Logic regression based methods in application to detection of gene-gene interactions

Magdalena Malina

Institute of Mathematics, University of Wrocław, Poland

Małgorzata Bogdan

Wrocław University of Technology, Poland

in cooperation with
Katja Ickstadt, Holger Schwender – Department of Statistics, University of Dortmund, Germany

Poznań, 18.05.2010

PURPOSE AND METHOD

- A quantitative trait locus (QTL) is a position on a chromosome (i.e.locus) that contributes to a quantitative trait.

Main purpose: finding the mutations in DNA sequence, that influence the feature of interests.

- second or higher order interactions between QTLs

Method: Logic regression (Ruczinski, Kooperberg, LeBlanc)

- Developed to identify complex interactions in genetic data;
- Designed for data with many binary predictors
- Predictors are Boolean combinations of binary variables - logic expressions:
  - Combinations of binary predictors X_i's obtained by using logical operators $\wedge$ (AND), $\vee$ (OR) and ^{c} (NOT) (ex. $L = (X_1 \wedge X_2) \vee X_3^c$)
  - Binary predictors

Genotype	X_d	X_r	
AA (0)	0	0	Homozygous Reference
aA (1)	1	0	Heterozygous
aa (2)	1	1	Homozygous Variant

$X_1, X_2, \ldots, X_k$ -binary explanatory variables, Y - feature of interest.

- In **logic regression** one considers the model of the form

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j,$$

L_j - logic expression consisted of predictors $X_i, i = 1, 2, \ldots, k$.

- A generalized linear regression model :

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j X_j + \sum_{(i,j)\in I} \gamma_{(i,j)} X_i \cdot X_j, \quad I = \{(i,j) : i,j = 1, 2, \ldots, k\}$$

  - $g(E(Y)) = E(Y)$ - linear regression model
  - $g(E(Y)) = \log(\frac{E(Y)}{1-E(Y)})$ - logistic regression model

Example

- **F**or $X_1, X_2, \ldots, X_k$ - binary variables and $Y \sim \mathcal{N}(0,1)$
  - Logic regression model:

$$g(E(Y)) = \beta_1 \cdot X_1 \vee (X_2 \wedge X_3)$$

  - Logistic regression model for products :
    $g(E(Y)) = \gamma_1 \cdot X_1 + \gamma_2 \cdot X_2 * X_3 - \gamma_3 \cdot X_1 * X_2 * X_3$

Logic Feature Selection (*Schwender, Ickstadt*(2007))

- Identification of single 'best' model made by *simulated annealing algorithm*
- Logic expressions in Disjunctive Normal Form - a $\vee$-combination of $\wedge$-combinations

$$L = (X_1 \vee X_2) \wedge (X_3 \vee X_4) = (X_1 \wedge X_3) \vee (X_1 \wedge X_4) \vee (X_2 \wedge X_3) \vee (X_2 \wedge X_4)$$

- B logic regression models on B bootstrap samples constructed
- Appropriate importance measure for each prime implicant (interaction) $P_g, g = 1, 2, \ldots, G$ computed

Importance measure

- If a model with few trees is considered:

$$VIM_{Multiple}(P_g) = \frac{1}{B} \sum_{b:P_g \in \alpha_b} (N_b - N_b^{-g}).$$

  - N_b- number of oob observations properly classified by the b-th regression model,
  - N_b^{-g}- number of oob observations properly classified by the b-th regression model , when P_g is removed from the model.
  - α_b -a set of all prime implicants in b-th iteration.

