Bioinformatics into TB surveillance: A new approach to an ancient foe

18 February 2025

David Ussery, a Professor in the Department of BioMedical Informatics at UAMS, and his Ph.D. student Brian Delavan discuss bioinformatics into TB surveillance, presenting a new approach to tackling this ancient foe

Tuberculosis (TB) is an ancient disease, with evidence of TB infections dating back over 4,000 years, from TB infections in Neolithic Italy and Denmark. [1] TB is caused by the bacteria Mycobacterium tuberculosis (Mtb) and is transmitted via saliva droplets of an infected person.

TB remains a public health threat. In 2019, an estimated 10 million people were infected worldwide, with approximately 1.2 million deaths. [2] TB was declared a global emergency in 2018 by the United Nations. [3]

The biology of TB infection

Once Mtb is inhaled, a complicated series of processes begin in the body to fight the bacteria. For 90% of those who inhale Mtb, the bacteria are successfully removed. For the remaining 10%, one of two outcomes occurs. For 5% of those exposed, active TB develops, and the person can transmit Mtb to others. For the other 5%, the Mtb is contained within granulomas, a collection of immune cells that form in response to Mtb infection. [4] The person with a granuloma does not show TB symptoms and cannot transmit Mtb. This person is classified as having Latent Tuberculosis Infection (LTBI).

Arkansas' pioneering role in TB treatment

Arkansas has a long history of pioneering approaches to dealing with TB. During the early 20th century, Arkansas was home to the largest TB sanatoriums in the United States. [5] Dr. Joseph Bates, an Arkansas physician, revolutionized TB treatment by showing TB patients could be treated as outpatients. [6,7] This change allowed TB treatment to occur away from sanatoriums. Eventually, this was adopted as the primary treatment plan around the world.

Arkansas continues to incorporate innovative approaches to TB control. Arkansas' newest innovation is the incorporation of bioinformatics into TB surveillance. Bioinformatics is the study of the flow of biological information and uses computational methods to analyze large amounts of biological sequences. In the case of TB, this includes Mtb genomes isolated from infected individuals.

## How is Arkansas incorporating bioinformatics into TB surveillance?

Bioinformatics is helping to provide new insights into TB. We are applying many bioinformatic techniques in Arkansas, including genomic analysis, spatial statistics, and machine learning. We are using spatial statistics to compare LTBI testing and TB cases to determine how they are aligned. Arkansas recently constructed the nation's first database of LTBI testing (manuscript submitted), allowing Arkansas' TB Control officer to obtain LTBI data much more efficiently.

We have used the XGBoost machine learning algorithms and SHAP values to determine what epidemiological and/or social vulnerabilities are driving TB cases in Arkansas that should have been detected through Arkansas' TB High-Risk screening program, but were not detected, allowing the TB program to emphasize these variables when conducting TB screening programs. [8,9]

Most impactful is incorporating genome sequences and machine learning to examine TB outbreaks. TB genomic sequences are combined with the Monte Carlo Markov Chain (MCMC) machine learning algorithm to predict TB transmission via a transmission tree. A transmission tree differs from a phylogenetic tree in two main ways.

A phylogenetic tree examines the relatedness of the samples and only considers the samples; it cannot infer any samples that may be missing, and the direction of transmission is not a consideration for a phylogenetic tree. [10,11] A transmission tree can infer missing cases and predict the transmission of disease, making a transmission tree a powerful tool for analyzing disease outbreaks.
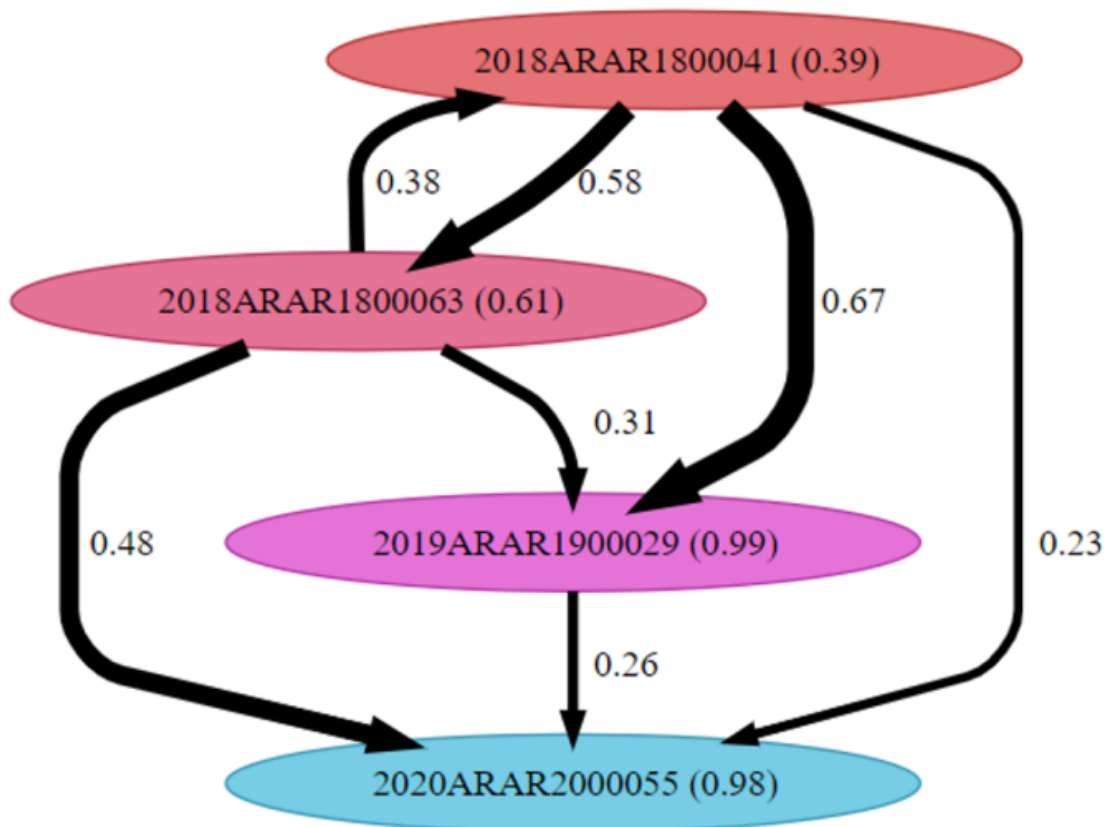
The use of genomes and transmission prediction

