

# SPOTCLUST: A Tool to Cluster Spoligotype Data for Tuberculosis Evolution and Epidemiology

Inna Vitol<sup>1</sup>, Jeff Driscoll<sup>2</sup>, Natalia Kurepina<sup>3</sup>, Barry Kreiswirth<sup>3</sup>,  
Kristin Bennett<sup>1</sup>

**Keywords:** clustering, mathematical models, spoligotyping

We present a novel clustering approach to advance global studies of *Mycobacterium tuberculosis* (TB) complex genotyping data. SPOTCLUST uses mixture models to identify families of TB based on their spacer oligonucleotide typing (spoligotyping) patterns. SPOTCLUST incorporates biological information on spoligotype evolution without attempting to derive the full phylogeny of TB. Our approach provides a simple and robust tool for tuberculosis epidemiology. Tuberculosis is one of the most widespread infectious diseases in the world, infecting more than 1 billion persons, and is dramatically expanding due to HIV/AIDS and the emergence of multi-drug-resistant TB strains. More than two million people die each year of tuberculosis. This despite the fact that it is curable with early detection and prompt treatment.

Differentiating between various patient isolates and using the data for contact investigations are major applications for TB genotyping. The direct repeat (DR) locus is a region of the TB chromosome used as the target for the spoligotyping assay [1]. The DR locus of TB complex consists of well-conserved direct repeats interspersed with unique spacer sequences. Spoligotyping differentiates isolates by determining the absence or presence of 43 specific spacer sequences in the DR locus. Spoligotyping is a fast, highly reproducible method and the resulting fingerprint has a simple binary format, which allows the data exchange between laboratories and facilitates construction of large spoligotype databases [2]. Previous studies have grouped spoligotypes into 9 major families [3] that can be further broken down into 36 families in the global database SpolDB3 using visual rules [2].

Prior methods for automatic classification of TB strains based on spoligotyping used decision trees induced from the DB1 spoligotype database labeled by a human expert [3]. “Uninformative” examples were removed using a prototype selection algorithm [3]. While producing interpretable results, the decision tree approach required labeling the data, an error-prone and labor-intensive process compounded by the fact that the phylogeny of TB complex is still under investigation. Our unsupervised generative mixture models can both identify potential TB families and create good predictive models for spoligotype classification without requiring labeling and preprocessing. Moreover, this technique can be customized to exploit prior information on TB bacteria.

Our underlying mixture model assumes that within a TB family or cluster, the spacers can be treated as independent Bernoulli variables (the Naïve Bayes approach). Simple multivariate Bernoulli mixture model produced clusters inconsistent with expert knowledge. It is known that spoligotypes evolve by losing one or more contiguous spacers and that spacer duplication is very unlikely. To incorporate this knowledge, we assumed that each cluster has an unobserved Hidden Parent and that the children of the Parent (the observed strains) may lose a spacer with small probability, but are extremely unlikely to ever gain one. The EM algorithm was used to find maximum likelihood estimates of the mixture model’s parameters; for each cluster this corresponds

---

<sup>1</sup> Computer Science Department, Rensselaer Polytechnic Institute, 110 8<sup>th</sup> Street, Troy, NY. E-mail: vitoli@rpi.edu; bennek@rpi.edu

<sup>2</sup> Division of Infectious Diseases, Wadsworth Center, New York State Department of Health, P.O. Box 22002, New Scotland Ave., Albany, NY. E-mail: jeffrey.driscoll@wadsworth.org

<sup>3</sup> TB Center, Public Health Research Institute, 225 Warren Street, Newark, NJ. E-mail: nkurep@phri.org; barry@phri.org

