Adaptive Discriminant Analysis for Microarray-Based Classification

YIJUAN LU, QI TIAN, JENNIFER NEARY, FENG LIU, and YUFENG WANG
The University of Texas at San Antonio

Microarray technology has generated enormous amounts of high-dimensional gene expression data, providing a unique platform for exploring gene regulatory networks. However, the curse of dimensionality plagues effort to analyze these high throughput data. Linear Discriminant Analysis (LDA) and Biased Discriminant Analysis (BDA) are two popular techniques for dimension reduction, which pay attention to different roles of the positive and negative samples in finding discriminating subspace. However, the drawbacks of these two methods are obvious: LDA has limited efficiency in classifying sample data from subclasses with different distributions, and BDA does not account for the underlying distribution of negative samples.

In this paper, we propose a novel dimension reduction technique for microarray analysis: Adaptive Discriminant Analysis (ADA), which effectively exploits favorable attributes of both BDA and LDA and avoids their unfavorable ones. ADA can find a good discriminative subspace with adaptation to different sample distributions. It not only alleviates the problem of high dimensionality, but also enhances the classification performance in the subspace with naïve Bayes classifier. To learn the best model fitting the real scenario, boosted Adaptive Discriminant Analysis is further proposed. Extensive experiments on the yeast cell cycle regulation data set, and the expression data of the red blood cell cycle in malaria parasite *Plasmodium falciparum* demonstrate the superior performance of ADA and boosted ADA. We also present some putative genes of specific functional classes predicted by boosted ADA. Their potential functionality is confirmed by independent predictions based on Gene Ontology, demonstrating that ADA and boosted ADA are effective dimension reduction methods for microarray-based classification.

This work is supported in part by San Antonio Life Science Institute (SALSI), ARO grant W91INF-05-1-0404, and DHS grant N0014-07-1-0151 to Q. Tian, and NIH grant 1R21AI067543–01, San Antonio Area Foundation and UTSA Faculty Research Award to Y. Wang. Y. Wang is partially supported by NIH RCMI grant 2G12RR013646 to UTSA. The project described is also supported by grant number 1SC1GM081068 from the National Institute of General Medical Sciences to Y. Wang. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. J. L. Neary is supported by the NIH MBRS-RISE program.
Authors' addresses: Y. Lu, Q. Tian (corresponding author): Department of Computer Science, University of Texas at San Antonio, TX 78249; emails: {lyijuan,qitian}@cs.utsa.edu; F. Liu, Department of Pharmacology, University of Texas Health Center at San Antonio, TX 78249; email: liuf@uthscsa.edu; J. Neary, Y. Wang (corresponding author); Department of Biology, University of Texas at San Antonio, TX 78249; email: jingraha@lonestar.utsa.edu, yufeng.wang@utsa.edu.

Categories and Subject Descriptors: I.5 [**Pattern Recognition**]: Design Methodology—*Pattern analysis*

General Terms: Algorithms

Additional Key Words and Phrases: LDA, BDA, ADA, boosted ADA, dimension reduction, microarray

ACM Reference Format:

Lu, Y., Tian, Q., Neary, J., Liu, F., and Wang, Y. 2008. Adaptive discriminant analysis for microarray-based classification. ACM Trans. Knowl. Discov. Data. 2, 1, Article 5 (March 2008), 20 pages. DOI = 10.1145/1342320.1342325 http://doi.acm.org/10.1145/1342320.1342325

1. INTRODUCTION

Microarray technology produces a large number of high-dimensional gene expression data with various temporal-spatial patterns. It is widely accepted that genes of correlated functions or components of the same biological networks often exhibit similar expression patterns. Unveiling the distinct expression patterns from microarray profiles may provide insights into the components and mechanisms of gene regulatory networks.

Data clustering has been used for decades in image processing and pattern recognition [Jain et al. 1999], and in the last several years it has become a popular data-analysis technique in genomic studies using gene-expression microarray [Brown et al. 2000; Chipman et al. 2003]. Each microarray provides expression measurements for thousands of genes, and clustering is a useful exploratory technique to analyze gene-expression data as it groups similar genes together and allows biologists to identify potentially meaningful relationships between them and to reduce the amount of information that must be analyzed. There are two types of clustering methods: unsupervised and supervised learning. Unsupervised learning method does not require any a priori information. The main assumption underlying unsupervised clustering analysis for gene-expression data is that genes that belong to the same biological process, and genes in the same pathway, would have similar expression over a set of arrays (be it time series or condition dependent).

The commonly used unsupervised clustering methods [Ringner et al. 2002; Cho et al. 2003] in gene expression space are hierarchical clustering [Eisen et al. 1998], K-means clustering [Tavazoie et al. 1999], and self-organizing maps (SOMs) [Tamayo et al. 1999]. Because unsupervised learning cannot utilize some prior information about which samples or genes are expected to group together and it cannot construct a classifier for predicting unknowns, supervised methods with given label information are likely more promising for gene classification and prediction in microarray analysis. In general, supervised learning methods are used to construct a robust classifier first, which accurately recognizes patterns from given training samples. Then testing samples would be classified into known phenotypes based on the trained classifier. Supervised classification algorithms have been applied to microarray-based classification [Cho et al. 2003] including Fisher linear discriminant analysis [Dudoit et al.

2000], k nearest neighbor [Li et al. 2001], decision tree, multi-layer perceptron [Khan et al. 2001] and support vector machines (SVM) [Brown et al. 2000] etc.

Despite some encouraging results, the problem of high dimensionality in microarray data remains unsolved. The machine learning is afflicted by the *curse of dimensionality* as the search space grows exponentially with the dimension. Despite the widely held view that high throughput approaches are swamping us with data, in fact much of the time *high dimensionality* obscures the details in the data. The problem of high dimensionality can be alleviated by dimension reduction. Linear discriminant analysis (LDA) [Fisher 1936, 1938; Duda et al. 2001] and Biased Discriminant Analysis (BDA) [Zhou and Huang 2001] are both effective techniques for feature dimension reduction. LDA is considered as one of the best known *Data Analysis* techniques and has found numerous applications in science and engineering including face recognition [Belhummeur et al. 1997; Etemad et al. 1997], image retrieval [Swet and Weng 1999; Wu et al. 2000], and bioinformatics [Ewans and Grant 2001]. BDA also plays a key role in content-based image retrieval [Zhou and Huang 2001].

LDA makes the equivalent (unbiased) effort to cluster negative and positive samples by attempting to minimize the Bayes error by selecting the feature vectors w which maximizes $\frac{|W^T S_B W|}{|W^T S_W W|}$, where S_B measures the variance between the class means, and S_W measures the variance of the samples in the same class. In LDA, both positive and negative samples are treated equally when finding the optimal projection subspace.

Compared to LDA, BDA is biased towards the positive examples [Zhou and Huang 2001]. BDA tries to find an optimal mapping that all positive examples are clustered and all negative examples are scattered away from the centroid of the positive examples. Studies have shown that BDA works very well in image retrieval especially when the size of the training sample set is small [Zhou and Huang 2001].