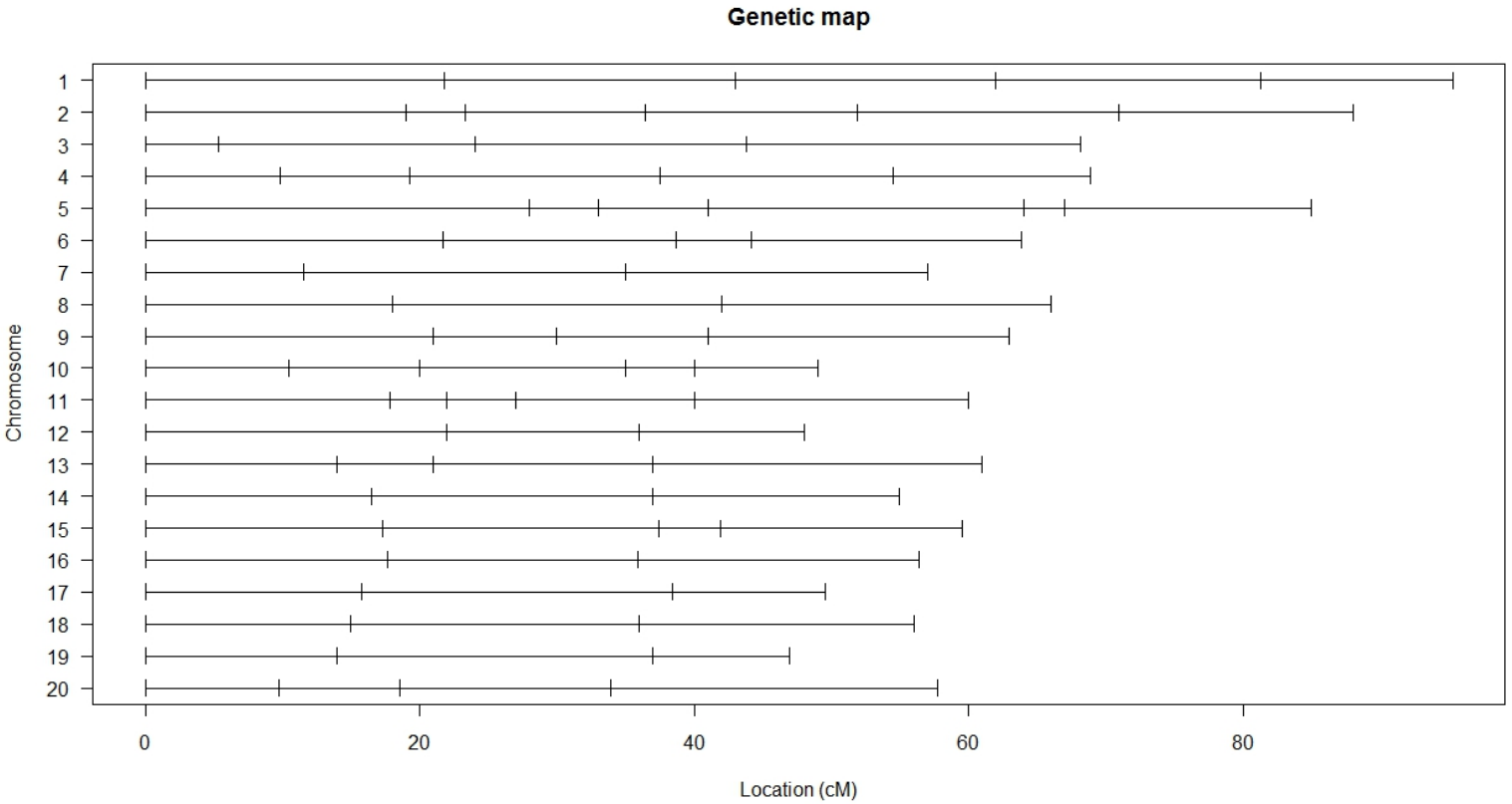
- If the response is a ***quantitative variable***

$$VIM(S_i) = \frac{1}{B} \sum_{b=1}^{B} (\log(MSPE_b^{(-S_i)}) - \log(MSPE_b)),$$

- a natural cutoff for calling S_i important, $i = 1, 2, \ldots, m$:
  - $1 - \frac{0.05}{m}$ quantile of the t distribution with $B - 1$ degrees of freedom (Bonferroni correction used)

Simulation study

- We simulate genotypes of 100 markers on the map given in http://phenome.jax.org/phenome/protodocs/QTL/QTL-Lyons3.xls



