©BRAND X, PHOTODISC

New Criteria for Selecting Differentially Expressed Genes

Filter-Based Feature Selection Techniques for Better Detection of Changes in the Distributions of Expression Levels

BY LIT-HSIN LOO,
SAMUEL ROBERTS,
LEONID HREBIEN,
AND MOSHE KAM

Complementary DNA (cDNA) microarrays [1] enable the screening of a biological sample for expressions of thousands of genes simultaneously under a variety of conditions [2], [3]. These genes contain information for making proteins that are essential to the structure and function of living cells. A major task in microarray studies is to select genes associated with specific physiological or pathological conditions. Given a set of candidate genes, there are two possible gene selection targets, namely *differentially expressed* genes and *relevant* genes. A gene is "differentially expressed" when it exhibits a certain distribution of expression levels under one condition and a significantly different distribution of expression levels under another condition. A gene is "relevant" in terms of a Bayes decision rule if removal of the gene alone will result in performance deterioration of an optimal Bayes rule used for classifying the rest of the candidate genes [4]. Although relevant genes are differentially expressed, differentially expressed genes are not necessary relevant.

In many applications of cDNA microarrays, such as the identification of biomarkers for clinical diagnosis, selecting relevant genes is sufficient. However, with the advent of better cDNA microarray technologies and the completions of whole genome sequencings for several species, scientists are no longer satisfied with the simple association or dissociation of a gene to a certain condition. One of the current interests in genomic research is to study the interaction between major genes involved in a biological process under a condition of interest by building genetic network or pathway [5]. Usually these major genes only represent a small subset of all the genes assayed in a microarray. Thus gene selection is potentially helpful to improve the quality of the network built and reduce the computational time needed to build the network. Although many of these genes have indispensable roles within a network, they may not be relevant. To illustrate how a differentially expressed gene can be irrelevant, an exemplary network consisting of three genes is shown in Figure 1(a). In this network, both Gene 1 and Gene 2 can activate Gene 3. Under condition A, none of the three genes is activated and no phenotype is thus observed. Under condition B, the activation of Gene 3 causes an observed phenotype. The network has two operating states under this condition. In the activated state 1, the activation of Gene 1 inhibits Gene 2. In the activated state 2, the deactivation of Gene 1 activates Gene 2. Expression-level distributions of Gene 3 and Gene 1, which conform to the given network structure, are shown in Figure 1(b), where the distribution of expression levels under condition A is shown by the dashed red line and the distribution of expression levels under condition B is shown by the solid green line. In this example, Gene 3 is a relevant gene and its distributions of expression levels obtained under different conditions have small overlap. Gene 1 is irrelevant by definition because it will give a high error rate on most classifiers. Yet, it is differentially expressed as evident from the significant change in the distributions of expression levels. All of the three genes, which participate in the network, should be selected. Since the complexity of interactions between genes and the number of layers of a real genetic network are usually higher than this exemplary network, we expect many genes in a real network will have overlapping distributions such as Gene 1. Thus, selecting differentially expressed genes is a more appropriate target in these applications of cDNA microarrays.

The two most common methods of gene selection are *wrapper-based* and *filter-based* [4]. In wrapper-based methods [Figure 1(b)], genes are selected by ranking subsets of genes through a classifier. The classifier estimates the predictive power of a subset and uses the estimate as a score to rank the subset. The process is quite often repeated iteratively to evaluate subsets of genes. Recently, Guyon et al. proposed a support-vector machine-based recursive feature elimination algorithm for ranking genes [6]. Although wrapper-based methods do not assume a specific data model, these methods are designed to select relevant genes. Depending on the gene selection threshold, they are unlikely to select differentially expressed but irrelevant genes (such as Gene 1).

A better way to select differentially expressed genes is to use filter-based methods. In filter-based methods, a criterion value based solely on the property of a gene is calculated for each gene and used to decide if the gene should be selected [Figure 1(b)]. In hypothesis-driven analyses, statistical tests are often used to perform the gene selection. A statistical test is a procedure for deciding whether a hypothesis about a quantitative feature of a population is true or false. In hypothesis testing, the criterion value is

0739-5175/07/$25.00©2007IEEE

also called a test statistic. Common gene selection criteria include the Welch t-statistic (WTS) and the Wilcoxon (or Mann-Whitney) rank sum. After obtaining the test statistic for each gene, the statistical significance of each test statistic is assessed in order to identify differentially expressed genes. The assessment requires the distribution of the test statistic under null hypothesis. For example, one often assumes that expression levels are normally distributed. Thus, the null distribution of a WTS can be approximated by the t-distribution. When the normality assumption may not be valid, one can still use the Wilcoxon rank sum (WRS), where the test statistic is z-distributed for a large number of samples.

Recently, various resampling-based tests have been used for estimating the null distributions of test statistics empirically without making any assumption on the data model. For instance, resampling-based tests of WTS [7], [8], the Fisher correlation score [2], and significance analysis of microarrays (SAM) [9] were proposed. Although these tests have different procedures to assess the significance of the obtained test statistics, most of them still use variants of the t-statistic as their criteria. Let us assume that the expression levels of a gene come from two classes of samples obtained from two different conditions: the positive class (+1) and the negative class (–1). The expression levels of gene G_i are represented by a vector $G_i = [\, x_{i,1} \quad x_{i,2} \quad \ldots \quad x_{i,n}\,]$, where $x_{i,j}$ is the expression level of the ith gene on the jth sample, and $i = 1, 2, \ldots m$. The gene vector G_i represents the gene expression levels of the ith gene in all n samples. We use $y_j \in \{-1, +1\}$ to label the class of the jth sample. The WTS for each expression level vector of a gene, G_i, is defined as [8]

$$\mathrm{WTS}(G_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}}. \tag{1}$$

Here n^+ and n^- are the numbers of expression levels, μ_i^+ and μ_i^- are the means of expression levels, and $(\sigma_i^+)^2$ and $(\sigma_i^-)^2$are the variances of expression levels in the positive and negative classes, respectively. By using resampling to estimate the distribution of WTS under null hypothesis, we can relax the assumptions on the distributions of the expression levels and make the procedure nonparametric [7], [8]. The WRS test may seem to be a better candidate for testing the null hypothesis without making any assumption about the data distribution. However, past studies on microarray data [8] have demonstrated that the resampling-based tests using WTS are often more powerful than the WRS test, even when the data are not normally distributed. In statistics, the *power of a test* is the probability that the test will reject the null hypothesis when the alternative hypothesis is true. In other studies [2], [10], the Fisher correlation score (FCS), a variant of WTS, was used to score genes. The FCS of a feature vector G_i is defined as

