Bayesian Nonparametric Clustering Analysis with an Incorporation of Biological Network for High-Dimensional Multi-Scale Molecular Data

Yize Zhao

Yale University
E-mail: yize.zhao@yale.edu

Abstract: Investigating cancer genome based on multi-type omics data and how it advances personalized medicine is a global medical issue. Though some of the existing clustering methods are capable to character certain degree of concordant and heterogeneity across data types, none of them has incorporated biological network information within and across molecular modalities under cancer subtype discovery. Meanwhile, it is biologically important to identify the core set of biomarkers that are informative to the similarity among samples in each subtype. In this work, with the goal to achieve cancer subtype discovery, we construct a unified clustering model with an incorporation of biological network within and across different molecular data types and simultaneously identifying informative molecular biomarkers for each subtype. Different from existing parametric methods, we adopt a nonparametric approach based on Bayesian Dirichlet process mixture (DPM) models, which is more adaptable to different data types, robust to statistical assumptions and has no constrain on the number of clusters. The performance of the proposed model has been assessed by extensive simulation studies and GCTA.