Cancerous Tissue Classification Using Microarray Gene Expression

Pei-Chun Chen, Victoria Popic and Yuling Liu
Stanford University

In this project, we apply machine learning techniques to perform tumor vs. normal tissue classification using gene expression microarray data, which was proven to be useful for early-stage cancer diagnosis and cancer subtype identification. We compare the results of both supervised learning (k-nearest-neighbors, SVMs, boosting) and unsupervised learning (k-means clustering, hierarchical clustering) routines on three datasets: GSE3 (renal clear cell carcinoma), GSE8054 (pancreatic cancer) and SRBCT (round blue-cell tumors). In order to eliminate the non-informative genes from the data sets, we apply feature selection using the t-test, differential expression test, and the pairwise correlation coefficient between the class labels and each gene, which boosted the classification accuracy for most of the methods that we carried out. We present the misclassification error rate for hold-out (0.3), 5-fold, and leave-one-out cross validation. In agreement with prior work, we found machine learning techniques to be very efficient with this classification task. For binary classification (cancer vs. normal) the highest accuracy (close to 95% for GSE3 and more than 99% for GSE8054) was achieved with AdaBoost and a linear kernel SVM. For multi-class classification (SRBCT tumor subtypes) we achieve an accuracy of 100% with a linear kernel SVM without feature selection and 98% after reducing the feature dimension by 4 using the correlation coefficient feature selection technique.

Introduction

As the central dogma of biology suggests, DNA makes RNA, which makes protein. Since gene expression (i.e the proteins created in the cell) controls the differentiation of tissues, decoding gene expression has become an important active research area in molecular biology and bioinformatics. The first step towards this goal is to provide an accurate measurement of gene expression levels in different cells. Powerful techniques, such as the microarray technology, have been developed to measure the abundance of mRNA sequences, which is the intermediate product in the gene expression process. This state-of-the-art method is able to simultaneously measure the transcribed mRNA of more than 20,000 genes in the whole human genome. These expression datasets are obtained by quantitatively measuring the hybridization, fluorescence, silverence, and so on, of sample mRNA to the immobilized cDNA on the microarray.

If there is a major difference in cell phenotypes, we can expect a very different gene expression profile. Therefore, gene expression has been widely used to classify different types of tissues. In particular, cancer, non-cancer, and cancer subtype tissue classification is very important since a reliable early-stage cancer detection is able to significantly improve the chances of surviving cancer and precisely identifying different stages of cancer can help plan a better therapy strategy.

There are two major challenges associated with tissue classification based on gene expression: the dimension of the input features and the noise in the data set (mainly due to sample contamination). Usually, the microarray data set is composed of more than 20,000 gene expression measurements. However, due to the high-cost of the microarray experiments, the number of tissue samples is very limited (several hundred at most). Thus the classification model is prone to suffer from overfitting. The second challenge is due to the fact that the cancer tissues extracted are unavoidably contaminated with normal tissues, which introduces significant noise. Since only a small number of genes show significant difference in expression levels between cancer and normal tissues, it is possible for this differences to be overwhelmed by to the noise and the high dimensionality of the feature space.

Significant prior work has been done in this field.[1–4] This work achieves high classification accuracy using various machine learning techniques (e.g., clustering, neural networks, SVMs). In our paper, we also investigate different machine learning methods and feature selection techniques and are also able to achieve high precision in classifying cancer and non-cancer tissues, as well as identify cancer subtypes.

Methods and Results

Data Sets

We evaluate the performance of the classification techniques on two gene expression data sets profiled using the microarray technology. The first data set (GSE3) contains gene expression levels of 36,864 genes reported in the renal clear cell carcinoma study by Boer et al.[5] and consists of 81 cancer tissue samples and 90 normal control samples. The second data set (GSE8054) comes from the allele-specific expression study of pancreatic cancer published by Tan et al.[6] and contains the expression levels of 928 genes in 317 cancer tissue samples and 175 normal tissue samples. We use the expression level of each gene as a separate input feature. Therefore, our input data can be described as an m x n matrix (m = # tissue samples, n = # features), where entry (i, j) represents the expression level of gene j in tissue i. Each sample (i.e. matrix row) is associated with a +1 or -1 label indicating whether the sample tissue is cancerous or normal, respectively.

