

Statistical Methods for Network-based Analysis of Genomic Data

Hongzhe Li

*Department of Biostatistics, University of Pennsylvania School of Medicine,
USA*

A central problem in genomic research is the identification of genes and pathways involved in diseases and other biological processes. The genes identified or the univariate test statistics are often linked to known biological pathways through gene set enrichment analysis in order to identify the pathways involved. However, most of the available procedures for identifying differentially expressed genes do not utilize the known pathway structure information. In this paper, we propose to develop Markov random field (MRF)-based methods for identifying genes and subnetworks that are related to disease or other phenotypes. Such procedures model the dependency of the differential expression patterns of genes on the networks using the MRF models. Simulation studies indicated that the methods quite are effective in identifying genes and subnetworks that are related to disease and have higher sensitivity and low false discovery rates than some commonly used procedures that do not use the pathway structure information. Application to three breast cancer microarray gene expression data sets identified several subnetworks on several of the KEGG transcriptional pathways that are related to breast cancer recurrence. Biological interpretation of these networks and some further extensions of the methods will also be discussed.

[Hongzhe Li, Department of Biostatistics, University of Pennsylvania School of Medicine, USA; hli@cceb.upenn.edu]