Controlling the false discovery rate at detection of biological aberrations in -omic data

Djork-Arné Clevert [a,b], Andreas Mayr [a], Andreas Mitterecker [a], Günter Klambauer [a], and Sepp Hochreiter [a,*]

[a] Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria

[b] Charité University Medicine, Berlin, Germany

Motivation: A low false discovery rate (FDR) at the detection of biological aberrations (mRNA, miRNA, copy numbers, methylation state) in integrative -omic studies ensures sufficient detection power and prevents failures. Studies based on -omic data face the problem of the combinatorial multiplicity of the number of hypotheses that are tested - all dependencies between data sources - leading to an increase of false discoveries and spurious correlations. Falsely discovered aberrations will fail at a subsequent association test, though correction for multiple testing must take them into account. Thus, a high FDR not only decreases the discovery power of studies but also the significance level of the remaining discoveries after correction for multiple testing.

Methods: We considerably reduce the FDR at the detection of biological aberrations by using probabilistic latent variable models. These models assess the reliability of detections by estimating data consistencies, noise levels, and signal strength. They are optimized by Bayesian maximum a posteriori approaches, where the priors prefer models, which represent the null hypothesis, e.g. a gene that is not differentially expressed or constant copy number 2 for all samples. The posterior can only deviate from this prior by high information content in the data which hints at an aberration, the alternative hypothesis. The information gain of the posterior over the prior gives the informative/non-informative (I/NI) call which serves as a filter for aberration candidates. It can be shown that the I/NI call filter applied to null hypotheses is independent of the test statistic which in turn guarantees that the type I error rate control by correction for multiple testing is still possible after filtering. I/NI-calls perform well on data set with unbalanced design, whereas variance-based filtering approaches fail.

Results: Probabilistic latent variable models have lower FDR than other methods without loss of sensitivity as shown at different data sets like for mRNA analysis or for copy number estimation based both on microarray and on next generation sequencing data.