Detecting Genetic Interactions in Disease

Ingo Ruczinski
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Many statistical approaches have been used in the literature for the analysis of SNP-SNP or SNP-environment interactions:

- Logistic Regression with Higher Order Interactions
- Multifactor Dimensionality Reduction (MDR)
- Classification and Regression Trees (CART)
- Random Forests
-Boosting
- Multivariate Adaptive Regression Splines (MARS)
- Neural Networks

- Logic (Boolean) Regression and Monte Carlo Logic Regression

References:

Overview

Interactions

[Lucek and Ott]

"Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect."

"...patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way ‘and’ patterns, ‘or’ patterns, and combinations of loci."

Biological and Statistical Interactions

Statistical interaction:
Deviation from additivity in a linear statistical model.

Epistasis:
Masking of phenotype expressed by one gene by the effects of another gene.


Logic Regression

- $X_1, \ldots, X_d$ are 0/1 (False/True) predictors.
- $Y$ is a response variable.
- Fit a model
  \[
  g(E(Y)) = b_0 + \sum_{j=1}^{d} b_j \cdot L_j,
  \]
  where $L_j$ is a Boolean combination of the covariates, e.g. $L_j = (X_1 \lor X_2) \land X_3^2$.
- Determine the logic terms $L_j$ and estimate the $b_j$ simultaneously.

SNPs are coded as dominant and recessive:

<table>
<thead>
<tr>
<th>SNP</th>
<th>X.R</th>
<th>X.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AT</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TT</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

References:
**The Move Set for Logic Regression**

<table>
<thead>
<tr>
<th>Possible Moves</th>
<th>Alternative Leaf</th>
<th>Regress Operator</th>
<th>Final Move</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>and</td>
<td>or</td>
<td>and</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

**Growing Logic Models**

- Initial Tree
- Parent Branch
- Input Leaf
- Terminal Leaf

- or
- and
- or
- and
- or
- and
- and so on...

**Simulated Annealing for Logic Regression**

We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^t b_j \cdot L_j$.

- Select a scoring function (RSS, log-likelihood, ...).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with $L_j = 0$ for all $j$.
- Carry out the Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending on the scores and the temperature.

**Model Selection**

We implemented two flavors for the required model selection. Both approaches require a definition of model size.

- Cross-validation:
  This is most applicable when prediction is the main objective, i.e., not necessarily the best option in SNP association studies.

- Permutation tests:
  This is a test for association, i.e., the preferred test in SNP association studies. The model size is chosen via a sequence of hypothesis tests.

Reference:

**Genetic Analysis Workshop GAW 12**

logit(affected) = $\beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{j=1}^K \beta_{w_j} \times L_j$
Monte Carlo Logic Regression

- Goal: identify all models and combinations of covariates that are potentially associated with the outcome.
- Use reversible jumps to implement an MCMC algorithm with priors on models and model size.
- The prior on model size does influence the total number of SNPs selected.
- The prior on model size has virtually no influence on the relative ordering of the SNPs or combinations thereof.

Reference:

Missing Data

- The most common approach for dealing with missing data is to omit the observations that have missing records in the model's covariates. This approach can have several shortcomings, including:
  - Loss of power.
  - Bias in the parameter estimates.

A good reference on this topic is Greenland and Finkle (1995).

- Some other used approaches are:
  - To impute a value from the marginal distribution of the covariate.
  - To create an extra level indicating missingness, if the covariate is a factor.

These choices tend to be not so great either.


Missing Data - Approaches

- Multiple imputation can be used to draw valid statistical inference from data with missing values when the data are missing at random (Little and Rubin 1987, Schaefer 1997).

  - In essence, multiple imputation acknowledges the uncertainty due to missing data, instead of simply ignoring it: several complete data sets are generated, and the uncertainty in the model parameter estimates incorporates the standard errors of the parameter estimates as well as the variability between the parameter estimates from the replicate data sets.

  - While the hypothesis of missing at random cannot formally be tested, it is a lot less stringent than the requirement of missing completely at random, which is the underlying assumption made when observations are omitted.

References:

Not Missing at Random

<table>
<thead>
<tr>
<th>Method</th>
<th>Confidence Threshold</th>
<th>Overall Call Rate</th>
<th>Hom Call Rate</th>
<th>Het Call Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>0.26</td>
<td>94.16%</td>
<td>97.24%</td>
<td>86.32%</td>
</tr>
<tr>
<td>DM</td>
<td>0.33</td>
<td>95.96%</td>
<td>98.24%</td>
<td>90.16%</td>
</tr>
<tr>
<td>BRLMM</td>
<td>0.3</td>
<td>97.40%</td>
<td>97.40%</td>
<td>97.75%</td>
</tr>
<tr>
<td>BRLMM</td>
<td>0.4</td>
<td>96.27%</td>
<td>96.30%</td>
<td>98.48%</td>
</tr>
<tr>
<td>BRLMM</td>
<td>0.5</td>
<td>96.79%</td>
<td>96.82%</td>
<td>98.93%</td>
</tr>
<tr>
<td>BRLMM</td>
<td>0.6</td>
<td>99.15%</td>
<td>99.18%</td>
<td>99.25%</td>
</tr>
</tbody>
</table>

From the “white paper” http://www.affymetrix.com/support/technical/product_updates/brlmm sounded_affx
Not Missing at Random

Multiple Imputation

We looked into two approaches:

1. Haplotype-based imputation
   - The idea here is to reconstruct the haplotypes (for example via the EM algorithm), and impute the missing values from the estimated haplotype frequencies.

2. Tree-based imputation
   - The idea here is to use decision trees to impute the genotype data, borrowing information from neighboring SNPs and other variables.


Tree-based Imputation

Example

<table>
<thead>
<tr>
<th>Number of Pairs</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPD Lys751Gln</td>
<td></td>
<td></td>
</tr>
<tr>
<td>original data</td>
<td>202</td>
<td>1.90</td>
</tr>
<tr>
<td>multiple imputations</td>
<td>321</td>
<td>1.45</td>
</tr>
<tr>
<td>XPD Gln751Gln</td>
<td></td>
<td></td>
</tr>
<tr>
<td>original data</td>
<td>202</td>
<td>2.18</td>
</tr>
<tr>
<td>multiple imputations</td>
<td>321</td>
<td>1.31</td>
</tr>
<tr>
<td>Positive Family History</td>
<td></td>
<td></td>
</tr>
<tr>
<td>original data</td>
<td>202</td>
<td>2.53</td>
</tr>
<tr>
<td>multiple imputations</td>
<td>321</td>
<td>2.53</td>
</tr>
</tbody>
</table>

Example

<table>
<thead>
<tr>
<th>Family History not complete</th>
<th>Family History complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>AC</td>
</tr>
<tr>
<td>raw numbers</td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>43</td>
</tr>
<tr>
<td>control</td>
<td>35</td>
</tr>
<tr>
<td>percentages</td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>40.2</td>
</tr>
<tr>
<td>control</td>
<td>32.7</td>
</tr>
</tbody>
</table>

Take Home Message

- Do something about the missing data!
  - They can introduce bias, and reduce the power in the analysis.

- Only using the marginal distributions for imputation doesn’t count!

- Haplotype based imputation slightly beats the tree based imputation if the genotype / phenotype relationship is best described by haplotypes, and you have few or no environmental variables.

- Tree based imputation is fast, accepts many environmental variables, and does remarkably well given its simplicity.

- Always take the response variable into account!

Reference: