Computational Methods for Gene Expression-Based Tumor Classification

BioTechniques 29:1264-1270 (December 2000)

Momiao Xiong, Li Jin, Wujun Li, and Eric Boerwinkle

University of Texas—Houston
Health Science Center, Houston,
TX, USA

ABSTRACT

Gene expression profiles may offer more or additional information than classic morphologic- and histologic-based tumor classification systems. Because the number of tissue samples examined is usually much smaller than the number of genes examined, efficient data reduction and analysis methods are critical. In this report, we propose a principal component and discriminant analysis method of tumor classification using gene expression profile data. Expression of 2000 genes in 40 tumor and 22 normal colon tissue samples is used to examine the feasibility of gene expression-based tumor classification systems. Using this method, the percentage of correctly classified normal and tumor tissue was 87.0%. The combined approach using principal components and discriminant analysis provided superior sensitivity and specificity compared to an approach using simple differences in the expression levels of individual genes.

INTRODUCTION

Diagnostic pathology has traditionally relied on macro- and microscopic histology and tumor morphology as the basis for classifying tumors. Current classification frameworks, however, are unable to discriminate among tumors with similar histopathological features but vary in clinical course and response to treatment (12).

Cancer is a dynamic disease. It has been estimated that a new genetic alteration occurs in one out of 10 000 tumor cells at each cell division (15). This leads to an increasing number of genetic aberrations in cancer cells, some of which result in altered gene expression (14). These changes in gene expression are at least partially responsible for the characteristic of a tumor or tumor-type. Genome-wide gene expression profiles provide detailed “tumor signatures” and can potentially be used for molecular diagnosis and classification of tumors (13).

In the past decade, array technologies have been developed that can be used to simultaneously assess the expression of thousands of genes (3,8,11,12,16,17). An early use of human microarrays in cancer biology used an 870 gene array to identify gene expression differences between a malignant melanoma cell line and a nontumorigenic variant of this line (3). It was found that 17% of the genes were expressed at significantly lower levels in the normal cell line than in the tumor cell line and that 7.3% of the genes were expressed at significantly higher levels. Since then, several studies used arrays to analyze gene expression in colon, breast, and other tumors and demonstrated the

potential power of molecular profiles for classifying tumors (1,4,5,9).

Information and data analysis are key components to the successful development of tumor classification systems based on gene expression profiles. Most existing gene expression data in cancer research focus on differential gene expressions in tumor and normal cells or clustering genes into groups based on similarity of gene expressions. However, the characteristics of gene expression profiles are yet to be fully explored for tumor classification.

Alon et al. (1) monitored expression of more than 6500 human genes in 40 tumor and 22 normal colon tissue samples. They used a deterministic annealing algorithm to separate tumor and normal tissues into two distinct clusters. Cluster analysis attempts to classify tissue samples in an unsupervised fashion and does not fully exploit prior knowledge of the tissue samples. In general, accuracy of classification by unsupervised clustering is lower than that by supervised learning algorithms. Golub et al. (4) proposed a neighborhood analysis method for class discovery and class prediction using gene expression profiles. Looking at the gene expression patterns from bone marrow samples of acute leukemia patients, Golub et al. (4) were able to distinguish between patients with acute lymphoblastic leukemia and those with acute myeloid leukemia. Their discriminant decision rules were based on the quantity of prediction strength (PS) and the threshold of the PS. However, they failed to give the procedure used for the automatic selection of the threshold of the PS for classification. Heuristic rules for selection of the PS threshold might

have been used with a certain level of subjectivity. In addition, their method was not formulated in terms of standard statistical frameworks for classification. Therefore, the most widely used statistical methods and software for classification cannot be exploited.

The purpose of this report is to develop an analytic framework for a gene expression-based tumor classification scheme that can allow us to analyze data in a formal, systematic, and integrated manner and investigate the impact of some factors on the accuracy of the proposed tumor classification systems. Expressions of 2000 genes in each of 40 tumor and 22 normal colon tissue samples will be used to show that gene expression profiles provide a precise cancer classification system.

MATERIALS AND METHODS

Discriminant Analysis

We will use Fisher's linear discriminant function for classifying tumor and normal tissues. Fisher's approach does not assume that the observations are normally distributed. It does, however, implicitly assume that the population covariance matrices are equal (6). Tumor and normal tissues are classified on the basis of k selected feature variables. The feature variables can represent the level of expression of k genes or k summary measures representing the level of gene expression of any number of genes or other measures. Suppose that n_N normal and n_T tumor tissue samples are examined. For tissue sample i , we have the vector $Y_i' = (Y_{i1}, Y_{i2}, \dots, Y_{ik})$. The Y_i s for normal (N) and tumor (T) samples constitute the following data matrix,

$$Y_N = [Y_{N1}, Y_{N2}, \dots, Y_{Nn_N}]_{(k \times n_N)} \quad [\text{Eq. 1}]$$

$$Y_T = [Y_{T1}, Y_{T2}, \dots, Y_{Tn_T}]_{(k \times n_T)} \quad [\text{Eq. 2}]$$

From these data matrices, the sample mean vectors and covariance matrices are determined by:

$$\bar{Y}_N = \frac{1}{n_N} \sum_{i=1}^{n_N} Y_{Ni},$$

$$S_N = \frac{1}{n_N - 1} \sum_{i=1}^{n_N} (Y_{Ni} - \bar{Y}_N)(Y_{Ni} - \bar{Y}_N)' \quad [\text{Eq. 3}]$$

$$\bar{Y}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} Y_{Ti},$$

$$S_T = \frac{1}{n_T - 1} \sum_{i=1}^{n_T} (Y_{Ti} - \bar{Y}_T)(Y_{Ti} - \bar{Y}_T)' \quad [\text{Eq. 4}]$$

$$S = \frac{(n_N - 1)S_N + (n_T - 1)S_T}{n_N + n_T - 2} \quad [\text{Eq. 5}]$$

Fisher's idea was to transform the multivariate observations Y_N and Y_T into univariate observations Z_N and Z_T such that the Z s were separated as much as possible. Fisher suggested taking linear combinations of the Y s to generate Z s, which can be easily manipulated mathematically. The midpoint, $\hat{m}$, between the two univariate sample means, $\bar{Z}_N = (\bar{Y}_N - \bar{Y}_T)' S^{-1} Y_N$ and $\bar{Z}_T = (\bar{Y}_N - \bar{Y}_T)' S^{-1} Y_T$, is given by:

$$\hat{m} = \frac{1}{2} (\bar{Y}_N - \bar{Y}_T)' S^{-1} (\bar{Y}_N + \bar{Y}_T) \quad [\text{Eq. 6}]$$

The classification rule based on Fisher's linear discriminant function for an unknown sample, Y_0 , is as follows:

Assign Y_0 to N if $(\bar{Y}_N - \bar{Y}_T)' S^{-1} Y_0 \geq \hat{m}$, and

Assign Y_0 to T if $(\bar{Y}_N - \bar{Y}_T)' S^{-1} Y_0 \leq \hat{m}$.

To evaluate the precision of Fisher's method for classifying tissue type based on its gene expression profile, we will use a cross-validation scheme (10). The tissue samples for classification will be randomly divided into two sets. Fisher's linear discrimination function will be trained on the first set (training set) and then tested on the second set (test set). There is a tradeoff involved in any data partitioning. A small training set will compromise the accuracy of prediction, while a small test set will jeopardize the accuracy of validation. In this report, we will use three different proportional partitionings of the tissue samples: (i) 50% for training and 50% for testing; (ii) 68% for training and 32% for testing; and (iii) 95% for training and 5% for testing.

Q-Mode Analysis

A genome-wide gene expression profile, which involves thousands of genes, constitutes a data set of high dimension. A practical objective is to reduce the number of variables while pre-

serving as much of the original information as possible. Although k variables (i.e., the expression of k genes) are required to reproduce the total system variability, often much of this variability can be accounted for by a small number, p , of principal components. Therefore, the p principal components can replace the initial k variables.

Due to the expense of each array experiment, the number of observations is generally smaller than the number of variables (genes). In addition, for tumor classification studies, biological sample characteristics rather than individual genes are of primary interest. Therefore, in this report, we use Q-mode principal component analysis (2), which is based on a $(n \times n)$ product-moment matrix of the observations, rather than the variables as preclassification data processing.

Let $X = Y' - \bar{Y}'$, where Y' is the transpose of the matrix Y , as similarly defined in Equations 1 and 2, whose elements are the variables of gene expression for tumor and normal tissues. $\bar{Y}'$ is the sample mean of the matrix Y' . Principal component analysis intends to replace the k variables $x_1, x_2, \dots, x_k$ by p principal components $z_1, z_2, \dots, z_p$ where p is smaller than k , while minimizing loss of information. The variables $z_1, z_2, \dots, z_p$ are the principal components if they are linear functions of $x_1, x_2, \dots, x_k$ with the following properties (10):

1. z_1 has the maximum possible variance among all possible linear functions of $x_1, x_2, \dots, x_k$

2. z_2 has the maximum possible variance among all possible linear functions of $x_1, x_2, \dots, x_k$ subject to z_2 being uncorrelated with z_1 .

3. In general, z_p has the maximum possible variance among all possible linear functions of $x_1, x_2, \dots, x_k$ subject to z_p being uncorrelated with $z_1, z_2, \dots, z_{p-1}$ for $2 \leq p \leq k$.

Some principal components may arise as a result of random sampling, measurement error, or individual variation. Such components contain no systematic variation in the original data and will not be used for classification. We include l principal components with the largest eigenvalues, which account for 90% of the summation of the eigenvalues (i.e., 90% of the total vari-

ance of gene expression) in the classification scheme.

Gene Expression Data of Tumor and Normal Colon Tissues

Feasibility of the aforementioned methods is explored using the expression profiles of 2000 genes using an oligonucleotide array (Affymetrix, Santa Clara, CA, USA) in 22 normal and 40 colon cancer tissues, which were originally retrieved from www.molbio.princeton.edu/colondata (1) and are now posted in www.sph.uth.tmc.edu/hgc.

RESULTS

Principal component analysis of the gene expression profiles of 2000 genes from the 62 colon samples revealed that the first seven principal components accounted for 90.44% of the total expression variation. The contribution of the remaining components was small and therefore ignored in subsequent analyses. The performance of Fisher’s linear discriminant function for tumor classification was therefore evaluated by using the loading scores of the seven principal components for each sample as predictor variables.

The total tissue specimens were resampled using jackknifing. The resampled specimens were partitioned into three proportions of the training and testing sets: 50%:50%, 68%:32%, and 95%:5%. The result is presented in Table 1. The accuracy of classification is defined as the percentage of correct classification for both normal and tumor tissues. The accuracy of classification throughout this study was averaged over the 62 times. Using training and test proportions of 68% and 32%, respectively, the accuracy of classification for both the cross-validation set and the test set was 87.0%.

Comparison of expression patterns between normal and tumor tissue samples revealed that a large number of genes were expressed at similar levels. However, the expression profiles also revealed that 42 genes were expressed at significantly different levels ($P < 0.01$). We next determined whether these 42 genes, which show different expression between normal and tumor

Table 1. Percentage of Correctly Classified Normal and Tumor Samples Using the First Seven Principal Components^a

Proportion of Training-Test Set	Accuracy for Training Set	Accuracy for Cross-Validated Set	Accuracy for Test Set
50%–50%	89.5%	84.7%	85.7%
68%–32%	89.8%	87.0%	87.0%
95%–5%	90.2%	87.8%	88.2%

^aResults are an average of 62 replicate partitions.

Table 2. Percentage of Correctly Classified Normal and Tumor Samples Using 42 Genes Showing Significant Differences in Gene Expression

Proportion of Training-Test Set	Accuracy for Training Set	Accuracy for Cross-Validated Set	Accuracy for Test Set
50%–50%	99.7%	54.7%	66.6%
68%–32%	99.8%	58.4%	64.7%
95%–5%	98.6%	65.2%	67.2%

samples, could be used for reliable classification. Fisher’s linear discriminant function was used for classification analysis (using SPSS software), and the results are presented in Table 2. As expected, the accuracies for the training sets using these 42 genes were generally higher than those using the seven principal components. The accuracies for both the cross-validated set and the test set based on the 42 genes were, however, much lower than the principal component analyses. The data presented here indicate that the combined approach of using principal components for data reduction and discriminant analysis for classification provides superior sensitivity and specificity compared to approaches that utilize simple differences in expression levels of individual genes.

DISCUSSION

An alternative approach to morphologic- or histologic-based tumor classification schemes is to use genomic expression profiles. Cancer occurs as a consequence of progressive accumulation of genetic aberrations, which lead to a deviation of gene expression profiles from that of the normal cells (7). In addition, a large number of genes are expressed differentially in tumor versus

normal cells. The pattern of total variations of gene expression across tissues distinguishes normal from tumor tissue and hence can be used for classification purposes. At present, the number of samples studied in a typical microarray gene expression study is much smaller than the number of genes whose expression is being monitored. From a mathematical point of view, it is infeasible to use the gene expression profiles of all of the genes for tumor classification using Fisher’s linear discriminant function because the rank of the covariance matrix describing gene expression is equal to the number of samples, and hence, the covariance matrix is singular. Therefore, the maximum number of variables allowed for any classification study using a linear discriminant function is equal to the number of samples ($n = 62$ for the colon cancer study). To circumvent this problem, a data reduction algorithm is necessary.

There are two ways in which one might attempt to reduce the number of variables and hence the dimensionality of the data set. First, one could simply choose a subset of genes that show different expression between tumor and normal cells. In this study, we chose 42 genes that show significant difference in their expression between tumor and normal tissues using a simple t -test ($P < 0.01$). Using these 42 genes, it was

found that the accuracy of tumor classification using Fisher's linear discriminant function was low. This finding suggests that the status of tissue samples is determined by expression of many genes that have a subtle yet systematic difference between tumor and normal tissues (13).

