The Korean Communications
in Statistics Vol. 10 No. 1, 2004
pp. 59-77

Statistical Methods for Gene Expression Data[1]

Choongrak KIM[2]

Abstract

Since the introduction of DNA microarray, a revolutionary high through-put biological technology, a lot of papers have been published to deal with the analyses of the gene expression data from the microarray. In this paper we review most papers relevant to the cDNA microarray data, classify them in statistical methods' point of view, and present some statistical methods deserving consideration and future study.

Keywords : classification, false discovery rate, hybridization, microarray.

1. INTRODUCTION

DNA microarray technology, cDNA microarray(Schena *et al.* 1995) or oligonucleotide chips(Lockhart *et al.* 1996), made it possible to monitor gene expression levels on a genomic scale. Since the introduction of high through-put technology, a lot of papers dealing with the analyses of gene expression data from the microarray have been published during the last 6-7 years. Now, it is a good time to review, summarise, and classify the statistical methods used in analysing gene expression data so far. Of course there were review papers in this area. For example, Bassett *et al.* (1999) argued the necessity for universal standards to make the data more suitable for comparative analysis and for inter-operability with other information resources. Duggan *et al.* (1999) reviewed technical aspects of cDNA micoarrays. On the other hand, Brazma and Vilo(2000) gave a mini-review mainly on the clustering analysis. Smyth *et al.* (2002) made an excellent review on statistical issues in gene expression data. Nguyen *et al.* (2002) focused on biological and technological aspects in DNA microarray experiments. Recently, Sebastiani *et al.* (2003) reviewed many statistical issues in functional genomics. In this paper we make more extensive and updated review on cDNA microarray data, and cover almost all the statistical tools used in relevant papers so far. Also, we try to make this paper easy and self-contained, so that statisticians without prior knowledge on biology or biologists with few statistical backgrounds can understand what's going on in this area. Even though many statistical methods for cDNA microarray data are not available to the oligonucleotide

[1] This research was done when the author visited the Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison during Dec. 2001 and Dec. 2002. This research was supported by the Korea Research Foundation Grant(KRF-2001-013-D00012).

[2] Department of Statistics, Pusan National University, Pusan, 609-735, KOREA

array data, we focus on the statistical methods for cDNA microarray data analysis in this papaer.

The central dogma(DNA replication, RNA transcription, and RNA translation) and the process of producing gene expression data are introduced in Section 2. Section 3 gives the details of generating gene expression data such as image processing, normalization, and two possible types of gene expression data. In Section 4 a lot of statistical methods which are applied to the analysis of gene expression data are introduced. Specifically they are parametric modelling, t-type statistics, issues in multiple testing, analysis of variance model, singular values decomposition, clustering, and discrimination. Finally, some remarks are given in Section 5.

2. BIOLOGICAL BACKGROUNDS

2.1 Gene, mRNA, and cDNA

Every organism has a genome containing biological information which is needed to maintain and construct that organism. Genomes are made of DNA(deoxyribonucleic acid). DNA is polymeric molecules made of nucleotides. Each nucleotide has three parts: a sugar, a phospate group, and a base. The four bases in DNA are adenine(A), cytosine(C), guanine(G), and thymine(T). DNA in living cells is double-stranded, and has double helix form. In the two strands, the base-pairing rules are that A pairs with T and G pairs with C. The two DNA molecules in a double helix are called complementary sequences to each other.

The biological information is contained in a gene which consists of a segment of DNA. The information contained in a gene is read by proteins and this process consists of two stages: transcription and translation - key process of central dogma(DNA replication, transcription, and translation). During transcription DNA is transcribed into mRNA(messenger ribonucleic acid), RNA copy of a gene, and during translation mRNA is translated to produce a protein. These series of biochemical reactions(gene -> mRNA -> protein) are called gene expression, and the corresponding genome sciences are often called genomics, transcriptomics, and proteomics, respectively. Therefore, the abundance of protein is highly dependent on the abundance of mRNA. While it is quite difficult to measure the abundance of protein, it is relatively easy to measure the abundance of mRNA by the method of DNA microarray. Hence measurement of mRNA level gives gene expression data.

2.2 Microarrays

2.2.1 Oligonucleotide Array

Oligonucleotide array, developed by Affymetrix GeneChip (Lockhart *et al.* 1996), has recently been used a lot in genomic research. There are usually 20 probe pairs to interrogate each gene, and each probe pair has a perfect match(pm) and mismatch(mm). The pm probe is made to match a subsequence of the gene(usually 25 bases long), and the mm probe is the same as pm except that the middle base is changed to its complement. Then, the average of the pm-mm differences for 20 probe pairs is used as a gene expression for the gene of interest. For the details of oligonucleotide array, see Lipshutz *et al.* (1999).

2.2.2 cDNA Microarray

cDNA microarray, developed by Brown Lab. of Stanford University (Schena *et al.* 1995), is a glass microarray slide, onto which tens of thousands of single-stranded DNA sequences are attached using a robotic arrayer at fixed spots. The main purpose of cDNA microarray is to compare mRNA abundance, which determines the abundance of a specific protein, in two different samples. Two different samples are usually called targets(or a sample and a control). Two samples are then reversed and transcribed into cDNA, labeled using different fluorescent dyes(red dye Cy5 for sample and green dye Cy3 for control). Both samples are mixed and washed over the microarray, and then hybridized with the arrayed DNA sequence, called probe. Finally, the relative abundance of the hybridized RNA is excited by a laser. The idea of hybridization was already developed in 1970s, and was used, for example, in RNA dot blot and Southern hybridization. At that time nylon or nitrocellulose are used instead of glass. cDNA microarray has much higher density than others and is the most powerful technology so far. The preparation of cDNA microarray must be carefully done depending on the study of interests. The researchers have to make careful decision, for example, on the number of slides, how to hybridize, choice of reference sample, etc.

