A Data Mining Approach To Spacer Oligonucleotide Typing of *Mycobacterium tuberculosis*
**Mycobacterium tuberculosis**

- Gram negative
- Variant subclass
- Multidrug-resistant tuberculosis (MDR-TB)
Tuberculosis (TB)

- Death: 4.2 people/day in Taiwan
- 1 person/sec infected in whole world
- WHO claimed “a global emergency”
- HIV patients are susceptible to TB
The way to resist TB

- Improvement of drug availability
  - New antibiotics
- Improvement of diagnosis abilities
  - DNA fingerprinting
- Analyzing the transmission pattern
  - Building of genotyping database
  - Classification and similarity analysis
Spacer oligonucleotide typing (Spoligotyping)

- Identifying the subclass of TB
- Detecting the existence of spacers between DRs
- Spacers are conserved and benefit primer design
Dealing with the spoligotyping data by human brain will be a tough work

- The amount of binary data are too huge to handle
- Separating signal or noise is subjective
- The processes are so routine and human will get tired

The best soluble way:
A genotyping database + automatic classification method
Flow chart of data-mining approach

- Spoligotyping Data
  - Notations and clade
  - Establishment of knowledge rules
  - Cross-validation procedure
  - Classification
TB Spoligotyping

- Purification of genomic mycobacterial DNA
- Performing PCR, DR conserved sequences as primer
- Performing southern hybridization, DR-spacer as probe
- Membrane exposed to x-ray film
TB Spoligotyping

- Database 1 (DB1)
  - 7352 spoligotypes classify to 342 pattern types
- Database 2 (DB2)
  - Using another 25 new published spacer (van Embden et al. 2000)
  - 323 spoligotypes
- Comparing DB1 with DB2
For machine learning, standard notations of DB is essential.

The 7352 instances of DB1 were clustered in 9 classes, defined by human experts using previously published results (Kremer *et al.*, 1999; van Embden *et al.*, 2000; Sola *et al.*, 2001)
## Notations and Clades

<table>
<thead>
<tr>
<th>Notation</th>
<th>Species</th>
<th>No.</th>
<th>Spoligotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afri</td>
<td><em>M. africanum</em></td>
<td>180</td>
<td>8,9,39 absent</td>
</tr>
<tr>
<td>T</td>
<td>Undefined</td>
<td>1590</td>
<td>At least 1-30, 21-24 present, 33-36 simultaneously absent, 31 present, 9 or 10 present</td>
</tr>
<tr>
<td>Beijing</td>
<td>Beijing</td>
<td>1268</td>
<td>1-34 are absent</td>
</tr>
<tr>
<td>EA-I</td>
<td>East-African-Indian</td>
<td>907</td>
<td>29-32 and 34 are simultaneously absent, at least 1-30 is present</td>
</tr>
<tr>
<td>Haarlem</td>
<td>Haarlem</td>
<td>1034</td>
<td>31,33-36 are simultaneously absent, at least 1-30 is present</td>
</tr>
<tr>
<td>LAM-1</td>
<td>Latin American and Mediterranean</td>
<td>819</td>
<td>21-24,33-36 are simultaneously absent, at least 1-30 is present</td>
</tr>
<tr>
<td>LAM-2</td>
<td>New family</td>
<td>294</td>
<td>9-10,33-36 are simultaneously absent, at least 1-30 is present</td>
</tr>
<tr>
<td>X</td>
<td>English-speaking countries</td>
<td>1186</td>
<td>18 or 33-36 absent, 18 absence sometimes be linked to absence of 39-42</td>
</tr>
<tr>
<td>M. bovis</td>
<td><em>M. bovis</em></td>
<td>74</td>
<td>39-43 are simultaneously absent, 33-36 are simultaneously present</td>
</tr>
</tbody>
</table>
Generating the decision tree by machine learning technique

- Using **SIPINA** software package
  - Collects the tree induction algorithm, like C4.5

According to knowledge rule, the input flowing in the decision tree will belong to a “leaf” by assigned a optimal criterion
Q: How to separate blue ball from red ball?
A: By the concept of “attribute:class”
How does C4.5 algorithm works?

- The concept of “Entropy”
  - The content of information
- Assigning each “attributes” with “classes”
  - The “ball” is “blue”
  - The Afri is “lack for spacers 8,9 and 39”
- Criteria
  - if ($E_{\text{attribute}} < E_{\text{max}}$) go to (subclass)
  - else ($E_{\text{attribute}} \geq E_{\text{max}}$) go to (leaf)
Fig. 1. Decision tree induced from DB1 in the 43-dimensional space.
Fig. 2. Decision tree induced from the subset of prototypes of $DB1$ in the 43-dimensional space.
Accuracy estimation by cross-validation procedure

- Assessment of the predictive ability of a model

\[ f \]

\[ f - 1 \]

\[ k-\text{NN} \]

Comparing the result generated by human
Assigning unknown attributes with a class

- Classifying unknown attributes by computer instead of labor work
- $k$-nearest-neighbor classifier ($k$-NN)
- Prototype selection
  - Eliminating the noises generate by $k$-NN
**k-nearest-neighbor classifier (k-NN)**

- Computing the similarity between test instance and training instance

```
1
2
3
...
n
```

- weight
- unknown
- classified
Prototype selection

- Combination of a group of similar records
- Keep only one of the data in the “edge” of the dataset
- Advantages:
  - Prevent overfitting (deep trees)
  - Reduce redundancy
  - Eliminate the noise
Knowledge discovery from DB1

- Human brain vs. machine
- Assessing by following criteria
  - The number of rules
    - Number of root to tree
  - The mean size of the rules
    - Mean of root to tree
  - The success rate on the learning sample
  - The spacers used in the model
    - Less is better
Knowledge discovery from DB1

Table 1. Results according to four performance criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Rules deduced from</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
</tr>
<tr>
<td>Number of rules</td>
<td>17.0</td>
</tr>
<tr>
<td>Mean size of the rules</td>
<td>25.4</td>
</tr>
<tr>
<td>Accuracy</td>
<td>76.6</td>
</tr>
<tr>
<td>Number of spacers</td>
<td>41.0</td>
</tr>
</tbody>
</table>
Prototype selection from the 7352 strains

- \( DB_{1\text{sub}} \) generated
- Decrease of the number of rules
  - Spacers reduced
- Increase of the accuracy
Contribution of the 25 additional spacers

- Without the profile of *M. africanum*
- Decrease of the number of rules
  - Spacers reduced
  - Model is smaller
- Total variance is increase
  - 48 spacers: 53%
  - 68 spacers: 60%
  - Resolution improved
Contribution of the 25 additional spacers

**Fig. 3.** Decision tree induced from DB2 in the 43-dimensional space after removing the 25 additional spacers.

**Fig. 4.** Decision tree induced from DB2 in the 68-dimensional space by the C4.5 induction algorithm.
Fig. 5. Correlation circle from the eight spacers used in the decision tree.
Conclusion

- Automatically discover simple knowledge rules from spoligotyping data
- Eliminating the use of uninformative spacers
- Future studies:
  - Other *M. tuberculosis* representations like Varible Number of Tandem DNA Repeats (VNTRs), Mycobacterial Interspersed Repetitive Units (MIRUs)
  - Improvement of data-mining methods