Gene Selection

In order to find genes with the most predictive power, we filtered/ranked the genes according to the following criteria: (1) t-test (with FDR), (2) significant differential expression (using the MATLAB Volcano plot (Figure 1 shows the volcano plots obtained for the two data sets: the top two quadrants show the genes that were significantly differentially expressed – significantly up or down regulated), and (3) pairwise correlation coefficient between the class labels and each gene (after sorting features according to their absolute correlation coefficient values, we run learning algorithms using different sizes of feature sets to find the optimum set of features as plotted in Figure 2). Due to the high dimensionality of the feature space, techniques such as forward or backward search are too prohibitive. The data set

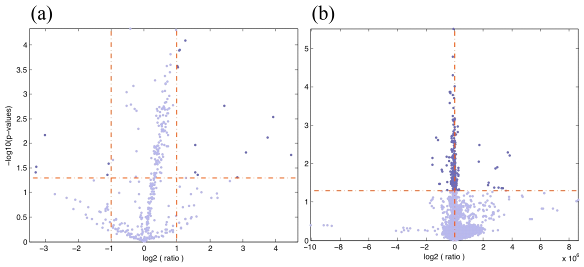

Fig. 1. Feature selection using MATLAB Volcano plots. (a) GSE8054 (b) GSE3.

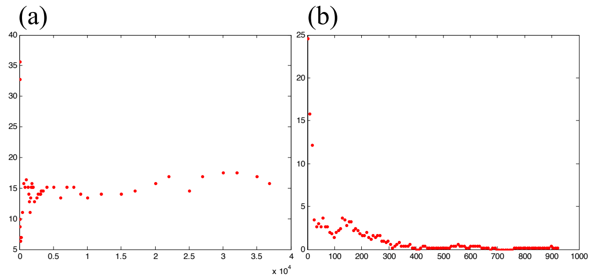

Fig. 2. (a) GSE3 Linear kernel SVM error rate vs. feature-label correlation. The lowest error of 6.43 was achieved for feature size of 150 (around 1/250 of the original feature size). With a small number of features, the error rate is around 35%, which is expected due to high bias. As the feature size grows, the minimum error rate (6.43%) is achieved with feature size 150. The error rate then increases again and levels out around 15% without improvement from using more features. (b) GSE8054 Linear kernel SVM error rate vs. correlation. The lowest error of 0.2 was achieved for feature size 313 (around 1/3 of the original feature size).

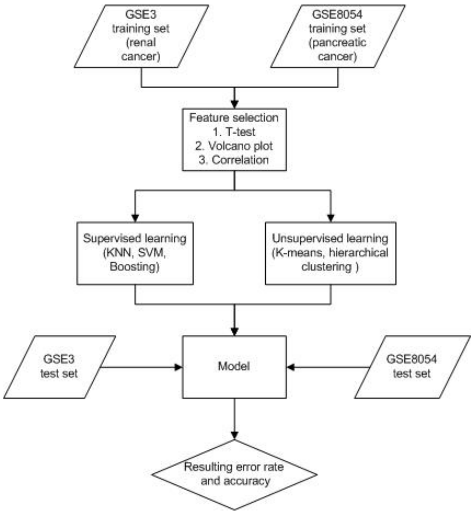
GSE3 training set (renal cancer)
GSE8054 training set (pancreatic cancer)
Feature selection
1. T-test
2. Volcano plot
3. Correlation
Supervised learning (KNN, SVM, Boosting)
Unsupervised learning (K-means, hierarchical clustering)
GSE3 test set
Model
GSE8054 test set
Resulting error rate and accuracy

Fig. 3. Classification pipeline.

feature dimensionality was reduced as follows: GSE8054 928 → 463 by (1) → 313 by (3) → 18 by (2); GSE3 36864 → 924 by (1) → 450 by (2) → 150 by (3). The p-value cutoff used for (1) and (2) was set to 0.05. We evaluated the learning algorithms on the original and each of the reduced data sets.

