Optimal definition of high and low risk cells in MDR analysis

Understanding the effects of genes and environmental factors on the development of complex diseases, such as cancer, is a major aim of genetic epidemiology. These kinds of diseases are controlled by complex molecular mechanisms characterised by the joint action of several genes, each having only a small effect. In this context traditional methods involving single markers have limited use and more advanced and efficient methods are needed to identify gene interactions and epistatic patterns of susceptibility (Figure 1).

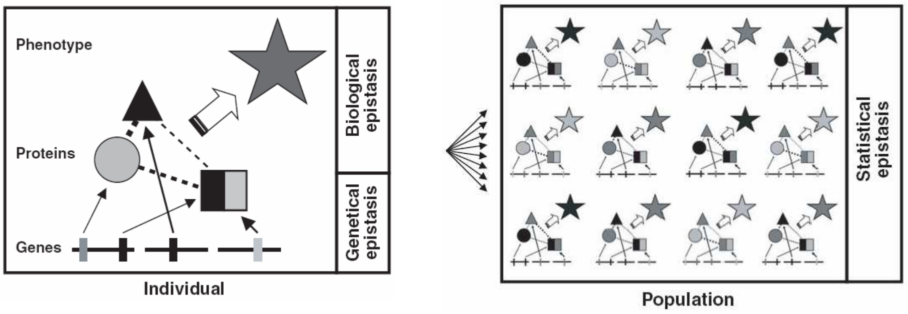

Figure1: Epistasis

Standard methods to analyse case-control data in this context broadly fall into two classes: parametric multi-locus methods including regression (e.g., Park and Hastie 2007) and (bagged) logic regression (Ruczinski et al., 2004) or non-parametric multi-locus techniques such as most machine learning and data mining approaches. Several data mining methods have been used for interaction detection such as tree-based methods (e.g., Recursive Partitioning and Random Forests), pattern recognition methods (e.g., Symbolic Discriminant Analysis, Mining association rules, Neural networks and Support vector machines), and data reduction methods (e.g., Detection of Informative Combined Effects, Multifactor Dimensionality Reduction and Logic regression). A nice overview is given by Onkamo and Toivonen (2006).

Whereas the aforementioned non-parametric approaches are appealing because no distributional assumptions are imposed on the genotype-phenotype effect, parametric approaches have severe limitations when there are too many independent variables in relation to the number of observed outcome events. However, when analyzing gene interactions in case-control studies adjustment for confounding variables and for main effects is usually required and parametric methods might be more flexible.

In this project you will focus attention on the Multifactor Dimensionality Reduction method, MDR, (Ritchie et al. 2001; Figure 2) that has recently achieved a great popularity. The MDR strategy tackles the dimensionality problem related to interaction detection and reduces the multiple dimensions to one by pooling multi-locus genotypes into two groups of risk: high and low. Although MDR has proven its usefulness and has some nice properties, it suffers from some major drawbacks including that some important interactions could be missed due to pooling too many cells together when defining high and low risk cells. Hence, the main goal is to come up with alternative definitions to define high and low risk cells, and to compare these via simulations in R with the existing definition. R code on particular steps in the classical MDR approach is already available.

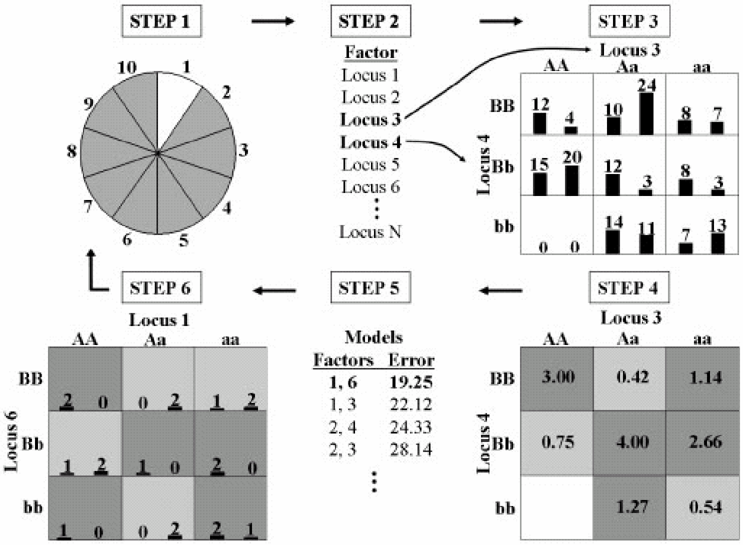

Figure 2: The six steps involved in MDR methodology. The dark shaded cells in step 4 are high risk cells, the light shaded cells are low-risk (Rithie et al 2001).

Bibliography

- Onkamo P and Toivonen H. A survey of data mining methods for linkage disequilibrium mapping (2006). *Human Genomics,* 2 (5): 336-340.
- Park MY, Hastie T (2007) Penalized logistic regression for detecting gene interactions. Biostatistics (advance access pub April 11, 2007).
- Ritchie, M. D.; Hahn, L. W.; Roodi, N.; Bailey, L. R.; Dupont, W. D.; Parl, F. F. & Moore, J. H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am.J.Hum.Genet.,* 2001, 69, 138-147.
- Ritchie, M. D.; Hahn, L. W. & Moore, J. H.
  Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet.Epidemiol., Wiley-Liss, Inc,* 2003, 24, 150-157.
- Ruczinski I., Kooperberg C. and LeBlanc M.L. Exploring interactions in high-dimensional genomic data: an overview of LogicRegression. Journal of Multivariate Analysis, 2004, 90, 178–195.

Renseignements: Kristel Van Steen

Promoteur: Kristel Van Steen