

**Department of Applied Mathematics and Statistics
The Johns Hopkins University**

SEMINAR

Donald Geman
Dept. of Applied Mathematics & Statistics
The Johns Hopkins University

April 22, 2004
304 Whitehead Hall
Refreshments: 3:30 p.m.
Seminar: 4:00 p.m.

RANK-BASED CLASSIFICATION OF GENE EXPRESSION PROFILES

ABSTRACT

Statistical inference from gene expression microarray data is difficult due to the small number of observations, typically tens, relative to the large number of genes, typically thousands. Consequently, standard methods in machine learning, such as support vector machines and random forests, may lead to over-fitting and inflated estimates of performance in detecting disease, identifying tumors, and predicting treatment responses. Moreover, the results may be difficult to interpret in biological terms. Working together with Daniel Naiman, Christian d'Avignon, and Raimond Winslow, we address these problems by a purely rank-based analysis, for instance, comparing the mRNA counts in selected pairs. The results so far are very promising; we obtain highly accurate and transparent decisions from small samples in standard classification tasks. However, there are many unanswered questions, both statistical and biological.