Evaluation of the minimum sample size for detecting SNPs interactions with classification tree models in case-control studies

Yolanda Benavente, Joan Valls, Silvia de Sanjose

[1]Catalan Institute of Oncology, Barcelona, Spain

Recently, a big effort has been made to analyze the effect of individual SNPs on chrinic diseases, such as cancer. Most studies focus their attention on selecting first a small subset of SNPs, which is potentially known to be related to the index disease.. Frequently, a case-control design is performed and therefore the association between SNPs and disease can be effectively measured using an indicator such as the odds ratio, and subsequent methods for classical inference can be applied. In addition, the analysis can be extended to assess the potential interactions between SNPs. However, most published studies limit themselves to the former analysis. This is mainly due to two facts. From one hand, it is widely known that the sample size has to be higher for detecting interactions and, from the other, there exist a variety of methods for the interaction analysis. Classification And Regression Tree (CART) models have attracted attention for these studies since interactions among the SNPs are easily to handle because CART models give an heuristic solution that is quite acceptable most times. The reason may be that it can capture the heterogeneity hidden in the observed data because subgroups representing SNP interactions are discovered. Finally, the cost in terms of parameters is low. We are interested in evaluating the minimum sample size for detecting true SNPs interactions, in case-control studies using classification tree models. To carry out this work, we have assumed that we know the true SNP interaction structure in a case-control study and the wild genotype SNP prevalence. Given an equal number of cases and controls, our aim is to assess how big has to be the sample size to detect all the richness of the “true tree”. To solve this question, we have implemented functions in R that generate data from a known “true tree” structure given a particular sample size. Generation is done according to the expected frequencies (deterministic approach). This process has been repeated systematically for different sample sizes. We have found that more depth in the tree and a low prevalence can spectacularly increase the sample size. This method allows us to know the sample size required for capturing the richness of true interactions and can be potentially useful for decision making.