- 50 iterations for each model

Simulated model results

We considered the model with $\varepsilon \sim \mathcal{N}(0,1)$:

$$y = \beta_0 + \beta_1(D1Mit21(1) \wedge D2Mit206(2)^C) + \beta_2(D3Mit57(1) \wedge D4Mit17(1)) +$$

$$+ \beta_3(D6Mit46(2) \wedge D6Mit38(2)) + \beta_4(D1Mit17(1) \wedge D4Mit42(1)) + \varepsilon,$$

- $\beta_0 = -0.45, \quad \beta_1 = 1.2, \quad \beta_2 = 1.3, \quad \beta_3 = 1.4, \quad \beta_4 = 1.5$

Interaction	Mean Imp.	No.	%	First positions
$D1Mit21(1) \wedge D2Mit206(2)^C$	0.015277	46	0.92	3(4),4(3),5(4)
$D3Mit57(1) \wedge D4Mit17(1)$	0.062278	49	0.98	1(12),2(11),4(8)
$D6Mit46(2) \wedge D6Mit38(2)$	0.000861	21	0.42	13(2),24,38
$D1Mit17(1) \wedge D4Mit42(1)$	0.1169	50	1	1(30),2(9),3(6)

Real data analysis : QTL-Lyons3 data

- We consider the data by Lyons et.al.(2003): phenotypes related to cholesterol gallstones formation in an intercross of CAST/Ei and 129S1/SvImJ inbred mice
  http://phenome.jax.org/phenome/protodocs/QTL/QTL-Lyons3.xls
- Phenotype : bile score characteristics (0-1)
- The significance level $\alpha = 0.05$; for additive effects: 0.000125 , for two-way interactions : $6.23441 \cdot 10^{-7}$ (Bonferroni corr.).

	Imp.	Prop.	Expression	p-value
1	0.846	0.237	!D2Mit113(1)	$2.85 \cdot 10^{-5}$
2	0.357	0.077	!D2Mit113(1) : !D16Mit65(2)	$1.34 \cdot 10^{-8}$

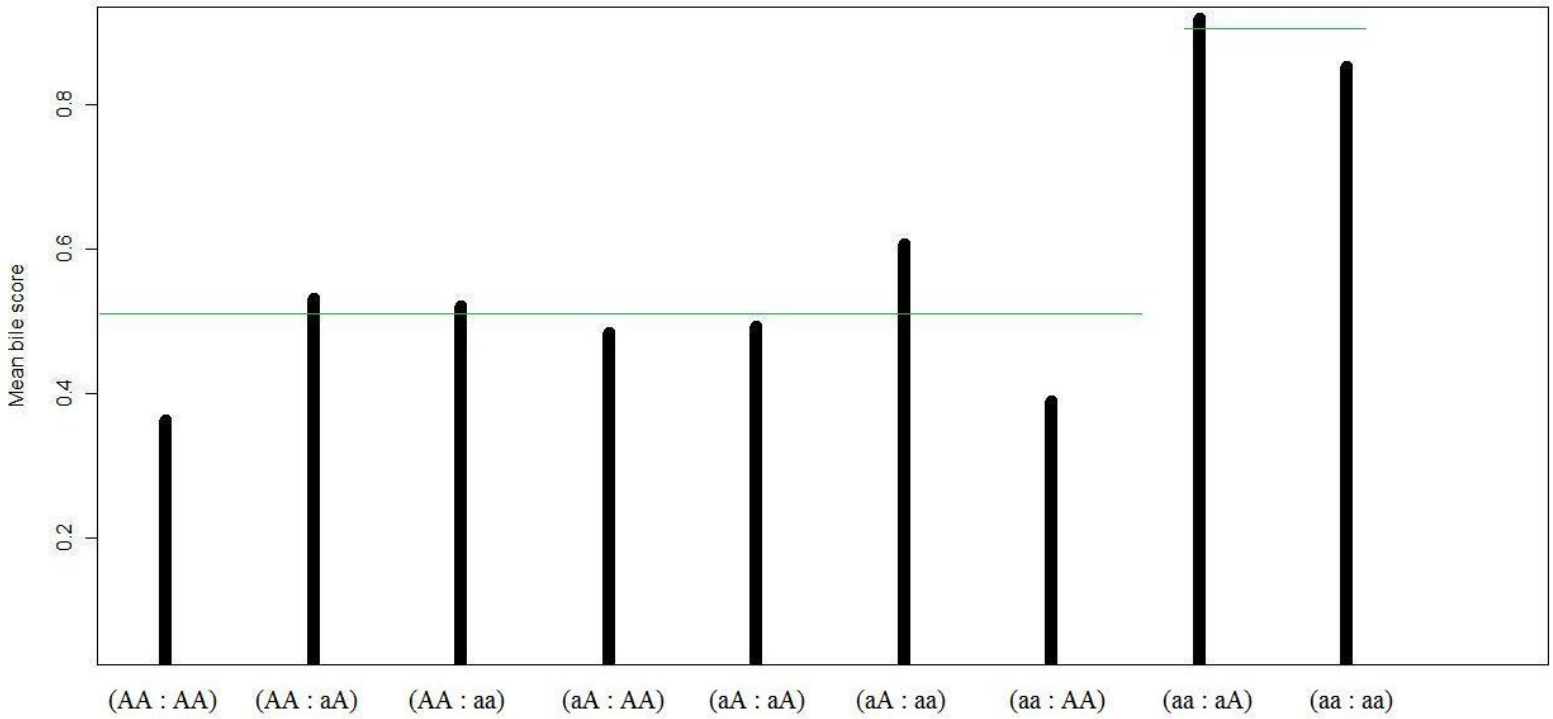
- One significant additive effect: $!D2Mit113(1)$
- One two-way interaction : $!D2Mit113(1) \wedge !D16Mit65(2)$
- One *'suggestive'* interaction: $!D2Mit113(1) \wedge !D15Mit79(2)$ (p-value of a single t-test $8.56 \cdot 10^{-6}$)

