A Forest Approach to Identification of Genes and Gene-Environment Interactions for Complex Diseases

Heping Zhang

School of Public Health
Yale University School of Medicine
Published Genome-Wide Associations through 12/2010, 1212 published GWA at p<5x10⁻⁸ for 210 traits

NHGRI GWA Catalog
www.genome.gov/GWASudies
<table>
<thead>
<tr>
<th>Abdominal aortic aneurysm</th>
<th>Cleft lip/palate</th>
<th>Homocysteine levels</th>
<th>Osteoarthritis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute lymphoblastic leukemia</td>
<td>Cognitive function</td>
<td>Hypospadias</td>
<td>Osteoporosis</td>
</tr>
<tr>
<td>Adhesion molecules</td>
<td>Conduct disorder</td>
<td>Idiopathic pulmonary fibrosis</td>
<td>Otosclerosis</td>
</tr>
<tr>
<td>Adverse response to carbamazepine</td>
<td>Colorectal cancer</td>
<td>IgA levels</td>
<td>Other metabolic traits</td>
</tr>
<tr>
<td>Adiponectin levels</td>
<td>Corneal thickness</td>
<td>IgE levels</td>
<td>Ovarian cancer</td>
</tr>
<tr>
<td>Age-related macular degeneration</td>
<td>Coronary disease</td>
<td>Inflammatory bowel disease</td>
<td>Pancreatic cancer</td>
</tr>
<tr>
<td>AIDS progression</td>
<td>Crohn's disease</td>
<td>Intracranial aneurysm</td>
<td>Pain</td>
</tr>
<tr>
<td>Alcohol dependence</td>
<td>Cutaneous nevi</td>
<td>iris color</td>
<td>Paget's disease</td>
</tr>
<tr>
<td>Alopecia areata</td>
<td>Dermatitis</td>
<td>Iron status markers</td>
<td>Pancreatic disease</td>
</tr>
<tr>
<td>Alzheimer disease</td>
<td>Drug-induced liver injury</td>
<td>Ischemic stroke</td>
<td>Parkinson's disease</td>
</tr>
<tr>
<td>Amyloid A levels</td>
<td>Endometriosis</td>
<td>Juvenile idiopathic arthritis</td>
<td>Pericarditis</td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis</td>
<td>Eosinophil count</td>
<td>Keloid</td>
<td>Peripheral arterial disease</td>
</tr>
<tr>
<td>Angiotensin-converting enzyme activity</td>
<td>Eosinophilic esophagitis</td>
<td>Kidney stones</td>
<td>Phosphatidylcholine levels</td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>Erectile dysfunction and prostate cancer treatment</td>
<td>LDL cholesterol</td>
<td>Phosphorous levels</td>
</tr>
<tr>
<td>Arterial stiffness</td>
<td>Erythrocyte parameters</td>
<td>Leprosy</td>
<td>Photic sneeze</td>
</tr>
<tr>
<td>Asparagus anemia</td>
<td>Esophageal cancer</td>
<td>Leptin receptor levels</td>
<td>Phytosterol levels</td>
</tr>
<tr>
<td>Asthma</td>
<td>Essential tremor</td>
<td>Liver enzymes</td>
<td>Platelet count</td>
</tr>
<tr>
<td>Atherosclerosis in HIV</td>
<td>Exfoliation glaucoma</td>
<td>Longevity</td>
<td>Polycystic ovary syndrome</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>Eye color traits</td>
<td>LP (a) levels</td>
<td>Primary biliary cirrhosis</td>
</tr>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>F cell distribution</td>
<td>LpPLA(2) activity and mass</td>
<td>Primary sclerosing cholangitis</td>
</tr>
<tr>
<td>Autism</td>
<td>Fibronogen levels</td>
<td>Lung cancer</td>
<td>PR interval</td>
</tr>
<tr>
<td>Basal cell cancer</td>
<td>Folate pathway vitamins</td>
<td>Magnesium levels</td>
<td>Progranulin levels</td>
</tr>
<tr>
<td>Behcet's disease</td>
<td>Follicular lymphoma</td>
<td>Major mood disorders</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>Fuchs' corneal dystrophy</td>
<td>Malaria</td>
<td>Protein levels</td>
</tr>
<tr>
<td>Biliary atresia</td>
<td>Freckles and burning</td>
<td>Male pattern baldness</td>
<td>PSA levels</td>
</tr>
<tr>
<td>Bilirubin</td>
<td>Gallstones</td>
<td>Matrix metalloproteinase levels</td>
<td>Psoriasis</td>
</tr>
<tr>
<td>Bitter taste response</td>
<td>Gastric cancer</td>
<td>MCP-1</td>
<td>Psoriatic arthritis</td>
</tr>
<tr>
<td>Birth weight</td>
<td>Glioma</td>
<td>Melanoma</td>
<td>Pulmonary funct. COPD</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>Glycemic traits</td>
<td>Menarche &amp; menopause</td>
<td>QT interval</td>
</tr>
<tr>
<td>Bleomycin sensitivity</td>
<td>Hair color</td>
<td>Meningococcal disease</td>
<td>Quantitative traits</td>
