Motivation

The odyssey cohort study consists of 8,394 participants who donated blood samples in 1974 and 1989 in Washington County, Maryland. The cohort has been followed until 2001, and environmental factors such as smoking and dietary intake are available. The goals of the study include finding associations between polymorphisms in candidate genes and disease (including cancer and cardiovascular disease). Particularly, gene-environment and gene-gene interactions associated with disease are of interest. Currently, SNP data from 51 genes are available.
Motivation

[With Brian Caffo, Steve Goodman, and Giovanni Parmigiani]

Associations between chromosomal deletions and stages of bladder cancer:

| STAGE | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| P4    | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Q5    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| P8    | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| P9    | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Q9    | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| P11   | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| Q13   | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| Q14   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| P17   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| Q18   | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |

Motivation

[With Tony Alberg]

Figure 2. DNA repair genotypes in relation to NMSC: cross-sectional and prospective cohort comparisons

2007

No history of melanoma or noncutaneous malignancies

Prospective cohort comparison of DNA repair genotypes in relation to NMSC risk

First time diagnosis of NMSC prior to other cancer N=493

No history of NMSC N=27,279

1989

Cross-sectional comparison of DNA repair genotypes in 1989

NMSC N=709

NMSC N=27,772

1989

2007
Table 1. Amino Acid Substitution Variants Identified in DNA Repair and Repair-Related Genes (Source: Mohrenweiser et al 2002, and Goode et al 2002)

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Exon</th>
<th>Codon</th>
<th>Common Residue</th>
<th>Variant Residue</th>
<th>Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base Excision Repair</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADPRT</td>
<td>17</td>
<td>761</td>
<td>Val</td>
<td>Ala</td>
<td>0.18</td>
</tr>
<tr>
<td>APE1</td>
<td>5</td>
<td>148</td>
<td>Asp</td>
<td>Glu</td>
<td>0.33</td>
</tr>
<tr>
<td>OGG1</td>
<td>7</td>
<td>326</td>
<td>Ser</td>
<td>Cys</td>
<td>0.15-0.45</td>
</tr>
<tr>
<td>OGG1</td>
<td>Nucleotide 7143*</td>
<td>A</td>
<td>G</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>OGG1</td>
<td>Nucleotide 11657*</td>
<td>A</td>
<td>G</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>POLD1</td>
<td>1</td>
<td>19</td>
<td>Arg</td>
<td>His</td>
<td>0.12</td>
</tr>
<tr>
<td>POLD1</td>
<td>3</td>
<td>119</td>
<td>Arg</td>
<td>His</td>
<td>0.15</td>
</tr>
<tr>
<td>POLD1</td>
<td>4</td>
<td>173</td>
<td>Ser</td>
<td>Asn</td>
<td>0.05</td>
</tr>
<tr>
<td>XRC1</td>
<td>6</td>
<td>194</td>
<td>Arg</td>
<td>Trp</td>
<td>0.13</td>
</tr>
<tr>
<td>XRC1</td>
<td>10</td>
<td>399</td>
<td>Arg</td>
<td>Gln</td>
<td>0.24</td>
</tr>
<tr>
<td><strong>Nucleotide Excision Repair</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERCC2</td>
<td>10</td>
<td>312</td>
<td>Asp</td>
<td>Asn</td>
<td>0.4</td>
</tr>
<tr>
<td>ERCC2</td>
<td>23</td>
<td>751</td>
<td>Lys</td>
<td>Gln</td>
<td>0.32</td>
</tr>
<tr>
<td>ERCC4</td>
<td>8</td>
<td>415</td>
<td>Arg</td>
<td>Gln</td>
<td>0.06</td>
</tr>
<tr>
<td>ERCC5</td>
<td>15</td>
<td>1104</td>
<td>Asp</td>
<td>His</td>
<td>0.18</td>
</tr>
<tr>
<td>RAD23B</td>
<td>7</td>
<td>249</td>
<td>Ala</td>
<td>Val</td>
<td>0.10</td>
</tr>
<tr>
<td>XPC</td>
<td>8</td>
<td>499</td>
<td>Ala</td>
<td>Val</td>
<td>0.24</td>
</tr>
<tr>
<td>XPC</td>
<td>15</td>
<td>939</td>
<td>Lys</td>
<td>Gln</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Double Strand Break/Recombination Repair</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBS1</td>
<td>5</td>
<td>185</td>
<td>Gln</td>
<td>Gln</td>
<td>0.34</td>
</tr>
<tr>
<td>XRCC2</td>
<td>3</td>
<td>188</td>
<td>Arg</td>
<td>His</td>
<td>0.05</td>
</tr>
<tr>
<td>XRCC3</td>
<td>7</td>
<td>241</td>
<td>Thr</td>
<td>Met</td>
<td>0.43</td>
</tr>
<tr>
<td>XRCC4</td>
<td>5</td>
<td>247</td>
<td>Ala</td>
<td>Ser</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Damage Recognition, Repair and Cell Cycle Check point</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDKN2A</td>
<td>2</td>
<td>148</td>
<td>Ala</td>
<td>Thr</td>
<td>0.05</td>
</tr>
<tr>
<td>RAD52</td>
<td>8</td>
<td>287</td>
<td>Ser</td>
<td>Asn</td>
<td>0.05</td>
</tr>
<tr>
<td>MLH1</td>
<td>8</td>
<td>219</td>
<td>Ile</td>
<td>Val</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Mismatch Repair</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH3</td>
<td>10</td>
<td>514</td>
<td>Gln</td>
<td>Lys</td>
<td>0.05</td>
</tr>
<tr>
<td>MSH3</td>
<td>21</td>
<td>940</td>
<td>Arg</td>
<td>Gln</td>
<td>0.1</td>
</tr>
<tr>
<td>MSH3</td>
<td>23</td>
<td>1036</td>
<td>Thr</td>
<td>Ala</td>
<td>0.3</td>
</tr>
<tr>
<td>MSH6</td>
<td>1</td>
<td>39</td>
<td>Gly</td>
<td>Gln</td>
<td>0.24</td>
</tr>
</tbody>
</table>