In supervised learning, both LDA and BDA have pros and cons. LDA assumes that positive and negative samples are from the same sources (distributions), respectively and they could be clustered in the projected space. BDA assumes that positive samples must be functionally similar while negative samples may be from different functional categories. Hence, when all negative samples are from the same distribution and clustered together, LDA will outperform BDA. However, BDA will outperform LDA when negative samples are from different classes and scattered. Figure 1 shows two classical examples, where LDA outperforms BDA as shown in Figure 1 (a) and BDA outperforms LDA as shown in Figure 1 (b). In reality, many applications do not fit exactly into either of the two assumptions, which means either LDA or BDA cannot find an optimal projection (as shown in Figure 2).

In this article, we propose a novel Adaptive Discriminant Analysis (ADA), which merges LDA and BDA in a unified framework that offers increased flexibility and a richer set of alternatives to LDA and BDA in the parametric space. ADA can find a good projection with adaptation to different sample distributions and discover the classification in the subspace with naïve Bayes classifier.

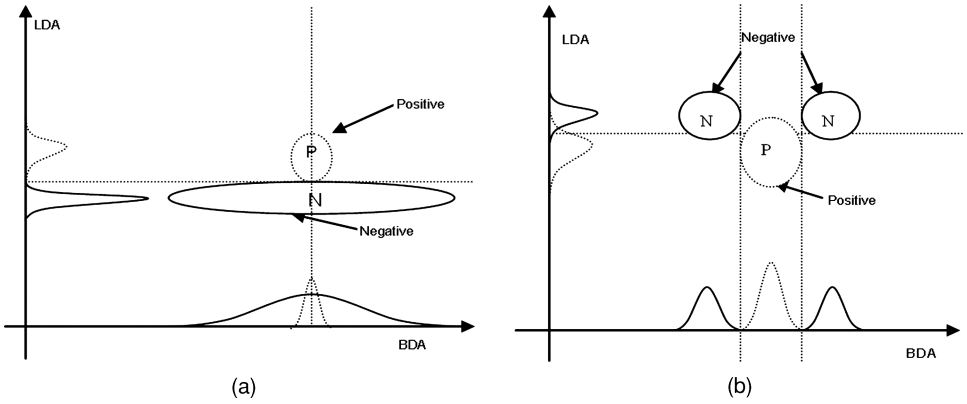

Fig. 1. LDA outperforms BDA in (a), and BDA outperforms LDA in (b).

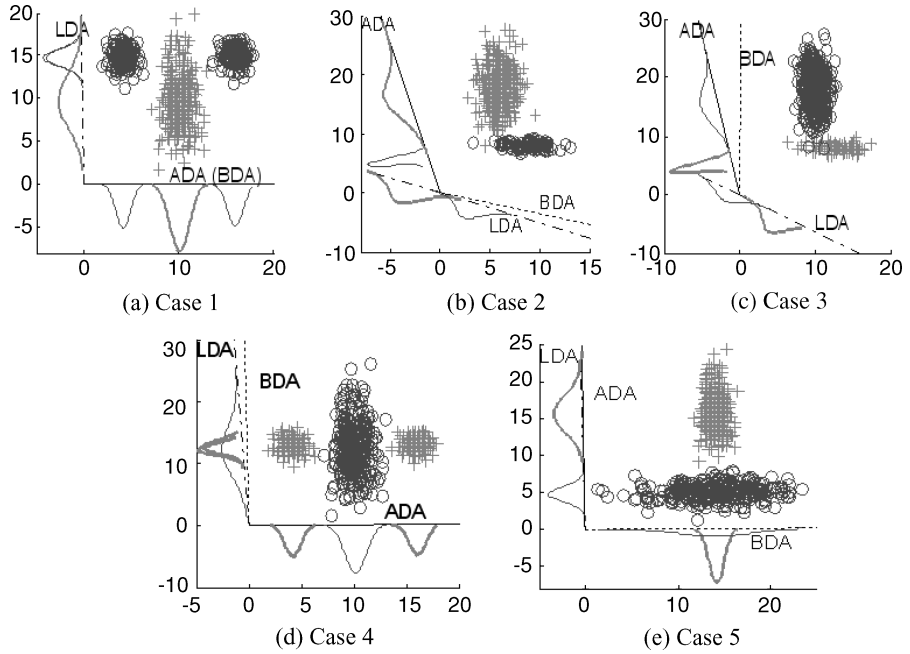

Fig. 2. Comparison of BDA, LDA and ADA on 2-D synthetic data.

ADA is a parametric method. Parameter optimization and selection are important but difficult. In this paper, we also propose a boosted ADA. Instead of searching the whole parametric space, the boosted ADA provides a unified and stable solution to find close to the optimal ADA prediction result.

Extensive experiments on the yeast cell cycle regulation data set and *Plasmodium falciparum* red blood cell cycle data set show that ADA and boosted ADA are effective dimension reduction methods for classification.

The rest of this article is organized as follows. In Section 2, we illustrate Adaptive Discriminant Analysis in detail. In Section 3, boosted Adaptive

Discriminant Analysis is proposed. In Section 4, we apply ADA and boosted ADA on gene classification and analyze and compare the results. Finally, our contributions and future work are presented in Section 5.

2. ADAPTIVE DISCRIMINANT ANALYSIS

2.1 Linear Discriminant Analysis

LDA is one of most widely used discriminant analysis techniques in classification and dimension reduction. LDA tries to find an optimal projection W from originally high d_1-dimensional space to a low d_2-dimensional space, which makes samples from the same class cluster to each other and samples from different classes separate from each other. The problem of finding the optimal W can be mathematically represented as the following maximization problem:

$$W_{opt} = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \tag{1}$$

$$S_B = \sum_{j=1}^{C} N_j \cdot (\boldsymbol{m}_j - \boldsymbol{m})(\boldsymbol{m}_j - \boldsymbol{m})^T \tag{2}$$

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} \left(\boldsymbol{x}_i^{(j)} - \boldsymbol{m}_j\right)\left(\boldsymbol{x}_i^{(j)} - \boldsymbol{m}_j\right)^T. \tag{3}$$

Here, the between-class matrix S_B measures the separability of class centers and the within-class scatter matrix S_W measures the within-class variance in the low-dimensional space. We use $\{\boldsymbol{x}_i^{(j)}, i = 1, \ldots, N_j\}$, $j = 1, \ldots, C$ to denote the feature vectors of training samples. C is the number of classes. When C = 2, it is Fisher Discriminant Analysis (FDA) and when C > 2, it is called Multiple Discriminant Analysis (MDA), a natural extension of FDA to multiple classes. N_j is the number of the samples of the j^{th} class, $\boldsymbol{x}_i^{(j)}$ is the i^{th} sample from the j^{th} class, $\boldsymbol{m}_j$ is mean vector of the j^{th} class, and $\boldsymbol{m}$ is grand mean of all examples. Since FDA and MDA are both linear techniques, they are also referred to as LDA.

To maximize the ratio of Equation (1), the optimal W are composed of the generalized eigenvector(s) $\boldsymbol{w}_i$ associated with the largest eigenvalue(s). $W_{opt} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{C-1}]$ contains $C - 1$ eigenvectors corresponding to $C - 1$ eigenvalues; that is, $S_B w_i = \lambda_i S_W w_i$ [Fisher 1938]. It should be noted that W maps the original d_1-dimensional data space $\boldsymbol{X}$ to a d_2-dimensional space Δ (where $d_2 \leq C - 1$).