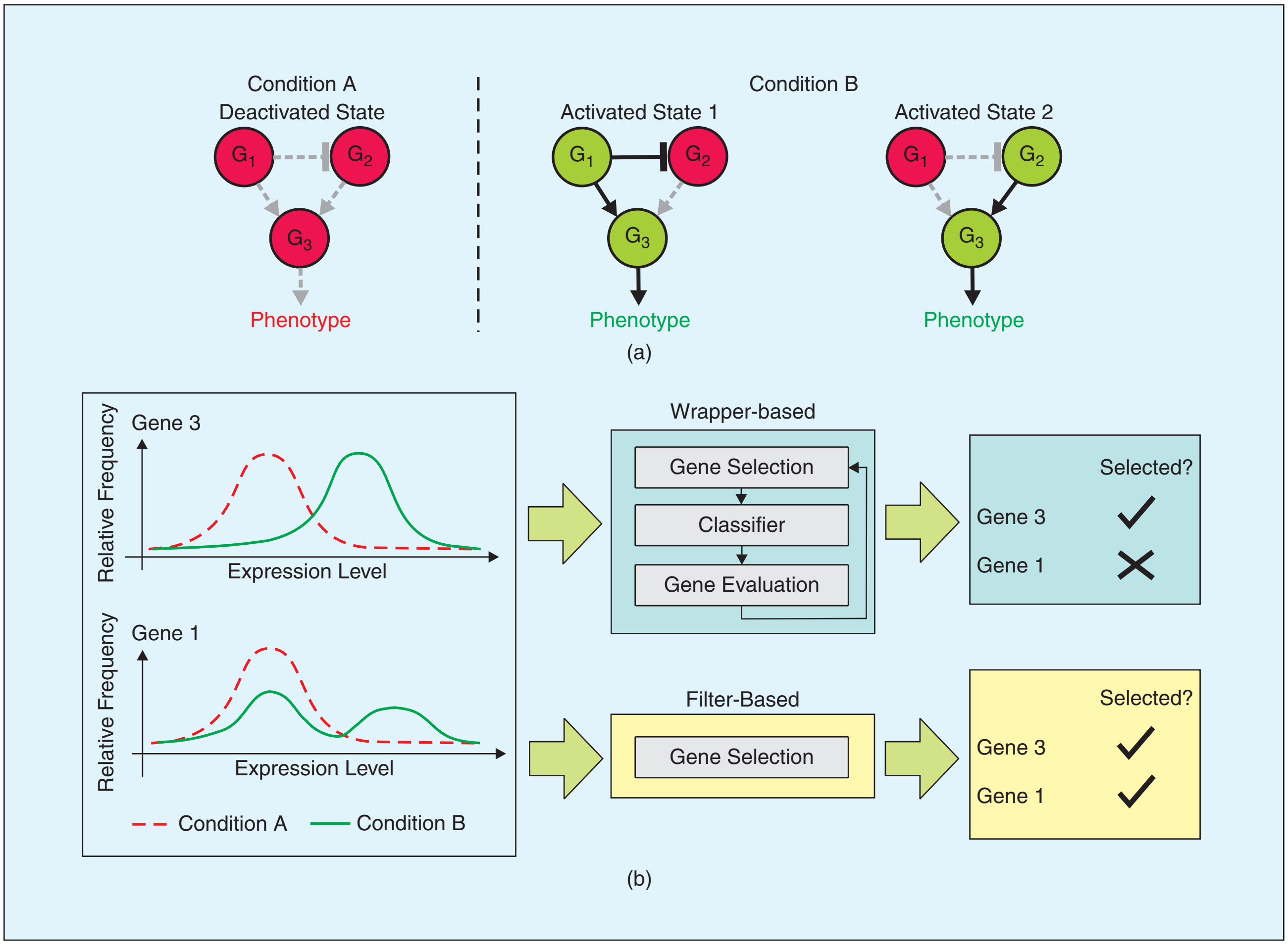

Fig. 1. Wrapper-based versus filter-based gene selection methods.

$$\mathrm{FCS}(G_i) = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sigma_i^+ + \sigma_i^-}. \qquad (2)$$

Approaches such as SAM [9] and normal mixture models [11] also use variants of the WTS as their criteria.

For the purpose of selecting differentially expressed genes, there are two potential difficulties with the existing family of filter-based gene selection criteria based on the WTS. First, *they assume that the variances of expressions of a differentially expressed gene are small*. WTS and its variants can be considered signal-to-noise ratio measurements of the expression levels of a gene. The numerators are signal estimators that measure the difference between means of expression levels from two classes; i.e., $|\mu_i^+ - \mu_i^-|$. The denominators are *serial noise estimators* that average two components of noise, σ_i^+ and σ_i^-. The use of a serial noise estimator assumes implicitly that the expression levels of "good" genes in *both* classes are relatively constant and small. Thus, both σ_i^+ and σ_i^- have to be small in order to give a high test statistic. If this is not the case, WTS and its variants provide low test statistics. Depending on the null distribution, the low test statistics are likely to lead to the acceptance of the null hypothesis. However, there are instances when the variance of the expression levels of a gene increases significantly under different conditions (such as Gene 1 in Figure 1). In these instances, WTS and its variants will assign (erroneously) low test statistics to a gene on account of the high variance of one class of expression levels, thereby failing to select the differentially expressed gene.

The second difficulty is that *criteria based primarily on means may not be powerful enough.* Tests using WTS and its variants use the difference between means of the positive and negative class expression levels to test if the null hypothesis should be rejected. There are many practical instances when the means remain approximately the same under different conditions while higher moments exhibit significant changes. Examples of these genes will be shown later in this article.

In this study, we propose the use of a *parallel* noise estimator (rather than the serial noise estimator employed by WTS and FCS) to address the first problem. The second problem is dealt with by introducing a signal estimator that calculates the average of the differences between expression levels in the two classes. Two new criteria for identifying differentially expressed genes, the average difference score (ADS) and the mean difference score (MDS), are formulated. We compare the performance of ADS and MDS to that of several commonly used criteria, including WTS, FCS, and the WRS on simulated and real biological datasets. We find that ADS and MDS outperform these existing criteria.

Understanding and Preparation of the Data

Real Biological Datasets

We considered two oligonucleotide array datasets in this study. The first dataset [2] consists of gene expression levels from bone marrow samples obtained from 27 adult patients with acute lymphoblastic leukemia (ALL) (the positive class) and 11 adult patients with acute myeloid leukemia (AML) (the negative class). The Hu6800 array used to obtain the data has 7,129 probe sets corresponding to 6,817 human genes and expressed sequenced tags (ESTs). This dataset has been studied widely in the literature [2], [7] to identify genes that are differentially expressed between AML and ALL samples.

The second dataset [3] is much larger and more diverse than the first dataset. It consists of gene expression levels from 190 primary tumor samples representing 14 classes of common human cancer (the positive class) and 90 normal tissue samples (the negative class). The 14 classes of human cancer include breast adenocarcinoma, prostate adenocarcinoma, lung adenocarcinoma, colorectal adenocarcinoma, lymphoma, bladder transitional cell carcinoma, melanoma, uterine adenocarcinoma, leukemia, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma, pleural mesothelioma, and central nervous system. The expression levels were measured from the Hu6800 and Hu35KsubA arrays containing a total of 16,063 probe sets. The dataset can be used to identify genes that are differentially expressed between cancer and normal samples. These genes may participate in a common genetic pathway that is shared by cancerous cells from diverse origins and tissues.

