Your Name: Chun Tang
Your E-mail address: chuntang@cse.buffalo.edu
Your Dissertation Proposal Title: Mining Phenotypes and Informative Genes Underlying Gene Expression Profiles
Your Committee:
Chair: Dr. Aidong Zhang
Members: Dr. Jian Pei, Dr. Murali Ramanathan (Department of Pharmaceutical Sciences)
Your Dissertation Proposal Abstract:

Recently introduced DNA microarray technology permits rapid, large-scale screening for patterns of gene expression and gives simultaneous, semi-quantitative readouts on the level of expression of thousands of genes for samples. The raw microarray data (images) can then be transformed into gene expression matrices where usually a row in the matrix represents a genes and a column represents a sample.
The numeric value in each cell characterizes the expression level of the particular gene in a particular sample. Microarray technology has a significant impact on the field of bioinformatics, requiring innovative techniques to efficiently and effectively extract, analysis, and visualize these fast growing data.

While most of the previous studies focus on clustering either genes or samples, it is interesting to ask whether we can partition the complete set of samples into exclusive groups (called phenotypes) and find a set of informative genes that can manifest the phenotype structure simultaneously. The mining of phenotypes and informative genes can provide valuable information for biologists to understand the roles of genes and the phenotypes of samples.

Most of the genes collected by microarray experiments may not necessarily be task-specific. A small percentage of genes which manifest the meaningful phenotype structure of the samples are buried in large amount of noise. Uncertainty about which genes are relevant makes it difficult to construct an informative gene space.
The number of genes and the number of samples are very different in a typical gene expression matrix. Usually, we may have tens or hundreds of samples but thousands or tens of thousands of genes.
Since the number of samples is usually limited, such data sets are very sparse in high-dimensional genes space. Unsupervised phenotype structure and informative gene discovery of such sparse high-dimensional data sets presents an interesting but also very challenging problem. No existing approaches can be effectively and efficiently used to mine phenotypes and informative genes of such data sets.

In this proposal, we propose the new problem of mining phenotypes and informative genes from gene expression data sets and a novel unsupervised analyzing framework to detecting phenotypes and informative genes underlying gene expression data sets.
A series of statistical measurements are proposed to measure the quality of the mining results. These measurements delineate local pattern qualities based on a partition of samples on a subset of genes to coordinate between sample phenotype discovery and informative space detection. Two interesting algorithms are developed: the heuristic search and the mutual reinforcing adjustment method. Iterative pattern adjustment strategies are presented to approach the optimal solution which the pattern quality is maximized. The methods dynamically measure

and manipulate the relationship between samples and genes while conducting an iterative adjustment of genes and samples to approximate the informative genes and the phenotypes of the samples simultaneously. We present an extensive performance study on both real-world data sets and synthetic data sets. Our results strongly suggest that the two proposed methods are effective and scalable. The mining results are clearly better than the previous methods. They are ready for the real-world applications. The mutual reinforcing adjustment method is in general more scalable, more effective and with better quality of the mining results.