3. GENERATION OF GENE EXPRESSION DATA

3.1 Gene Expression Profiles

We can build up gene expression profiles under various conditions, for example, under different environmental stress conditions, under different tumor types from cancer patients, and under normal and abnormal cells. Therefore the gene expression profiles can be written in a matrix form with rows representing genes, columns representing samples corresponding to various conditions described above. This gene expression matrix is quite different from the usual matrix appeared in general situations in the sense that the dimension of rows(p) is so big, say, thousands, and that of columns(n) is relatively very small, say, tens. This "tall and skinny" matrix, therefore, has so called "small n, large p" problem.

3.2 Image Processing

The raw data produced by microarrays are monochrome images, so that they are not ready to use until the images are transformed into real values. This transformation process is called image processing. This process is quite complicated and full description of it is very lengthy, and therefore we describe the image process very briefly. The image process mainly consists of 3 parts: addressing, segmentation, and intensity extraction. Addressing is identification of the location of each spot. The segmentation procedure is the classification of pixels as either foreground or background. Foreground is a spot of interest, and background is contribution which is not due to the hybridization. Finally, the intensity extraction step is computing intensities for each spot. The foreground intensity is mean, median, or mode of the pixel intensities within the spot, and there are two ways of computing background intensities; perimeter method, computing median of the pixel intensities in a region surrounding the spot , and local valley method, computing median of the pixel intensities in the local valleys in between spots. There are a lot of methods which are variations of these two methods. See Smyth *et al.* (2002) for details. The final fluorescence of a spot is given by the background correction, subtracting background intensities from the foreground intensities. There are possibilities that the corrected intensities are negative values. This phenomenon can cause reducing the quality and reliability of imaging process, and deserves further studies.

Chen *et al.* (1997) proposed a pixel selection method in cDNA microarray data based on the Mann-Whitney test, and Schadt *et al.* (2000) discussed the process of image analysis, background correction, and normalization problem in oligonucleotide array. Yang *et al.* (2002) reviewed existing image analysis methods and proposed a new image processing method, and implemented it in a software package named Spot. Kooperberg *et al.* (2002) noted that the standard background correction, foreground intensity - back-ground intensity, can cause problems when the foreground intensity is low, and proposed a Bayesian approach for background correction. Also, Chen *et al.* (2002) suggested improved version of image analysis in Chen *et al.* (1997). On the other hand, Bozinov and Rahnenfuhrer (2002) introduced a new method for intensity assessment of gene spots based on clustering, and argued that their approach performed superior to other existing methods and highly robust against various types of artifacts. Recently, Edwards (2003) suggested a smooth correction for the negative background-corrected intensity.

Let R and G be the background corrected red and green intensities, respectively. Then, the most intuitive statistic for the relative red intensity to the green intensity is the ratio of R to G, i.e., $T = R/G$. T could be greater or less than 1 depending on the relative intensity of R to G, and either case is referred to induced ($T > 1$) or repressed ($T < 1$). Discussions based on T will be introduced in section 4.1. But, most authors prefer log-transformed value $X = \log_2(R/G)$ to T as raw data, and Smyth *et al.* (2002) noted some reasons for log-transformation. The base 2 is used for convenience and interpretability. In fact, most of statistical analyses on gene expression profiles are based on data X.

3.3 Normalization

It is not guaranteed that intensities of two dyes are always equal, and it is known that often the intensities are higher for the green dye. Also, if multiple slides are used, there are variations between slides. Further, we cannot avoid biological variations in each RNA extract. Other variations could result from spatial positions of the slide, the process of hybridization, and other microarray technology. To adjust these biases normalization is necessary before carrying out further analysis. If a standard subset of control spots, called housekeeping genes, is available, they can be used in normalization. Schadt *et al.* (2002) suggested to use invariant genes, genes whose ranks remain the same for both the red and green intensities, for normalization. But it is difficult to choose housekeeping genes or invariant genes. Most of normalization methods suggested so far are global normalization by subtracting a constant c from the log-ratio value X. (Chen *et al.* 1997, Kerr *et al.* 2000). On the other hand, Yang *et al.* (2001) proposed intensity-dependent c by using the nonparametric regression such as LOWESS. They also discussed normalization problem on between-slide variations. Tseng *et al.* (2001) showed that normalization between fluorescent labels is necessary and that the normalization is slide dependent and nonlinear. They used residuals after normalization to provide prior information on variance components in the analysis of comparative experiments. Coombes *et al.*(2002) explored various sources of variation in microarray data using high-density cDNA membrane array. By using the non-linear method Wilson et al. (2003) suggested two normalization method, and by using discrimination method Benito *et al.* (2004) adjusted systematic microarray data biases, and called it distance weighted discrimination.

3.4 Two Types of Gene Expression Data

Let $X_{ij}= \log_2(R_{ij}/G_{ij}), i=1,\cdots,p\ ;j=1,\cdots,n$ be the gene expression data for the i-th gene and the j-th sample, then $\boldsymbol{X}=(X_{ij})$ is $p\times n$ gene expression matrix. Sometimes we have additional information for each sample $y_j, j=1,\cdots,n$ to X_{ij} . For example, y_j could be death or survival(binary), tumor categories(polytomous), or survival times(continuous). If only X_{ij} are available, it is called unsupervised, and if both X_{ij} and y_j are available, it is called supervised. The terms, supervised and unsupervised, are usually used in the literature of machine learning. Typical statistical methods for analyzing unsupervised data are clustering analysis and singular values decomposition. On the other hand, discriminant analysis is usually used for the supervised data. The discriminant analysis entails class prediction problem(classification and prediction). These methods will be discussed in detail in Section 4.