Following [7], two preprocessing steps were applied to the normalized datasets: 1) thresholding—floor of 100 and ceiling of 16,000; 2) filtering—removal of genes with max/min ≤ 5 or (max-min) ≤ 500, where *max* and *min* correspond to the maximum and minimum expression levels of a gene across all samples. Next, the expression levels for each sample were normalized by standardization (subtract the mean and divide by the standard deviation of the expression levels). Each gene in the normalized dataset was then checked for potential outliers. Outliers were defined as expression levels that are three times the interquartile range above the third quartile or below the first quartile. A gene is discarded if an outlier is detected. This whole preprocessing procedure reduces the number of genes from 7,129 to 2,415 for the leukemia dataset and from 16,063 to 9,651 for the multicancer dataset.

Simulated Datasets

In real biological data, the actual differentially expressed genes are usually unknown. Thus, the actual performance of a gene selection criterion cannot be evaluated directly. The problem can be overcome by using simulated data with known differentially expressed genes. We generated n samples of expression levels. Each sample consisted of m simulated genes. Of the m genes, the number of differentially expressed genes was m_d. We considered two different models for the differentially expressed genes:

- *Normal model with uniformly-distributed noise*. Each differentially expressed gene is generated from two different normal distributions with additive uniformly distributed noise in the range of [$-a$, a]. One distribution is used for expression levels of the positive class samples and one for expression levels of the negative class samples.
- *Mixture of normal model with uniformly distributed noise*. Each differentially expressed gene is generated from six different normal distributions with additive uniformly distributed noise. The sum of three of the distributions is used to generate the expression levels of the positive class samples, and the sum of another three distributions is used to generate the expression levels of the negative class samples. The model is used to simulate non-normally distributed expression levels while not perfectly matched to any particular dataset.

The rest of the $m - m_d$ nondifferentially expressed genes are generated by a single normal distribution with additive uniformly distributed noise. The means of all the normal distributions we used were generated randomly from another normal distribution with zero mean and σ standard deviation. A higher value of σ will increase the probability of generating more dispersed means and hence increasing the chance of generating expression levels of a more differentially expressed gene.

By using the simulated data, we are able to assess the performance of a criterion when the data are normally distributed and when they deviate from normality. We expect WTS to give the best performance for data generated from the normal model. However, it is still of interest to assess performance of the other criteria compared to WTS on normal data, especially if these criteria are more effective than WTS on data that are not distributed normally.

Data Mining

New Feature Selection Criteria

We propose the replacement of the serial noise estimator $(\sigma_i^+ + \sigma_i^-)$ in (2) with a parallel noise estimator $(\sigma_i^+\sigma_i^-)/(\sigma_i^+\sigma_i^-)$. The terms *serial* and *parallel* reflect the resemblance of these expressions to those used in calculating the resistance of serial and parallel combinations of Ohmic resistors. The parallel noise estimator will still give a high test statistic if either one or both variances of the expression distributions are relatively low. It may be able to detect changes that are missed by WTS and its variants when one of the variances is relatively high compared to the other. The MDS is thus formulated as:

$$\begin{aligned}\mathrm{MDS}(G_i) &= \frac{\left|\mu_i^+ - \mu_i^-\right|}{\left(\frac{\sigma_i^+\sigma_i^-}{\sigma_i^+ + \sigma_i^-}\right)} \\ &= \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sigma_i^+} + \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sigma_i^-}. \end{aligned} \tag{3}$$

If the expression distributions have close means but different higher moments, the signal estimator $|\mu_i^+ - \mu_i^-|$, which measures the difference in means, is not sufficient to distinguish between distributions. In this case, $|\mu_i^+ - \mu_i^-|$ can be replaced by the average difference between expression levels from one class to the mean of expression levels from another class. We use

$$d_i^+ = \frac{1}{n^+}\sum_{j=1}^{n}\left|x_{i,j} - \mu_i^-\right|\left(\frac{1+y_j}{2}\right)$$

to measure the average difference between all expression levels of the positive class samples to the mean expression levels of negative class samples, and

$$d_i^- = \frac{1}{n^-}\sum_{j=1}^{n}\left|x_{i,j} - \mu_i^+\right|\left(\frac{1-y_j}{2}\right)$$

to measure the average difference between all expression levels of the negative class samples to the mean expression levels of positive class samples. A new criterion, the ADS, can then be formulated as

$$\begin{aligned}\mathrm{ADS}(G_i) &= \frac{d_i^+ + d_i^-}{\left(\frac{\sigma_i^+\sigma_i^-}{\sigma_i^+ + \sigma_i^-}\right)} \\ &= \frac{d_i^+ + d_i^-}{\sigma_i^+} + \frac{d_i^+ + d_i^-}{\sigma_i^-}. \end{aligned} \tag{4}$$

The average of differences, $d_i^+ + d_i^-$, has the advantage of being able to detect changes in higher moments, such as variance, skewness, kurtosis, etc. However the extra sensitivity comes with a price. Outliers have more influence in $d_i^+ + d_i^-$ than the $|\mu_i^+ - \mu_i^-|$. If the data may consist of outliers, MDS is preferable to ADS.