2.2 Biased Discriminant Analysis

In two-class LDA, the equivalent (unbiased) effort has been made to cluster negative and positive samples. Intuition suggests that clustering the negative genes may be difficult and unnecessary because they may be from functionally different classes (Figure 1(a)). Zhou and Huang (2001) proposed Biased Discriminant Analysis (BDA). The intuition behind the BDA is that "*all positive examples are alike, and each negative example is negative in its own way.*" That means that the positive samples are visually similar (or functional similar in

gene-expression data) and should be clustered in the projected space. On the other hand the negative samples might be from different classes (functions) and it is difficult to find a mapping to make them close to each other. For classification, it would be sufficient enough to make the negative samples far away from the center of positive ones.

BDA differs from traditional LDA defined in equations (1)–(3) in a modification on the computation of between-class scatter matrix S_B and within-class scatter matrixS_W. They are replaced by $S_{N\to P}$ and S_P, respectively.

$$W_{opt} = \arg\max_W \frac{|W^T S_{N\to P}\, W|}{|W^T S_P\, W|} \tag{4}$$

$$S_{N->P} = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{5}$$

$$S_P = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T, \tag{6}$$

where $\boldsymbol{m}_p$ is the mean vector of the positive examples. $S_{N\to P}$ is the scatter matrix between the negative examples and the centroid of the positive examples, and S_P is the scatter matrix within the positive examples. $N \to P$ indicates the asymmetric property of this approach, which means the user's biased opinion towards the positive class, thus the name of biased discriminant analysis [Zhou and Huang 2001].

Although the idea of BDA is simple and it is very successful in content-based image retrieval, we find that its assumption is still inappropriate in some scenarios which will be explained in part C. The complex nature of gene expression data requires a classification method that can adaptively fit the distribution of genes from different functional classes and discover a good classification.

2.3 Adaptive Discriminant Analysis

Given that LDA and BDA have their own assumptions and pay attention to different roles of the positive and the negative examples in finding the optimal discriminating subspace, it is our expectation that they can be unified. In addition, there are many cases that both LDA and BDA are not applicable.

To provide a better model fitting the complex distributions for positive and negative genes, we propose an Adaptive Discriminant Analysis (ADA), which finds an optimal projection

$$W_{opt} = \arg\max_W \frac{|W^T[(1-\lambda)\cdot S_{N\to P} + \lambda\cdot S_{P\to N}]\, W|}{|W^T[(1-\eta)\cdot S_P + \eta\cdot S_N]\, W|}, \tag{7}$$

in which

$$S_{N->P} = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{8}$$

$$S_{P->N} = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{9}$$

$$S_P = \sum_{i\in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{10}$$

$$S_N = \sum_{i\in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T, \tag{11}$$

Table I. Special Cases of ADA

(λ,η)	Optimal Projection	Note
(0,0)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{N->P}W^T\rvert}{\lvert WS_PW^T\rvert}$	Case 1 (BDA)
(0,1)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{N->P}W^T\rvert}{\lvert WS_NW^T\rvert}$	Case 2
(1,0)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{P->N}W^T\rvert}{\lvert WS_PW^T\rvert}$	Case 3
(1,1)	$W_{ADA} = \arg\max_W \frac{\lvert WS_{P->N}W^T\rvert}{\lvert WS_NW^T\rvert}$	Case 4 (Counter-BDA)
$(\frac{1}{2}, \frac{1}{2})$	$W_{ADA} = \arg\max_W \frac{\lvert W(S_{P->N}+S_{N->P})W^T\rvert}{\lvert W(S_P+S_N)W^T\rvert}$	Case 5 (LDA-like)

x_i is the i$^{\text{th}}$ gene expression vector. The m_P and m_N are the means of positive and negative gene expressions, respectively. S_P (or S_N) is the within-class scatter matrix for the positive (or negative) examples. $S_{N\to P}$ (or $S_{P\to N}$) is the between-class scatter matrix from the negative (or positive) examples to the centroid of the positive (or negative) examples. The two parameters λ and η control the bias between positive and negative genes and range from $(0, 0)$ to $(1, 1)$. Proper setting of parameters may fit the real distribution of gene expression data.

Table I summarizes five special cases of ADA. From Table I, we can find that the ADA recovers BDA when λ and η are set to be 0 and 0 in Case 1. Case 5 corresponds to a LDA-like projection with λ and η set to 0.5 and 0.5. Case 4 finds a projection that is on the contrary side of BDA, which is called Counter-BDA. Case 2 and Case 3 is a couple of contrary distribution scenarios, which assume that the negative (positive) genes are similar and positive (negative) genes might be from different classes. All these five cases fit certain gene feature distributions and have correspondence with some scenarios as illustrated in Figure 2.

In order to show the advantages of ADA over BDA and LDA, we use synthetic data to simulate different sample distributions as shown in Figure 2. Original data are simulated in 2-D feature space, and positive examples are marked with "+" s and negative examples are marked with "o" s, as shown in the figure. In each case, we apply BDA, LDA, and ADA to find the best projection direction by their own criterion functions. The resulting projection lines are drawn in dotted, dash-dotted, and solid lines, respectively. In addition, the distributions of positive and negative samples along these projections are also drawn like bell-shaped thicker and thinner curves along projection line, assuming Gaussian distribution for each class.

From Figure 2, we can see these five cases actually represent several typical data distribution scenarios. Case 1 best fits the distribution where all positive gene expressions are alike while negative ones may be irrelevant (functional dissimilar) to each other and from different distributions (Figure 2(a)). Case 4 is on the opposite side of Case 1, in which negative genes share strong correlations while positive genes may be quite different (Figure 2(d)). Case 2 and Case 3 represent the imbalanced data set, where the size of positive (negative) genes is much larger than that of negative (positive) genes. Case 5 is the scenario where the major descriptive directions of positive genes and negative genes are upright.

From projection results, we can see LDA treats positive and negative samples equally. This makes it a bad choice in Case 1 and Case 4. Similarly, since BDA assumes all positive samples are projected together, it fails in Case 4 and Case 5. In Case 2 and Case 3, BDA and LDA are found not applicable for imbalanced data sets. The reason for this is that LDA or BDA tends to severely bias to the dominating samples.

In all five cases, ADA yields good projection with positive samples and negative samples well separated and outperforms BDA and LDA. Note in Case 4, both BDA and LDA totally fail while ADA still produces a good projection. It clearly demonstrates that no matter whether it is an imbalanced data set or samples are from different subclass clusters, ADA can adaptively fit into different distributions of samples and find a balance between clustering and separating, which are embedded in the criterion function.

In general, the above five cases illustrate several representative scenarios, where our ADA framework could model more feature distributions than LDA or BDA alone. Here, we only show five special cases of ADA. More accurate data model fitting could be achieved by fine parameter tuning.

3. BOOSTING ADAPTIVE DISCRIMINANT ANALYSIS

In Section 4, ADA analysis will show promising performance. However, the optimal classifier often lies not only between but also beyond BDA and LDA in the parametric space of (λ, η). In general it is hard to tell which parameters are best. Searching the whole parametric space will result in heavy computational complexity. Often the best pair we found for one particular data set was different from that of other data sets and therefore no generalization can be made.

AdaBoost [Freund and Schapire 1999], developed in the computational machine learning area, has emerged as a competitive technique that has a theoretically justified ability to improve the performance of any weak classification algorithm in terms of bounds on the generalization error.