3.5 Frequently Quoted Datasets

So far a huge number of gene expression data are generated, and here we list frequently quoted data sets, with short explanation and available web site.

(1) Breast Cancer(Perou *et al.* 1999)
Four groups of human breast cancer (luminal A, luminal B/C, normal, basal /ERBB2); 5531 genes and 85 samples; http://genome-www.stanford.edu/sbcmp

(2) Leukemia (Golub *et al.* 1999)
Two types of acute leukemias (acute lymphoblastic leukemia(ALL) and acute myeloid leukemia(AML)); 6817 genes and 72 samples(43 ALL(38 B-cell ALL, 9 T-cell ALL)and 25 AML); http://www.genome.wi.mit.edu/MPR

(3) Lymphoma (Alizadeh *et al.* 2000)
Three most prevalent adult lymphoid malignancies (chronic lymphocytic leukemia(CLL), follicular lymphoma(FL), and diffuse large B-cell lymphoma (DLBCL)); 4682 genes and 81 samples (29 CLL, 9 FL, 43 DLBCL); http://genome-www.stanford.edu/lymphoma

(4) NCI60 (Ross *et al.* 2000)
60 cell lines from the National Cancer Institute's anti-cancer drug screen; 5244 genes and 61 samples (7 breast, 5 central nervous system, 7 colon, 6 leukemia, 8 melanoma, 9 non-small-cell-lung-carcinoma, 6 ovarian, 2 prostate, 9 renal, 1 unknown); http://genome-www.stanford.edu/nci60

4. STATISTICAL METHODS

4.1 Parametric Modelling

One of the main goals in microarray data analysis is to determine whether gene expression differs significantly for red and green fluorescent intensity levels. Then, the most intuitive statistic for the gene expression is the ratio statistic $T=R/G$ or the log-transformed ratio $X=\log_2(R/G)$, and the naive decision rule for detecting the significant gene is finding genes with high or low values of the ratio statistic. This decision is called k-fold change method: For T, if $T>k$ or $T<1/k$ and for X, $|X|>k$. However, the k-fold change method is not desirable in the sense that the usual pattern of the gene expression data reveals that variance for low intensities are much larger than that of high intensities. Therefore, the k-fold method, not considering the dispersion aspect of intensities, is quite dangerous to use, and this fact was noted by many authors.

Early works include studies of gene expression patterns in human cancerand in yeast during metabolic shift from fermentation to respiration. The first statistical approach to the microarray data analysis is done by Chen *et al.* (1997). They assumed that R_{ij} and G_{ij} are independent and normally distributed with constant coefficient of variation, and derived approximate density for the ratio statistic T. Based on the density they suggested confidence intervals for true ratio ρ and the maximum likelihood estimation of the coefficient of variation parameter. . On the other hand, Newton *et al.* (2000) suggested a Bayesian approach. First, they assumed that R and G follow independent gamma distribution with common scale parameter. For the prior of scale parameter, they assumed common gamma distribution, and obtained the Bayes estimator for the true ratio ρ. Based on the oligonucleotide array, Ibrahim *et al.* (2002) assumed log-normal distribution for the normalized gene expression data and took normal and inverse gamma priors for the mean and variance parameter, respectively. Chen *et al.* (2002) noted that the constant coefficient of variation assumption of Chen *et al.* (1997) is incorrect especially when the signal-to-noise ratio is low, and refined their original results.

4.2 t-type statistics

Lee *et al.* (2000) noted the importance of replication in microarray gene expression analysis. They argued that by pooling data from replicates a more reliable results can be provided since any single microarray output is subject to substantial variability. As mentioned in section 4.1, the use of k-fold change method is very dangerous because it does not consider the dispersion. Therefore, if replicates are available, the easiest and simplest statistic for the gene expression is t-statistic. First we consider the simplest case: one-sample with n replications. Let $\overline{x_i} = \sum_{j=1}^{n} x_{ij} / n$ and $s_i^2 = \sum_{j=1}^{n} (x_{ij} - \overline{x_i})^2 / (n-1)$ be the sample mean and sample variance of the i-th gene, respectively, and let

$$t_i = \frac{\overline{x_i}}{s.e.(\overline{x_i})}, \quad i = 1, \cdots, p$$

Dudoit *et* al. (2002) used a usual type of t-statistic, i.e., they used *s.e.* $(\overline{x_i}) = s_i / \sqrt{n}$. But, some authors noted that the usual t-statistic give too much weight for unusually small values for s_i, while the k-fold change method give too much weight for unusually large values for $\overline{x_i}$ regardless of s_i. In fact, some authors tried to attenuate the usual t-statistic by adding a penalty term in *s.e.* $(\overline{x_i})$. Tusher *et al.* (2001) and Efron *et al.* (2001) suggested *s.e.* $(\overline{x_i}) = a + s_i / \sqrt{n}$ where a is a penalty term. Tusher *et al.* (2001) chose a to minimize the coefficient of variation of t values, and Efron *et al.* (2001) suggested the 90th percentile of s values as a by an empirical Bayes method. On the other hand, Lonnstedt and Speed(2002) suggested *s.e.* $(\overline{x_i}) = \sqrt{(a + s_i^2) / n}$ using a parametric empirical Bayes approach. They estimated a as a function of mean and variance of s^2 values.

