Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

Comparative Study of Microarray Data for Cancer Research

John. H. Phan, Chang. F. Quo, and *May D. Wang**
The Wallace H. Coulter Department of Biomedical Engineering,
Georgia Institute of Technology/Emory University, 313 Ferst Drive, Atlanta, GA 30332, USA

ABSTACT

In comparison to traditional "single-gene" study method such as reverse transcriptase-polymerase chain reaction (RT-PCR), microarray technology can produce high-throughout gene expression data simultaneously. 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. In the paper, we are reporting the results of the first phase development of our novel system, to use unsupervised clustering methods to discover gene relationship and to use knowledge-based supervised classification to get highly accurate prediction in cancer diagnosis and prognosis study. This work sets up foundation for our next step drug target study.

Keywords— microarray gene expression analysis,

I. INTRODUCTION

Microarray technology has dramatically increased the amount of data available to biologists in the last few years. The rate at which scientists can analyze and understand this data, however, has not grown as fast as the incoming data. Therefore, biologists have become more and more dependent on computational tools to quickly process and apply this data. Microarray data are generally analyzed so that functionally similar genes or biomarkers can be grouped into categories to help reconstruct a larger picture of signaling pathways, the biological mechanisms that drive life. The grouping methods fall into two types, unsupervised clustering methods and supervised classification methods.

Unsupervised techniques attempt to analyze data without prior knowledge of how the data should be classified or clustered. Methods of unsupervised clustering are hierarchical clustering and self organizing maps (SOM). These algorithms are based on the theory that, through iterations, features eventually emerge from the data to mediate an organization into reasonable groups. Unsupervised hierarchical methods are able to analyze microarray data in two ways. First, microarray samples, each of which can consist of thousands of genes, can be clustered. This can elucidate relationships between genetic expression of different cells or organisms. For example, samples of cancerous tissue may be extracted from a number of different control and test patients to determine the prognosis of the patients, since similar stages of severity are expected to express similar genes, thus will cluster together. Second, genes may be clustered together using expression data from a number of samples to learn which genes are expressed at similar levels.

In contrast, supervised classification methods are based on learning machines that rely on data for which specific classifications are already known. These algorithms typically learn how 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 is expected, at least for a good algorithm, to accurately classify a related test data set. Methods for supervised learning include neural networks and support vector machines (SVMs). For microarray data, SVMs have been shown to outperform neural networks for several reasons [1]. SVMs will be the main focus for supervised clustering methods in this paper.

This paper will first outline methods in both unsupervised and supervised classification. It will then discuss more detailed algorithms to analyze microarray data and how unsupervised and supervised, specifically hierarchical clustering and SVMs can be combined to improve sample classification and feature selection. The test data used in this paper is breast cancer data taken from [12] and ovarian cancer data from Ovarian Cancer Institute, University of Georgia at Athens.

II. BACKGROUND

A. Unsupervised Clustering

Clinical situations require the accurate prediction of disease with minimal cost to the patient. Unsupervised clustering mimics the clinical dilemma of differentiating healthy patients from cancer patients. Using microarray data, a group of samples is classified into 2 or more classes based on similarities between individual samples with no prior information provided. Furthermore, this problem of classification is compounded by technical and biological variation in gene expression.

Similarity between samples can be defined using a variety of distance metrics. Consequently, different distance metrics reveal different clustering patterns. As we combine the analysis of these clustering patterns with prior knowledge of the samples, we achieve 2 objectives – (1) to extract marker genes that are significant in distinguishing

0-7803-8439-3/04/$20.00©2004 IEEE

the various classes of samples and (2) to identify the most efficient distance metric in revealing underlying cluster patterns.

The success of agglomerative unsupervised clustering depends largely on the selected distance metric. A variety of distance metrics exist in literature and are applied to yield various results [8]. Ideally, a well-defined sample dataset will yield similar results regardless of the distance metric. However, clinical data is usually far from ideal. Consequently, the choice of an appropriate distance metric is critical in order to reveal underlying expression patterns beneath the samples. Here, we discuss and compare 2 major classes of distance metrics: discrete against continuous.

