

## Assessing Interactions in High-Dimensional Genomic Data

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Email: ingo@jhu.edu. The slides and software used for this presentation are available at <http://biostat.jhsph.edu/~iruczins>

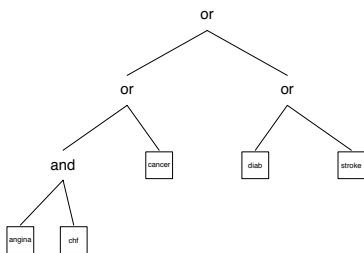
## A Public Health Related Example

- The Women's Health and Aging Study (WHAS) began in 1992 to study the causes and the course of disability in moderately to severely disabled older women living in the community.
- The WHAS is a population-based longitudinal study of women with at least mild disability, 65 years of age or older, living at home in eastern Baltimore city or county.
- 1002 women agreed to participate and provided written informed consent.
- The major chronic diseases at baseline were ascertained by using complex algorithms. Follow-up evaluations were conducted every 6 months for 3 years.
- There is evidence that disability results from chronic diseases, and that interactions between diseases (comorbidities) are of importance in causing disability.
- The chronic diseases recorder included cancer, congestion heart failure, diabetes, degenerative disc disease, hip fracture, myocardial infarction, arthritis, osteoporosis, Parkinson's disease, pulmonary disease, stroke.

## A Public Health Related Example

$$p = \Pr(\text{death in round } j \mid \text{survival to round } j-1, X, \text{age})$$

$$\text{logit}(p) = -9.01 + 0.06 \cdot \text{age} + 1.07 \cdot L(X)$$



## Motivation

[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."

## Logic Regression

- $X_1, \dots, X_k$  are 0/1 (False/True) predictors.

- Y is a response variable.

- Fit a model

$$g(E(Y)) = b_0 + \sum_{j=1}^t b_j \cdot L_j$$

where  $L_j$  is a Boolean combination of the covariates, e.g.  $L_j = (X_1 \vee X_2) \wedge X_3^c$ .

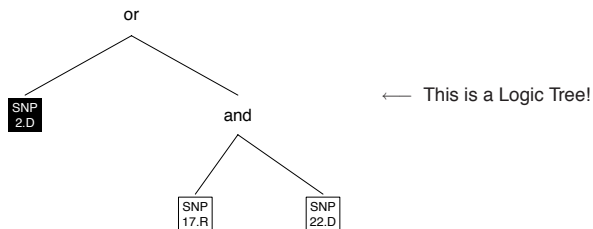
- Determine the logic terms  $L_j$  and estimate the  $b_j$  simultaneously.

SNP X	X.R	X.D
AA	0	0
AT	0	1
TT	1	1

- SNPs are usually coded as dominant and recessive:

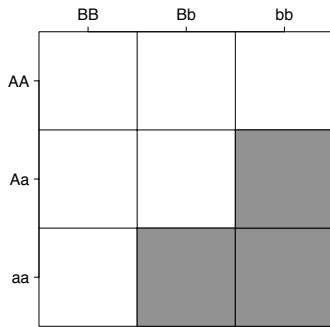
## Logic Trees

An equivalent representation of  $\text{SNP2.D}^c \vee (\text{SNP17.R} \wedge \text{SNP22.D})$  :



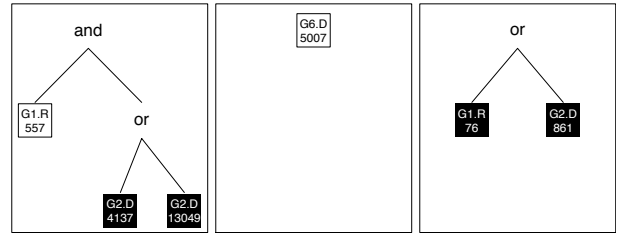
SNP 2	2.R	2.D	2.D <sup>c</sup>	SNP 17	17.R	17.D	SNP 22	22.R	22.D
AA	0	0	1	CC	0	0	GG	0	0
AT	0	1	0	CT	0	1	GT	0	1
TT	1	1	0	TT	1	1	TT	1	1

## An Example

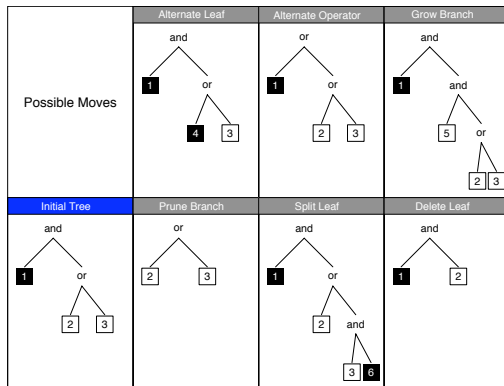


## Example: GAW 12

$$\text{logit}(\text{affected}) = \beta_0 + \beta_1 \times X_{\text{ENV}_1} + \beta_2 \times X_{\text{ENV}_2} + \beta_3 \times X_{\text{Gender}} + \sum_{i=1}^K \beta_{i+3} \times L_i$$