Classification Pipeline

We explored both supervised and unsupervised learning algorithms for the classification task. In particular, we considered the following techniques: k-nearest neighbors, SVMs, boosting, K-means clustering, and hierarchical clustering. Our data analysis pipeline is summarized in Figure 3. The performance of the learning algorithms is evaluated using hold-out, 5-fold, and leave-one-out cross validation.

Table 1. SVM accuracy with linear, quadratic, and radial basis kernels. Best results are seen for the linear kernel.

Dataset	Validation	Linear	Polynomial	Radial
GSE3	Holdout	0.7368	0.7018	0.5088
	5-fold	0.8421	0.7953	0.5263
	LOOCV	0.7895	0.7995	0.5263
GSE8054	Holdout	0.9756	0.8659	0.8997
	5-fold	0.9980	0.9980	0.9289
	LOOCV	0.9980	0.9959	0.9980

K-Nearest Neighbors (KNN)

The k-nearest neighbor classifier labels a test example with the label of the closest (most similar) example to it in the training set (when $k = 1$) or with the label of the majority of its closest k training examples. The k-nearest neighbor method is non-parametric and simple to implement. We have varied the parameter k $(1 : 10)$, as well as tried several metrics for the similarity measure (available as MATLAB options); in particular: the Euclidean distance, L1 distance (sum of absolute differences), and 'correlation' (1-sample correlation between points). This simple algorithm performs quite well on our data sets (see Figure 4), achieving highest accuracies for smaller values of k. It performed best on the data set GSE8054, which had fewest (and more relevant) features. The misclassification error rate of the classifier also decreased significantly on the GSE3 data set after feature selection was applied.

Support Vector Machines (SVM)

We used the libsvm tool with various options to train and test the datasets.[7] We tried three different kernels types: (1) linear kernel, (2) polynomial kernel with dimension two (quadratic), and (3) radial basis kernel. Table 1 summarizes the classification accuracy of each kernel type. The linear kernel has the best performance. Since the feature vector is high-dimensional comparing to the number of samples in the dataset, mapping features to an even higher dimension increases the variance and thus lowers the testing accuracy. By exploring various feature selection techniques (see Table 2), we acquire better accuracy with a much smaller size of the feature set (around 1/100 for the GSE3 dataset). The computation time and memory requirements are also greatly reduced. More specifically, using the 450 features selected using the volcano plot (out of the 36864 features), the accuracy of GSE3 goes up from 73.68% to 94.15%. The accuracy of GSE8054 stays high using both the t-test and the correlation method, with 463 and 313 features out of the original 928 features, respectively.

Boosting (with AdaBoost)

Adaptive boosting algorithms, such as AdaBoost, build one strong classifier by iteratively adding weak learners that can help classify the examples misclassified so far (this is achieved by maintaining a weight associated with each example s.t. a