B. Supervised Classification

Since supervised clustering depends on predefined classifications of data, the data can be partitioned in several ways to form test and training datasets. Supervised algorithms generally depend on random partitions of data in a process of alternating training and validating to gauge the performance of the algorithm. Another method of partition is called the leave-one-out method. Suppose, for example, there are n data points in a particular set. Using leave one out, the algorithm would train with n-1 data points and validate with the single left-out point. If the algorithm were to perform this partition n times, validating with a different left-out point for each iteration, it would have a measure of its performance from the number of correct validations. This is called the complete leave-one-out method.

C. Support Vector Machines

Support vector machines (SVM) are a supervised classification method. They are similar to neural networks in that they are iteratively trained to result in a mapping of data from an input to an output space. The optimization results in a hyperplane that can separate multiple classes of data. For neural networks, the perceptron is the typical single layer network that produces a separating plane [7]. SVMs have an advantage over neural networks in that, for case where data is not linearly separable, functions can be used to map the data from the input space to a higher dimensional feature space. Kernel functions can be engineered so that they are specific to a problem. Possible kernel functions are polynomial, radial basis, and sigmoid [1].

D. Multi-class data

Although the typical SVM formulation only able to separate two classes of data, some datasets may contain more than two classes. The ovarian cancer data, for example, contains four classes. The multi-class formulation (more than two classes) is implemented by creating multiple two-class problems. For each class, the classification problem is one of separating that class from each of the other classes. As an example, for four groups of data, group 1 is separated from 2, 3, and 4, group 2, from 1, 3, and 4, etc. Using this method, for four groups, six independent hyperplanes are generated which, together, should effectively be able to classify data into the four groups.

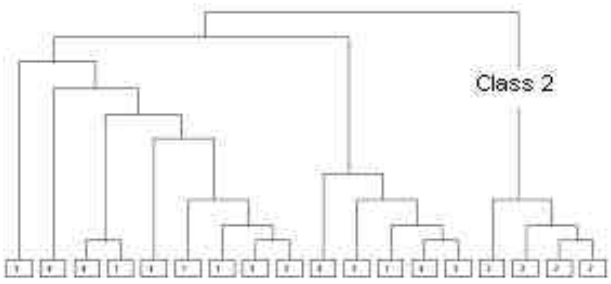

Figure 1 Unsupervised clustering of four class ovarian cancer data using three features

E. Rank comparison algorithm

The rank comparison algorithm computes the weights of the SVM using a leave-one-out dataset for each element of the dataset. Therefore, for the samples in a particular dataset, there is a number of weights for each sample that correspond to genes or biomarkers. Each of the weight sets is then sorted, resulting in a set for each sample ordered from smallest to largest. The theory behind sorting the weights is that each of the weight values corresponds to a particular gene. The magnitude of the weight value determines the significance of the corresponding gene. Weight values with larger magnitudes, for example, are more important for the classification of the dataset than are weight values with small magnitudes. The sorted weights, therefore, will have significant genes at the beginning and end components corresponding to large negative and large positive. This implies that significant gene indexes will exist at the beginning and the end.

Since there are n vectors of weights, one for each sample, the correspondence of gene indexes are not expected to match with 100% accuracy, although this would be ideal. In reality, the significant genes should fall within the ends of the weight vector. In order to check this assumption, all of the vectors were compared and a histogram analysis is conducted for each gene.

F. Individual feature analysis

This algorithm uses a more brute force method to test each gene individually. For each of the genes, a complete leave-one-out was conducted on the n samples resulting in a prediction rate with n being a 100% prediction rate. This algorithm is very computationally intense because of the number of optimizations required is m*n, where m is the number of genes/biomarkers and n is the number of samples. In order to complete the calculations in a reasonable amount of time, the data could be divided into blocks and analyzed separately. For both the breast cancer data and the ovarian cancer data, the significant genes selected using this algorithm should result in a higher

prediction rate when those genes are used as features in the complete leave-one-out.

