Fighting the Tuberculosis Pandemic using Machine Learning

Kristin P. Bennett
Rensselaer Polytechnic Institute
TB-Insight Team

http://www.cs.rpi.edu/~bennek/tbinsight/index.html
This work is supported by NIH R01LM009731.
1/3 of World Latently Infected with TB

Estimated tuberculosis (TB) incidence rates, 2011

2.4 Million Deaths per Year
Tuberculosis cases and rates, New York City, 1982-2012

1. Rates are based on official Census data.
Number of TB Cases in U.S.-born vs. Foreign-born Persons
United States, 1993–2011*

*Updated as of June 25, 2012.
Drug Resistance Threat

- Susceptible
- Drug Resistant
- MDR-TB - Multi-Drug Resistant
- XDR-TB - Extremely-Drug Resistant
- TDR-TB? - Totally Drug Resistant
Modern TB Control

TB Controller: Find source(s) of infection in order to identify people who need treatment and stop future transmission.

Tools: Contact Investigation
DNA Fingerprints of TB bacteria
Two or More DNA fingerprints gathered for every TB Patient in USA

- Spoligotype
  111111111111111000000000000000000000000000
- MIRU
  22531b153321
- RFLP
  BW90
Major Phylogeographic Lineages of the MTBC

Determined by RD’s
Predictable by Spoligotype

[Gagneux S et al. PNAS 2006]
Spoligotype Genetic Diversity

Spoligoforests labeled by CDC Expert Rules

Classification Models

TB-Lineage
- Bayesian Network: 2010.
Sublineages-varying opinions

62 sublineages based on spoligotype signatures
[Brudey et al BMC Microbiol. 2006]

10 groups based on 212 single nucleotide polymorphism (SNP) markers
[Filliol J. Bacteriol 2006]

8 groups based on 230 sSNPs
[Gutacker Genetics 2002]
What’s the story with sublineages?

Spoligoforest – drawn using GraphViz Twopi
Data labels – SpolDB4 [Brudey et al 2006]
SPOTCLUST: Hidden-Parent Bayesian Network for Spoligotypes (Vitol et al, 2006)

Standard resource for tuberculosis sublineage identification used in over 96 publications.

Strains of *Mycobacterium tuberculosis* from Western Maharashtra, India, Exhibit a High Degree of Diversity and Strain-Specific Associations with Drug Resistance, Cavitary Disease, and Treatment Failure.

Characterization of multiple and extensively drug resistant *Mycobacterium tuberculosis* isolates with different ofloxacin-resistance levels.

Distinct clinical and epidemiological features of tuberculosis in New York City caused by the RD*<sub>Rio</sub> Mycobacterium tuberculosis* sublineage.

Mycobacterium bovis infection in livestock workers in Ibadan, Nigeria: evidence of occupational exposure.

High prevalence of subclinical tuberculosis in HIV-1-infected persons without advanced immunodeficiency: implications for TB screening.

Whole cell & culture filtrate proteins from prevalent genotypes of *Mycobacterium tuberculosis* provoke better antibody & T cell response than laboratory strain H<sub>37</sub>Rv.
Direct repeats (DR) separated by variable spacers

Contiguous on chromosome, order well conserved

Forty three spacers used

Presence of a spacer is detected: 1 - present (✓), 0 - absent (☐)

<table>
<thead>
<tr>
<th>Strain</th>
<th>Binary description of spoligotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. tuberculosis</em></td>
<td><img src="image1" alt="Binary description" /></td>
</tr>
<tr>
<td><em>Beijing</em></td>
<td><img src="image2" alt="Binary description" /></td>
</tr>
<tr>
<td><em>M. bovis</em></td>
<td><img src="image3" alt="Binary description" /></td>
</tr>
<tr>
<td>Indo-Oceanic</td>
<td><img src="image4" alt="Binary description" /></td>
</tr>
</tbody>
</table>
Rule-Based Method: 
(Sub)Lineage Visual Rules

- Determined by human experts.
- Ill-defined.
- Incomplete.
- Frequently ambiguous.
- No precedence.
- May or may not correspond to actual evolutionary groups.