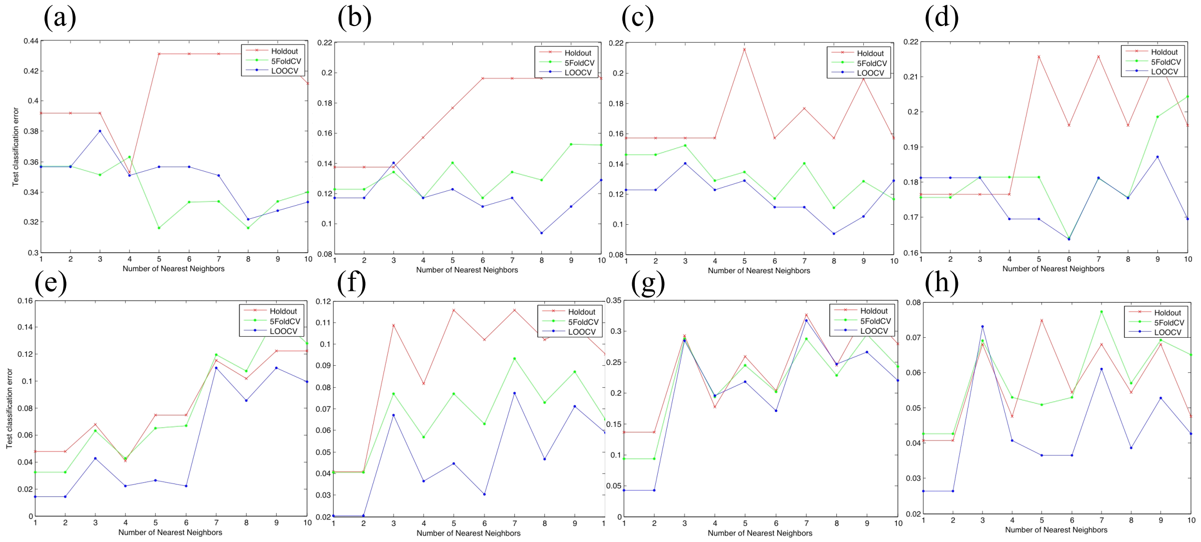

Fig. 4. KNN misclassification rate. Distance metric: euclidean. GSE3: (a) - (d). GSE8054: (e) - (h). (a) and (e) are obtained with original dataset containing all genes, (b) and (f) with t-test gene selection, (c) and (g) with t-test and showing differential expression and (d) and (h) with correlation gene selection.

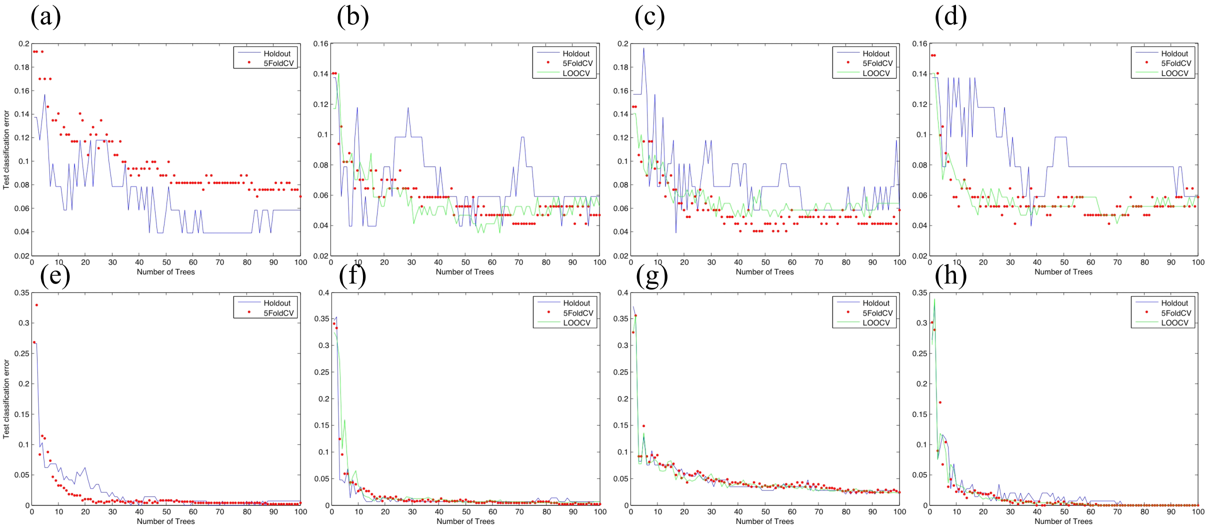

Fig. 5. Boosting misclassification rate. AdaBoost with Tree weak learners. GSE3: (a) - (d). GSE8054: (e) - (h). (a) and (e) are obtained with original dataset containing all genes, (b) and (f) with t-test gene selection, (c) and (g) with t-test and showing differential expression and (d) and (h) with correlation gene selection.

Table 2. Linear kernel SVM accuracy with gene selection. Best accuracy improvements are seen for the t-test based selection.

Dataset	Validation	t-Test	Volcano Plot	Correlation
GSE3	Holdout	0.9415	0.9415	0.9240
	5-fold	0.9298	0.9240	0.9357
	LOOCV	0.9064	0.8947	0.92.40
GSE8054	Holdout	0.9898	0.8435	0.9878
	5-fold	0.9939	0.8435	0.9878
	LOOCV	0.9939	0.8496	0.9939

larger weight is given to the harder to classify data points). We used AdaBoost for our classification problem with the default 'Tree' weak learner. Figure 5 shows the misclassification rate vs. the number of weak learners added to the classification ensemble. As can be seen, this classification algorithm performs very well on our data (the error goes down to 0 on GSE8054 and 0.04 on GSE8054).