</tr>
<tr>
<td>Blond or brown hair</td>
<td>Hair morphology</td>
<td>Metabolic syndrome</td>
<td>Recombination rate</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>Handedness in dyslexia</td>
<td>Migraine</td>
<td>Red vs. non-red hair</td>
</tr>
<tr>
<td>Blue or green eyes</td>
<td>HDL cholesterol</td>
<td>Moyamoya disease</td>
<td>Refractive error</td>
</tr>
<tr>
<td>BMI, waist circumference</td>
<td>Heart failure</td>
<td>Multiple sclerosis</td>
<td>Renal cell carcinoma</td>
</tr>
<tr>
<td>Bone density</td>
<td>Heart rate</td>
<td>Myeloproliferative neoplasms</td>
<td>Renal function</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Height</td>
<td>N-glycan levels</td>
<td>Response to antidepressants</td>
</tr>
<tr>
<td>C-reactive protein</td>
<td>Hemostasis parameters</td>
<td>Narcolepsy</td>
<td>Response to antipsychotic therapy</td>
</tr>
<tr>
<td>Calcium levels</td>
<td>18th platelet</td>
<td>Nasopharyngeal cancer</td>
<td>Response to hepatocit C treat</td>
</tr>
<tr>
<td>Cardiac structure/function</td>
<td>Hepatitis</td>
<td>Neuroblastoma</td>
<td>Response to metformin</td>
</tr>
<tr>
<td>Carcinine levels</td>
<td>Hepatocellular carcinoma</td>
<td>Nicotine dependence</td>
<td>Response to statin therapy</td>
</tr>
<tr>
<td>Carotenoid/coqopher levels</td>
<td>Hirschsprung's disease</td>
<td>Obesity</td>
<td>Restless legs syndrome</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>HIV-1 control</td>
<td>Open angle glaucoma</td>
<td>Retinal vascular caliber</td>
</tr>
<tr>
<td>Cerebral atrophy measures</td>
<td>Hodgkin's lymphoma</td>
<td>Open personality</td>
<td>Rheumatoid arthritis</td>
</tr>
<tr>
<td>Chronic lymphocytic leukemia</td>
<td>Ribavirin-induced anemia</td>
<td>Optic disc parameters</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Serum metabolites</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Skin pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Smoking behavior</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Speech perception</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sphingolipid levels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Statin-induced myopathy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stroke</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Systemic lupus erythematosus</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Systemic sclerosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T-tau levels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tau Aβ1-42 levels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Telomere length</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testicular germ cell tumor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thyroid cancer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tooth development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total cholesterol</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Triglycerides</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tuberculosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Type 2 diabetes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ulcerative colitis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Urease</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Venous thromboembolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ventricular conduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vertical cup-disc ratio</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vitamin B12 levels</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vitamin D insufficiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vitiligo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Warfarin dose</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weight</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>White cell count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YKL-40 levels</td>
</tr>
</tbody>
</table>
Challenges

The identified markers or genes explained a small fraction of the diseases.