## The Move Set for Logic Regression

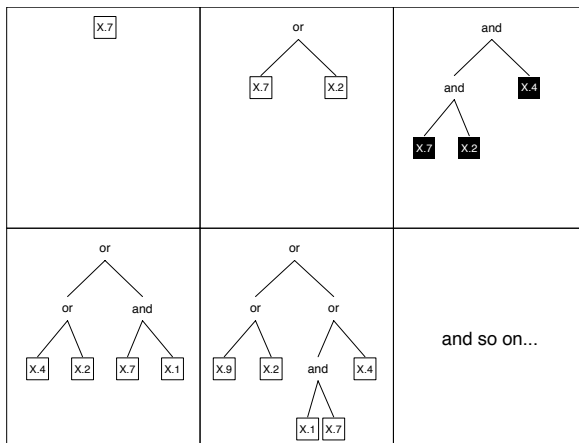


## Simulated Annealing for Logic Regression

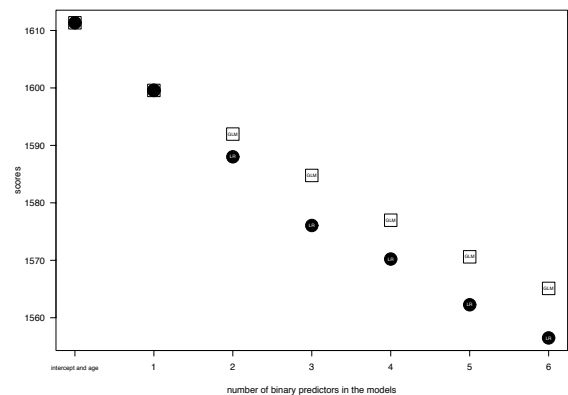
We try to fit the model  $g(E(Y)) = b_0 + \sum_{j=1}^L b_j \cdot L_j$ .

- 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.

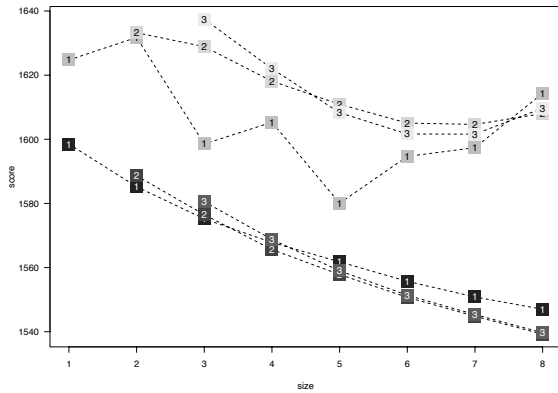
## Growing Logic Models



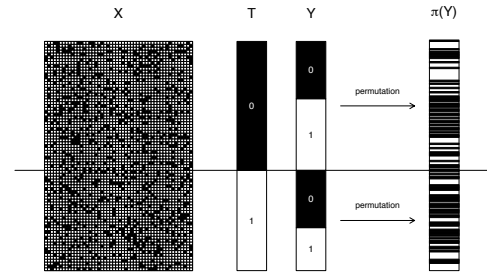
## A Public Health Related Example



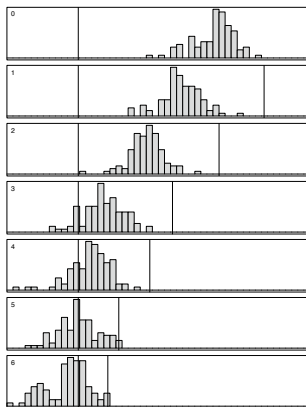
## Model Selection 1 : Cross Validation



## Model Selection 2 : Permutation Tests



## Model Selection 2 : Permutation Tests

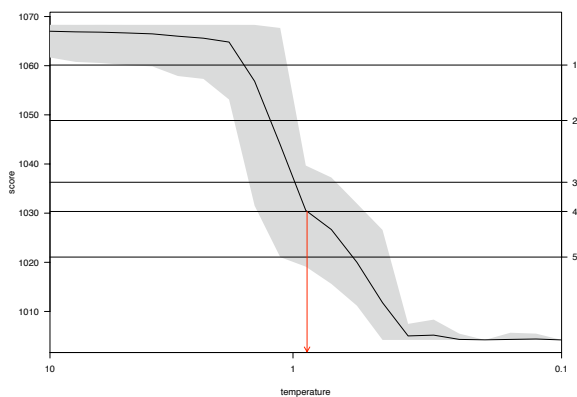


## Multiple Models 1 : Monte Carlo LR

- 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.

Kooperberg C, Ruczinski I. Identifying Interacting SNPs using Monte Carlo Logic Regression, Genet. Epidemiol, 28(2): 157-170, 2005.

## Multiple Models 2 : Metropolis-Hastings



## Multiple Models 2 : Metropolis-Hastings

Let  $\gamma_S$  be the score of a certain state  $S$ .

