CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, December 5

12:15 – 12:45 PM

Davidson Hall, Room 207

MINING GENE EXPRESSION DATA GENERATED BY NEXT-GENERATION SEQUENCING TECHNOLOGY

JYOTA SNYDER

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Next generation DNA sequencing (NGS) technology has emerged as a result of genomic research progress made related to the Human Genome Project. A protocol of this technology, called RNA-seq, was designed to sequence RNA transcripts from biological samples. The quantification of gene expression levels using RNA-seq has been predicted to replace microarrays for gene expression profiling. Multivariate data mining methods have been used with microarray gene expression data for determining parsimonious biomarkers that can accurately discriminate between biological classes. At the time of this writing, there was scarce research literature related to multivariate data mining based approaches for biomarker discovery with RNA-seq gene expression data.

This study investigated whether current RNA-seq technology (including preprocessing steps that transform raw data into a gene expression matrix) provides data that can be successfully used for multivariate biomarker discovery (that is, to provide multivariate biomarkers with high sensitivity and specificity). The preprocessing performed for this study involved applying upper quartile normalization and $\log_2$ transformation to a training set of 452 lung adenocarcinoma samples obtained from The Cancer Genome Atlas and examining the effect of such transformation on the data. Afterwards, an eight gene biomarker was found from the training data using multivariate data mining methods, and the marker was validated with the use of an independent test data set obtained from a study in which 164 biological samples were sequenced using the same NGS technology as utilized for the training data set.

The classification of the independent test data set indicated that the biomarker had 97.7% sensitivity and 94.8% specificity. Since these classification metrics were relatively high and the training and test data sets were collected by different labs in different countries, this provided evidence that such data can be adequately prepared for use with the advanced data mining methods for biomarker discovery that are used to analyze microarray gene expression data.

For further information:

gotchevi@ccsu.edu 860-832-2839

http://www.math.ccsu.edu/gotchev/colloquium/