A Threshold Independent Method for the Identification of Differentially Expressed Sets of Genes

Dr Frances Turner, Dr Michael Stumpf

Revolutions in high throughput technologies allow large sets of genes or even whole genomes to be studied simultaneously. While experiments such as microarrays efficiently produce large datasets, biological interpretation of such data can be challenging. The use of microarrays to identify genes that are differentially expressed between two conditions is becoming a widely applied technique. Such an approach can be used to identify genes that are important in an organism's response to a particular environment, or to identify genes that could potentially underlie an observed phenotype, for instance a diseased state. The resulting data gives a measurement for each individual gene, but assessment of which genes should be considered to have significantly changed expression, and derivation of biological understanding from this is not necessarily straightforward. A common approach is to look for enrichment of a set of functionally related genes among those genes showing differential expression. While a number of algorithms have been developed to perform statistical tests for this enrichment they have a range of practical limitations. There maybe difficulty in defining a level of change in gene expression which may be considered to potentially reflect a biologically significant response. Another major challenge is to achieve meaningful p-values where multiple tests are performed. A common approach is to look for enrichment of any Gene Ontology (GO) annotation among genes showing differential expression. However this requires a large number of separate statistical tests that, due to the structure of the Gene Ontology, will not all be independent, making corrections for multiple tests problematic.

Here we present a method for the detection of sets of genes that are enriched among differentially expressed genes. The method does not rely on an arbitrarily pre-defined threshold to determine which genes are considered to be differentially expressed, allowing sets of functionally related genes showing small but coordinated changes in expression to be identified. Unlike existing approaches this method can generate meaningful p-values when a large number of overlapping sets (for example sets of genes sharing each GO term) are tested. Whilst this method has been developed to aid the analysis of microarray data measuring differential expression of genes, it could be easily applied to other quantitive genome wide datasets.