It is not difficult to extend the above results to the two-sample case. Assume that the microarray samples consist of treatment and control, and each has n_1 and n_2 replications, respectively, where $n = n_1 + n_2$. Let $\overline{d_i} = \overline{x}_{1i} - \overline{x}_{2i}$ be the difference between treatment and control means, and let s^2_{1i} and s^2_{2i} be the sample variance of treatment and control, respectively. Dudoit *et al.* (2002) used sum of each variance in estimating the standard error, i.e.,

$$t_i = \frac{\overline{d_i}}{\sqrt{\frac{s^2_{1i}}{n_1} + \frac{s^2_{2i}}{n_2}}}$$

while Tusher *et al.* (2001) used the pooled variance with the penalty term as

$$t_i = \frac{\overline{d_i}}{a + s_{pi}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s^2_{pi} = ((n_1 - 1)s^2_{1i} + (n_2 - 1)s^2_{2i})/(n_1 + n_2 - 2)$ is the pooled variance and a a is penalty term.

On the other hand, Wang and Ethier (2004) suggested a generalized likelihood ratio test to identify differentially expressed genes, and showed that this test is more powerful than the fold change method and t-test.

4.3 Multiple Testing Problem and Significance

One of difficult problems involved in microarray data is multiple testing. For example, two-sample t-test might be used to test for significant evidence of differential expression of every individual gene between two samples. Two-sample t-test for each gene with a Type I error of 5% will produce approximately $.05 \times m$ false positives, where m is the number of hypotheses, i.e., the number of genes to be tested.

Table 1. Possible Outcomes from m Statistical Hypotheses

	Do Not Reject Null	Reject Null	Total
Null is True	U	V	m_0
Alternative is True	T	S	m_1
	W	R	m

The most commonly used measure in multiple hypothesis testing is the family-wise error rate (FWER, also called family-wise Type I error), which is the probability of at least one error(false positive) over the collection of tests, i.e., $\Pr(V \geq 1)$. To achieve the FWER, one of the best known method is Bonferroni's correction, but as is well-known it is too conservative. In fact, it controls each test at a/m for a given level of significance a so that the FWER is guaranteed to be less than or equal to a. A less conservative one is Sidak procedure, but it still makes stringent requirements for the rejection of any one of the null hypothesis. These methods are guaranteed to control the level, however, the power is relatively small. To overcome these disadvantages, some step-down methods, improving power with the level preserving, such as Westfall and Young(1993)'s step-down method and permutation method have been suggested. Shaffer(1995) made an excellent review on these methods.

On the other hand, Benjamini and Hochberg(1995) introduced a new measure in multiple testing. They introduced a measure called *false discovery rate*(FDR), the expected proportion of errors committed by falsely rejecting null hypotheses. This concept is very appealing when one is more interested in the rate of false positives among all rejected hypotheses rather than the probability of making at least one Type I error. By using the notations given in Table 1, FDR can be expressed as $E\left[\frac{V}{R}I(R>0)\right]$. When R = 0, the FDR is set to 0. Benjamini and Hochberg(1995) presented an algorithm based on p-values to control the level when the tests are independent. Throughout the simulation study they showed that the FDR is less stringent than the FWER, and is therefore more powerful. Benjamini and Yekutieli(2001) extended the FDR to the dependent tests case, and Genovese and Wasserman(2002) introduced a dual quantity to the FDR, the false non-discovery rate(FNR) which is defined as the expected proportion of the false negatives among not rejected hypotheses, i.e., $FNR=E\left[\frac{T}{W}I(W>0)\right]$. While the FDR is computed by the given level α, Storey(2001) suggested using *positive* false discovery rate(pFDR), which is defined as $pFDR=E\left[\frac{V}{R}\mid R>0\right]$. He interpreted FDR as "the rate that false discoveries occur" and pFDR as "the rate that discoveries are false". He gave a Bayesian interpretation of pFDR such that pFDR is equal to Pr(null is true | test statistic is contained in the rejection region) which is so called "posterior Bayesian Type I error". Storey(2002) discussed estimation issues about pFDR and compared between FDR and pFDR. Also, Reiner, Yekutieli, and Bejamini (2003) suggested a procedure for identifying differentially expressed genes using false discovery rate.

4.4 Cluster Analysis

4.4.1 Methods in Cluster Analysis

Cluster analysis is a very important and most widely used tool for unsupervised type data. Cluster analysis partitions a set of objects into groups or clusters, and within each cluster objects are supposed to be as similar to each other as possible in some sense. There are many different clustering algorithms, but they can be classified into two basic types: hierarchical clustering and non-hierarchical(or flat) clustering. A hierarchical clustering is a hierarchy that each node means a subclass of its mother's node. The tree of a hierarchical clustering can be made either by bottom-up(starting with individual objects and grouping the most similar ones) or top-down(starting with all the objects and dividing them into similar groups). On the other hand, non-hierarchical clustering consists of a number of clusters and the relation between clusters is undetermined. K-means clustering is the simplest and the best-known non-hierarchical clustering. K-means, where K is usually predetermined, is an iterative algorithm that defines clusters by the mean of objects. First, we give a set of initial clusters, and then repeat assigning each object to the cluster whose center is closest. After all objects have been assigned, we calculate the mean of each cluster.