III. RESULTS

A. Unsupervised Clustering

Analysis of the breast cancer data using the unsupervised clustering method did not produce expected results. Even after filtering features using gene distance from mean of all the samples, clustering was erratic. However, analysis of the four class ovarian cancer data using features discovered by the individual feature analysis algorithm resulted in correct clustering. Clustering using class two features is shown **in Figure 1.**

B. Rank comparison algorithm

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 (end data not shown). Figure 2 shows the

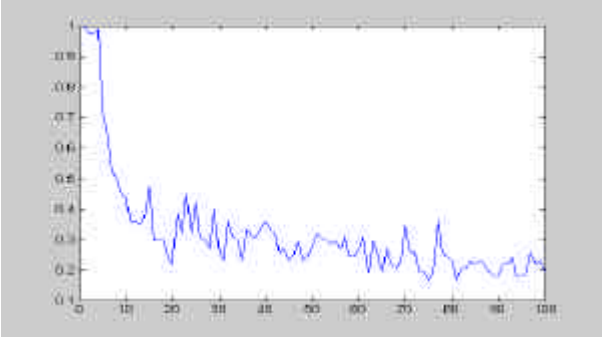

Figure 2 First 100 genes from rank comparison algorithm. The x-axis shows the gene number while the y-axis shows occurrence of that gene.

beginning 100 genes (most negative) out of 24481 genes.

Using all 24481 features of the breast cancer data in the complete leave one out algorithm on 78 samples, a prediction rate of 88.46% was calculated. Using the top 7 genes from the rank comparison algorithm (genes shown in table 1) resulted in a decrease in the prediction rate at 75.64%. The top 1 gene showed a 71.79% prediction rate. These results are not as expected and may be due to inherent problems with the rank comparison algorithm.

TABLE I
SIGNIFICANT GENES FROM RANK COMPARISON

Index	Gene Number	Occurrence (out of 78 samples)
1	9936	78
2	2206	76
3	21896	76
4	5862	77
24479	22788	57
24480	20265	76
24481	24194	76

C. Individual feature analysis

Using the individual feature analysis algorithm resulted in slightly better, but still unsatisfactory numbers. The top 10 genes obtained from this algorithm (data not shown) resulted in a prediction rate of 88.46% using complete leave-one-out, identical to the control of 24481 genes. The top 5 genes showed a slight improvement at 89.74% and the top single gene fell to 84.62%.

D. Four class ovarian cancer data

The individual feature analysis algorithm was run on 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, in which the best genes were found to separate the samples with only around 80% accuracy, a significant amount of the ovarian cancer genes were able to accurately separate 100% of the samples. Each of the four classes were separated from three other classes, therefore three sets of rankings were computed for each class. For example, the data describing class one genes were from the separation of classes 1 / 2, 1 / 3, and 1 / 4. Each of the three sets for each class was screened for similar gene indexes, resulting in the following biomarkers.

TABLE II
INDIVIDUAL FEATURE ANALYSIS OF OVARIAN CANCER DATA

Class	Gene IDs	Gene Names
1	3459	CHGB
2	4128 6619 7651	KIAA1030 SNCG AFAP
3	7684	FLJ10803
4	395 11297	UBC TNFRSF25

IV. DISCUSSION

A. Possible problems with SVM algorithms

A problem with the rank comparison algorithm is that it assumes a strict ordering of the genes, but in reality, significant genes may still fall within the ends of the vector but in a slightly different order. In order to test this, the range of vector components to test gene frequency was widened to include more than one component, as shown in Figure 3. The resulting data show a slower drop-off of the rate with more genes at 100%.

Genes ranked using the individual feature analysis resulted in low prediction rates of around 80% using the breast cancer data. The low prediction rates may be due to the low dimensionality of the data. The complete leave-one-out algorithm may not be an accurate prediction of gene

rank because each SVM optimization was 1-dimensional and it is unlikely that 1-dimensional data would be linearly separable. The issue with dimensionality in the slow algorithm should be addressed. It may be possible to use a fuzzy method to separate the data in 1 dimension to avoid the problem of non-separable data. An alternate method would be to analyze a range of genes around the gene of interest in a "blocking" method similar to the method mentioned for the fast algorithm. In this case the blocks could be either selected based on initial gene ordering or strategically selected based on the fast algorithm or some other pre-ranking algorithm.

B. Identified genes linked to ovarian cancer

Although the dimensionality of the feature selection algorithm fails for the breast cancer data, analysis of the ovarian cancer data produced decent results. The problem of linear separability may not have affected this data because of the significantly smaller number of samples, 18 vs. 78 in the breast cancer data.

