A maximum average power test for large scale time-course data of counts with applications to RNA-Seq analysis

Wen Zhou

Colorado State University
E-mail: wzhou70@asu.edu

Abstract: Experiments that longitudinally collect RNA sequencing (RNA-seq) data can provide transformative insights in biology research by revealing dynamic patterns of genes. Such experiments create great demands for new analytic approaches to identify differentially expressed (DE) genes based on large-scale time-course count data. Existing methods, however, are sub-optimal with respect to power and may lack theoretical justification. Furthermore, most existing tests are designed to distinguish among conditions based on overall differential patterns across time, though in practice, a variety of composite hypotheses are of more scientificc interest. Lastly, some current methods may fail to control the false discovery rate (FDR). In this paper, we propose a new model and testing procedure to address the above issues simultaneously. Specifically, conditional on a latent Gaussian mixture with evolving means, we model the data by negative binomial distributions. Motivated by Storey (2007) and Hwang and Liu (2010), we introduce a general testing framework based on the proposed model and show that the proposed test enjoys the optimality property of maximum average power. The test allows not only identification of traditional DE genes but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and demonstrate its good performance via simulation studies. The procedure reveals interesting biological insights when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of the marine diatom Phaeodactylum tricornutum.