There are many other clustering methods which are versions of the hierarchical and non-hierarchical clusterings. Self-organizing maps(SOMs), developed by Kohonen(1997), is quite similar to K-means clustering, and it is well illustrated in Tamayo *et al.* (1999). Tibshirani *et* al. (1999) investigated two-way clustering to simultaneously cluster both genes and samples, and proposed a new method called gene shaving. Lazzeroni and Owen (2002) suggested plaid models, a form of two-sided cluster analysis that allows clusters to overlap, i.e., it allows a gene to be in more than one cluster or in none at all.

4.4.2 Applications of Cluster Analysis

Eisen *et al.* (1998) seems to be the first paper applying clustering methods to identify groups of co-regulated genes from two sets of data: a single time course of a canonical model of the growth response in human cells and an aggregation of data from experiments on the budding yeast S. cerevisiae. Also, Chu *et al.* (1998) explored assay changes of the budding yeast during sporulation using the clustering method. Spellman *et al.* (1998) used DNA microarray data from yeast cultures to create a comprehensive catalog of yeast genes whose transcript levels vary periodically within the cell cycle.

Tamayo *et al.* (1999) pointed out a number of shortcomings of hierarchical clustering, K-means clustering, and Bayesian clustering, and advocated the use of SOMs in gene expression data. They applied SOMs method to the gene expression data from hematopoietic differentiation, and developed a computer package, GENECLUSTER, to produce and display SOMs of gene expression data. Based on the hierarchical clustering Scherf *et al.* (2000) analyzed gene expression patterns for their relationship to drug sensitivity using NCI60 data set. On the other hand, Kerr and Churchill (2001) applied bootstrapping to assess the stability of results from a cluster analysis. Dougherty *et al.* (2002) developed a congruency model to analyze the inferential precision of clustering algorithms such as K-means, fuzzy C-means, SOMs, and hierarchical clustering. Recently, Ding (2003) suggested two-way ordering to select informative genes from unsupervised data, and applied it to gene expression data.

4.5 Classification

For the supervised type data, where responses are available in addition to the gene expression matrix X, we can predict sample classes based on the data, and this problem is usually called classification which can be also termed as "classification and class prediction". Let be $\mathbf{x}_j = (x_1, \cdots, x_{pj})$ the gene expression profile of the j-th sample of the gene expression matrix $X=(x_{ij})$, and y_j be the class label. Then, x_j and y_j, $j=1,\cdots,n$. correspond to the predictor variable and the response, respectively. The classification step is done by the learning data, data set with known class labels, and using the results of the classification step the class prediction step is done by the test data (or validation data), data with either class labels known or not. The method used in the classification step is called classifier or predictor, and there are numerous classifiers. The classical classifiers are the linear discriminant analysis(Fisher 1936), the nearest neighbor method(Fix and Hodges 1951), and the classification trees(Breiman *et al.* 1984). Also, in the field of machine learning, there are aggregation methods such as bagging(Breiman 1996) and boosting(Freund and Schapire 1997), neural networks(Ripley 1996), and support vector machine(Vapnik 1998).

Using the hierarchical clustering algorithm Perou *et al.* (1999) classified breast carcinomas based on variations in gene expression patterns derived from cDNA microarrays. Alizadeh *et al.* (2000) used hierarchical clustering to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: Diffuse large B-cell lymphoma, follicular lymphoma, and chronic lymphocytic leukemia. Also, using the hierarchical clustering, Ross *et al.* (2000) studied gene expression variation in 60 cancer cell lines(NCI60) and found association between gene expression patterns as well as other properties such as growth rates.

While all the above papers are concerned about the classification only, Golub *et al.* (1999) first studied both the classification and the class prediction. They studied the automatic procedure of discovering the distinction between two types of leukemia: acute myeloid leukemia(AML) and acute lymphoblastic leukemia(ALL). In fact, they used the self-organizing maps for the classification and "weighted gene voting scheme"(a version of the linear discrimination) for the class prediction. Ben-Dor *et al.* (2000) developed a clustering-based classification rule and applied it to several cancer data sets. Hastie *et al.* (2001) proposed a new method, called "tree harvesting" which starts with a hierarchical clustering of genes, and modelled the response as a sum of the average expression profiles of chosen clusters. Tibshirani *et al.* (2002) proposed a new method of class prediction, called "nearest shrunken centroid", which has the similar form to the t-statistics of Tusher *et al.* (2001) and turned out to be very close to the linear discriminant. They developed a computing package called PAM(Prediction Analysis of Microarray). By using a compound covariate prediction classifier, Radmacher *et al.* (2002) argued the use of leave-one-out cross-validation for the computation of misclassification rate and permutation test for the assessment of the significance of the prediction result. Olshen and Jain (2002) suggested using the nearest neighbor classifier and permutation test to derive quantitative conclusions from microarray expression data. On the other hand, Lee and Lee (2003) extended support vector machines(SVM) to the multicategory SVM and applied it to the classification of multiple cancer types.

Some authors compared the performance of classifiers. Brown *et al.* (2000) compared performance of SVM, Parzen windows, linear discriminant, and two decision tree learners, and found that SVM outperform others. Dudoit *et al.* (2002) compared the performance of linear discriminant, nearest neighbor, classification trees, and aggregating classifiers using three well-known gene expression data, and they concluded that linear discrimination or nearest neighbors performed as well as or better than others. Dudoit and Fridlyand (2003) suggested a method using bagging to improve the accuracy of clustering, and Dettling and Buhlmann (2003) discussed a more robust boosting method for tumor classification.