The basic idea of boosting is to iteratively reweight the training examples based on the outputs of some weak learners. The intention is to increase the weights of the incorrectly classified examples and decrease the weights of the correctly classified examples. This forces the classifier to focus more on the incorrectly classified examples in the next iteration. The final prediction is the combination of the prediction from each classifier weighted by its classification performance, that is, the smaller the training error rate, the larger the weight.

AdaBoost thus provides a general way of combining and enhancing a set of ADA classifiers in the parametric space. With affordable computational cost, AdaBoost can provide a unified and stable solution to find close to optimal ADA prediction results. The reweight and retraining mechanism is expected to enhance each classifier's performance. Unlike most of the existing approaches that boost individual features to form a composite classifier, our scheme boosts both the individual features and a set of weak classifiers. Our algorithm is as follows:

Algorithm AdaBoost with ADA
Given: Training Sample set X, corresponding label Y and K ADA classifiers with different (λ, η)

Initialization: weight $w_{k,t=1}(x) = 1/|\,\mathrm{X}\,|$

AdaBoost:
For $t = 1, \ldots, T$
For each classifier $k = 1, \ldots, K$ do
i. Train the classifier on weighted mean for all the samples, positive samples and negative samples and weighted scatter matrices in the following way. Note that $\sum_{x\in X} w_{k,t}(x) = 1$.
(a) Update weighted mean μ_{all}, μ_p, and μ_n

$$\mu_{all} = \sum w_{k,t}(x) \cdot x / \sum w_{k,t}(x)$$
$$\mu_p = \sum_{x\in p} w_{k,t}(x) \cdot x / \sum_{x\in p} w_{k,t}(x)$$
$$\mu_n = \sum_{x\in n} w_{k,t}(x) \cdot x / \sum_{x\in n} w_{k,t}(x)$$

(b) Update within-class scatter matrix S_w and between-class scatter matrix S_B
ii. Get the probability-rated prediction on each sample $h_{k,t}(x) \in (-1,\ 1)$
iii. Compute the weight of classifiers $\alpha_{k,t}$:

$$r_{k,t} = \sum_{x\in X} w_{k,t}(x) \cdot h_{k,t}(x) \cdot y \quad \alpha_{k,t} = \tfrac{1}{2} \ln\left(\tfrac{1+r_{k,t}}{1-r_{k,t}}\right)$$

iv. Update the weight of each sample

$$w_{k,t+1}(x) = w_{k,t}(x) \exp(-\alpha_{k,t} \cdot h_{k,t}(x) \cdot y)/Z_t$$

where Z_t is chosen such that $\sum_{x\in X} w_{k,t}(x) = 1$.
End for each classifier
End for t
The final prediction $H(x) = sign\left(\sum_{k=1..K} \sum_{t=1..T} \alpha_{k,t} \cdot h_{k,t}(x)\right)$

4. EXPERIMENTS AND RESULTS

4.1 Adaptive Discriminant Analysis on Yeast Cell Cycle Regulation Data Set

In order to evaluate ADA on gene expression data, we applied ADA, BDA, LDA, and Principle Component Analysis (PCA) [Jolliffe 2002] with Bayes classifier and Support Vector Machine (SVM) on the same data set and compared their classification results.

4.1.1 *Data Set.* We chose to use the baker's yeast (*Saccharomyces cerevisiae)* cell cycle expression data [Eisen et al. 1998] as our first test dataset. A total of 80 different DNA microarray hybridization experiments on 6,221 yeast open reading frames (ORFs) were included. This data set was chosen because the *S. cerevisiae* genome has been sequenced and annotated. Additionally, the large existing knowledge-base for the yeast genome and accumulating experimental supporting evidence makes this organism an ideal test bed for estimating the accuracy of our proposed methods. According to the Comprehensive Yeast Genome Database (CYGD), 4449 among 6221 genes have predicted or characterized functions.

Various cell cycle conditions have been included in the 80 microarray experiments, including α factor-based synchronization, Cdc15-based synchronization,

Table II. Functional Classes and Distribution of Member Genes Used in Our Evaluation

Class ID	Functional Class	Number of genes
1	TCA Cycle	18
2	Respiration	68
3	Cytoplasmic ribsome	171
4	Proteasome	77
5	Histone/Chromosome	51
6	Other classes	1939
Total		2324

elutriation synchronization, Cln3 and Clb2 experiments, and the conditions under nitrogen deficiency, glucose depletion, mitotic cell division, spore morphogenesis and diauxic shift. This data set has served as a benchmark in numerous microarray studies and is publicly available at http://rana.lbl.gov/EisenData.htm.

To compare the performance of classification techniques, we focused on five representative functional classes (Table II) that have been previously analyzed and demonstrated to be learnable by Brown et al. [2000] and Mateos et al. [2002]. Genes involved in the TCA cycle, respiration, cytoplasmic ribosomes, proteasomes, and histone/chromosome were chosen. Many of these genes are tightly regulated in an orchestrated manner, and biologically, these categories of genes might be expected to show coexpression patterns [Ng et al. 2003].

For better evaluation, data were preprocessed by removing annotated genes with incomplete expression, resulting in a data set of 2324 annotated genes. Of those, 385 genes belong to the five functional classes and the remaining 1939 genes are related to other cellular processes (Table II).

4.1.2 *Experiments.* In a widely cited microarray classification study, Brown et al. [2000] compared the performance of SVM, two decision tree learners (C4.5 and MOC1) and Parzen windows, etc. in gene classification to the same data set. SVM, especially SVM with kernel functions, consistently and significantly outperformed the other algorithms for functional classification. Therefore in our experiments, we focused on comparing ADA with Bayes classifier and SVM using different polynomial and radial basis kernel (RBF) functions. Here polynomial kernel functions were $K(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X} * \boldsymbol{Y} + 1)^d$, with $d = 1, 2, 3, 4$, and RBF functions used were $K(\boldsymbol{X}, \boldsymbol{Y}) = \exp(-||\boldsymbol{X} - \boldsymbol{Y}||^2/2\alpha^2)$. In this work, α[1] was set to be a widely used value, the median of the Euclidean distances from each positive example to the nearest negative example [Brown et al. 2000]. Besides, for the sake of showing ADA outperforms the traditional dimension reduction methods such as LDA *etc.,* we also compared the performance of single LDA, BDA with ADA.

In order to compare with SVM, we performed a two-class classification with positive genes from one functional class and the negative genes from the remaining classes. Each gene could be classified in one of the four ways: *true positive* (TP), *true negative* (TN), *false positive* (FP), and *false negative* (FN),

[1]We also searched a range of α for an optimal performance. The best performance is similar as the one in [Brown et al. 2000].