- We use the acceptance function
 
$$\alpha(\gamma_{old}, \gamma_{new}, t) = \min\{1, \exp([\gamma_{old} - \gamma_{new}]/t)\}$$
- 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  $\pi_t(S)$ .
- If we know the model size where the signal ends and the noise starts, we can read off the corresponding temperature from the diagnostic plot!

## Multiple Models 2 : Metropolis-Hastings

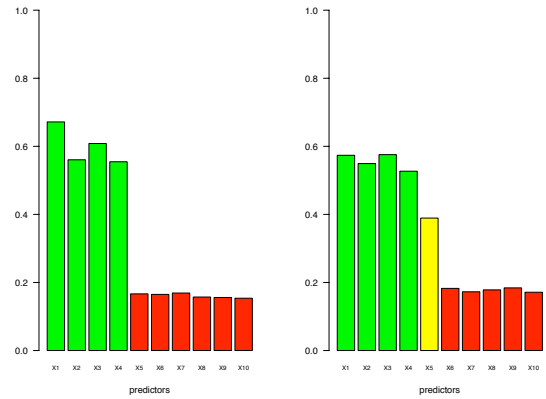
Example: Simulate 10 binary predictors  $X_1, \dots, X_{10}$ .

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

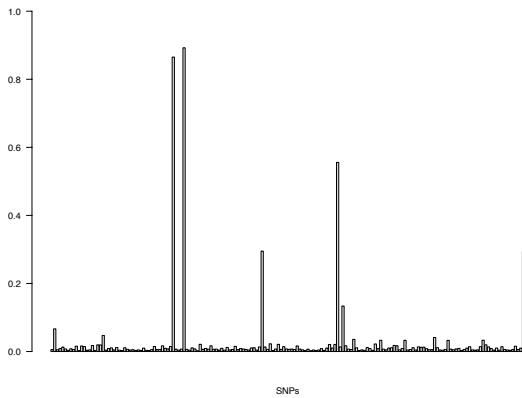
Run a homogeneous Markov chain during "crunch time" for two separate cases:

Case 1: All  $X$  are independent.

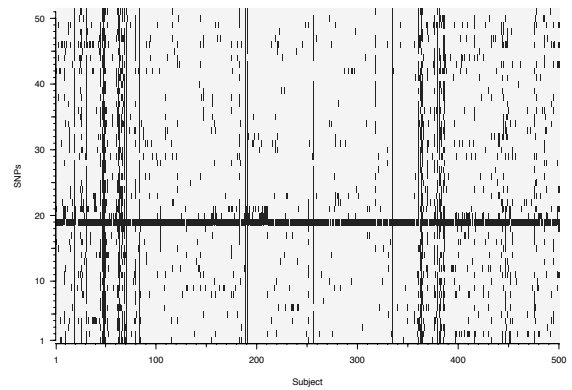
Case 2: All  $X$  are independent, except  $X_4$  (in the signal) and  $X_5$  (not in the signal), which are heavily correlated.



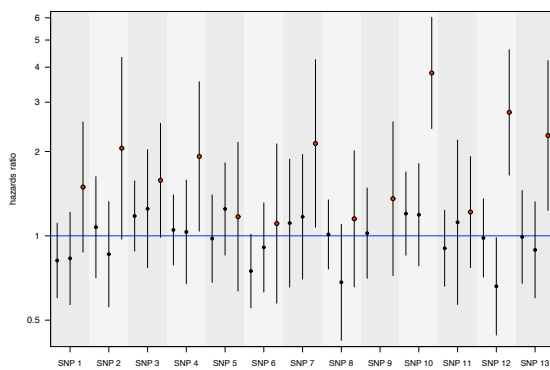
## Multiple Models 2 : Metropolis-Hastings



## Missing Data



## Missing Data



## Missing Data

The polymorphisms of the DNA repair gene XPD (751) for case/control pairs and family history reporting status, in a breast cancer study:

	missing				not missing			
	AA	AC	CC	na	AA	AC	CC	na
raw numbers								
case	43	54	5	5	61	121	25	7
control	35	57	12	3	90	102	22	0
percentages								
case	40.2	50.5	4.7	4.7	28.5	56.5	11.7	3.3
control	32.7	53.3	11.2	2.8	42.1	47.7	10.3	0.0

## Missing Data

---

Number of Pairs    Odds Ratio    Confidence Interval

### XPD Gln751Lys

original data set	202	1.90	( 1.20 – 3.00 )
multiple imputations	321	1.45	( 1.00 – 2.10 )

### XPD Gln751Gln

original data set	202	2.18	( 1.08 – 4.40 )
multiple imputations	321	1.31	( 0.74 – 2.34 )

### Positive Family History

original data set	202	2.53	( 1.43 – 4.50 )
multiple imputations	321	2.53	( 1.58 – 4.03 )

<http://biostat.jhsph.edu/~iruczins>