- For the interaction !D2Mit113(1) : !D16Mit65(2)

	Estimate	Std.Error	t value	$Pr(>\|t\|)$
(Intercept)	0.45238	0.07175	6.305	$1.16 \cdot 10^{-9}$
!D2Mit113(1)	-0.06349	0.13100	-0.485	0.62831
!D16Mit65(2)	0.07703	0.08101	0.951	0.34248
!D2Mit113(1):!D16Mit65(2)	0.42471	0.14817	2.866	0.00448

	Estimate	Std.Error	t value	$Pr(>\|t\|)$
(Intercept)	0.50235	0.03187	15.764	$< 2 \cdot 10^{-16}$
!D2Mit113(1):!D16Mit65(2)	0.38828	0.06630	5.857	$1.34 \cdot 10^{-8}$

- (aa):(aA) or (aa):(aa) - significant increase in the mean bile score:



QTL-Stylianou1 data

- QTL-Stylianou1 - obesity in an intercross of SM/J and NZB/BINJ http://churchill.jax.org/datasets/qtlarchive/fatpads.shtml
- 151 markers on 20 chromosomes, 513 individuals
- Phenotype : gonadal fat pad weight
- Significance level $\alpha = 0.05$; for additive effects: $8.278146 \cdot 10^{-5}$, two-way interactions : $2.736577 \cdot 10^{-7}$; three-way : $1.361459 \cdot 10^{-9}$

Significant effects found:

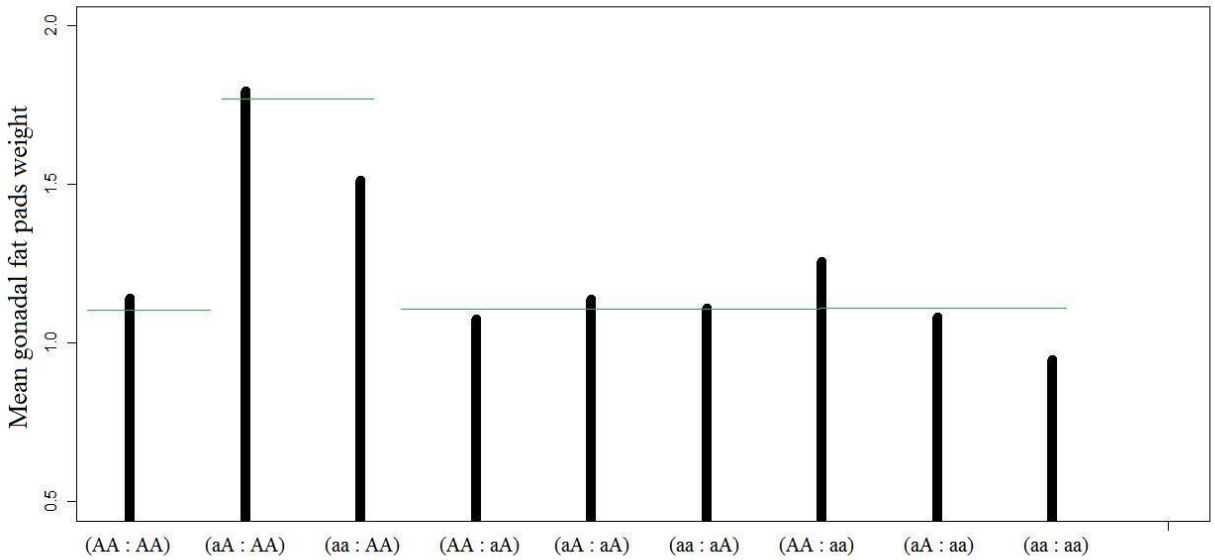
- 6 main effects on 17^{th} and 19^{th} chromosomes (3 effects on each)
- 18 two way interactions : 6 between 3^{rd} and 19^{th}, 6 between 5^{th} and 19^{th}, 3 between 17^{th} and 19^{th}, 3 between 12^{th} and 19^{th};
- One three way: $!D19Mit71(1) \wedge D3Mit131(1) \wedge DXMit1(1)$

- !D19Mit71(1):D16Mit32(1) - new significant two-way interaction

	Estimate	Std.Error	t value	$Pr(>\|t\|)$
(Intercept)	1.135346	0.080777	14.055	$< 2 \cdot 10^{-16}$
!D19Mit71(1)	0.006393	0.176652	0.036	0.97115
D16Mit32(1)	-0.042055	0.091609	-0.459	0.64638
!D19Mit71(1):D16Mit32(1)	0.628552	0.196995	3.191	0.00151

	Estimate	Std.Error	t value	$Pr(>\|t\|)$
(Intercept)	1.10482	0.03697	29.887	$< 2 \cdot 10^{-16}$
!D19Mit71(1):D16Mit32(1)	0.62342	0.08415	7.408	$5.32 \cdot 10^{-13}$