Class one data can be separated from each of the other classes using only one gene, CHGB or chromogranin B/secretogranin 1. Similar genes such as chromogranin A have been linked directly to ovarian and other types of cancer [4] [13]. The group two data can be distinguished by three genes, KIAA1030, SNCG, and AFAP. KIAA1030 is a gene fragment that is not yet well characterized. SNCG has also been linked directly to ovarian cancer [6]. AFAP has not been directly linked to ovarian cancer, but has been implicated in colon cancer [9]. Class three can be distinguished by an uncharacterized fragment FLJ10803. Class four is related to UBC, an ubiquitin gene, and TNFRSF25, in the TNF receptor family of genes. Ubiquitin and TNF (Tumor necrosis factor) have both been linked to ovarian cancer [11] [14].

V. CONCLUSION

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 can significantly improve unsupervised clustering methods. Unsupervised clustering of the gene expression of microarray data may also help to identify related features to optimize 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 the dimensionality and linear separability issues. Both of these issues may be resolved by application of various kernel functions to the SVM optimization problem.

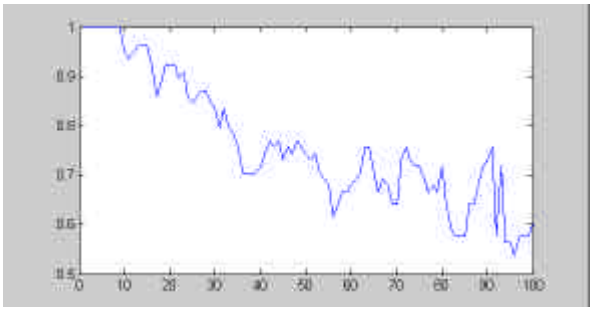

Figure 3 Percent occurrence of genes with a +/- 5 component margin of error. The x-axis shows the gene number while the y-axis shows occurrence of that gene.

ACKNOWLEDGEMENT

The authors want to thank Dr. John McDonald for providing the ovarian cancer microarray data for doing algorithm validation.

REFERENCES

[1] Brown, M. P. S., W. N. Grundy, et al. (1999). Support Vector Machine Classification of Microarray Gene Expression Data. Santa Cruz, CA, Department of Computer Science, University of California.

[2] Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. U. Fayyad. Boston, Bell Laboratories, Lucent Technologies: 1-43.

[3] Dudoit, S., J. Fridlyand, et al. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." J Amer Stats Assoc 97: 457.

[4] Fukunaga, M., Y. Endo, et al. (1997). "Small cell neuroendocrine carcinoma of the ovary." Virchows Arch 430(4): 343-348.

[5] Gunn, S. R. (1998). Support Vector Machines for Classification and Regression, University of Southampton: 1-53.

[6] Gupta, A., A. Godwin, et al. (2003). "Hypomethylation of the synuclein gamma gene CpG island promotes its aberrant expression in breast carcinoma and ovarian carcinoma." Cancer Research 63(3): 664-673.

[7] Ham, F. M. and I. Kostanic (2001). Principles of Neurocomputing for Science & Engineers. New York, McGraw-Hill, Inc.

[8] Johnson, R. A. and D. W. Wichern (1998). Applied Multivariate Statistical Analysis, Prentice Hall.

[9] Knudsen, A., M. Bisgaard, et al. (2003). "Attenuated familial adenomatous polyposis (AFAP). A review of the literature." Fam Cancer 2(1): 43-55.

[10] Quackenbush, J. (2001). "Computational analysis of microarray data." Nature Reviews Genetics 2: 418-427.

[11] Starita, L. and J. Parvin (2003). "The multiple nuclear functions of BRCA1: transcription, ubiquitination and DNA repair." Curr Opin Cell Biol 15(3): 345-350.

[12] Veer, L. J. v. t., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature 415: 530-536.

[13] Wu, J., A. Erickson, et al. (2000). "Elevated serum chromogranin A is detectable in patients with carcinomas at advanced disease stages." Ann Clin Lab Sci 2: 175-178.

[14] Yang, W., A. Godwin, et al. (2004). "Tumor necrosis factor-alpha-induced matrix proteolytic enzyme production and basement membrane remodeling by human ovarian surface epithelial cells: molecular basis linking ovulation and cancer risk." Cancer Research 64(4): 1534-1540.