More markers & GxG?

Environment variables & GxE?

Incorporation of biologic knowledge?

Better characterization and use of traits?
In classic statistical modeling, we tend to have an adequate sample size for estimating parameters of interest. Often, we have hundreds or thousands of observations for the inference on a few parameters. We can try to settle an “optimal” model.

In genomic studies, we have more and more variables (gene based) but the access to the number of study subjects remains the same. One model can no longer provide an adequate summary of the information.
Outline

- Background
- Challenges
- Methods
  - Trees and Forests
  - Forest Size
  - Feature Importance
  - Uncertainties in Predictors
  - Interactions
- Acknowledgement
Complex Traits

Diseases that do not follow Mendelian Inheritance Pattern

Genetic factors, Environment factors, G-G and G-E interactions

Interactions: effects that deviate from the additive effects of single effects
SNP and Complex Traits

Regression Approach

\[
\begin{align*}
\{M_i|\gamma_i\} &= \frac{1}{P(\gamma_i)} \prod_j [P(\gamma_j|\alpha_j = 0) P(\gamma_i|M_i, \alpha_j) - P(\gamma_i) P(\alpha_j|M_i)] + P(\gamma_i) P(\alpha_j)
\end{align*}
\]

\[3; k, c) = P(\gamma_k = k | \gamma_c = c) = \gamma(\beta; k, K - 1, \gamma(\beta; k, \gamma, 0, \gamma_0, 0))
\]

\[P(\gamma_i) = \prod_j [P(\gamma_j|\alpha_j = 0) P(\gamma_i|\alpha_j, 0)]
\]

\[P(\gamma_i) = \prod_j [\pi(\beta; \gamma_j, 0) P(\gamma_i)]
\]

Note to see that \((\partial / \partial \beta) \pi(\beta; k, c) = c \gamma(\beta; k, K - 1, \gamma(\beta; k, \gamma, 0, \gamma_0, 0))
\]

\[\log(P\{M_i|\gamma_i\}) = \frac{\partial}{\partial \beta} \log(P(\gamma_i)) + \sum_j \frac{\partial}{\partial \beta} \log(\pi(\beta; \gamma_j))
\]

The null hypothesis that \(\beta = 0\), we have

\[\frac{\partial}{\partial \beta} \log(\pi(\beta; \gamma_j, 0) P(\gamma_j|M_i))
\]

\[= [1 - \gamma(0; \gamma_j, 0, \gamma_0, 0)]
\]

\[\frac{\partial}{\partial \beta} \log P(\gamma_i)|_{\beta=0} = \sum_j \frac{\partial}{\partial \beta} \log P(\gamma_j)|_{\beta=0}
\]

\[= \sum_j [1 - \gamma(0; \gamma_j, 0, \gamma_0, 0)]
\]

\[= \sum_j 1 - \gamma(\gamma_j, 0, \gamma_0, 0)
\]

\[\frac{1}{P(\gamma_j)}
\]

The coefficient of linkage disequilibrium,

\[\frac{P(AA) - P(AA|DL)}{P(AA)}
\]
Recursive Partitioning

A technique to identify heterogeneity in the data and fit a simple model (such as constant or linear) locally, and this avoids pre-specifying a systematic component.
Leukemia Data

Source: [http://www-genome.wi.mit.edu/cancer](http://www-genome.wi.mit.edu/cancer)

Contents:

- 25 mRNA - acute myeloid leukemia (AML)
- 38 - B-cell acute lymphoblastic leukemia (B-ALL)
- 9 - T-cell acute lymphoblastic leukemia (T-ALL)
- 7,129 genes

Question: are the microarray data useful in classifying different types of leukemia?
$\{M_i|\gamma_i\} = \frac{P\{\gamma_i\}}{P\{\gamma_i\}} \prod_j \left[ \frac{P\{\gamma_i|\gamma_j = 0\} P\{M_i|\gamma_j\}}{P\{M_i\}} \prod_j \left[ \frac{P\{|\beta|; \gamma_j; 0\} P\{\gamma_i\}}{P\{|\beta|; \gamma_j; 0\}} \right] \right]$.

