Feature selection as pre-screening tool for multifactor dimensionality reduction

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 or epistatic patterns.

The Multifactor Dimensionality Reduction method, MDR, (Ritchie et al. 2001) 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. It is an attractive technique to detect gene-gene interaction in case-control studies because it allows for the detection of multiple genetic loci jointly associated with a discrete clinical endpoint in the absence of a main effect, it is non-parametric in nature, no assumptions need to be formulated about the underlying genetic inheritance model, it generates low false positive rates.

Figure 1 shows the six steps involved in a classical MDR analysis. In step 2, the starting point is a subselection of N genetic markers from the initial pool that can be as large as 1,000,000 markers. Taking a subselection is essential because of computational considerations: it makes a large difference to investigate all possible couples or trios of markers in a group of N=250 or N=1,000,000 markers! However, whether or not being successful in detecting higher-order genetic interactions, using N markers only, may heavily depend on the choice of the subset. The topic of this project is to investigate several existing strategies to select favorable combinations of features (for instance wrapper models or filter models including the TuRF method of Moore and White 2007 and Bayesian modeling of genetic associations as performed by Sebastiani et al 2008).

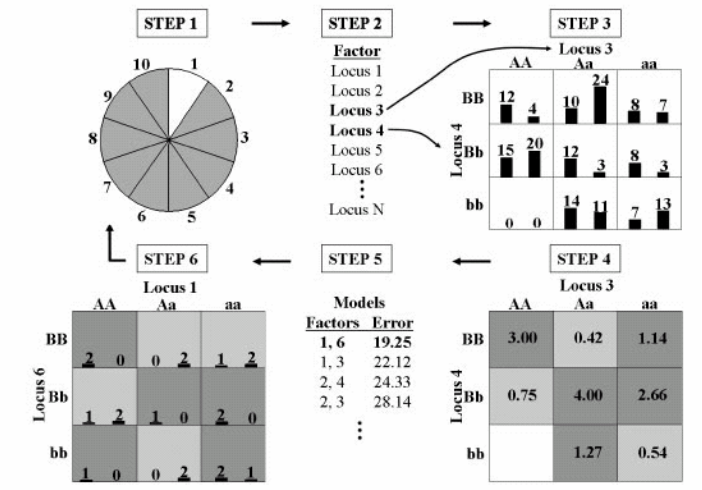


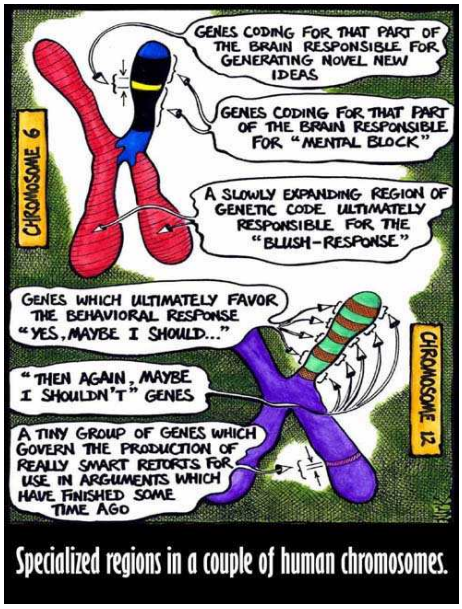
Figure 1: The six steps involved in MDR methodology. The dark shaded cells in step 4 are high risk cells, the light shaded cells are low-risk (Ritchie et al 2001).

Bibliography

- Moore J and White. *Tuning ReliefF for Genome-Wide Genetic Analysis*. Lecture notes in computer science. Springer Berlin/Heidelberg 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.
- Sebastiani, P.; Ramoni, M. F. and Kohane, I. Machine Learning in the Genomics Era. Editorial to the special issue Methods in Functional Genomics. *Machine Learning Journal*., 2003, 52, 5–9.
- Sebastiani, P; Wang, L.; Nolan, V. G.; Melista, E.; Ma, Q.; Baldwin C. T. and Steinberg, M. H. Fetal Hemoglobin in Sickle Cell Anemia: Bayesian Modeling of Genetic Associations. *American Journal of Hematology*, 2008, 83(3):189–95.

Renseignements: Kristel Van Steen

Promoteur: Kristel Van Steen



(Source: http://www.nearingzero.net)