Detecting Multivariate Gene Interactions in RNA-Seq Data Using Optimal Bayesian Classification

Jason M. Knight[1], Ivan Ivanov[2], Robert S. Chapkin[3], Edward R. Dougherty[4]
[1] Electrical & Computer Engineering, TAMU
Jason@jasonknight.us
[2] Veterinary Physiology & Pharmacology, TAMU
IIvanov@cvm.tamu.edu
[3] Program in Integrative Nutrition & Complex Diseases, TAMU
r-chapkin@tamu.edu
[4] Electrical & Computer Engineering, TAMU
edward@ece.tamu.edu

RNA-Seq data, multivariate interactions, optimal Bayesian classfication.

RNA-Seq is a high-throughput technique for measuring the gene expression profile of a target tissue or even single cells. Due to its increased accuracy and flexibility over microarray technologies, it is widely applied in biological fields to uncover the transcriptional mechanisms at play in a given physiology or phenotype. Typically, this analysis involves mapping the RNASeq reads to a reference genome, quantifying transcript expression, and then performing testing for differential gene expression to determine which genes are expressed at significantly different levels in the phenotypes being compared. Tools such as Cufflinks, edgeR, and DESeq2 provide these univariate statistical tests using well characterized univariate statistical models of gene expression. However, one is often interested in phenotypes which can only be differentiated by the state of several genes simultaneously. These multivariate relationships cannot be detected using univariate testing procedures only. Instead, it is necessary to consider the joint expression patterns between multiple genes simultaneously. While this problem can be approached using the setting of multivariate statistical testing, we instead opt to utilize the theory of statistical classification for two primary reasons. First, translational medicine aims to apply scientific knowledge to improve medical practice, and classifications prediction of phenotypes from gene expression data is well aligned with this goal. Second, the model-based approach used in optimal Bayesian classification allows for the use of prior biological knowledge to improve results in the small number of samples typically available in biological studies.