Using Random Forest to Predict a Complete Operon Map of the *Mycobacterium tuberculosis* Genome

Author: Chee Lee
Sponsor Instructor: Paul Roback
St. Olaf College
Tuberculosis: The Epidemic

- Affects 1/3 of the world’s population

↑ The causative agent of tuberculosis is *Mycobacterium tuberculosis* (MTB)
Project Aims

- To construct a complete MTB operon map
- To build a model with high accuracy and prediction power for future use
- To incorporate predictors with missing values
Why Predict Operons?

- A better understanding of the MTB genome is the first step to understanding the bacteria’s biological regulatory processes.
- Operons are sets of gene involved in these processes.
- Identify which processes can be disrupted.
- Find new and more effective ways to treat tuberculosis.
Operon pairs (OPs) are two genes that are
• Located in the same operon
• Located on the same DNA strand
• Adjacent to one another
• Regulated and transcribed together

Figure 1. Illustration of the operon pair definitions. Grey boxes indicate genes that are part of a known operon, white boxes indicate genes of unknown status. Arrows show the direction of transcription of each gene.
The Dataset

Table 1. Numeric summary of the dataset

<table>
<thead>
<tr>
<th>Known Operon Pairs</th>
<th>86</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Non-Operon Pairs</td>
<td>1353</td>
</tr>
<tr>
<td>Potential (Unclassified) Operon Pairs</td>
<td>2560</td>
</tr>
<tr>
<td>Total Gene Pairs</td>
<td>3999</td>
</tr>
</tbody>
</table>
The Dataset

- Collected from numerous studies
- Four parts:
  1. Known OPs and NOPs
  2. Intergenic distance
    - Number of base pairs separating gene pairs
  3. Coexpression correlations calculated from 45 microarray experiments
  4. Presence of 19 promoters and 2 terminators
    - Promoters are bindings sites for initiation of DNA transcription. Terminators signal end of transcription.
- Missing data
  - Exists in the coexpression data
  - 245 of the 2,560 unclassified potential operon pairs had incomplete data (10% missing)
  - 184 of the 1439 known operon pairs and non-operon pairs had incomplete data (13% missing)
- How to handle missing data?
Application of Random Forest

- Random forest classifies using many decision trees

Example of a tree that may be found in one of the random forests (this one is not fully grown yet)
Application of Random Forest

- A random forest is grown using 2/3 of the data (training set)
- The other 1/3 are out-of-bag (OOB) cases
- All trees are grown to the fullest
- Gives OOB error, which is misclassification of OOB cases
- Classification of POPs (test set) is determined by votes from trees
Application of Random Forest

- Three random forests were grown:
  1. Using **complete cases**, i.e. known OPs and NOPs that have no missing data (1,181 NOPs and 74 OPs)
  2. Using **proximities to impute missing values** on the known OPs and NOPs (1,353 NOPs and 86 OPs)
     - Imputed values are weighted using proximities, which is the probability that two gene pairs are classified exactly the same in a tree
  3. Using the ‘**rough fix**’ method to fill in missing values on the known OPs and NOPs (1,353 NOPs and 86 OPs)
     - Use medians of variable in the class (OP or NOP) if continuous, or most frequent value if categorical
- Each random forest is grown using the training set (i.e. known OPs and NOPs)
- 3,000 decision trees were grown in each forest
Advantages of Random Forest

- Does not overfit
- Resistant to noise (i.e. adding more predictors)
- Incorporates interactions of multiple variables
- Measures variable importance
- Gives measures of sensitivity (% correctly classified OPs) and specificity (% correctly classified NOPs) as well as OOB error
## Results: Model Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity (% true OPs)</th>
<th>Specificity (% true NOPs)</th>
<th>OOB error (% OOB cases misclassified)</th>
<th>% OPs overall (known OPs + predicted OPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF with complete cases</td>
<td>86.5</td>
<td>90.3</td>
<td>9.96</td>
<td>26.7</td>
</tr>
<tr>
<td>RF with imputations</td>
<td>87.2</td>
<td>89.9</td>
<td>10.28</td>
<td>28.0</td>
</tr>
<tr>
<td>RF with rough fix fills</td>
<td>87.2</td>
<td>90.5</td>
<td>9.66</td>
<td>29.8</td>
</tr>
</tbody>
</table>
% Overlapping POP Predictions

- Complete cases random forest
- Random forest, missing values imputed
- Random forest, missing values rough fixed

*Diagram not to scale
Important Variables

• In top 20 most important variables, all three random forests included the same 14 microarray experiments, same 2 promoters, and intergenic distance.
• Rerun of the random forests with 20 most important variables, and 10 most important variables show minimal decrease (1%-3%) in sensitivity and specificity.
Table 3. List of variables appearing in 20 most important variables of all three random forests. Letters are microarray experiments. ‘*’ indicates variable is important in all three random forests.

<table>
<thead>
<tr>
<th>Code</th>
<th>Actual</th>
<th>Code</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*</td>
<td>ethanol</td>
<td>dist*</td>
<td>intergenic distance</td>
</tr>
<tr>
<td>AJ</td>
<td>GSE1642chlorpromazine(13)</td>
<td>E*</td>
<td>hypo</td>
</tr>
<tr>
<td>AB</td>
<td>GSE1642Natural(5)</td>
<td>G*</td>
<td>KCN</td>
</tr>
<tr>
<td>BG*</td>
<td>GSE1642mtm(19)</td>
<td>H*</td>
<td>nitea</td>
</tr>
<tr>
<td>BU*</td>
<td>GSE1642procept 6776(5)</td>
<td>K*</td>
<td>sigB.inhibit</td>
</tr>
<tr>
<td>BV*</td>
<td>GSE1642procept 6778(9)</td>
<td>L*</td>
<td>sigE.inhibit</td>
</tr>
<tr>
<td>BZ</td>
<td>GSE1642roxithromycin(6)</td>
<td>R*</td>
<td>GSE1642#111895(8)</td>
</tr>
<tr>
<td>C*</td>
<td>h2o2o</td>
<td>sigAu</td>
<td>Sig A promoter</td>
</tr>
<tr>
<td>CF*</td>
<td>GSE1642triclosan(16)</td>
<td>sigDc</td>
<td>Sig D promoter</td>
</tr>
<tr>
<td>CL*</td>
<td>GPL280H2O2(9)</td>
<td>sigG*</td>
<td>Sig G promoter</td>
</tr>
<tr>
<td>D*</td>
<td>hypa</td>
<td>sigM*</td>
<td>Sig M promoter</td>
</tr>
<tr>
<td>DE</td>
<td>GSE1642deferoxamine(4)</td>
<td>U</td>
<td>GSE1642#241(8)</td>
</tr>
</tbody>
</table>
Conclusions

- Random forest is able to predict the operon status of the 2,560 POPs with high sensitivity and specificity
- By imputing values for missing data using proximities and the rough fix method, POPs that are more difficult to classify can be identified
- Important variables allows for future use of models as a starting point to predict OPs on phylogenetically related organisms (e.g. other prokaryotes)