The Centers for Disease Control and Prevention (CDC) defines a TB outbreak as two or more TB cases sharing a genotype. Once an outbreak is declared, the genomes from the TB samples are downloaded and cleaned. Arkansas uses the BEAUTi/BEAST to analyze the genomes and perform the MCMC algorithm. [12]

The MCMC algorithm approximates complex probability distributions and uses the Monte Carlo Simulation to randomly sample the probability space and estimate quantities of interest.

The Markov Chain generates a sequencer of states when the next state depends only on the current state, which ensures the algorithm efficiently navigates the space. Arkansas analysis uses 100,000 MCMC runs.

The files produced by BEAUTi/BEAST are then analyzed by the BREATH program. [13] The BREATH program visualizes whom-infected-whom while also providing two probability values. The probability value on the arrow leading from one case to another is

the probability that the node at the head of the arrow infected the node at the tail of the arrow. The probability value within the node indicates the probability that the node was infected by an unsampled case.

We applied the BEAUTi/BEAST/BREATH to a current TB outbreak in Southwest Arkansas. The resulting predicted transmission tree is shown in the figure.

This tree branch reveals much information to help us understand the Southwest Arkansas TB outbreak. The transmission tree predicts whom-infected-whom within this branch. This allows the Arkansas TB Control officer to examine where this group interacted, whom they interacted with, and other epidemiological questions pertaining only to this group. This saves the outbreak investigators time, effort, and resources and may result in a better outcome, since interventions will be targeted to this group.

Arkansas has a long tradition of innovation in the TB control space. From large sanitoriums to new TB treatment, Arkansas has played an outsized role in revolutionizing TB care and control. Through applying bioinformatics to TB care and control, Arkansas continues innovating and pushing towards eliminating TB in Arkansas.

References

1. Smith I. Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence. Clin Microbiol Rev. 2003;16(3):463-496. doi:https://doi.org/10.1128/CMR.16.3.463-496.2003
2. World Health Organization. Global Tuberculosis Report 2020.; 2020.
3. Long B, Liang SY, Koyfman A, Gottlieb M. Tuberculosis: a focused review for the emergency medicine clinician. Am J Emerg Med. 2020;38(5):1014-1022. doi:https://doi.org/10.1016/j.ajem.2019.12.040
4. Ehlers S, Schaible UE. The granuloma in tuberculosis: dynamics of a host-pathogen collusion. Front Immunol. 2012;3:411. doi:https://doi.org/10.3389/fimmu.2012.00411
5. Petersen S. Arkansas State Tuberculosis Sanatorium: The Nation's Largest. Ark Hist Q. 1946;5(4):311. doi:https://doi.org/10.2307/40018309
6. Bates J. Ambulatory Treatment of Tuberculosis – An Idea Whose Time Is Come. Am Rev Respir Dis. 1974;109(3):317-319.
7. Floyd L, Bates, Joseph. Stalking the Great Killer: Arkansas's Long War on Tuberculosis. University of Oklahoma Press; 2023.
8. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016:785-794. doi:https://doi.org/10.1145/2939672.2939785
9. Marcílio-Jr WE, Eler DM. Explaining dimensionality reduction results using Shapley values. Expert Syst Appl. 2021;178:115020. doi:https://doi.org/10.1016/j.eswa.2021.115020

10. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Leitner T, ed. Mol Biol Evol. 2019;36(3):587-603. doi:https://doi.org/10.1093/molbev/msy242

11. Ayabina D, Ronning JO, Alfsnes K, et al. Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. Microb Genomics. 2018;4(10). doi:https://doi.org/10.1099/mgen.0.000219

12. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969-1973. doi:https://doi.org/10.1093/molbev/mss075

13. Colijn C, Hall M, Bouckaert R. Taking a BREATH (Bayesian Reconstruction and Evolutionary Analysis of Transmission Histories) to simultaneously infer phylogenetic and transmission trees for partially sampled outbreaks. Published online July 15, 2024. doi:https://doi.org/10.1101/2024.07.11.603095

Contributor Details

- Article Categories
- Health

- Article Tags
- Biology
- Genetic Research
- North America Analysis

- Publication Tags
- OAG 046 – April 2025

- Stakeholder Tags
- SH - Department of Biomedical Informatics - University of Arkansas for Medical Sciences