K-Means Clustering

We applied the unsupervised k-means clustering approach to partition the samples into tumor and normal cells. Partitioning the samples into only 2 clusters (cancer/normal) can be expected not to perform very well because a particular tumor class can consist of significantly different sub-types and we can expect different gene expression levels to correspond to different cancer stages. We tested several values for the cluster count k (1:5). To assign a label to a test data point, we first assigned a label to each cluster center based on the majority label of the training examples assigned to this center and then used the label of the cluster center closest to the test data point to determine its label. Figure 6 shows the misclassification rate on the original data set and the filtered data set that resulted in the lowest error

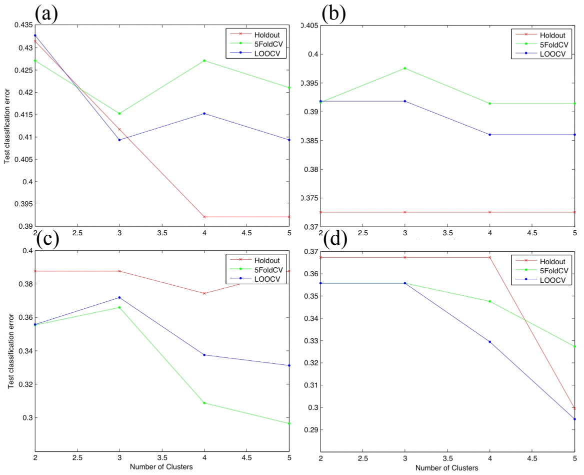

Fig. 6. K-means misclassification rate. (a) GSE3 all features, (b) GSE3 gene selection using correlation (best error rate), (c) GSE8054 all features (d) GSE8054 gene selection using the t-test (best error rate).

rates for this problem. Better results can be seen for higher values of k; however, the classification accuracy of this technique is significantly lower than that of other examined algorithms. Feature selection does improve the performance slightly.

Hierarchical Clustering

Hierarchical clustering is a technique that builds a binary tree by iteratively merging groups of points (individual data points or cluster centers) that are similar to each other by some given distance metric. Therefore, it can efficiently cluster the dataset into a desired number of clusters and, at the same time, provide a visually interpretable distance measure between the data points or cluster centers, which makes it a good choice to visualize and analyze the structure of our data, as some cancer types can contain an arbitrary number of subtypes and usually it is unknown how many or what subtypes a specific cancer has. Thus it is possible to discover or validate the structure of specific cancer types based on hierarchical clustering. We clustered both datasets using this method and the dendrogram of the gene expression levels is shown in Figure 7 (the genes correspond to the columns and the samples correspond to the rows). The green, black, and red colors in the heat maps indicate a low, medium, and high expression of the corresponding gene in this sample, respectively.

Informative patterns can be observed in the graphs. For GSE8054, we can clearly see the two distinct classes (corresponding to cancer and normal tissues) for which the expression of the genes flips; namely, the same gene is down-regulated (green) in samples of one class and up-regulated in the samples of the other class (red). The samples are split into roughly $2/3$ and $1/3$, which corresponds to the number of the cancerous and normal samples in the set, respectively. Comparing the graphs before and after feature selection (both according to the t-test and the pairwise correlation coefficient), we can see fewer genes that have similar expression across the two classes since these genes will be filtered out according to our criteria for feature selection. The dendrogram for GSE3 seems much more complicated, indicating possibly many subtypes in the cancer tissues; we can also find a lot of outliers that don't cluster well with the rest of the samples, which can be the cause of the lower classification accuracy values as compared to GSE8054.

Multiple Class Classification Extension

After running the learning routines for binary classification, we tackle the multi-class classification problem and test the efficiently of gene selection techniques using the data set from the small round blue cell tumor study[3] that contains the expression levels of 2307 genes corresponding to 7 different cancer subtypes: neuroblastoma (NB-C, NB-T), rhabdomyosarcoma (RMS-C and RMS-T), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS-T and EWS-C).