Note to see that $(\partial / \partial \beta) \pi(\beta; \gamma, \epsilon) = \epsilon$

$$\log(P\{M_i|\gamma_i\}) = \frac{\partial}{\partial \beta} \log(P\{\gamma_i\})$$

$$+ \sum_j \frac{\partial}{\partial \beta} \log(\pi(\beta; \gamma_j; 0)) P\{dd|M_i\}$$

The null hypothesis that $\beta = 0$, we have

$$\frac{\partial}{\partial \beta} \log(\pi(\beta; \gamma_j; 0) P\{dd|M_i\}) = [1 - \gamma(0; \gamma_j, 0)] - \gamma(0; \gamma_j; 0)$$

$$= \frac{\partial}{\partial \beta} \log P\{\gamma_i\}|_{\beta=0} = \sum_j \frac{\partial}{\partial \beta} \log P\{\gamma_i\}|_{\beta=0} = \sum_j [1 - \gamma(0; \gamma_j, 0) - \gamma(0; \gamma_j; 0)]$$

Convenience, we drop the two irrelevant terms.

$$\log(P\{M_i|\gamma_i\})|_{\beta=0} = \sum_j [1 - \gamma(\gamma_j) - \gamma(0; \gamma_j; 0)]$$

The coefficient of linkage disequilibrium:

$$P(\text{AA}) - P(\text{dd}, \text{AA}) - P(\text{AA})P(\text{DL})$$
\[
\{y_i[M_i]\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j \left( \frac{P\{y_i|\delta_j = 0\} P(M_i|\delta_j = 0)}{P\{y_i\} \prod_j (\pi(\beta_j; \gamma, 0) P(\varepsilon_i|\beta_j))} \right) 
\]

Note to see that \( \frac{\partial}{\partial \beta} \pi(\beta; \gamma, 0) = \varepsilon \)

\[
\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} \log(P\{y_i\}) 
+ \sum_j \frac{\partial}{\partial \beta} \log(\pi(\beta_j; \gamma)) 
\]

The null hypothesis that \( \beta = 0 \), we have:

\[
\frac{\partial}{\partial \beta} \log(\pi(\beta; \gamma, 0) P(\delta|M)) 
= [1 - \gamma(0, \gamma, 1) - \gamma(0, \gamma)] 
\]

\[
\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j \left[ 1 - \gamma(\gamma) \right] 
\]

For convenience, we drop the two irrelevant terms:

\[
\log(P\{M_i|y_i\})|_{\beta=0} = \sum_j \left[ 1 - \gamma(\gamma) \right] - \gamma
\]

The coefficient of linkage disequilibrium:

\[
P(AA) - P(dd, AA) - P(AA)|P(DL) 
\]
Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data.

To identify a constellation of models that collectively help us understand the data. For example, in GWAS, we can select and rank the genes that may be highly associated with a trait.
Bagging (Bootstrap Aggregating)

The Collaborative Center for Statistics in Science
How Big a Forest?

A general practice in using random forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient.

- Furthermore, random forests are viewed as a “black-box” because of its sheer size.
Forest Size?

- Explore whether it is possible to find a common ground between a forest and a single tree
  - retain the easy interpretability of the tree-based methods
  - avoid the problems that the tree-based methods suffer from.

- Does a forest have to be large, or how small can a forest be?
Shrink a Forest

- Shrink the forest with two objectives
  - maintain a similar (or even better) level of prediction accuracy
  - reduce the number of the trees in the forest to a manageable level
Criteria

- Two measures are considered to determine the importance of a tree in a forest
  - by prediction
  - by similarity
Prediction Based Criterion

- “by prediction” method
  – focuses on the prediction
  – A tree can be removed if its removal from the forest has the minimal impact on the overall prediction accuracy.
Prediction Based Criterion

- “by prediction” method
  
  - For tree $T$ in forest $F$, calculate the prediction accuracy of forest $F_{(-T)}$ that excludes $T$.
  