Table III. Comparison of *Precision*, *Recall*, *F_measure* for Various Classification Methods on Yeast Cell Cycle Regulation Data Set

		SVM%					PCA	BDA	LDA	ADA
Class (#)	Method	D-p 1	D-p 2	D-p 3	D-p 4	RBF	(%)	(%)	(%)	(%)
TCA Cycle (class 1)	*Precision*	0.0	**60.56**	65	28.89	3.33	35.24	0.0	59.10	42.32
	Recall	0.0	13.33	16.67	5.56	0.56	50.56	0.0	27.22	**48.89**
	f_measure	0.0	21.15	25.64	9.10	0.95	40.38	0.0	34.98	**44.48**
Respiration (class 2)	*Precision*	0.0	71.21	61.64	47.31	**90.17**	0.0	12.11	40.73	50.56
	Recall	0.0	20.28	22.22	11.39	11.94	0.0	2.50	33.33	**40.83**
	f_measure	0.0	31.13	32.26	17.84	20.68	0.0	3.92	36.37	**44.73**
Cytoplasmic ribsome (class 3)	*Precision*	88.27	89.06	86.12	85.97	**96.28**	0.0	57.38	68.89	70.90
	Recall	47.84	46.55	45.85	43.27	45.67	0.0	40.23	56.67	**58.65**
	f_measure	61.8	60.89	59.6	57.31	61.72	0.0	46.98	62.00	**64.00**
Proteasome (class 4)	*Precision*	0.0	1.667	0.0	0.83	**72.5**	0.0	0.0	37.74	49.72
	Recall	0.0	0.123	0.0	0.12	5.56	0.0	0.0	15.06	**18.02**
	f_measure	0.0	0.23	0.0	0.22	10.17	0.0	0.0	21.04	**26.17**
Histone/ Chromosome (class 5)	*Precision*	10	**90.28**	65.29	52.93	86.67	0.0	0.0	26.54	26.76
	Recall	0.59	11.57	11.18	8.824	9.02	0.0	0.0	15.88	**16.47**
	f_measure	1.11	**20.2**	18.52	14.66	16.16	0.0	0.0	19.44	20.06

according to the CYGD annotation and classifier results. The yeast gene data set is a typical imbalanced data set with a large number of negative genes versus positive genes. For example, a striking difference was observed in the Histone/chromosome class: only 51 positive instances were included, but the number of negative instances reached 2273. In such an imbalanced data, accuracy and single *precision* are not suggested as good evaluation metrics given that FN is more important than FP [Brown et al. 2000].

In order to measure the overall performance of each classifier, we chose to employ *f_measure* = 2*(*Recall***Precision*)/(*Recall* + *Precision*) which takes both *Precision* and *Recall* factors into account [Van Rijsbergen 1979]. By definition, *Precision* = (number of TP instances)/(number of TP + FP predictions), and *Recall* = (number of TP instances)/(number of TP + FN instances). *Recall* is a measure of the completeness of the retrieved set (i.e., the percentage of retrieved objects in the correct answer set), while *Precision* measures the purity of the retrieved set (i.e., the percentage of relevant objects among those retrieved). Usually, there must be a trade-off between these two measures because improving one sacrifices the other. In our case, where imbalanced data with negative instances dominates, *Recall* is the more important measure because it focuses more on FN predictions.

The expression of 2324 annotated genes during the cell cycle listed in Table II served as the ground truth data set. In our experiments, each method classified the genes in the test set to five learnable functional classes. When classifying one class, we set all the genes belonging to that class positive and the remaining genes negative. For each class, we randomly selected 2/3 positive genes and 2/3 negative genes as a training set and the remaining gene data as a testing set for classification. This procedure was repeated 100 times. Finally, the average values of *Recall*, *Precision*, and *f_measure* of 100 rounds were obtained.

Precision, *Recall*, and *f _measure* for five different classifiers on the yeast five functional classes are listed in Table III. The first five methods are SVM using

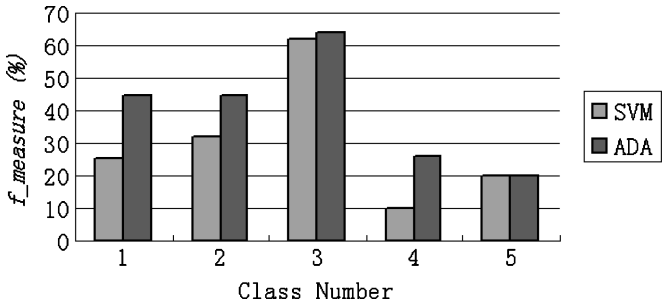

Fig. 3. Comparison of the best *f_measure* value obtained by SVM and ADA on yeast's five functional classes (corresponding class number in Table III).

different polynomial kernels and RBF kernels. Here, D-$p1$ to D-p 4 represents four types of polynomial kernel functions with d varying from 1 to 4. Others are BDA, LDA and ADA. For ADA, we searched (λ, η) from (0, 0) to (1, 1) with step size 0.1, in order to find its optimal value to obtain the best result for each class.

Good and stable performance of ADA is shown in Table III. Using *Recall* or *f_measure* as criteria which are more important evaluation factors than *Precision* when working for an imbalanced data set, ADA outperformed all other methods (except in Histone/Chromosome, where the results are very comparable with SVM),

PCA failed for most classes and gave zero *Precision* and zero *Recall*, because PCA can not determine most discriminant features for these classes. SVM failed for most classes with small size and yielded low *f_measure*. For example, with the Proteasome class, almost zero *Precision*, *Recall*, and *f_measure* were obtained by SVM, indicating little use for SVM in this case. It is likely due to the failure of SVM to find sufficient labeled data to train classifiers well when sample size is small. By contrast, ADA substantially increased the performance of classification, especially on *Recall* and *f_measure*. For example, in the functional class Proteasome, when compared to the best *recall* and *f_measure* obtained by the SVM method (5.56% and 10.17%), ADA improved the results to 18.02% for *Recall* and 26.17% for *f_measure*. In the class TCA Cycle, the best *Recall* and *f_measure* of SVM method were 16.67% and 25.64%, while ADA achieved a significantly better 48.89% for *Recall* and 44.48% for *f_measure*, again significantly better.

Comparison of the best *f_measure* value obtained by SVM and ADA on the five yeast functional classes is shown in Figure 3. Clearly ADA showed superior classification over SVM, demonstrating its promise for classifying microarray gene expression data.

As expected, SVM demonstrated fairly good and stable performance on all the five classes, as indicated by its relatively high *Precision* value, but some exceptions were observed. For instance, the D-p 1 SVM achieved zero P*recision* and zero *Recall* for four of the five classes, suggesting that all the positive instances recognized were incorrect. Despite a seemingly better classification

with higher-dimensional dot product kernel, it is hard to find the dimension (i.e., d) with the best performance.

Compared to ADA, all single dimension reduction methods demonstrated inferior performance on the five classes. For example, BDA yielded unsatisfactory performance for two classes, Proteasome and Histone. This is likely due to the positive genes of these two classes being largely different, and BDA attempting to map all positive examples clustered. The results for all of our yeast experiments show that ADA has its own capability of emphasizing different aspects for the alternative schemes, and offers more flexibility than one single method.

4.2 Validation on *P. falciparum* Microarray Data Set

Given the demonstrated positive performance of ADA on the yeast benchmark data, we next applied it to the microarray data of the erythrocytic (blood) stage in the life cycle of *P. falciparum*, the causative agent of malaria.

4.2.1 *Data Set.* Malaria is a historically established global infectious disease that kills two million people yearly. Over the past decades, the causative agent of malaria, the protozoan parasite *Plasmodium* has developed increased resistance to the antimalarial drugs that were once effective, which poses a threat to public health and underscores the need for a robust pipeline of new drug targets. The completion of the genome sequencing for *P. falciparum*, the most prevalent form of *Plasmodium*, has set the stage for a quantum leap in our understanding of the fundamental processes of the parasite life cycle and mechanisms of drug resistance and immune evasion and opens a new direction of genomics-based drug discovery. However many genes are annotated as hypothetical due to insufficient homology to any known functional proteins [Gardner et al. 2002]. Some 60% of the predicted genes may code for proteins currently lacking valid functional assignments. This high level of ambiguity significantly limits the comparative genomics approach and impedes efforts to achieve a system level understanding of parasite biology.

