Fighting the Tuberculosis Pandemic using Machine Learning

Kristin P. Bennett
and the TB-Insight Team
Departments of Mathematical and Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
bennek@rpi.edu
http://tbinsight.cs.rpi.edu

Tuberculosis (TB) stubbornly persists as a leading cause of death worldwide. According to the World Health Organization, one third of the human population is infected, either latently or actively, with TB. The emergence of drug resistant TB strains remains a constant threat since the mortality rate associated with extensively drug resistant TB is over 80%. *Mycobacterium tuberculosis* complex (MTBC) is the causative agent of tuberculosis. High-throughput genotyping combined with modern machine learning are fundamentally changing how TB is tracked and controlled for public health purposes. MTBC isolates from TB patients are routinely genotyped using multiple biomarkers, which include spacer oligonucleotide types (spoligotypes), Mycobacterial Interspersed Repetitive Units - Variable Number Tandem Repeats (MIRU-VNTR), and IS6110 Restriction Fragment Length Polymorphism (RFLP). Today in the United States (US), isolates from every single TB patient are routinely genotyped creating massive and somewhat overwhelming databases of patient and pathogen data that are not full exploited [CDC Guide, 2004]. Machine learning techniques are a proven success story for unlocking the information captured in large scale national and international public-health databases for TB control. The key to success is building the known biology of MTBC into the machine learning models. SPOTCLUST, an online tool for determining MTBC sublineages using spoligotypes based on Bayesian Networks customized to capture the evolution of spoligotypes, has become a standard tool in TB molecular epidemiology used by public health groups world-wide with results reported in over 60 publications [Vitol, 2006; Sintchenko,2007]. This is a testimony to the robustness and generalization capacity of Bayesian Networks since SPOTCLUST was trained in 2005 on a small unlabeled dataset of about 400 spoligotypes collected by the New York State Department of Health.

The next generation of MTBC lineages models and data visualization methodologies are being created using over 40,000 patients as part of the TB-Insight project at RPI [Shabbeer, 2012; Aminian, 2010]. Novel knowledge-based Bayesian networks (KBBN) capture the knowledge of MTBC obtained from expert-defined rules and large DNA fingerprint databases to classify strains of MTBC into fifty-one genetic sublineages. The model uses two high-throughput biomarkers: spoligotypes and MIRU that are the new standard for public MTBC genotyping in the US and Europe. KBBN provides an elegant and simple way to incorporate existing widely accepted visual rules for MTBC sublineages into a classifier designed to capture known properties of the MTBC biomarkers. Unlike prior knowledge-based SVM approaches which require rules expressed as polyhedral sets, KBBN directly incorporates the rules without any modification. Computational results show that KBBN achieves much higher accuracy than methods based purely on rules, and than Bayesian networks trained on biomarker data alone. The

resulting sublineages and putative evolutionary histories are visualized using novel optimization-based graph visualization methodologies which capture both the evolutationary distances between isolates and aesthetic design criteria such as minimization of edge crossing overlaps. Analysis of New York City Patient Data by sublineage shows how MTBC genomic clusters and patient characteristics are closely related, helping to answer the question of which clusters of patients are associated with ongoing uncontrolled transmission versus latent reactivation of disease previously acquired, one of the greatest challenges in TB control in the developed world. A key value of the Bayesian Network approach is that the model can be released on the WWW without compromising patient confidentiality providing a valuable resource to TB public health care workers and researchers in both the developing and developed world.

References

DX uŵ]v] U X Z U X< WX v U ^ } v{ u o ˙ v] v ` } l { o] v }(
Mycobacterium tuberculosis compl ex l inea P_ U *BMC Bioinformatics*, 11: 3: S4, 2010.

^ŵ] Z] o }v }(v} v P } µ W] Àv] w }v }o ,v } }l (} d -
Controll ers, Epidemiol ogies, Laboratorians, and Other Program Staff, United States Centers for Disease
Control and Prevention, 2004.

X Z^ P ÌKX U X U Z v < XWX U ^t } } o } µ }o
] u] } P ˙ }(µ µ } _] U *Infection, Genetics and Evolution*, in press, 2012.

X ^ Z o U / X U o U ^ Z P { P for disease management and surveill ance U
Nature Reviews Microbiol ogy, 5: 6: 464-470, 1740-1526, 2007.

J. Vitol , J. Driscol , B. Kreiswirth, N. Kurepina, K. P. Bennett, "I dentifying Mycobacterium tubercul osis
Compl ex Strain Famil ies using Spol igotypes", *Infection, Genetics and Evolution*, 6: 6: 491-504, 2006.

M. Aminian, A. Shabbeer, K. Hadley, C. Ozcaglar, S. Vandenberg K. P. Bennett
^ v] -based Bayesian network for the d assi fication of *Mycobacterium tuberculosis* compl ex
µ] P U D U } v(î ì í í X

Topic: Bioinformatics
Preference: Oral

This work is supported by NI H R01-ALM009731.