

STATISTICS DEPARTMENT



SEMINAR

TITLE: ON THE STATISTICS OF GENE SET ENRICHMENT ANALYSIS

SPEAKER: Michael Newton

TIME: 4:00 P.M.

DATE: Wednesday, September 13, 2006

ROOM: 140 BARDEEN

ABSTRACT:

A prespecified set of genes may be enriched, to varying degrees, for genes that have altered expression levels relative to two or more states of a cell. Knowing the enrichment of gene sets defined by functional categories, such as gene ontology (GO) annotations, is valuable for analyzing the biological signals in microarray expression data. A common approach to measuring enrichment is by cross classifying genes according to membership in a functional category and membership on a selected list of significantly altered genes. A small Fisher's exact test p-value, for example, in this 2 x 2 table indicates significant enrichment. Other approaches retain the quantitative gene-level scores and calibrate enrichment by referring a category-level statistic to a permutation distribution associated with the original differential expression problem. In studying limitations of these approaches we uncover a class of random-set enrichment scoring methods. The class includes Fisher's test as a special case that uses selected genes, but it also includes enrichment scores that average gene-level evidence across the category. Averaging and selection methods are compared empirically using Affymetrix data on expression in nasopharyngeal cancer tissue, and theoretically using a location mixture model of differential expression. We find that each method has a domain of superiority in the state space of enrichment problems, and that both methods have benefits in practice. Our analysis also reveals a power balance problem when comparing enrichment scores across differently-sized categories, for which we provide a partial solution.



Coffee and Cookies at 3:30 p.m. in Room 1210 MSC