We found that ADS actually generalizes the *independently consistent expression* (ICE) discriminator that was proposed by Bijlani et al. [12], namely:

$$\begin{aligned}\mathrm{ICE}(g_i) = {} & \frac{1}{\sigma_i^+ n^-}\sum_{j=1}^{n}\left|x_{i,j} - \mu_i^+\right|\left(\frac{1-y_j}{2}\right) \\ & + \frac{1}{\sigma_i^- n^+}\sum_{j=1}^{n}\left|x_{i,j} - \mu_i^-\right|\left(\frac{1+y_j}{2}\right). \end{aligned} \tag{5}$$

Identifying Differentially Expressed Genes

The problem of identifying differentially expressed genes can be stated as a multiple hypothesis testing problem [7]. For each gene, we test the null hypothesis that the gene is not differentially expressed. If there are m genes, m hypothesis tests are performed. The significance of each test statistic is determined by calculating its p-values. The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the observed value, assuming that the null hypothesis is true. Differentially expressed genes are those genes with p-values lower than a predetermined p-value threshold.

In order to determine if a test statistic is significant, its distribution under the null hypothesis is required. However, this information is generally not available. We can estimate the distribution empirically by using resampling methods, such as permutation [7], [8], [13].

In multiple hypothesis testing, the probability of committing a false alarm increases quickly with the number of tested hypotheses. A small p-value for a test may occur simply by chance when a large enough number of hypotheses are tested. Since typical microarray experiments monitor expressions for thousands of genes simultaneously, they are prone to this deficiency. The remedy is to adjust the raw p-values, obtained for each gene, to account for the large number of hypotheses. A detailed comparison of various p-value adjustment procedures for multiple hypothesis testing can be found in [7]. In particular, Benjamini and Hochberg's step-up procedure for controlling the *false discovery rate* [14] has been shown to retain substantially more power than other family-wise error rate controlling procedures [7], [13], [14]. After the p-values have been adjusted through the Benjamini and Hochberg's procedure, differentially expressed genes are identified as those with adjusted p-values smaller than a predetermined threshold.

Evaluation of Discovered Knowledge

We implemented the following six criteria for identifying differentially expressed genes: ADS (4); MDS (3); (FCS) (2);

ICE (5); WTS (1); and WRS [15]. The distributions of the test statistics under the null hypothesis were estimated by resampling for all criteria, except for WRS. For WRS, the normal approximation for the distribution of the null hypothesis was used [15]. Evaluations of these six criteria were performed on both simulated and real biological datasets.

Evaluation Criteria

On simulated datasets, the performances of the criteria were evaluated using the gene selection true positive rate (TPR) and false positive rate (FPR), defined as:

$$TPR = \frac{\text{number of differentially expressed genes selected}}{\text{total number of differentially expressed genes}}, \tag{6}$$

and

$$FPR = \frac{\text{number of non-differentially expressed genes selected}}{\text{total number of non-differentially expressed genes}}. \tag{7}$$

In medical diagnosis, TPR is also called *sensitivity* and 1-FPR is called *specificity*.

When we compare the performance of two criteria, we measure the difference between the TPRs and the difference between FPRs of these two criteria. These differences are random variables that depend on the statistics of the simulated data. By repeating the performance comparison procedure several times on different randomly generated data, we can obtain average values of the differences. Usually the magnitudes of the differences are very small, and it is difficult to tell if the differences are significant just by measuring their magnitudes. A one-sample *t*-test [16] is used to determine if the differences are statistically significant.

The evaluation of these criteria on real biological datasets is more challenging because the identity of real differentially expressed genes is usually unknown. Thus, the TPR and FPR of the gene selection cannot be calculated directly. Although one can use a learning machine to classify the selected genes and use the estimated TPR and FPR of the classification as a performance indicator of the selected genes, this approach is biased by the chosen learning machine. Furthermore, one of the main objectives of our new criteria is to select differentially expressed but irrelevant genes that usually give poor classification performances. For these reasons, we use a more conservative approach by using the biological functions of the selected genes, whenever these functions are known, as an indicator of the relative performance of a gene selection criterion. By querying the National Center for Biotechnology Information (NCBI) PubMed database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) for abstracts related to a gene, we can find many diseases known to be associated with a gene. For example, if a gene was selected to be related to leukemia, we can query the PubMed database to see if any previous study has associated this gene with leukemia. Although a zero hit does not preclude the association of this gene to leukemia, a multiple-hits result greatly enhances our confidence about the association. We are especially interested in finding out whether there are biologically meaningful differentially expressed genes that were missed by some criteria but selected by others.

Performance on Simulated Dataset

In our study, we used $m = 5,000$, $m_d = 40$, $n = 120$, and $n^+ = 60$ to generate a dataset. The testing process, including the generation of data and means, was repeated 20 times to obtain average TPR and FPR, and to calculate the statistical significance of their differences. We have considered adjusted *p*-value thresholds of 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 and studied changes in TPR and FPR as a function of noise levels, *p*-value thresholds, and standard deviations of the mean.