4.6 Analysis of Variance Model

Instead of using the ratio data X_{ij} , Kerr *et al.* (2000) suggested an analysis of variance model based on the raw data when replications are available. In fact, they considered the following model to account for the multiple sources of variation:

$$\log(Y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ij} + (VG)_{kg} + \varepsilon_{ijkg},$$

where Y_{ijkg} is the background corrected intensity for the i-th array, the j-th dye, the k-th sample, and the gth gene, μ is the overall mean signal, A_i is the effect of the i-th array, D_j is the effect of the j-th dye, V_k is the effect of the k-th sample(they used the term "variety" instead of sample), G_g is the effect of the g-th gene, and $(AG)_{ig}$ and $(VG)_{kg}$ represent interaction terms. Also, they assumed that ε_{ijkg} is iid random variables with mean 0 and variance σ^2. The most important term of interest is $(VG)_{kg}$ which captures departures from the overall averages that are attributable to the specific combination of the k-th sample and g-th gene. Also, they argue that this model combine the normalization process with the data analysis because the A, D and V terms effectively normalize the data without preliminary data manipulation. They advocate the log transformation is natural because the effects in the data are believed to be multiplicative. Kerr *et al.* (2000) applied this model to several designs like the Latin square design.

The ANOVA model approach is possible only when replications are done in either within or between arrays. In fact, the importance of replications has been noted by Lee *et al.* (2000) who fit a normal linear mixture model. Recently, Wolfinger *et al.* (2001) suggested two interconnected ANOVA models, the normalization model and the gene model.

4.7 Singular Values Decomposition

As another method for the analysis of gene expression data, the singular values decomposition(SVD) has recently been used. The SVD of the $p \times n$ gene expression matrix **X** of rank r gives **X** = **UDV'**, where **U** is $p \times p$ orthogonal matrix, **D** is $p \times n$ matrix with all zeros except the first r diagonal elements $D_{ii} = d_i > 0$, called singular values, and **V** is $n \times n$ orthogonal matrix. As immediate consequences of SVD by assuming $d_1 \geq \cdots \geq d_r \geq 0$, we have the following: The r nonzero eigenvalues of $\mathbf{XX}'$ and $\mathbf{X}'\mathbf{X}$ are the same and they are squares of singular values d_i. The column vectors of $\mathbf{V}$ are eigenvectors of $\mathbf{X}'\mathbf{X}$. Holster *et al.* (2000, 2001) defined the vectors $\mathbf{a}_i,\ i=1,\cdots,r$ to be the first r rows of $\mathbf{DV}'$, and call it characteristic modes associated with the matrix X. Then, it is clear that the i-th gene expression can be written as a linear combination of r characteristic modes, i.e., $\mathbf{x}_i = \sum_{j=1}^{r} u_{ij}\mathbf{a}_i$. Since $\sum_{j=1}^{r}(u_{ij}d_j)^2 = 1$, for any i, the contribution of the first k modes to the i-th gene is $c_i^{(k)} = \sum_{j=1}^{k}(u_{ij}d_j)^2$. Also, its average over all genes is $\overline{c}^{(k)} = \sum_{i=1}^{p} c_i^{(k)}/p$.

By the SVD analysis of real data sets, Holster *et al.* (2000) argued that the actual gene expressions yield singular values of sufficiently different magnitude so that only the first few modes are required to capture the essential features of the expression data in most cases. In most cases, the first two modes captures many of the essential features, and the average contribution of the first two modes turned out as 62%, 69%, and 72% in the three data sets considered by Holster *et al.* (2000). In fact, the measure of the contribution of modes is very similar to the Shannon entropy. Define the contribution of the i-th singular value as $\varepsilon_i = d_i^2 / \sum_{j=1}^{4} d_j^2$, then the Shannon entropy is defined as $s = \log(1/r)\sum_{i=1}^{r}\varepsilon_i \log(\varepsilon_i)$. Note that s is always between 0 and 1. If s is close to 0, then one or two singular values and the corresponding eigenvectors contain most of the information on the gene expression. On the other hand, if s is close to 1, the singular values are quite uniform. Holster *et al.* (2001) applied the SVD analysis to describe the time evolution of gene expression levels by using a time translational matrix(gene expression matrix with columns representing the time translation) to predict future expression levels of genes based on their expression levels at initial times. This analysis corresponds so called to dynamic modelling instead of static modelling in gene expression data. Also, Alter *et al.* (2000) used the SVD analysis to the gene expression data, and represented the first few eigenvectors as sinusoidal functions.

5. REMARKS

During the last 7-8 years a lot of papers about the statistical analyses of gene expression data have been published since the introduction of the microarray technology. In this paper we review, summarise, and classify statistical methods used in those papers. Broadly speaking those papers can be classified into two categories: (1) How to get reliable and accurate gene expression data? and (2) How to find the significantly expressed genes and how to interpret the results?. The first category contains issues in image processing and normalization, and the second one contains many statistical methods or models such as parametric modelling, t-type statistics, multiple testing problems, cluster analysis, discrimination, analysis of variance model, singular values decomposition, and etc.

Even with big recent improvements in this area there are still lots of rooms for further researches deserving careful consideration. Here we list our personal thoughts on some possible future research area. (1) In the process of image analysis and normalization a measurement error model approach could be an alternative method because the observed intensity entails a lot of variations which can be regarded as measurement errors. (2) Also, a wide class of transformation should be considered instead of log transformation of the ratio of the red intensity to the green intensity. (3) Future research must be done in applying the nonparametric regression to the normalization step since the dispersion varies a lot along with the amount of intensities. Nonparametric variance function estimation would be worth consideration. (4) More extensive comparisons between many methods of discrimination are also worth pursuing. So far, comparisons are done for limited classifiers. (5) "Small *n*, large *p*" problem make it impossible to regard finding the significantly expressed genes as a variable selection problem in regression, however, the approach could be possible if we can choose a moderate sized subset of coregulated genes.