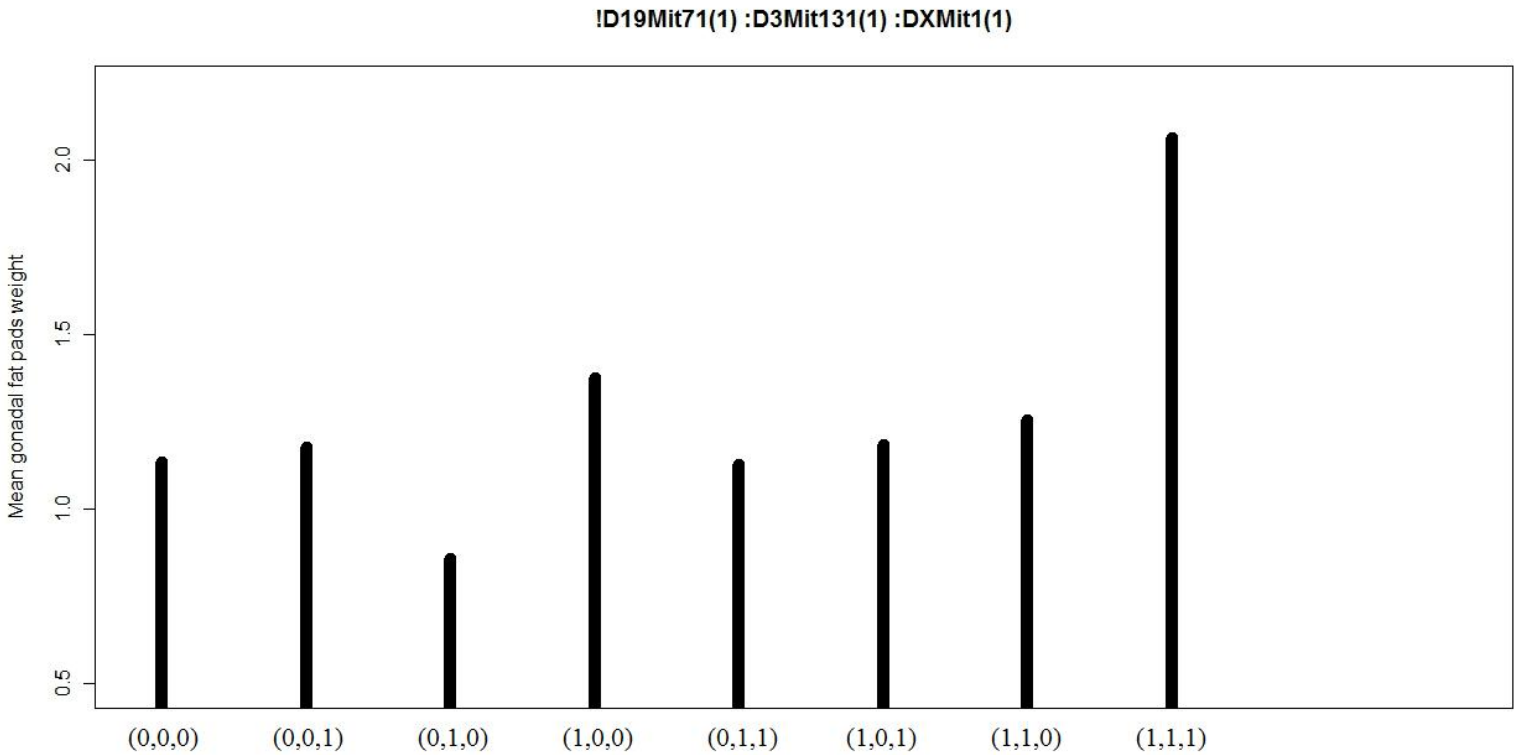
- (aA):(AA) or (aa):(AA) - significant increase in phenotype



QTL-Stylianou1 data

- The three-way interaction !D19Mit71(1) : D3Mit131(1) : DXMit1(1)

	Estimate	t value	$Pr(>\|t\|)$
(Intercept)	1.1263	32.740	$< 2 \cdot 10^{-16}$
!D19Mit71(1):D3Mit131(1):DXMit1(1)	0.9385	8.851	$< 2 \cdot 10^{-16}$



Bibliography

[1]Lyons MA, Wittenburg H, Li R, Walsh KA, Leonard MR, Churchill GA, Carey MC, Paigen B. *New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice.* Physiol Genomics. 2003 Aug 15;14(3):225-39. PMID 12837957

[2]Stylianou IM, Korstanje R, Li R, Sheehan S, Paigen B, Churchill GA. *Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci.*, Mamm Genome. 2006 Jan;17(1):22-36. PMID 16416088

[3] Schwender, H., Ickstadt, K. (2008). *Identification of SNP interactions using logic regression.* Biostatistics, 9. 187-198

[4]Ruczinski I., Kooperberg C., LeBlanc M., *Logic regression*, J. Comput. Graphical Statist. 12 (3),(2003),474-511

[5]Kooperberg C., Ruczinski I., *Identifying Interacting SNPs Using Monte Carlo Logic Regression*, Genetic Epidemiology 28, 157-170 (2005)

R packages used :

[6] Karl W Broman and Hao Wu, *R/qtl Analysis of experimental crosses to identify genes contributing to variation in quantitative traits.*

[7] Holger Schwender, *R/logicFS , Identification of SNP Interactions*