For the datasets generated from the normal model, the effects of *p*-value threshold on TPR for all criteria are shown in Figure 2(a). As indicated before, we expect WTS to provide the best result, since it was designed for the normal model. The effects on the FPR are shown in Figure 2(b). We show averages collected over 20 runs. The noise level was $a = 0.01$ and the standard deviation of the mean was $\sigma = 1$. Ideally, a criterion will exhibit high TPR and low FPR over a range of *p*-value thresholds. The slope of the TPR curves is relatively high for $p < 0.10$ for all criteria and the rate of removing genes falls sharply after $p = 0.10$. Thus, $p = 0.10$ was selected to retain most of the genes that are significantly differentially expressed. At this threshold, WTS, ADS, MDS, and FCS have TPR performances that cannot be significantly differentiated, while ICE and WRS appear to have significantly

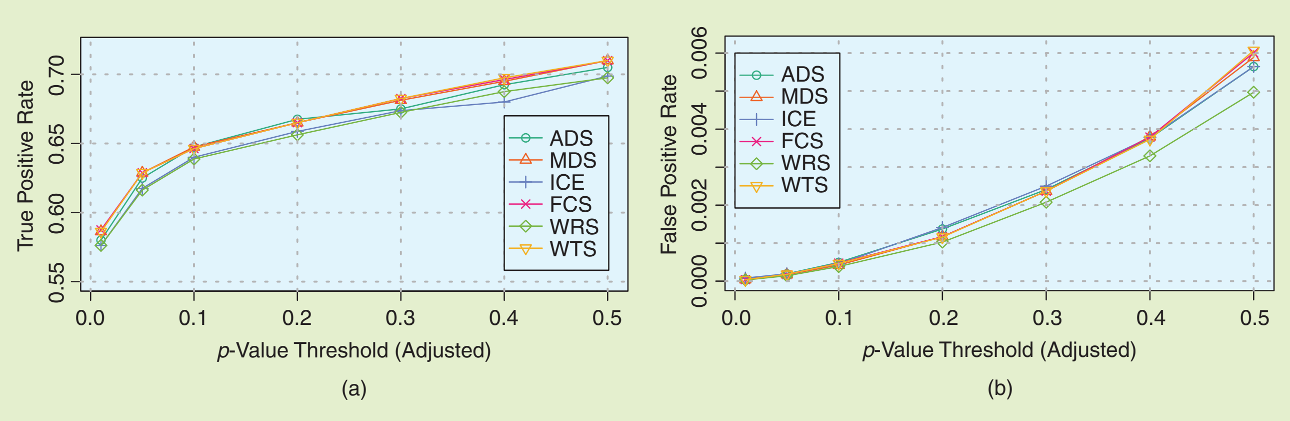

Fig. 2. Effect of *p*-value thresholds on the performance of the criteria tested on simulated datasets generated from the normal model. (a) Effect on the true positive rate. (b) Effect on the false positive rate.

lower TPR performances than ADS. At higher p–value thresholds (>0.2), the TPRs of ADS become less than WTS and FCS, while the TPRs of MDS remain similar to WTS and FCS [Figure 2(a)]. All criteria appear to have similar FPRs at p-value threshold $= 0.10$, except for WRS, which has significantly lower FPR than ADS. At higher p-value thresholds (>0.20), the FPRs of WRS become significantly less than all other criteria [Figure 2(b)]. Our result for WTS and WRS is similar to the results obtained by others [25]. Overall, ADS, MDS, FCS, and WTS have similar performance in terms of both TPR and FPR for commonly used p-value thresholds (<0.2). ICE has lower TPR but similar FPR with this group of criteria. WRS has both the lowest TPR and the lowest FPR. The results show that the parallel noise estimator-based criteria (ADS and MDS) are *not inferior* to WTS in both TPR and FPR when the data were generated by the normal model.

For the datasets generated from the mixture model, the effects of p-value threshold on TPR for all criteria are shown in Figure 3(a). Effects on the FPR are shown in Figure 3(b) (again, these are averages over 20 runs). The noise level was $a = 0.01$ and the standard deviation of mean was $\sigma = 1$. As in Figure 2, we found that the slope of the TPR curves is relatively high for $p < 0.10$ for all criteria. Thus, $p = 0.10$ was selected again as the p-value threshold. Figure 3 demonstrates the advantage of criteria based on parallel noise estimators: MDS has significantly higher TPR than serial noise estimator-based criteria (WTS and FCS) over all p-value thresholds [Figure 3(a)], while having similar FPR [Figure 3(b)]. Although both ICE and ADS outperform other criteria in terms of TPR, they also have higher FPRs. Among these two criteria, ADS has lower TPR but also lower FPR than ICE. Similar to the results of the normal model, WRS has the lowest FPR and TPR. Overall, ICE has the highest TPR but the worst FPR. ADS has the second highest TPR, but significantly better FPR than ICE. MDS has the third highest TPR and its FPR is similar to WTS and FCS. The results show that parallel noise estimator-based criteria have higher TPR than other criteria when the data are generated by a mixture model.