*M. africanum*  

*Haarlem 2*
First Try – Naïve Bayes

*M. tuberculosis* Haarlem2 Family

- Prototype = probabilities

- Bernoulli Mixture Model

- Biology is wrong!!!
Unsupervised Hidden Parent Multivariate Bernoulli Mixture Model

Model child spacer $S$, given unobserved parent spacer $H$

- With very high probability child matches parent
- Children are much more likely to lose spacer than gain

- $P(S=1|H=1) = 0.99$
  $P(S=0|H=1) = 0.01$
- $P(S=0|H=0) = 1-1e^{-7}$
  $P(S=1|H=0) = 1e^{-7}$
SPOTCLUST (2006)

- Unsupervised except 34 SPOLDB2 Visual Rules used to initialize clusters
- Trained using 535 spoligotypes
- Number of sublineages (36) picked by MCCV
New Challenges

• More data: 119,684 isolates from US CDC, NYDOH, NY State DOH, and Institut Pasteur de Guadeloupe, MIRUVNTRPlus
• More types of DNA fingerprints – Spoligotypes and MIRU
• More proposed sublineages (70?)
• Putative labels from multiple experts
• Missing Data
## Who’s Right?

<table>
<thead>
<tr>
<th>Ctop</th>
<th>Cmid</th>
<th>Csub</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indo-Oceanic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>EAI6-BGD1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EAI7-BGD2</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td></td>
<td>EAI3-IND</td>
</tr>
<tr>
<td>Manila</td>
<td></td>
<td>EAI2-Manila</td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td>EAI-Mexico</td>
</tr>
<tr>
<td>Nonthaburi</td>
<td></td>
<td>EAI2-nonthaburi</td>
</tr>
<tr>
<td>Vietnam</td>
<td></td>
<td>EAI4-VNM</td>
</tr>
<tr>
<td><strong>Unknown Mid-level</strong></td>
<td></td>
<td>EAI1-SOM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAI2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAI8-MDG</td>
</tr>
<tr>
<td><strong>Mycobacterium africanum</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>West African 1</td>
<td></td>
<td>AFRII2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFRI3</td>
</tr>
<tr>
<td>West African 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mycobacterium bovis</strong></td>
<td><strong>Mycobacterium bovis</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BOV.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BOV.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BOV.3</td>
</tr>
<tr>
<td><strong>Mycobacterium canettii</strong></td>
<td><strong>Mycobacterium canettii</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Canettii</td>
</tr>
<tr>
<td><strong>Mycobacterium caprae</strong></td>
<td><strong>Mycobacterium caprae</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caprae</td>
</tr>
<tr>
<td><strong>Mycobacterium microti</strong></td>
<td><strong>Mycobacterium microti</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microti</td>
</tr>
<tr>
<td><strong>Mycobacterium mungi</strong></td>
<td><strong>Mycobacterium mungi</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M. mungi</td>
</tr>
<tr>
<td><strong>Mycobacterium pinnipedii</strong></td>
<td><strong>Mycobacterium pinnipedii</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pini1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pini2</td>
</tr>
</tbody>
</table>
Semi-supervised Hierarchical Lineage Model

A: 12 Major Lineages
B: 22 Mid-level Lineages
C: 70 + 9 Sub-lineages

• Estimated 92% Cross-validated Accuracy
Major Lineage Results

- Balanced Classification Rate about 98%
- No changes in major lineages
- MANU Modern?
Model Adds Sublineages but not Mid or Major Lineages
Ten New Putative Sublineages

- Discovers new sublineages and assigns them mid-level lineage
- New lineage characterized by “long” deletions frequently covering “typical” lineage deletions
- Covers many previously unlabeled isolates.

<table>
<thead>
<tr>
<th>Sublineage 1</th>
<th>Sublineage 2</th>
<th>Pos1</th>
<th>Pos2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.161</td>
<td>X1</td>
<td>0.362</td>
</tr>
<tr>
<td>X</td>
<td>0.161</td>
<td>X2</td>
<td>0.371</td>
</tr>
<tr>
<td>X</td>
<td>0.161</td>
<td>X3</td>
<td>0.202</td>
</tr>
<tr>
<td>X</td>
<td>0.161</td>
<td>LAMB</td>
<td>0.015</td>
</tr>
<tr>
<td>X</td>
<td>0.161</td>
<td>OtherSub5</td>
<td>0.01</td>
</tr>
<tr>
<td>X</td>
<td>0.161</td>
<td>OtherSub7</td>
<td>0.033</td>
</tr>
</tbody>
</table>
Visualizing X Lineage Spoligoforest