  - $\Delta_{(-T)}$ represents the difference in prediction accuracy between $F$ and $F_{(-T)}$.
  
  - The tree with the smallest $\Delta_{(-T)}$ is the least important one and hence subject to removal.
Similarity Based Criterion

- “by similarity” method
  - is based on the similarity between two trees.
  - A tree can be removed if it is “similar” to other trees in the forest.
• “by similarity” method
  – The correlation of the predicted outcomes by two trees gives rise to a similarity between the two trees.
  – For tree $T$, the average of its similarities with all trees in $F(\cdot)$, denoted by $\rho_T$, reflects the overall similarity between $T$ and $F(\cdot)$.
  – The tree with the highest $\rho_T$ is the most similar to the trees in $F(\cdot)$ and hence subject to removal.
Critical Point

- Select the optimal size sub-forest
  - Let $h(i)$, $i=1,...,N_f-1$, denote the performance trajectory of a sub-forest of $i$ trees
    - $N_f$ is the size of the original random forest.
  - If we have only one realization of $h(i)$, we select the optimal size sub-forest by maximizing $h(i)$ over $i=1,...,N_f-1$.
  - If we have multiple realizations of $h(i)$, we select the optimal size sub-forest by using the 1-se rule.
- The size of this smallest sub-forest is called the critical point of the performance trajectory.
Simulation Designs

- For each data set, we generated 500 observations, each of which has one response variable and 30 predictors from Bernoulli distribution with success probability of 0.5.

- Chose $\nu$ of the 30 variables to determine the response variable.

$$y = \begin{cases} 
1, & \text{if } \sum_{i=1}^{\nu} X_i / \nu + \sigma > 0.5, \\
0 & \text{Otherwise.}
\end{cases}$$

- Where $\sigma$ is a random variable following the normal distribution with mean zero and variance $\sigma^2$.

- Considered two choices for $\nu$ (5 and 10) and two choices of $\sigma$ (0.1 and 0.3).
Simulation Designs

- To perform an unbiased comparison of the three tree removal measures, we simulated three independent data sets
  - The training set is used to train the initial random forest
  - The execution set is used to delete trees from the initial forest to produce sub-forests
  - The evaluation set is used to evaluate the prediction performance of the sub-forests
- The generation and use of these three data sets constituted one run of simulation, and we replicated 100 times.
Simulation Results

- Randomly selected one run of simulation and presented the stepwise change in the prediction performance in Figure 1.

The “by prediction” method is preferable — It can identify a critical point during the tree removal process in which the performance of the sub-forest deteriorates very rapidly.

- The performance of the sub-forests may begin to improve before the critical point.
Prediction performance of sub-forests produced from different datasets and methods
Simulation Designs

• In practice, we generally have one data set only.

• May not have the execution and evaluation data sets as in previous simulation.

• How do we select the optimal sub-forest with only one data set?
Simulation Designs

• After constructing an initial forest using the whole data set as the training data set
  – use one bootstrap data set for execution and the out-of-bag (oob) samples for evaluation.
  – use the oob samples for both execution and evaluation.
  – use the bootstrap samples for both execution and evaluation.
  – re-draw bootstrap samples for execution and re-draw bootstrap samples for evaluation.
A performance summary plot of the “by prediction” method.
Simulation Results

• The performance trajectories of the four bootstrap-based approaches may not overlap with the “golden” standard.

• For the selection of the optimal sub-forest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-forest.
Simulation Results

- Using the bootstrap samples for execution and the oob samples for evaluation is an effective sample-reuse approach to selecting the optimal sub-forest.
• **Dataset**

  – the microarray data set of a cohort of 295 young patients with breast cancer, containing expression profiles from 70 previously selected genes.

  – previously studied by van de Vijver *et al.*

• **The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not.**
Application

- Method used
  - The “by prediction” measure
  - The original data set to construct an initial forest
  - A bootstrap data set for execution
  - The oob samples for evaluation.