Similar results were also obtained from the comparisons of criteria using other values of standard deviation of mean (σ) and noise level (a), and thus are omitted.

In summary, parallel noise estimator-based criteria (ADS and MDS) have significantly higher TPRs than traditional serial noise estimator-based criteria (WTS and FCS) on mixture model-generated data. The FPRs of MDS are also similar to WTS and FCS, while the higher TPRs of ADS come with a cost in higher FPRs. For normal model-generated data, no significant performance decrease was observed for ADS and MDS in either TPR or FPR. Although ICE has the highest TPR for data generated by the mixture model, its FPR is also relatively high and it does not perform well for data generated by the normal model. For this reason ICE appears to be less desirable than ADS and MDS. In general, ADS is preferred to MDS due to the former's higher sensitivity. However, MDS should be used when an FPR similar to WTS is needed. WRS is the most conservative among all criteria. It should be used when the lowest FPR is needed, even at the expense of the lowest TPR.

Performance on Real Biological Datasets

Similar to the simulated data, the resampling testing procedure is used to select differentially expressed genes. We found that the number of selected genes dropped significantly for adjusted p-value threshold ≤ 0.10; thus, the threshold was selected to be 0.10. This selection of threshold matches the p-value threshold used for analyzing the simulated data in this article.

We then compared all the criteria with WTS. WTS was selected as the baseline for comparison because it is one of the most commonly used criterion. The composition of genes selected by all criteria compared to WTS is listed in Table 1(a) for the leukemia dataset and Table 1(b) for the multicancer dataset. For each criterion, the total number of selected genes, the number of these genes that are also selected by WTS, the number of these genes that are missed by WTS, and the number of genes missed by the criterion but selected by WTS are listed in the tables. For both datasets, all criteria were able to select most of the genes selected by WTS but also missed some of them. Among them, ADS, ICE and WRS selected relatively more genes but also missed relatively larger numbers of WTS genes than other criteria. This result is in general agreement with our simulated data results.

Next, we turned our attention to the genes that were selected by other criteria but missed by WTS (the "additional" genes) and the genes that were missed by other criteria but selected by WTS (the "missing" genes). We would like to determine how biologically significant are these additional genes and missing genes. These genes were queried on the NCBI PubMed database for all publications related to the genes. For the leukemia dataset, a hit is a gene that has at

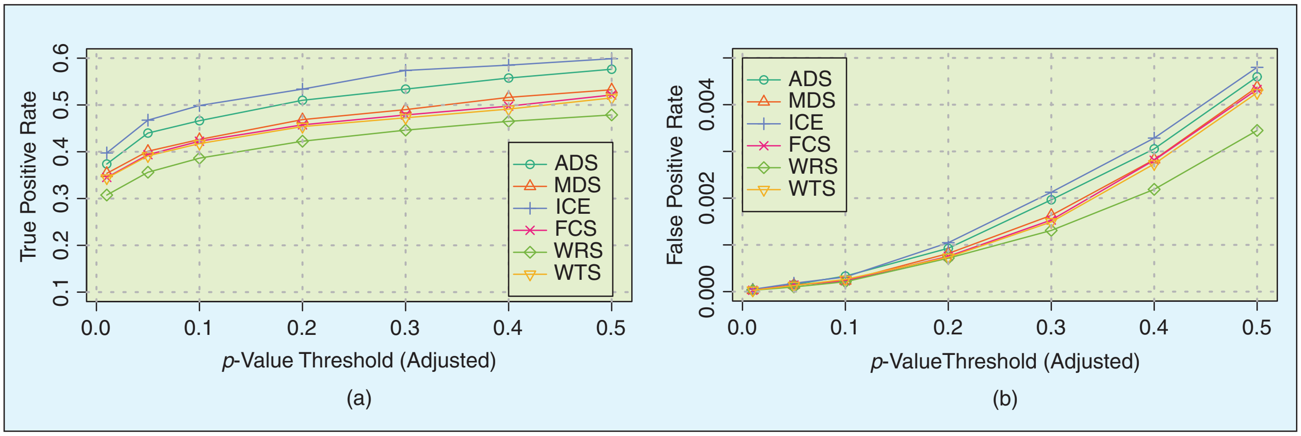

Fig. 3. Effect of p-value thresholds on the performance of the criteria tested on simulated datasets generated from the mixture model. (a) Effect on the true positive rate. (b) Effect on the false positive rate.

least two publications found with any of the following terms listed in their abstracts: “AML,” “ALL,” “Myeloblast,” “Lymphoblast,” and “Leukemia”; and at least one publication found with any of the following terms listed in its abstract: “Cancer,” and “Tumor.” For the multicancer dataset, a hit is a gene that has at least one publication with the term “Oncogene” listed in its abstract, or at least one publication with the term “Cancer” or “Tumor,” and at least one publication with the term “Adenocarcinoma” listed in their abstracts. The number of hits provides an estimate of the importance of the selected genes based on past publications. A high number of hits among additional genes

Table 1(a). Composition of genes detected by the studied criteria compared to WTS for the leukemia dataset.

