Bayesian Combinatorial Partitioning For Detecting Interactions Among Genetic Variants

Shyam Visweswaran, MD, PhD[1] and An-Kwok Ian Wong, BS[1]
[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract

Detecting epistatic (nolinear) interactions among single nucleotide polymorphisms (SNPs) at multiple loci is important in the analysis of genomic data in association studies. We developed a Bayesian combinatorial partitioning (BCP) for detecting such interactions among SNPs that are predictive of disease. When compared with multifactor dimensionality reduction (MDR), a widely used combinatorial partitioning method for detecting interactions, BCP has significantly greater power and is computationally more efficient.

Background

The development of high-throughput genotyping technologies to simultaneously assay many thousands of single nucleotide polymorphisms (SNPs) has led to a flurry of studies with the aim of uncovering SNPs associated with common diseases. Such studies have uncovered more than 50 common variants that have been found to be associated with disease such as type 2 diabetes, cardiac and immunological diseases. In addition, interactions among genetic variants at multiple loci are likely to play an important role in such diseases, and an important challenge in the analysis of SNP data is the identification of *epistatic* loci that interact in a nonlinear fashion in their association with disease. Biologically, epistasis refers to gene-gene interaction when the action of one gene is modified by one or several other genes. Statistically, epistasis refers to interaction between genetic variants at multiple loci in which the net effect on disease from the combination of genotypes at the different loci is not accurately predicted by a simple linear combination of the individual genotype effects. The detection of statistical epistasis has the potential to identify interacting genetic loci that interact biologically.

Methods for detecting epistasis

While traditional statistical methods like logistic regression can identify interactions, it is unable to identify interactions among SNPs that do not possess significant univariate effects. For identifying epistatic interactions that may not be detected by traditional methods, data mining techniques such as set association analysis, genetic programming, neural networks, random forests and combinatorial methods have been applied [1]. In particular, combinatorial methods search over all possible combinations of loci to find combinations that are predictive of disease. A widely used combinatorial partitioning method is the multifactor dimensionality reduction method (MDR) that has been successfully applied in identifying epistatic interactions in several diseases [2]. MDR exhaustively evaluates single SNPs, 2-SNPs, 3-SNPs, up to *n*-SNPs in their ability to accurately predict disease by using cross-fold validation. We have developed a novel combinatorial method called Bayesian combinatorial partitioning (BCP) that uses a Bayesian score to evaluate single SNPs, 2-SNPs, 3-SNPs, up to *n*-SNPs in their ability to predict disease.

Evaluation

We evaluated BCP and MDR on a synthetic dataset for combinations of up to 4-SNPs. BCP had significantly greater power (i.e., higher accuracy in correctly identifying interacting SNPs) and was 50-140 times faster than MDR. This is likely due to BCP's ability to use the entire dataset for evaluating a combination compared to MDR which performs cross-fold validation to evaluate a combination.

Even though BCP is substantially faster than MDR, exhaustive evaluation of all combinations is infeasible for high dimensional datasets like those generated in genome-wide association studies. In future work we plan to develop heuristics to reduce the number of combinations that will have to be evaluated.

References

1. Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der AD, Feskens EJ. The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. BMC Genet 2006;7:23.
2. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 2003;19(3):376-82.