Graphvis Two-PI
Multi-Objective Embedding Methods

Good visualizations minimize stress and edge crossings.
Edge Crossing Constraints as Classification

No Crossing = Correct Classification

Crossing = Incorrect Crossing
Multi-Objective Graph Embedding

Add classifier for each potential edge/node crossing parameterized by $U$

- embedding error on $X$ (MDS stress)
- + misclassification error based on $U$ (SVM max margin)

$$\min_{X, U} \text{Stress}(X) + \sum_{i=1}^{m} \rho_i \left[ \| ( - A_i(X)u_i + 1)_+ \|_i + \| ( B_i(X)u_i + 1)_+ \|_i \right]$$

Optimize by an alternating algorithm on $X, U$ using scalable classification and embedding algorithms
Comparison: LAM sublineages
Visualizing X Lineage adding MIRU and Spoligotype distances

Shabbeer et al, 2012 Multiobjective Embedding, MDS +SVM
Spoligotype Genetic Diversity within 4.3K TB Patients in NYC – 2001-2007
Host-Pathogen Graphs

Patients = Circles
DNA Fingerprint = Box

Boxes nested to indicate multiple DNA fingerprints
• Other box = spoligotypes
• Inner box = RFLP.

Color by Patient property = Region of Birth

Split by lineage
Euro American
Disease as Stock Market

Companies = Bacteria

Buyers = Patients
Stock Market Tree Map

S&P 500 Stock Market Treemap

size represents 'weight' of company
color represents 'change' in stock price

Price Change
- > 8
- 6 - 8
- 4 - 6
- 2 - 4
- 0 - 2
- 0 - 2
-2 - 4
-4 - 6
-6 - 8
< -8

Mouse over tiles to see detailed info. Click on tiles to see current price (NYSE)

Split by Lineage - East African Indian

<table>
<thead>
<tr>
<th>Patient</th>
<th>ID</th>
<th>Biomarker 1</th>
<th>Biomarker 2</th>
<th>TB continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>105</td>
<td>S00669</td>
<td>MY8</td>
<td>Indian Subcontinent</td>
</tr>
<tr>
<td>Patient 2</td>
<td>2443</td>
<td>S00210</td>
<td>GD139</td>
<td>Indian Subcontinent</td>
</tr>
<tr>
<td>Patient 3</td>
<td>2452</td>
<td>S00210</td>
<td>MY44</td>
<td>Indian Subcontinent</td>
</tr>
<tr>
<td>Patient 4</td>
<td>2487</td>
<td>S00247</td>
<td>NO12</td>
<td>East Africa</td>
</tr>
</tbody>
</table>
NYC - East Asian
Euro American
NYC - Euro-American
Ongoing transmission in the US that became resistant to INH antibiotic
NYC - M. bovis
M. Bovis with Age of Patient/Country

NYC *M. bovis* (2001-2007)

- Extra pulmonary *M. bovis* strikes
  - Mexican Immigrants
  - US-born children of Mexican Immigrants
- Hypothesized caused: Unpasteurized cheese
Indo-Oceanic with Time in US

Large clusters with few US patients and no found epi-links.
Indo-Oceanic Anomaly

Hypothesized cause: IO strains have longer latency phenotype
Survival curves for all lineages

Proportion of cases not yet activated
Surveillance Data can reveal novel phenotypes and genotypes.
Challenges of Disease Control using Molecular Epidemiology

- Coupling human intelligence and analytics to help public health care workers control disease
- Informing local decisions with global data
- Allocating scarce control resources effectively by predicting disease dynamics
- Incorporating rapidly evolving data
  - Contact Investigations
  - New biomarkers for pathogen/host
  - Electronic Medical Records
  - Social media
- Getting the biology right
- Preserving Privacy
Personalized Medicine based on Host and Pathogen DNA

- Discovery of Host/Pathogen Coadaptation
- Control and treatment efforts guided by host and pathogen DNA
- Better models for drug development, etc

Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype

Thorax, 2013
Other Diseases/Pathogens

MRSA

During this talk approximately
998 people developed active TB
276 people died of TB