Séminaire – Math For Genomics

Séance du mercredi 23 mai 2018. 10h30.
Evry. IBGBI. LaMME.

Post hoc inference via multiple testing

Bengamin SADACCA (Institut Curie)

Title: Motif enrichment analysis based on post-hoc inference of large-scale multiple testing.

B. Sadacca^{1,2}, O. Saulnier^{3}, F. Reyal1^{1,4}, P. Neuvial^{2}.

RNA-binding proteins (RBPs) are proteins that bind to specific DNA sequences (called motifs) to regulate post-transcription of messenger RNAs. Motif enrichment analysis aims at identifying the RBPs involved in the regulation of a given set of exons. For this, one wants to identify motifs that are found significantly more often in a set of given nucleotide sequences than expected by chance. Further, it has been shown that the binding position of the RBP (before or after the exon) has a key role in splicing regulation (Zong et al. 2014).

A method called rMAPS (Park et al. 2016) has recently been introduced to identify the binding positions of RBPs around skipped exons. The purpose of rMAPS is to identify known RBP motifs that are significantly enriched in differentially regulated exons between two sample groups as compared to control (background) events. rMAPS analyzes each set of 300 nt length sequences, with a sliding window of 50 nt, and counts the number of times the motif matches each sequence. The resulting "enrichment score" is then used to compare local enrichment in the window between significant exons and background exons by the Wilcoxon rank sum test. This process results in a set of 250 highly correlated p-values, which rMAPS summarizes by the minimum (raw) p-value.

We extend rMAPS by proposing a method to identify intervals significantly enriched for a given RBP. Here, "significantly" means that with high probability, the proportion of false positives among any of the selected intervals does not exceed a user-defined threshold. This method is based on the concept of post-hoc inference, as introduced by Goeman and Solari (Statistical Science, 2011) and further studied by Blanchard, Neuvial and Roquain (https://arxiv.org/abs/1703.02307). Importantly, the threshold on the FDP may be chosen by the user post hoc, ie after the data analysis. Compared with rMAPS, this approach reduces the number of identified false positives and allows the identification of their precise binding site.

1. RT2Lab Team, Translational Research Department, Institut Curie, PSL Research University, Paris, France.

2. Institut de Mathématiques de Toulouse, UMR5219 Université de Toulouse, CNRS UPS IMT, Toulouse, France.

3. Inserm U830, SIREDO Oncology Center, Institut Curie, PSL Research University, Paris, France.

4. Department of Tumor Biology, Institut Curie, Paris, F-75248, France.

Pierre NEUVIAL (IMT, CNRS)

Title: Post hoc inference via multiple testing; application to differential expression.

Multiple testing is a common issue in genomics. In particular, this is the case in differential expression (DE) studies, which aim at finding subsets of genes whose expression differs "significantly" between two conditions. In practice, statistical testing strategies are often complemented by biological knowledge. For example, one will only consider genes whose fold change is larger than some threshold, or that belong to a specific pathway. However, classical multiple testing procedures such as the Benjamini-Hochberg procedures for controlling the False Discovery Rate do not give guarantees on the proportion of true/false positives for such user-refined sets of genes.

Goeman and Solari (Statistical Science, 2011) have introduced "post hoc" procedures to bridge this gap. Such procedures yield statistical guarantees on the number or proportion of false positives in any number of (possibly data-driven) sets of genes.
In practice however, the available procedures rely on assumptions on the dependence between tests which (i) cannot be directly assessed, and (ii) may lead to few discoveries (conservativeness). We propose an alternative construction of post hoc procedures, which can naturally adapt to the dependency between tests. We illustrate the flexibility of the obtained procedures and their application to permutation-based DE studies.

This is joint work with Gilles Blanchard, Etienne Roquain and Benjamin Sadacca. A preprint (for the statistical part) is available at: https://arxiv.org/abs/1703.02307