The level of expression of each gene across tissues was correlated (1). Therefore, it is likely that the information contained in a large number of genes could be captured by a much smaller number of factors without appreciable loss of information. It is well known that principal component analysis is a linear transformation of a set of correlated variables into a set of uncorrelated variables ordered by their variance. Thus, data reduction can be achieved for classification by including only those variables with the greatest variances. This approach is perhaps less appealing, intuitively, than that of

retaining a subset of genes with significant difference in gene expression between normal and tumor tissues. However, it has the advantage that, for the same or less dimensionality, one could retain a greater proportion of the original gene expression information by using principal components. In this report, we demonstrate that, using seven principal components as classification variables, the accuracy of classification is more than 84% for all three partition schemes, which is better than that of using 42 genes selected based on expression difference between normal and tumor tissues.

It is interesting to note that misclassification may primarily be due to tissue sample composition. It has been reported that normal tissues have a high muscle index, while tumors have a low muscle index (1). For the partitioning of the tissue samples used in this report, three normal samples n_8 , n_{34} , and n_{36}

with a muscle index 0.2, 0.2, and 0.1, respectively, were classified as tumors, while three tumor samples t_{23} , t_{27} , and t_{34} with a muscle index of 0.2, 0.2, and 0.2, respectively, were classified as normal tissues. However, muscle index is not a unique factor that determines the type of tissue. Tumor samples with a much higher muscle index (e.g., t_{33} and t_{36} with a muscle index equal to 0.7) were correctly classified as tumors by the cluster algorithm proposed here.

One of the key issues in principal component analysis is to determine how great a reduction in dimensionality can be achieved while retaining much of the original variation. It was shown that seven principal components sufficed to describe the variance structure and accounted for 90.44% expression variations of 2000 genes. To evaluate the impact of the number of principal components used for classification on the percentage of correctly

Figure 1. Classification accuracy for varying numbers of principal components. Training and test proportions were 68% and 32%, respectively.

classified normal and tumor samples, we repeated the classification exercise using varying numbers of principal components. The results are plotted in Figure 1. Fisher's linear classifier using these seven principal components had the highest percentage of correctly classified normal and tumor samples, and using too many principal components for classification reduces the accuracy rather than improves the precision of tumor classification.

In summary, we have demonstrated that molecular classification of tumors is feasible and that a moderate number of tissue samples suffice to achieve reasonable precision of tumor classification. Development of microarray expression technology is revolutionizing tumor classification and may have far-reaching future implication for the diagnosis and treatment of cancer.

ACKNOWLEDGMENTS

M.M.X. is supported by National Institutes of Health grant nos. GM56515 and MH59518. The authors thank two anonymous reviewers for their helpful comments.

REFERENCES

- Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745-6750.
- Basilevsky, A. 1994. Statistical factor analysis and related methods. Theory and Applications, p. 278. John Wiley & Sons, New York.
- DeRisi, J., L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, and J.M. Trent. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14:457-460.
- Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-537.
- Iyer, V.R., M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87.
- Johnson, R.A. and D.W. Wichern. 1982. Applied Multivariate Statistical Analysis, p. 503. Prentice-Hall, Englewood Cliffs, NJ.
- Khan, J., L.H. Saal, M.L. Bittner, Y. Chen, J.M. Trent, and P.S. Meltzer. 1999. Expression profiling in cancer using cDNA microarrays. *Electrophoresis* 22:223-229.
- Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675-1680.
- Perou, C.M., S.S. Jeffrey, M.V.D. Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams et al. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96:9212-9217.
- Roecker, E.B. 1991. Prediction error and its estimation for subset-selected models. *Technometrics* 33:459-468.
- Spellman, P.T., G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297.
- Stephenson, J. 1999. Human genome studies expected to revolutionize cancer classification. *JAMA* 282:927-928.
- Strausberg, R.L. and M.J.F. Austin. 1999. Functional genomics: technological challenges and opportunities. *Physiol. Genomics* 1:25-32.
- Theillet, C. 1998. Full speed ahead for tumor screening. *Nat. Med.* 4:767-768.
- Tlsty, T.D., B.H. Margolin, and K. Lum. 1989. Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbruck fluctuation analysis. *Proc. Natl. Acad. Sci. USA* 86:9441-9445.
- Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15:1359-1367.
- Yang, G.P., D.T. Ross, W.W. Kuang, P.O. Brown, and R. Weigel. 1999. Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. *Nucleic Acids Res.* 27:1517-1523.

Received 24 February 2000; accepted 9 August 2000.

Address correspondence to:

Dr. Momiao Xiong
Human Genetics Center
University of Texas—Houston
P.O. Box 20334
Houston, TX 77225, USA
e-mail: mxiong@utshp.sph.uth.tmc.edu