* Amino acid substitution variants of these SNPs have not been published. However, these nucleotide substitutions occur in a gene of particular interest (see section B.2.c.), and have been found to associate strongly with risk of prostate cancer. (Goode et al 2002)

Motivation

Lucek and Ott (1997):

“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.”
Trees versus Rules


Classification versus Regression

Classification: ID3, C4.5, CN2, SLIQ, RIPPER, SLIPPER, SWAP1, and a gazillion of derivations of the former methods (Apte, Damerau, and Weiss 1994).

Regression: SWAP1R (Weiss and Indurkhya 1993b plus an extension by Torgo and Gama 1996), R2, M5, treed models (Chipman, George, and McCulloch 2002), CUBIST (Quinlan).

Both: CART, MARS (Friedman 1991).
Objectives for Format of Solutions


Worth having a look at: Wnek and Michalski (1994) compare a decision tree learning method (C4.5), a rule-learning method (AQ15), a neural net trained by a back-propagation algorithm (BpNet) and a classifier system using a genetic algorithm (CFS) with respect to their predictive accuracy and simplicity of solutions.

Alternatives to Straight Greedy Searches

Probablistic searches:


Alternatives to Straight Greedy Searches

Statistical approaches:

Bayesian CART (Chipman et. al. 1998, Denison et. al. 1998), EM algorithm (Jordan and Jabocs 1994), bagging (Breiman 1996), bumping (Tibshirani and Knight 1999), boosting (Freund and Schapire 1996), randomized decision trees (Amit and Geman 1997; Dietterich 1999), PRIM (Friedman and Fisher 1999).

Comparisons of some of those methods were for example carried out by Quinlan (1996), Dietterich (1999), and Breiman (1999).

Logic Regression

\[ X_1, \ldots, X_k \text{ are 0/1 (False/True) predictors.} \]

\[ Y \text{ is a response variable.} \]

Fit a model \( g(E(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j \), where \( L_j \) is a Boolean combination of the covariates, e.g. \( L_j = (X_1 \lor X_2) \land X_4^c \).

Determine the logic terms \( L_j \) and estimate the \( \beta_j \) simultaneously.
Logic Trees

An equivalent representation of \((X_1 \land X_2^c) \lor (X_3 \land (X_1^c \lor X_4))\) is the following:

```
  or
  /   \
and    and
  /     /
1  2   3

or
  /   \
and    and
  /     /
1  4
```

This is a Logic Tree!

Comparison to Decision Trees

```
Decision Tree
  C
  /\  \/
 A  B
  /\  \/
 D  A
  /\  \/
0  B
  /\  \/
0 1

Logic Tree
  or
  /   \
and    and
  /     /
A  B  C  D
```

A Decision Tree (CART) is something different!
Simulated Annealing for Logic Regression

We try to fit the model
\[ g(E(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j. \]

- Select a scoring function (RSS, log-likelihood, \ldots).
- 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.
Bladder Cancer Example

The diagram shows a graph with size on the x-axis and distance on the y-axis. The graph includes several lines and points, indicating relationships between size and distance for different conditions or questions labeled as Q5, Q14, Q18, and P4. The relationships are depicted with logical operators such as 'and' and 'or'. The graph illustrates how size and distance vary across different conditions, possibly indicating a trend or pattern in the data related to bladder cancer.
A Global Randomization Test of Association

A Sequential Randomization Test for Model Size
A Sequential Randomization Test for Model Size

Sequential Randomization Test for 2 Trees:

<table>
<thead>
<tr>
<th>X</th>
<th>T1</th>
<th>T2</th>
<th>Y</th>
<th>Perm(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>permutation</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>permutation</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>permutation</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>permutation</td>
</tr>
</tbody>
</table>

Diagram showing sequential randomization test with histograms and permutation arrows.
logit(affected) = $\beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^{K} \beta_{i+3} \times L_i$
Let $\gamma_S$ be the score of a certain state $S$.

- We use the acceptance function
  \[ \alpha(\gamma_{\text{old}}, \gamma_{\text{new}}, t) = \min\{1, \exp([\gamma_{\text{old}} - \gamma_{\text{new}}]/t)\} \]

- If we keep the temperature constant, this defines a homogeneous Markov chain.

- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$. 
Simulate 10 binary predictors $X_1, \ldots, X_{10}$.

Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon$, $\epsilon \sim N(0,1)$.

Run a homogeneous Markov chain during “crunch time” for two separate cases:

Case 1: All $X$ are independent.

Case 2: All $X$ are independent, except $X_4$ (in the signal) and $X_5$ (not in the signal), which are heavily correlated.
Multiple Models

Statistical Issues: Missing Values
Statistical Issues: Power

$n_1 = 709 \quad n_2 = 27772$

$n = 254$

$n = 493$

$n = 2304$

Statistical Issues: Power
References


The Bibliography of this paper contains all the references in this presentation.