From the previous experiments, we observed that linear-kernel SVM and boosting do well on the classification task. We thus use these two algorithms to train and test on the multiclass dataset with different sets of features. As shown in Figure 8(a), after using correlation-based feature selection for SVM, we get 1.2% error rate using only 500 out of 2307 features, and 0% with 1500 features. For boosting, we get 26.51% error rate with 500 features and 20.83% with 1500 features. Although the boosting error rate is quite high, comparing to the 33.73% error rate using all 2307 features, feature selection technique still helps lowering the error rate. In order to gain graphical insight on the clustering result, we employ the hierarchical clustering algorithm and show the results in Figure 8(b). With the rows and columns corresponding to genes and samples, respectively, we can see the samples are correctly classified into seven clusters.

Conclusion

Gene selection was effective (especially on the GSE3 data set) in improving the accuracy (as well as the speed and memory consumption) of the investigated algorithms; for example, the accuracy of the linear kernel SVM using holdout increased from 73% to 94%. On the GSE8054 data set, the t-test and correlation methods give better results than the differential expression test. This is because only 18 genes are both statistically significant and differentially expressed, making the test too selective.

Highest classification accuracy was achieved using the linear kernel SVM (99% on GSE8054 and 94% on GSE3), boosting (99% on both data sets with 50 weak learners), and KNN (99% on GSE8054 and 88% on GSE3 with small values for k); lowest accuracy was obtained using k-means clustering. The structure of the microarray data itself may be complex so that without enough domain knowledge, it is difficult to decide the number and the initial values of clusters that can give us higher accuracy while utilizing k-means.

We compared the genes selected using the three different techniques for both GSE3 and GSE8054. For GSE3, 59 genes were selected in both the t-test and the correlation methods, and 35 in both the differential expression and the correlation methods. This shows that the gene selection methods agree with each other quite well on the significance of genes. (Note, genes selected by the differential expression test are selected by the t-test as well, since it basically adds expression level constraints on the t-test.) For GSE8054, 147 genes were selected in both the t-test and the correlation methods, and 5 in both the differential expression and the correlation methods. In order to know if the selected genes make biological sense, we examined the 5 genes selected by all three methods in GSE8054. DAPK1, one

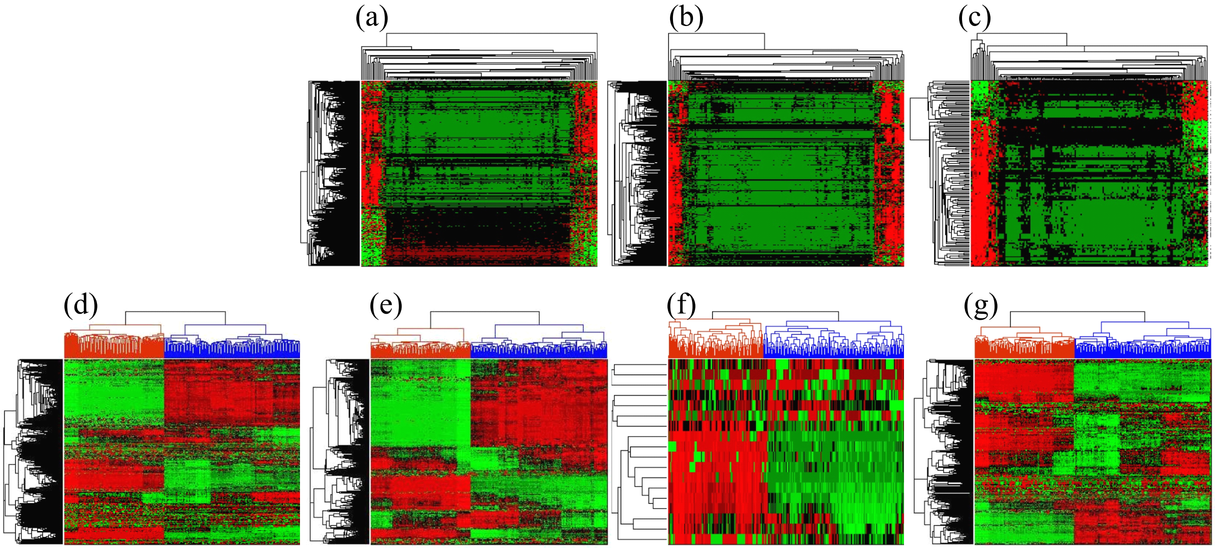

Fig. 7. Hierachical clustering. GSE3: (a) - (c). GSE8054: (d) - (g). (d) are obtained with original dataset containing all genes, (a) and (e) with t-test gene selection, (b) and (f) with t-test and showing differential expression and (c) and (g) with correlation gene selection.

