRxiv*. https://doi.org/10.1101/2024.01.02.23300347 (Preprint)