REFERENCES

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Brown, P.O., Botstein, D. and Štaudt, L.M.(2000) Different types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, 403, 503-511.

[2] Alter, O., Brown, P.O., and Botstein, D. (2000) Singular value decomposition for genome -wide expression data processing and modeling, *Proceedings of the National Academy of Science*, 97, 10101-10106.

[3] Basset, D.E., Eisen, M.B., Boguski, M.S. (1999) Gene expression informatics - it's all in your mine. *Nature Genetics*, 21, 51-55.

[4] Ben-Dor, A., Bruhn, L.K., Friedman, N., Nachman, I., Schummer, M., and Yakini, Z. (2000) Tissue classification with gene expression profiles, *Journal of Computational Biology*, 7, 559-583.

[5] Beito, M., Parker, J., u, Q., 졔, J., Xiang, D., Perou, C.M., and Marron, J.S. (2004) Adjustment of systematic microarray data biases, *Bioinformatics*, 20, 105-114.

[6] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Ser. B., 57, 289-300.

[7] Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency, *The Annals of Statistics*, 29, 1165-1188.

[8] Bozinov, D. and Rahnenfuhrer (2002) Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering, *Bioinformatics*, 18, 747-756.

[9] Brazma, A. and Vilo, J. (2000) Gene expression data analysis, *Federation of European Biochemical Societies Letters*, 480, 17-24.

[10] Breiman, L. (1996) Bagging predictors, *Machine Learning*, 24, 123-140.

[11] Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J. (1984) *Classification and Regression Trees*, Wadworth, Belmont, CA.

[12] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, Jr., M., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Science*, 97, 262-267.

[13] Chen, Y., Dougherty, E.R., and Bittner, M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics*, 2, 364-374.

[14] Chen, Y., Kamat, V., Dougherty, E.R., Bittner, M.L., Meltzer, P.S., and Trent, J.M. (2002) Ratio statistics of gene expression levels and applicatins to microarray data analysis, Bioinformatics, 18(9), 1207-1215.

[15] Chu, S., DeRisi, j., Eisen, M., Mulholland, J., Boststein, D., Brown, P.O., and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast, *Science*, 282, 699-705.

[16] Coombes, K.R., Highsmith, W.E., Krogmann, T.A., Baggegly, K.A., Stivers, D.N., and Abruzzo, L.V. (2002) Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays, *Journal of Computational Biology*, 9, 655-669.

[17] Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data, *Bioinformatics*, 19, 1061-1069.

[18] Ding, C.H.Q. (2003) Unsupervised feature selection via two-way ordering in gene expression analysis, *Bioinformatics,* 19, 1259-1266.

[19] Dougherty, E.R., Barrera, J., Brun, M., Kim, S., Cesar, R.M., Chen, Y., Bittner, M., and Trent, J.M. (2002) Inference from clustering with application to gene-expression microarrays, *Journal of Computational Biology*, 9, 105-126.

[20] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays, *Nature Genetics Supplement*, 21, 10-14.

[21] Dudoit, S., Fridlyand, J., and Speed, T. (2002) Comparison of methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97, 77-87.

[22] Dudoit, S. and Fridlyand, J. (2003) Bagging to improve the accuracy of a clustering procedure, *Bioinformatics*, 19, 1090-1099.

[23] Edwards, D. (2003) Non-linear normalization and background correction in one-channel

cDNA microarray studies, *Bioinformatics*, 19, 825-833.

[24] Efron, B., Tibsirani, R., Storey, J.D., and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, 96, 1151-1160.

[25] Eisen, M., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Science*, 95, 14863-14868.

[26] Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7, 179-188.

[27] Fix, E. and Hodges, J. (1951) Discriminatory analysis, nonparametric discrimination: consistency properties, Technical Report, School of Aviation Medicine, U.S. Air Force.

[28] Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55, 119-139.

[29] Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate problem, *Journal of the Royal Statistical Society, Ser.* B., 64, 499-517.

[30] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander E.S. (1999) Molecular classification of cancer : class discovery and class prediction by gene expression monitoring, *Science*, 286, 531-537.

[31] Goryachev, A.B., Macgregor, P.F., and Edwards, A.M. (2001) Unfolding of microarray data, *Journal of Computational Biology*, 8, 443-461.

[32] Hastie, T., Tibshirani, R., Botstein, D., and Brown, P.O. (2001) Supervised harvesting of expression trees, *Genome Biology*, 2, 1-12.

[33] Holster, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., and Fedoroff N.V. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *Proceedings of the National Academy of Science*, 97, 8409-8414.

[34] Holster, N.S., Maritan, A., Cieplak, M., Fedoroff N.V., and Banavar, J.R. (2001) Dynamic modeling of gene expression data, *Proceedings of the National Academy of Science*, 98, 1693-1698.

[35] Ibrahim, J.G., Chen, M-H., and Gray, R.J. (2002) Bayesian models for gene expression with DNA microarray data, *Journal of the American Statistical Association*, 97, 88-99.

[36] Kerr, M.K., Martin, M., and Churchill, G.A. (2000) Analysis of variance for microarray data, *Journal of Computational Biology*, 7, 819-837.

