Title: On two statistical elements of gene expression data analysis: differential expression and enrichment

Speaker: Michael A. Newton

Abstract

Two-sample comparison is a classical problem, though new and interesting statistical issues arise when the inference task is to accomplish a large number of such comparisons simultaneously. The canonical example comes from the analysis of gene expression; a particular case that I will present concerns the expression of human genes in nasopharyngeal cancer cells comparing cells grouped according to expression of the Epstein-Barr virus. After reviewing some basic issues and popular approaches to data analysis, I will turn to the problem of assessing enrichment of predefined gene sets (categories) for differentially expressed (DE) genes. Gene Ontology (GO) annotations provide a case in point; each category is a collection of genes that are associated with a common biological process, molecular function, or cellular localization. The hypergeometric distribution has been used to measure enrichment of DE genes in a GO annotation. One considers the cross classification of genes according to whether or not they are on the DE list and whether or not they have the annotation. A problem with this selection approach is that it reduces quantitative information about DE to indicators of whether or not genes are on the DE list. I will present an alternative averaging approach that uses additional information from the DE analysis. Neither approach is uniformly superior. Selection is preferred when the DE effect is large and the DE extent (proportion of affected genes in the annotation) is small. Averaging is preferred when the effect is small but the extent is large. In support of these claims I will present evidence from theoretical and empirical calculations, and I will review how these calculations play out in the nasopharyngeal cancer example.