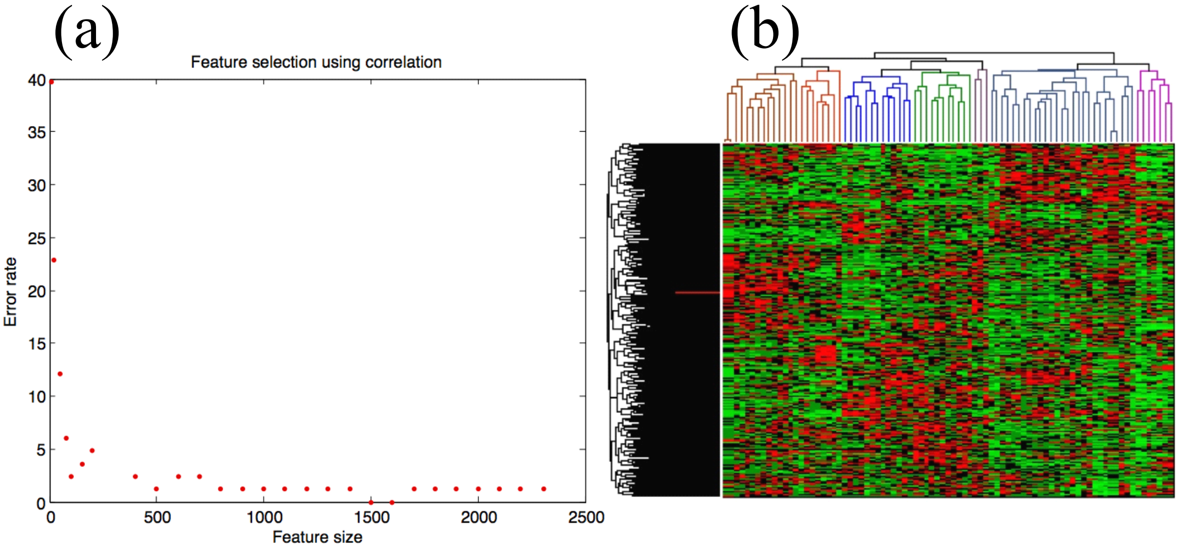

Fig. 8. Multi-Class Classification Results. (a) Linear kernel SVM error rate with correlation-based feature selection. (b) Hierarchical clustering into 7 classes. The binary tree is colored according to the class of the branch.

of the selected genes, is found to be very related to cancer. It is called Death-associated protein kinase 1 and is a tumor suppressor candidate. We also examined the 18 genes selected by the differential expression test in GSE8054 and found TGFA, Transforming growth factor alpha (TGF-α), which is highly related to cancer in its upregulation in some human cancers. These validate the results of our gene selection implementation.

[1] Ben-Dor, A., Shamir, R., and Yakhini, Z. *Journal of computational biology* **6**(3-4), 281–297 (1999).

[2] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. *Journal of Computational Biology* **7**(3-4), 559–583 (2000).

[3] Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., et al. *Nature medicine* **7**(6), 673–679 (2001).

[4] Li, T., Zhang, C., and Ogihara, M. *Bioinformatics* **20**(15), 2429–2437 (2004).

[5] Boer, J., Huber, W., Sültmann, H., Wilmer, F., Von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Füzesi, L., et al. *Genome research* **11**(11), 1861–1870 (2001).

[6] Tan, A., Fan, J., Karikari, C., Bibikova, M., Garcia, E., Zhou, L., Barker, D., Serre, D., Feldmann, G., Hruban, R., et al. *Cancer biology & therapy* **7**(1), 135–144 (2008).

[7] Chang, C. and Lin, C. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 27 (2011).