[37] Kerr, M.K., and Churchill, G.A. (2001) Bootstrapping cluster analysis : Assessing the reliability of conclusions microarray experiments, *Proceedings of the National Academy of Science*, 98, 8961-8965.

[38] Kohonen, T. (1997) *Self-Organizing Maps*, Springer, Berlin.

[39] Kooperberg, C., Fazzio, T.G., Delrow, J.J. and Tsukiyama, T. (2002) Improved background correction for spotted DNA microarrays, *Journal of Computational Biology*, 9(1), 55-66.

[40] Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data, *Statistica Sinica*, 12, 61-86.

[41] Lee, M.T., Kuo, F.C., Whitemore, G.A., and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations, *Proceedings of the National Academy of Science*, 97, 9834-9839.

[42] Lee, Y. and Lee, C-K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, 19, 1132-1139.

[43] Lipshutz, R. J., Fodor, S., Gingeras, T., and Lockhart, D. (1999) High-density synthetic oligonucleotide arrays. *Nature Genetics,* supplement 21, 20-24.

[44] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M.,Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 14, 1675-1680.

[45] Lonnstadt, I. and Speed, T. (2002) Replicated microarray data, *Statistica Sinica*, 12, 31-46.

[46] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2000) On differential variability of expression ratios : Improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology*, 8, 37-52.

[47] Nguyen, D.V., Arpat, A.B., Wang, N., and Carroll, R.J. (2002) DNA microarray experiments: Biological and technological aspects, *Biometrics*, 58, 701-717.

[48] Olshen, A.B. and Jain, A.N. (2002) Deriving quantitative conclusions from microarray expression data, *Bioinformatics*, 18, 961-970.

[49] Perou, C.M., Jeffrey, S.S., Van de Rijin, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschkov, A., Williams, C.F., Zhu, S.X., Lee, J.C.F., Lashkari, D., Shalon, D., Brown, P.O., and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proceedings of the National Academy of Science*, 16, 9212-9217.

[50] Radmacher, M.D., McShane, L.M., and Simon, R. (2002) A paradigm for class prediction using gene expression profiles, *Journal of Computational Biology*, 9, 505-511.

[51] reiner, A., Yekutieli, D., and Benjamini, Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics*, 19, 368-375.

[52] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, U.K..

[53] Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Shalon,

D., Meyers, T.G., Weinstein, J.N., Botstein, D., and Brown, P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, 24, 227-234.

[54] Schadt, E.E., Li, C., Su, C., andWong, W.H. (2000) Analyzing high-density oligonucleotide gene expression array data, *Journal of Cellular Biochemistry*, 80, 192-202.

[55] Schadt, E.E., Li, C., Ellis, B., and Wong, W.H. (2002) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data, *Journal of Cellular Biochemistry*, 84, 120-125.

[56] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467-470.

[57] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Kohn, K.W., Reinhold, W.C., Meyers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., and Weinstein, J.N.(2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, 24, 236-244.

[58] Sebastiani, P., Gussoni, E., Kohane, I.S., and Ramoni, M.F. (2003) Statistical challenges in functional genomics (with discussion), *Statistical Science*, 18, 33-70.

[59] Shaffer, J.P. (1995) Multiple hypothesis testing, *Annual Review of Psychology*, 46, 561-576.

[60] Smyth, G.K., Yang, Y.H., and Speed T. (2002) Statistical issues in cDNA microarray data analysis, Functional Genomics: *Methods and Protocols*, To appear.

[61] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Andres, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccaromyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, 9, 3273-3297.

[62] Storey, J.D. (2002) A direct approach to false discovery rates, *Journal of the Royal Statistical Society Ser.* B. 64, 479-498.

[63] Tamayo, P., Slonim, T., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation, *Proceedings of the National Academy of Science*, 96, 2907-2912.

[64] Tibshirani, R., Hastie, T., Eisen, M., Ross, D.T., Botstein, D., and Brown, P.O. (1999) Clustering methods for the analysis of dna microarray data, Technical Report, Department of Health Research and Policy, Stanford University.

[65] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G.(2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Science*, 99, 6567-6572.

[66] Tseng, G.C., Oh, M-K., Rohlin, L., Liao, J.C., and Wong, W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research*, 29, 2549-2557.

[67] Tusher, V.G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Science*, 98, 5116-5121.

[68] Vapnik, V.N. (1998) *Statistical Learning Theory*, Wiley, New York.

[69] Wang, S. and Ethier, S. (2004) A generalized likelihood ratio test to identify differentially expressed genes from microarray data, *Bioinformatics*, 20, 100-104.

[70] Westfall, P.H. and Young, S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*, Wiley, New York.

[71] Wilson, D.L., Buckley, M.J., Helliwell, C.A., and Wilson, I.W. (2003) New normalization methods for cDNA microarray data, *Bioinformatics*, 19, 1325-1332.

[72] Wolfinger, R. D., Gibson, G., Wolfinger, E.D., Bennett, L., Madadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology*, 8, 625-638.

[73] Yang, Y.H., Buckley, M.J., Dudoit, S., and Speed, T.P. (2002) Comparison of methods for image analysis on cDNA microarray data, *Journal of Computational and Graphical Statistics*, 11, 108-136.

[74] Yang, Y.H., Dudoit, S., Luu, P., and Speed, T.P. (2001) Normalization for cDNA microarray data, in *Microarrays : Optical Technologies and Informatics*, Proceedings of SPIE.

[Received December 2003, Accepted February 2004]