- The procedure is replicated for a total of 100 times.
  - The oob error rate is used to compare the performance of the initial random forest and the optimal sub-forest.
  - The sizes of the optimal sub-forests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26 and 61, respectively.
The Collaborative Center for Statistics in Science

Application

- The smallest optimal sub-forest in the 100 repetitions with the size of 7 is selected.
- As a benchmark, we used the 70-gene classifier proposed by Vijver, et al.
• Next table presents the misclassification rates based on the oob samples.

  – The initial forest and the optimal sub-forest achieve almost the same level of performance accuracy.

  – The 70-gene classifier has an out-of-bag error rate which is much higher than those of the forests.
Comparison of prediction performance of the initial random forest, the optimal sub-forest, and a previously established 70-gene classifier

<table>
<thead>
<tr>
<th>Method</th>
<th>Error rate</th>
<th>True Predicted</th>
<th>Good</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>26.0%</td>
<td>Good</td>
<td>141</td>
<td>17</td>
</tr>
<tr>
<td>Sub-forest</td>
<td>26.0%</td>
<td>Good</td>
<td>146</td>
<td>22</td>
</tr>
<tr>
<td>70-gene Classifier</td>
<td>35.3%</td>
<td>Good</td>
<td>103</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: The table shows the error rate, true positive predictions, and the number of good and poor predictions for each method.
The top three layers of the optimal sub-forest consisting of seven trees
What Did We Learn?

- It is possible to construct a highly accurate random forest consisting of a manageable number of trees.
  - the size of the optimal sub-forest is in the range of tens
  - some sub-forests can even over-perform the original forest in terms of prediction accuracy
- The key advantage
  - the ability to examine and present the forests.
- The limitation
  - future samples and studies are needed to evaluate the performance of the forest-based classifiers.
\[ \{ M_i | y_i \} = \frac{P \{ y_i \} \prod_j P \{ y_j | e_j = 0 \} P \{ M_i | y_i \}}{P \{ y_i \} \prod_j \left[ \pi(\beta; y_j; e) \right] P \{ e \} } \]

Note that

\[ P \{ y_i \} = \prod_j \left[ P \{ y_j | e_j = 0 \} \right] P \{ e \} \]

Log likelihood

\[ \log(P \{ M_i | y_i \}) = -\frac{\partial}{\partial \beta} \log(P \{ y_i \}) \]

The null hypothesis that \( \beta = 0 \), we have

\[ \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i; e) P \{ dd | M_i \}] = [1 - \gamma(0; y_i; 1) - \gamma(0; y_i; 1)] \]

\[ \frac{\partial}{\partial \beta} \log P \{ y_i \} |_{\beta=0} = \sum_j \frac{1}{P \{ M_i \}} \]

Convenience, we drop the two irrelevant

\[ \log(P \{ M_i | y_i \}) |_{\beta=0} = \sum_j \left[ 1 - \gamma(y_i) - \gamma(y_i) \right] \]

The coefficient of linkage disequilibrium

\[ P \{ AA \} - P \{ dd, AA \} - P \{ AA \} | P \{ DL \} \]
Variable Importance

Permutation importance (Breiman): For each tree in the forest, we count the number of votes cast for the correct class. Then, we randomly permute the values of variable $k$ in the oob cases and recount the number of votes cast for the correct class in the oob cases with the permuted values of variable $k$. The permutation importance is the average of the differences between the number of votes for the correct class in the variable-$k$-permuted oob data from the number of votes for the correct class in the original oob data, over all trees in the forest.
Permutation Importance

- Not necessarily positive
- Unbounded
- The magnitudes and relative rankings can be unstable when the number, \( p \), of predictors is large relative to the sample size.
- The magnitudes and relative rankings vary according to the number of trees in the forest and the number, \( q \), of variables that are randomly selected to split a node.
Permutation Importance

\[ \{M_i[y_i]\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j [P\{y_j|\psi_j = 0\}P\{\psi_j\} - P\{y_j|\psi_j = 1\}P\{\psi_j\}] \]

\[ = \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_j, 0)P\{\psi_j\} - \pi(\beta; y_j, 1)P\{\psi_j\}] \]