The release of genome data has allowed scientists to perform microarray expression studies to explore transcriptional profiles of parasites at various developmental stages and/or subcellular locations. A groundbreaking study by the DeRisi lab [Bozdech et al. 2003] using *P. falciparum*—specific microarray of long oligonucleotides (70mers) revealed the expression profiles of every hour for the entire 48-hour red blood cell stage, which manifests clinical symptoms in the host. The original data are downloadable from http://malaria.ucsf.edu/SupplementalData.php, and include the profiles of 46 consecutive time points (excludes the 23-hour and 29-hour time points). With standard quality control filtering and normalization, the complete data set consists of expression for 7091 oligonucleotides corresponding to over 4000 ORFs [Gardner et al. 2002]. Note that the spots with array features with a sum of median intensities smaller than the local background plus two times the standard deviation of the background were recorded as empty; hence their $\log_2(Cy5/Cy3)$value could not be calculated. For our purpose, these empty values were set as zero because they are very small nonnegative values.

Table IV. Functional Classes and Number of Member Genes Reported in [Bozdech et al. 2003]

Group ID	Functional Class	Number of genes
1	Transcription machinery	23
2	Cytoplasmic Translation machinery	159
3	Glycolytic pathway	14
4	Ribonucleotide synthesis	18
5	Deoxynucleotide synthesis	7
6	DNA replication	40
7	TCA cycle	11
8	Proteasome	35
9	Plastid genome	20
10	Merozoite Invasion	87
11	Actin myosin motors	17
12	Early ring transcripts	34
13	Mitochondrial	19
14	Organellar Translation machinery	39
Total		523

The original paper [Bozdech et al. 2003] reported 14 functional classes of proteins which exhibited distinct developmental profiles by Fourier Transform. Classes included components for genetic information flow (DNA replication, transcription, and translation), metabolic pathways (glycolysis, TCA cycle, ribonucleotide, and deoxynucleotide synthesis), protein-protein interaction networks (proteasome), organellar activities (plastid, mitochondria, and organellar translation machinery), and activities related to pathogenesis (merozoite invasion, actin-myosin motility, and early ring activity) (Table IV).

4.2.2 *Experiments.* In our second experiment, we utilized the same two-class classification by SVM, PCA, BDA, LDA and ADA for the *P. falciparum* data set, in order to see whether ADA also performed well for this new data. The polynomial and RBF kernels were the same as in our experiment on yeast data.

We used the classified genes in Table IV as our ground truth (for classification) and their corresponding 46-hour expressions as their feature representation. Because the number of genes in groups 3 and 7 were too small to train, we combined groups 3 and 7 to form a large group, which was acceptable considering that glycolysis and the TCA cycle are naturally consequential in metabolic pathways. Group 4 was also combined with group 5, given that they both represent nucleotide synthetic pathways. Some data with low expressions was also filtered from the data set. In total, 12 groups consisting of 472 genes were assembled after our combinations and filtering.

Because the malaria data set is also an imbalanced data set, we continued to use *f_measure* to measure the overall performance of each classifier. For each class, we randomly selected 2/3 positive genes and 2/3 negative genes as training set and the remaining gene data as testing set for classification. This procedure was repeated 100 times. Finally, we obtained the average values of *Precision*, *Recall* and *f_measure* for 100 rounds for each class.

Table V lists *f_measure* values for these eight different classifiers on the twelve functional classes. From this table, two trends are clearly identifiable.

Table V. Comparison of *F_measure* for Various Classification Methods on *P falciparum* Data Set

	SVM (%)					PCA	BDA	LDA	ADA
Functional Class	D-p 1	D-p 2	D-p 3	D-p 4	RBF	(%)	(%)	(%)	(%)
Transcription	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.57	**21.22**
Cytoplasmic Translation	82.9	86.1	86.3	86	**87.5**	64.68	84.12	79.88	84.12
Glycolysis pathway and TCA cycle	0.0	0.0	0.0	0.0	1.33	0.0	0.0	22.42	**22.79**
Nucleotide synthesis	0.0	0.0	0.0	0.0	0.0	0.0	0.0	21.35	**31.7**
DNA replication	0.0	0.0	0.0	0.0	17.6	0.0	34.3	59.38	**62.2**
Proteasome	0.0	0.0	0.0	0.0	**71**	0.0	0.0	31.45	29.23
Plastid genome	0.0	0.0	0.0	0.0	57.9	0.0	0.55	73.22	**73.22**
Merozoite invasion	79	77.9	75.1	74.1	84.1	30.48	76.92	84.4	**84.95**
Actin myosin motors	0.0	0.0	2.06	7.98	0.0	0.0	6.46	35.34	**39.96**
Early ring transcripts	84.9	86	85.8	85.2	**91.3**	0.0	79.26	89.38	90.65
Mitochondrial	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.97	**35.01**
Organellar Translation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	26.03	**36.37**

First, ADA significantly outperformed SVM for eight classes and was comparable to SVM for three classes (Cytoplasmic Translation, Merozoite invasion, and Early ring transcripts). SVM yielded zero *f_measure* on the majority of classes with small sample size. By contrast, ADA substantially improved the performance of classification. Second, ADA also performed better than BDA for all classes and performed better than LDA on eleven classes. These results demonstrate the capability of ADA to emphasize different aspects for the alternative schemes, thus offering more flexibility than any single method and validate its robust performance for gene classification. Similar as in Table III, PCA failed for most classes and gave zero Precision and zero Recall due to the lack of discriminant features.

4.3 Validation of Boosting Adaptive Discriminant Analysis

We employed this algorithm to the Cytoplasmic ribosome class of yeast cell regulation data set. For comparison, we also boosted individual features for a single classifier LDA and the best pair (λ^*, η^*)classifier found in the parametric space (e.g., the best one among 36 classifiers when the step size is 0.167). Figure 4 shows the results. We can find that: (i) As iteration goes on, the boosting improves the *f_measure* for all algorithms; (ii) Boosted ADA starts with a set of weak classifiers (e.g., 36 classifiers with step size 0.167), but after one iteration, the boosted ADA outperforms the boosted best classifier($\lambda^* = 0.5$, $\eta^* = 1$). This is because not only individual features are boosted but also a set of weak classifiers is combined into a strong one.

As we can imagine, for simply searching the parametric space, the larger the searched space, the better the performance of the best single classifier is. However, the exhaustive search means more computational costs. Table VI shows the boosted ADA classifier and the best single classifier of ADA analysis on Cytoplasmic ribsome class of yeast cell cycle regulation data set in the different search space. The range of (λ, η) is between (0, 0) and (1, 1). The searching step size of λ and η is 0.25, 0.2, 0.167, and 0.1 resulting in the searching space size

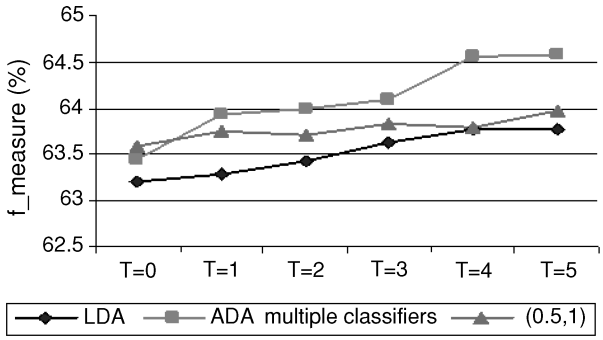

Fig. 4. Adaboosting on Cytoplasmic ribsome class.

