

# Development of a Knowledge-based Multi-scheme Cancer Microarray Data Analysis System

John H. Phan<sup>1</sup>, Chang F. Quo<sup>1</sup>, Kejiao Guo<sup>1</sup>, Weimin Feng<sup>1</sup>, Geoffrey Wang<sup>2</sup>, May D. Wang<sup>1\*</sup>

<sup>1</sup>Wallace H. Coulter Department of Biomedical Engineering,  
Georgia Institute of Technology and Emory University

<sup>2</sup>School of Biology, Georgia Institute of Technology  
313 Ferst Drive, Atlanta, GA 30332, USA

## Abstract

Comparing genes expressed in normal and diseased states assists the understanding of cancer pathophysiology, detection, prognosis, and therapeutic target study. Many existing expression analysis papers show that microarray data are usually case dependent, have small sample (patients) sizes, and have large gene dimensions. Thus, we have been developing a robust multi-parameter, multi-scheme knowledge-based optimization system that integrates the strengths of statistics, pattern-recognition, and support vector machines (SVM). The optimization logic identifies optimal cancer signature genes by utilizing different analysis models based on unsupervised and supervised clustering. Our system is being finalized by testing over public and in-house datasets with the intention of validation through clinical knowledge feedback.

## 1. Introduction

Microarray technology has dramatically increased the amount of data available to biologists in the last few years. The advancement of this technology also presents a big challenge. In cancer research, the issue is how to identify the signature genes, or biomarkers associated with particular cancer to perform precise, objective and systematic cancer diagnosis and treatment. More specifically, the goal is how to accurately analyze and interpret the resulting large amount of gene expression data with relatively small patient sample size. As such, we have been developing a novel multi-scheme system that can derive optimal decision based on the best utilization of gene expression data features and clinical, and biological knowledge. This process consists of grouping functionally similar genes or biomarkers into categories, revealing genetic relationships, and validating the grouping by cross-validation methods.

In the early phase development of our novel system, we have used unsupervised clustering methods

to discover gene relationship and used knowledge-based supervised classification to get highly accurate prediction in cancer diagnosis and prognosis study. Specifically, hierarchical clustering and SVMs are combined to improve sample classification and feature selection. The test data used in this paper is a previously published breast cancer data [1] and ovarian cancer data. This work sets up foundation for our next step drug target study.

## 2. Background

Unsupervised techniques analyze data without prior knowledge of how the data should be classified. Methods of unsupervised clustering include hierarchical clustering and self-organizing maps (SOM). Unsupervised methods can analyze microarray data in two ways – (1) clustering microarray samples, comprising up to thousands of genes, to elucidate relationships between genetic expressions of different groups and (2) clustering genes to reveal expression profiles. Similarities between samples can be defined using a variety of distance metrics [2]. Consequently, the choice of an appropriate distance metric is critical in order to reveal true underlying expression patterns beneath the samples. By combining the analysis of these clustering patterns with prior knowledge of the samples, we achieve two objectives – (1) to extract marker genes that are significant in distinguishing the various classes of samples and (2) to identify the most efficient distance metric in revealing underlying cluster patterns.

In contrast, supervised classification methods are based on learning machines that rely on data for which specific classifications are already known. These algorithms “learn” to classify data points provided by a limited training set comprised of a fraction of all available data. Once the algorithm has learned from the training set, it can then accurately classify a related data set. Methods for supervised learning include

neural networks and support vector machines (SVMs). SVMs have been shown to outperform neural networks[3], thus SVMs will be the main focus for supervised clustering methods in this paper.

SVM optimization results in a hyperplane that can separate classes of data. SVMs have an advantage over neural networks in that, for the case where data is not linearly separable, kernel functions can be used to map the data to a higher dimensional feature space [4].

## 2.1. Algorithms

The rank comparison algorithm computes sets of weights for the SVM using a leave-one-out dataset for each element (sample) of the dataset. Each weight set is then sorted, resulting in an ordered vector for each sample. The weight values correspond to particular genes; furthermore, the magnitude of the weight value determines the significance of the corresponding gene. Significant genes should remain reasonably within the ends of the weight vector.

The individual gene analysis algorithm performs slightly better than the rank comparison algorithm. For each gene, a complete leave-one-out validation is conducted on the samples. This algorithm, however, is computationally intense because the number of optimizations is equal to the number of samples multiplied by the number of genes. In order to complete calculations in a reasonable amount of time, data can be divided into blocks and analyzed separately. Significant genes selected using this algorithm should result in higher prediction rates when validated.

## 3. Results

Analysis of the breast cancer data using unsupervised hierarchical clustering produced erratic results. However, analysis of the four class ovarian cancer data using features discovered by the individual feature analysis algorithm resulted in proper clustering.

As expected, the rank comparison algorithm for the breast cancer data resulted in higher occurrence of significant genes at the beginning and end of the sorted weight vector. Using all 24,481 features of the breast cancer data in the complete leave-one-out algorithm on 78 samples resulted in a prediction rate of 88.46%. Using significant genes from the rank comparison algorithm unexpectedly decreased the prediction rate and may be due to inherent problems with the rank comparison algorithm.

Using the individual feature analysis algorithm resulted in slightly better performance. The top 10 genes obtained from this algorithm resulted in a

prediction rate of 88.46%, but the top single gene dropped the prediction rate to 84.62%.

The individual feature analysis algorithm was applied to each of the six pairs of classes for the ovarian cancer data resulting in six different sets of gene ranks. In contrast to the breast cancer data a significant number of the ovarian cancer genes were able to accurately separate 100% of the samples.

## 4. Discussion

Genes ranked using the individual feature analysis resulted in low prediction rates of around 80% using the breast cancer data. The individual feature analysis algorithm may not be an accurate prediction of gene rank because each SVM optimization is 1-dimensional and is unlikely to be linearly separable. It may be possible to use a fuzzy method or kernel functions to overcome the non-linear problem.

Although the dimensionality of the feature selection algorithm fails for the breast cancer data, analysis of the ovarian cancer data produced decent results. For several pairs of the four class data, a small number of genes were identified that could separate the classes with 100% accuracy. The results from clustering the four-class ovarian cancer data shows that the combination of supervised and unsupervised clustering can provide a powerful tool for classification. Features discovered using SVM methods could significantly improve unsupervised clustering methods. Unsupervised clustering of the gene expression of microarray data may also help to identify related features to improve SVM algorithms. Future work will include further analysis of unsupervised clustering of the breast cancer data using SVM-identified features and modification of the SVM algorithms to address non-linear issues.

## 5. References

- [1] L. J. v. t. Veer, H. Dai, et al, "Gene Expression Profiling Predicts Outcome of Breast Cancer," *Nature*, vol. 415, pp. 530-536, 2002.
- [2] R. A. Johnson, Wichern, D.W., *Applied Multivariate Statistical Analysis*: Prentice Hall, 1998.
- [3] M. P. S. Brown, Grundy, W.N. et al., "Support Vector Machine Classification of Microarray Gene Expression Data.," Department of Computer Science, University of California, Santa Cruz, CA 1999.
- [4] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.