Note to see that \( \frac{\partial}{\partial \beta} \pi(\beta; k, \epsilon) = \epsilon \)

\[ \log(P\{M_i[y_i]\}) = \frac{\partial}{\partial \beta} \log(P\{y_i\}) \]

The null hypothesis that \( \beta = 0 \), we have

\[ \frac{\partial}{\partial \beta} \log(\pi(\beta; y_j, 0))P\{dd|M_i\} = [1 - \gamma(0; y_j, 1) - \gamma(0; y_j, 0)] \]

\[ = 1 - \gamma(0; y_j) - \gamma(0; y_j) \]

\[ \frac{\partial}{\partial \beta} \log P\{y_i\}\big|_{\beta=0} = \sum_j [1 - \gamma(0; y_j) - \gamma(0; y_j)] \]

For convenience, we drop the two irrelevant terms

\[ \log(P\{M_i[y_i]\})\big|_{\beta=0} = \sum_j [1 - \gamma(0; y_j) - \gamma(0; y_j)] \]

\[ = \sum_j \frac{1 - \gamma(0; y_j) - \gamma(0; y_j)}{P\{M_i\}} \]

The coefficient of linkage disequilibrium:

\[ C^2 = \frac{P\{dd, AA\} - P\{dd, AA\} - P\{AA\}\}P\{DL\} \]
Wang et al. (2010) introduced a maximal conditional chi-square (MCC) importance by taking the maximum chi-square statistic resulting from all splits in the forest that use the same predictor.
Chen et al. (2007): Whenever node t is split based on variable k, let L(t) be the depth of the node and S(k,t) be the chi-square test statistic from the variable, then $2^{-L(t)} S(k,t)$ is added up for variable k over all trees in the forest.
SNPs and Haplotypes

The Collaborative Center for Statistics in Science
Haplotype Certainty

SNPs

✓ Directly observed
✓ No uncertainty
✓ Less informative
✓ Tree approaches

Haplotypes

✗ Inferred from SNPs
✗ Uncertain
✓ More informative
✗ Forest approaches

\[
\{M_0, y_0, \ldots, y_n\} = \prod_i \left[ P\{M_i\} \prod_i \left[ P\{y_i\} \mid \beta_i = a \right] \right] = \prod_i \left[ P\{M_i\} \prod_i \left[ \pi\{\beta_i; y_i, 0\} \right] \right]
\]

\[
3: \quad (\partial / \partial \beta) \pi(\beta; y, 0) = \gamma(\beta; y, 0, 0), \quad \gamma(\beta; 0, 1) = 0,
\]

\[
\log(P\{M_i|y_i\}) = \frac{\partial}{\partial \beta} \log(P\{y_i\}) + \sum_i \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\]

The null hypothesis that \( \beta = 0 \), we have:

\[
\frac{\partial}{\partial \beta} \log[\pi(\beta; y_i, 0)] = \begin{cases} 
1 - \gamma(0; y_i, 1) - \gamma(0; y_i, 1) 
\end{cases}
\]

\[
\left. \frac{\partial}{\partial \beta} \log P\{y_i\} \right|_{\beta=0} = \sum_i \left[ 1 - \gamma(\beta; y_i) \right]
\]

For convenience, we drop the two irrelevant

\[
\left. \log(P\{M_i|y_i\}) \right|_{\beta=0} = \sum_i \left[ 1 - \gamma(\beta; y_i) \right] - \gamma
\]

The coefficient of linkage disequilibrium:

\[
P\{AA\} - P\{dd, AA\} - P\{AA\}P\{DL\}
\]

