

Jelle J. Goeman
LUMC
Medical Statistics and Bioinformatics
Postzone S5-P
Postbus 9600
2300 RC Leiden
j.j.goeman@lumc.nl

and

Livio Finos
University of Padua
Statistical Sciences
livio@stat.unipd.it

Multiple testing of graph-structured hypotheses

Hypotheses tests in bioinformatics can often be set in a tree or graph structures in a very natural way. This is the case for example if tests are performed at different levels of generality, for example, in genomic studies, if simultaneously testing at probe, gene and chromosome level. Similar multi-level tests arise if genes are clustered in a hierarchical clustering graph, and hypothesis tests are not only performed at the single gene (leaf) level, but also at levels higher up in the graph. Other hypotheses tests can be structured in graph structures with a more general structure, e.g. a directed acyclic graph if each hypothesis corresponds to a gene ontology term, or a more general graph if the hypotheses correspond to proteins in a protein-protein interaction graph.

We present a general approach to constructing multiple testing procedures for hypotheses structured in a graph. The procedures that are constructed using this method are guaranteed to control the family-wise error rate: the probability of making any false rejection in the graph can be bounded by a pre-specified probability. An important feature of the proposed procedures is sequential rejection. Sequential rejection allows the procedure to follow up on rejected hypotheses by focusing on other interesting hypotheses nearby in the graph to already rejected ones, without violating strict multiplicity control. This property makes procedures more powerful as well as more interpretable, as the collection of rejected hypotheses tends to remain relatively coherent.

We demonstrate the method with tree-structured hypotheses arising from hierarchical clustering of genes belonging to a gene ontology term, testing genes and sets of genes for association with a phenotype. This method has the potential to find interesting subpathways associated with the phenotype. This procedure has been implemented in the globaltest R package which is available on www.bioconductor.org.