True discovery guarantees in differential gene expression analysis

Anna Vesely*†[1]

[1]University of Bologna – Italy

Abstract

In real, potentially high-dimensional, data analysis, researchers are often interested in detecting and quantifying signal within subsets of features. For instance, in the comparison between two populations, one may want to determine how many genes in a biological pathway are differentially expressed, or how many edges in a network module exhibit differential co-expression. Methods for selective inference for True Discovery Proportions (TDPs) address these questions by providing lower confidence bounds for the number or proportion of active features within subsets, simultaneously over all possible subsets. These inferences are obtained without any additional adjustment of the significance level. Moreover, the simultaneity allows for an exploratory approach, as it ensures valid inference on the TDP even when the subsets of interest are selected post hoc, after seeing the data. In this talk, I will introduce a general class of post-hoc TDP procedures built on sum-based global tests, such as p-value combinations, the sequence kernel association test, and Goeman's global test. The framework allows for any choice of the underlying sum test, selected according to the problem at hand and the desired power properties, and adapts to unknown dependence structures through permutation testing. The procedure is iterative, providing increasing statistical power with additional iterations while maintaining valid TDP control at every stage. It is implemented in the R package sumSome, with a C++ backend for efficient computation. The methodology will be applied in practice to the analysis of differential gene expression data.

Keywords: Next generation hypothesis testing

*Speaker

†Corresponding author: anna.vesely2@unibo.it