The Collaborative Center for Statistics in Science
Forest Forming Scheme

Unphased data

Reconstructed phased data 1

Tree 1

Importance index for haplotype 1

Reconstructed phased data 2

Tree 2

Importance index for haplotype 2

Reconstructed phased data 3

Tree 3

Importance index for haplotype 3

Reconstructed phased data 4

Tree 4

Importance index for haplotype k

The Collaborative Center for Statistics in Science
Inference from the Forest

Importance of haplotype $h$ in tree $T$

$$V_h = \sum_{t \in T, t \text{ is split by } h} 2^{-L_t} \chi_t^2,$$

where $L_t$ is the depth of node $t$ and $\chi_t^2$ is the value of the $\chi^2$-test statistic of independence.
Distribution of the maximum haplotype importance under null hypothesis is determined by permutation. First, disease status is permuted among study subjects while keeping the genome intact for all individuals. Then, each of the permuted data set is treated in the same way as the original data.
Simulation Studies (2 loci)

- 300 cases and 300 controls
- Each region has 3 SNPs
- 12 interaction models from Knapp et al. (1994) and Becker et al. (2005)
- 2 additive models with background penetrance
- 3 scenarios
  - Neither region is in LD with the disease allele
  - One of the regions is in LD (D’ = 0.5) with the disease allele
  - Both regions are in LD (D’ = 0.5) with the disease allele
Result for Scenario II

Identify the correct haplotype (Forest)

Identify SNPs in the correct region (FAMHAP)

Identify SNPs in the neutral region (FAMHAP)

Identify an incorrect haplotype (Forest)
Result for Scenario III

Identify at least one haplotype (Forest)
Identify both haplotypes (Forest)
Identify SNPs in at least one region (FAMPHAP)
Identify SNPs in both regions (FAMHAP)
Age-related macular degeneration (AMD)
Leading cause of vision loss in elderly
Affects more than 1.75 million individuals in the United States
Projected to about 3 million by 2020

Klein et al. (2005)
Case-control (96 AMD cases, 50 controls)
~100,000 SNPs for each individual
CFH gene identified
Analysis Procedure

**Willows program**

Each SNP is used as one covariate

Two SNPs identified as potentially associated with AMD (rs1329428 on chromosome 1 and rs10272438 on chromosome 7)

**Hapview program**: LD block construction

6-SNP block for rs1329428

11-SNP block for rs10272438
Two haplotypes are identified

Most significant: ACTCCG in region 1
(p-value = 2e-6)
Identical to Klein et. al. (2005)
Located in CFH gene

Another significant haplotype:
TCTGGACGACGA, in region 2 (p-value = 0.0024)
Not reported before
Protective
Located in BBS9 gene
Expected Frequencies

Haplotype 1

Haplotype 2
\[ \{M_i|y_i\} = \frac{P\{y_i\}}{P\{y_i\}} \prod_f \frac{P\{M_i|y_i, \gamma_j = 0\}}{P\{M_i\}} \prod_f \frac{P\{M_i|y_i, \beta_j, \gamma_j = 0\}}{P\{M_i\}} \prod_f \frac{P\{M_i|y_i, \beta_j, \gamma_j = 0\}}{P\{M_i\}} \] 

\[ = P\{M_i\} \prod_f \pi(\beta_j, \gamma_j = 0) P\{M_i\}\]

Note to see that \((\partial_f/\partial \beta) \pi(\beta_j, K, \gamma) = \gamma\) 

\[ \log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} \log(P\{y_i\}) + \sum_f \frac{\partial}{\partial \beta} \log(\pi(\beta_j, \gamma_j)) \]

The null hypothesis that \(\beta = 0\), we have 

\[ \frac{\partial}{\partial \beta} \log(\pi(\beta_j, \gamma_j)) \bigg|_{\beta = 0} = \sum_f [1 - \gamma(0; y_j, 1) - \gamma(0; y_j, 0)] \]

Convenience, we drop the two irrelevant 

\[ \log(P\{M_i|y_i\}) \bigg|_{\beta = 0} = \sum_f [1 - \gamma(0; y_j) - \gamma(0; y_j)] - \gamma(0; y_j) \]

The coefficient of linkage disequilibrium: 

\[ P\{AA\} - P\{dd, AA\} - P\{AA\} P\{DD\} \]
Acknowledgement

Minghui Wang, University of Science and Technology in China
Xiang Chen, St. Jude Hospital
Thank You!