Table VI. Comparison of the Boosted ADA Classifiers and Best Single Classifier of ADA Pair on Cytoplasmic Ribsome Class of Yeast Cell Cycle Regulation Data Set

Search space size	*f_measure* (%) of the best single classifier (λ^*, η^*)	Boosted ADA		
		$T = 1$	$T = 2$	$T = 3$
16	63.42 (0.33,0.66)	63.64	63.85	63.92
25	63.44 (0.75, 0.5)	63.61	63.92	63.89
36	63.56 (0.4, 0.8)	63.94	63.98	64.09
100	63.72 (0.5,1)	63.83	63.94	64.45

16, 25, 36, and 100, respectively. We find that the boosted ADA classifier is not sensitive to the size of the search space, for example, the boosted ADA classifier from a weak set of 16 single classifiers achieves the better performance (i.e., 63.85%) than the best single classifier (i.e., 63.72%) of search space size 100 after two iterations. Therefore, instead of searching a large parametric space to find the best single classifier, the boosted ADA classifier provides a more efficient way to combine a small set of classifiers into a more powerful one.

4.4 Putative Genes of Specific Functional Classes Identified by Adaptive Discriminant Analysis

After validating the performance of boosted ADA, we applied it to classify putative malaria genes into six functional categories based on their distinct developmental profiles across the 48-hour erythrocytic cycle. Our resultant list of genes (predicted by boosted ADA to belong to the same class) was compared to independent estimations of their potential functionality based on Gene Ontology [The Gene Ontology Consortium 2000]. Several representative genes that were predicted to belong to six functional classes are shown in Table VII, demonstrating that ADA and boosted ADA are effective expression classification methods.

From a systems biology perspective, the genes in these six selected classes are components of three types of cellular networks with different chemical properties:

(A) Transcription, translation, and DNA replication machineries represent genetic information processing networks with DNA/RNA-protein and

Table VII. Representative Genes Predicted by Boosted ADA That Are Consistent with Gene Ontology Classification

Class	Oligo_ID	Gene_ID	Annotation
Transcription	b182	PFB0290c	Zn-ribbon transcription factor (TFIIS family)
	e1086_1	PFE0305w	transcription initiation factor TFiid, TATA-binding protein
	n176_3	PF14_0469	transcription factor IIIb subunit
	c527	PFC0805w	DNA-directed RNA polymerase II
	i16046_2	PFI1130c	DNA-directed RNA polymerase II
	opfj12798	PF10_0269	DNA-directed RNA polymerase I
Translation	d16785_20	PFD1070w	eukaryotic initiation factor
	j73_18	PF10_0077	eukaryotic translation initiation factor 3, subunit 7
	l2_201	PFL0625c	eukaryotic translation initiation factor 3 subunit 10
	opfk12898	PF11_0447	translation initiation factor eIF-1A
	n150_48	PF14_0104	eukaryotic translation initiation factor 2 gamma
	j346_4	PF10_0103	eukaryotic translation initiation factor 2, beta
	m45177_10	PF13_0178	translation initiation factor 6
	c430	PFC0635c	translation initiation factor E4
	ks259_4	PF11_0245	translation elongation factor EF-1, subunit alpha
	f19787_1	PFI0645w	translation elongation factor 1 beta
	m37794_18	PF13_0214	translation elongation factor 1-gamma
	m45339_1	MAL13P1.243	elongation factor Tu
	opfm60512		
	B379	PFB0550w	peptide chain release factor
	opff72453	MAL6P1.210	nascent polypeptide associated complex alpha c
DNA replication	f18417_1	PFI0235w	replication factor A-related protein
	f57777_2	MAL6P1.125	DNA polymerase epsilon, catalytic subunit a
	i11401_2	PFI0530c	DNA primase, large subunit
	j248_9	PF10_0165	DNA polymerase delta catalytic subunit
	km3535_1	PF10_0362	DNA polymerase zeta catalytic subunit
	kn9335_1	PFI0530c	DNA polymerase alpha subunit III
	m57341_2	PF13_0251	DNA topoisomerase III
	m45918_6	MAL13P1.22	DNA ligase 1
	n157_4	PF14_0254	DNA mismatch repair protein Msh2p
	m446_3	PF13_0291	replication licensing factor
Nucleotide synthesis	d10917_2	PFD0830w	DNA polymerase epsilon
Glycolysis TCA	a1411_1	MAL6P1.160	pyruvate kinase
	opff72425		
	i1689_2	PFI1105w	Phosphoglycerate kinase
	m11919_1	PF14_0425	fructose-bisphosphate aldolase
	m48835_1	PF14_0598	glyceraldehyde-3-phosphate dehydrogenase
	n132_40	PF14_0378	triose-phosphate isomerase
	opfn0252		
	opff72413	MAL6P1.189	hexokinase
	f20989_1	PF08_0045	orotate phosphoribosyltransferase
	f692_1	PFL0630w	dihydroorotate dehydrogenase
	m12812_7	PF13_0141	L-lactate dehydrogenase
	i16629_1	PF13_0229	Inosine-5′-monophosphate dehydrogenase
	j151_11	PF13_0242	adenosine deaminase
	j22_5	PF13_0121	adenylate kinase
	j483_2	PFI1340w	hypoxanthine phosphoribosyltransferase
	j70_2	PF13_0121	GMP synthetase
Proteasome	n147_25	PF14_0716	Proteosome subunit alpha type 1
	opff72439	MAL6P1.88	proteasome subunit alpha type 2
	f960_4	PF07_0112	proteasome subunit alpha type 5
	f960_5		
	e27723_3	PFE0915c	proteasome subunit beta type 1
	n137_50	PF14_0676	20S proteasome beta 4 subunit
	a14680_4	PFA0400c	Beta 3 proteasome subunit

protein-protein interactions. In addition to the key enzymes (DNA-directed RNA polymerase complex), a number of transcription-association factors (TAF) may be important regulators for transcription (Table VII). Very little is known about the transcriptional regulation of *P. falciparum*, and the upstream motifs of many effector proteins are not yet discovered Furthermore, the translation machinery consists of various initiation factors, elongation factors, and peptide release factors.

(B) Glycolysis/TCA cycle and Nucleotide synthesis are typical metabolic networks that include cascades of enzyme-metabolite reactions. Not only does the presence of a series of co-expressed enzymes containing pyruvate kinase, hexokinase, glycerol-3-phosphate dehydrogenase, lactate dehydrogenase (Table VII) suggest that the crucial components in carbohydrate metabolism are conserved in the protozoan, it also portrays the various co-factors and metabolites involved in the activity of each enzyme.

