An Adaptive Alpha-spending Algorithm Improves the Power of Statistical Inference in Microarray Data Analysis

Jacob (Jaap) P.L. Brand,
Pennington Biomedical Research Center

Abstract:

Microarrays have recently become one of the most widely used and effective research tools in modern biology. They can produce a snapshot of the expression levels of thousands of genes simultaneously. Because the total flux through a pathway can change dramatically through subtle changes in expression levels, the identification of genes with subtle differential expression became increasingly more important. However, the identification of these genes with subtle differential expression is challenging due to the noisiness of the data, the very small number of replicates per treatment group – often no more than 5, and the huge number of genes involved. Multiplicity, which refers to the simultaneous testing of many null-hypotheses, has recently been recognized as a serious problem in identification of differentially expressed genes (Dudoit et al, 2003). Most approaches for multiplicity control may be of limited power because they neither take into account nor capitalize on the dependencies among genes. A general re-sampling framework for controlling Family Wise Error Rate (FWER) proposed by (Westfall and Young, 1993), which takes dependencies among expression levels of genes into account may result in a higher power and has been extended to controlling the False Discovery Rate (FDR) (Yekutieli and Benjamini, 1999). However, a limitation of this approach is it only implicitly accounts for the dependencies among genes through re-sampling of the data. This precludes the incorporation of additional contextual evidence for differential expression that can be provided through modeling biological relationships among the genes or through pooling of results from similar microarray studies.

As an alternative we proposed an adaptive alpha-spending algorithm that takes explicitly dependencies among genes into account by assigning gene-specific significance levels to each gene. The current version of the alpha-spending algorithm relies on a key biological principle: if the differential expression levels of two genes are positively/negatively correlated, then one of the two genes is an activator/repressor of the other gene. We have shown that the Bonferroni correction applied to the alpha-spending adjusted p-values approximately controls the Family Wise Error Rate under the complete null. Under certain conditions it is plausible that the method in (Benjamini and Hochberg, 1995) applied to the alpha spending adjusted p-values also controls the False Discovery Rate (FDR). We compared the power between the original p-values and post-processed p-values for 700 genes simulated under a variety of conditions including: different correlations among differentially expressed genes, different sample sizes and different levels of differential expression. The simulation study confirms that application of the method in (Benjamin and Hochberg, 1995) controls the FDR. We found that on average the adaptive alpha-spending algorithm yielded higher power and that it allows yielding greater benefits with increasing sample sizes and co-regulation. Applications of alpha-spending algorithm are not limited to microarray data analysis; this algorithm may be helpful in any situation in High Dimensional Biology involving multiple hypothesis testing.