ABSTRACT

The selection and analysis of differentially expressed gene profiles (markers) helps associate a biological phenotype with its underlying molecular mechanisms and provides valuable insights into the structure of pathways and cellular regulation. However, analyzing and interpreting a given list of gene markers to glean useful biological insights can be extremely challenging. This is in part due to the difficulty of objectively evaluating how well members of a given pathway or functional class of interest (Gene Set) are represented in the markers list. To address this problem we introduce a statistical methodology called Gene Set Enrichment Analysis (GSEA) for determining whether a given Gene Set is over-represented or enriched in a Gene List of markers ordered by their correlation with a phenotype or class distinction of interest. The method is based upon a score computed as the maximum deviation of a non i.i.d. Brownian Bridge (in the same spirit as the Kolmogorov-Smirnov statistic) and uses permutation testing to assess significance. When multiple Gene Sets are tested simultaneously we propose two approaches to address the multiplicity: Validation GSEA which controls the Familywise error rate (FWER) and Discovery GSEA which controls the False Discovery rate (FDR). The utility of this procedure will be illustrated on a variety of biological examples.