(C) Proteasomes are tightly-wrapped protein complexes combining threonine proteases with regulatory proteins. This protein amalgamation mediates protein-protein interactions involving cell cycle control and stress response. Previously we predicted a catalogue of the threonine proteases and ubiquitin hydrolases as the core elements of malarial proteasome [Wu et al. 2003]. This study reveals an ATP-dependent ubiquitin-proteasome pathway may exist in the malaria parasite.

5. DISCUSSIONS AND CONCLUSIONS

This article proposes a novel Adaptive Discriminant Analysis (ADA) for microarray-based classification. This approach addresses the high dimensionality problem by applying adaptive discriminant projection in an optimal linear discriminant subspace. In order to reduce the computational complexity and combine multiple classifiers into a single more powerful one, boosted Adaptive Discriminant Analysis is also proposed. Our approach is applied to gene classification of yeast cell cycle regulation data and on the *Plasmodium falciparum* data set. The superior performance of our method demonstrates that ADA is a promising and efficient approach to microarray data analysis.

The main contributions of this work are:

(1) ADA provides a richer set of dimension reduction schemes beyond LDA and BDA. It not only compensates for regularization that is afflicted by all sample-based estimation methods, but also finds an optimal projection with adaptation to different sample distributions.

(2) In order to reduce the searching time of parameter space, we propose a boosted Adaptive Discriminant Analysis. It boosts the individual features and a set of weak classifiers. The weighted training scheme in AdaBoost adds indirect nonlinearity and adaptivity to the linear methods, thus enhancing it by iterations. AdaBoost can provide a unified and stable solution to find close to optimal ADA prediction result with affordable computational cost.

(3) ADA provides insights into the components and dynamics of gene regulatory networks [The Gene Ontology Consortium 2000]. A significant limitation for using genome-centric data to decipher gene networks in pathogens is an

inability to infer gene functionality. This study may present an effective method to circumvent this problem: by classifying co-expressed genes in a specific temporal/spatial condition, we may identify co-regulated network modules. Some previously uncharacterized genes have been identified via the proposed ADA, and validation is ongoing (Lu et al., unpublished results). It is well recognized that traditional reductionist studies in which potential targets are selected as single, isolated entities are inefficient because important information is missing. The selected target may function in a time or place that makes it useless for targeting, or it may be possible for it to be replaced by some other redundant protein(s) upon inactivation. Network views of a cell's biology reveal these kinds of information while presenting an intuitive and informative summary of what is known of each protein. Even if no prior information is available for a protein, as in over 60% of the ORF in the *P. falciparum* genome, the contextual information supplied by a network view allows us to tentatively assign a function to the protein. The resultant network view may allow us to uncover potential vulnerabilities in the pathogen, possibly leading to new prevention and therapeutic strategies.

Important and challenging issues in expression data await further investigations. The immediate direction of our future work will be focused on the extension of the 1D vector model in ADA to 2D image model and kernel framework, which is capable of retaining temporal information of time-series expression data and dealing with data that are non-linearly separated.

REFERENCES

Belhummeur, P. N., Hespanha, J. P., and Kriegman, D. J. 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Patt. Analy. Mach. Intelli. 19*, 7, 711–720.

Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. 2003. The transcriptome of the intraerythrocytic development cycle of plasmodium falciparum. *Plos Biol. 1*, 1, 1–16.

Brown, M. P. S., Grundy, W. N., Lin D., and N. Cristianini, et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. In *Proceedings of the National Academy of Science.* 262–267.

Chipman, H., Hastie, T., and Tibshirani, R. 2003. Clustering microarray data. In *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall, Boca Raton, FL.

Cho, S. B. and Won, H. H. 2003. Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the 1st Asia-Pacific Bioinformatics Conference*, 189–198.

Duda, R., Hart, P., and Stork, D. 2001. *Pattern Classification*, 2nd Ed., John Wiley & Sons, Inc.

Dudoit, S., Fridlyand, J., and Speed, T. P. 2000. Comparison of discrimination methods for the classification of tumors using gene expression data. Tech. Rep. 576, Department of Statistics, University of Calofornia Berkeley.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Bostein, D. 1998. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Science.* 14863–14868.

Etemad K. and Chellapa, R. 1997. Discriminant analysis for recognition of human face images. *J. Optical Soci. Amer. 14*, 8, 1724–1733.

Ewans, W. J. and Grant, G. R. 2001. *Statistical Methods in Bioinformatics*, Springer-Verlag.

Fisher, R. A. 1936. The use of multiple measurement in taxonomic problems. *Annals of Eugenics 7*, 179–188.

Fisher, R. A. 1938. The statistical utilization of multiple measurements. *Annals of Eugenics 8*, 376–386.

FREUND, Y. AND SCHAPIRE, R. E. 1999. A short introduction to boosting. *J. Japan. Soc. AI 14*, 5, 771–780.

GANTT, S. M., MYUNG, J. M., BRIONES, M. R., LI, W. D., COREY, E. J., OMURA, S., NUSSENZWEIG, V., AND SINNIS, P. 1998. Proteasome inhibitors block development of Plasmodium spp. *Antimicrob Agents Chemother. 42*, 2731–2738.

GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., AND HYMAN, R. W., ET AL. 2002. Genome sequence of the human malaria parasite. *Plasmodium falciparum. Nature 419*, 498–511.

THE GENE ONTOLOGY CONSORTIUM. 2000. Gene Ontology: Tool for the unification of biology, *Nature Genet. 25*, 25–29,

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. 1999. Data clustering: A review. *ACM Comput. Surv. 31*, 3, 264–323.

JOLLIFFE I. T. 2002. *Principal Component Analysis*, 2nd Ed., Springer-Verlag.

KHAN, J., WEI, J. S., RINGNER, M., SAAL, L. H., LADANYI, M. ET AL. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine 7*, 6, 673–679.

LI, L., WEINBERG, C. R., DARDEN, T. A., AND PEDERSEN, L. G. 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinform. 17*, 12, 1131–1142.

MATEOS, A., DOPAZO, J., JANSEN, R., ET AL. 2002. Systematic learning of gene functional classes from dna array expression data by using multiplayer perceptrons. *Genomes. Resear. 12*, 11, 1703–1715.

NG, S., TAN, S., AND SUNDARARAJAN, V. S. 2003. On combing multiple microarray studies for improved functional classification by whole-dataset feature selection. *Genome Inform. 14*, 44–53.

RINGNER, M., PETERSON, C., AND KHAN, J. 2002. Analyzing array data using supervised methods. *Pharmacogenomics 3*, 403–415.

SWETS, D. AND WENG, J. 1999. Hierarchical discriminant analysis for image retrieval. *IEEE Trans. Patt. Analy. Mach. Intell. 21*, 5, 396–401.

TAMAYO, P., SLONIM, D., MESIROV, J., ET AL. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science.* 2907–2912.

TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J., AND CHURCH, G. M. 1999. Systematic determination of genetic network architecture. *Nat. Genet. 22*, 281–285.

VAN RIJSBERGEN, C. 1979. *Information retrieval*. 2nd Ed., Butterworths.

WU, Y., TIAN, Q., AND HUANG, T. S. 2000. Discriminant EM algorithm with application to image retrieval. In *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*.

WU, Y., WANG, X., LIU, X., AND WANG, Y. 2003. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. *Genome Resear. 13*, 601–616.

ZHOU, X. AND HUANG, T. S. 2001. Small sample learning during multimedia retrieval using bias Map. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Received November 2007; accepted December 2007