Clipper: p-value-free FDR control on high-throughput data from two conditions

Jingyi Jessica Li

High-throughput biological data analysis commonly involves identifying “interesting” features (e.g., genes, genomic regions, and proteins), whose values differ between two conditions, from numerous features measured simultaneously. The most widely-used criterion to ensure the analysis reliability is the false discovery rate (FDR), the expected proportion of uninteresting features among the identified ones. Existing bioinformatics tools primarily control the FDR based on p-values. However, obtaining valid p-values relies on either reasonable assumptions of data distribution or large numbers of replicates under both conditions, two requirements that are often unmet in biological studies. To address this issue, we propose Clipper, a general statistical framework for FDR control without relying on p-values or specific data distributions. Clipper is applicable to identifying both enriched and differential features from high-throughput biological data of diverse types. In comprehensive simulation and real-data benchmarking, Clipper outperforms existing generic FDR control methods and specific bioinformatics tools designed for various tasks, including peak calling from ChIP-seq data, differentially expressed gene identification from RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data. Notably, our benchmarking results for peptide identification are based on the first mass spectrometry data standard with a realistic dynamic range. Our results demonstrate Clipper’s flexibility and reliability for FDR control, as well as its broad applications in high-throughput data analysis.

Keywords: false discovery rate; p-value-free; two condition comparison; multiple testing