Comparing the influence of the sample size on the differential expression of pathways on univariate and multivariate methods

Shailesh Tripathi and Frank Emmert-Streib*
Computational Biology and Machine Learning Lab
Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast
97 Lisburn Road, Belfast, BT9 7BL, UK

Abstract

Interactions among genes constitute biological networks (pathways) responsible for the emergence of biological functions. Gene expression data from microarray experiments have the potential to capture the dynamic of the cell's state as a reflection of the underlying interconnected gene networks. Therefore, pathway-based approaches are essential to derive a meaningful understanding from microarray data in order to establish a robust relationship between molecular processes and their phenotype. So far, many statistical tests based on various null hypotheses and test statistics have been suggested to analyze groups of genes in order to detect pathological pathways. The performance of these methods depends critically on the sample size, the size of the pathway and the correlation structure among the genes. In this study, we are presenting a comparison of GSEA, Hotelling's T^2 and *sum of t-square* for real expression data from prostate cancer and acute lymphoid leukaemia (ALL). Further, we compare these methods for simulated data based on a Gaussian graphical model with an underlying network structure taking from the protein interaction network of *Saccharomyces cerevisiae*, in order to enforce realistic correlation structures. Our comparison reveals the effectiveness and limitation of different pathway-based methods.

*Corresponding Author. Email: v@bio-complexity.com