Criterion	Detected Genes	Common Genes with WTS	Genes not Detected by WTS	Genes Detected by WTS but Missed by Criterion
WTS (baseline)	676	676	—	—
ADS	719	631	88	45
ICE	698	601	97	75
MDS	732	660	72	16
FCS	723	650	73	26
WRS	682	594	88	82

Table 1(b). Composition of genes detected by the studied criteria compared to WTS for the multicancer dataset.

Criterion	Detected Genes	Common Genes with WTS	Genes not Detected by WTS	Genes Detected by WTS but Missed by Criterion
WTS (baseline)	4786	4786	—	—
ADS	5082	4213	869	573
ICE	4999	3931	1068	855
MDS	4886	4653	233	133
FCS	4935	4727	208	59
WRS	6132	4474	1658	312

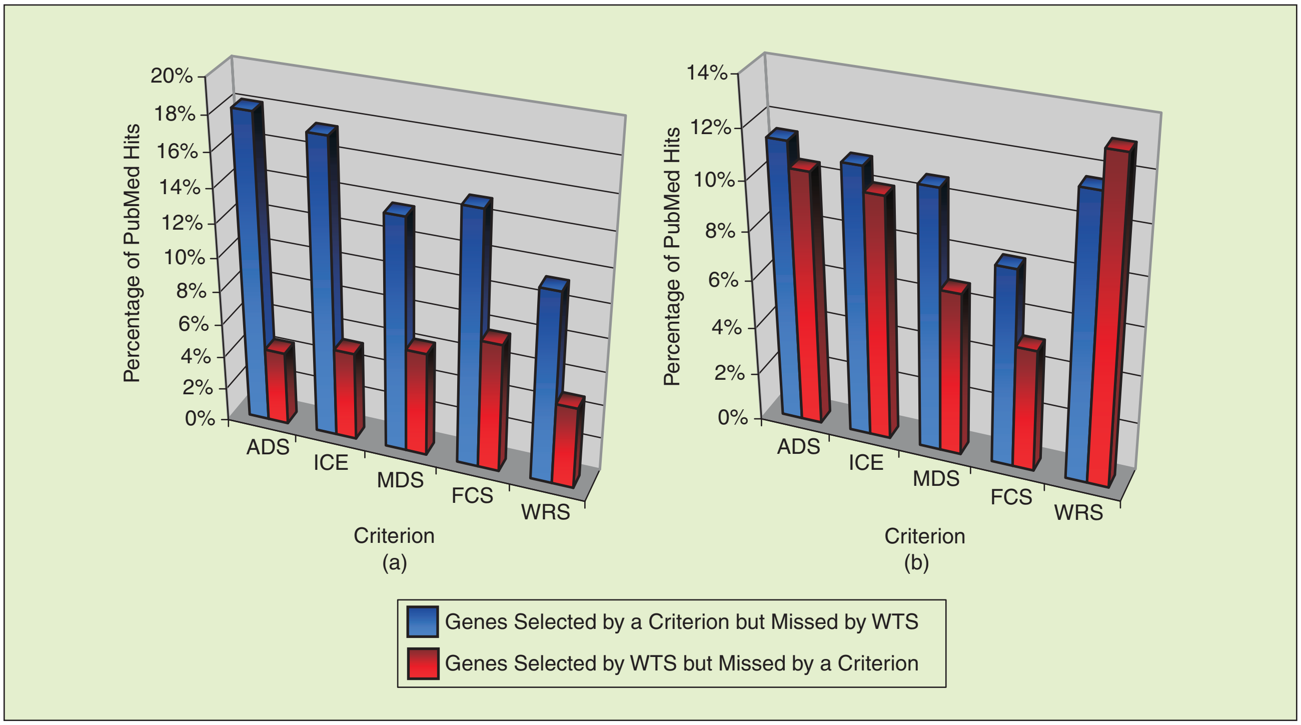

Fig. 4. Percentage of PubMed hits among genes selected by a criterion but missed by WTS and among genes selected by WTS but missed by a criterion for (a) the leukemia dataset and (b) the multicancer dataset.

indicates that the criterion is selecting extra genes that are likely to be biologically significant, while a low number of hits among missing genes indicates that the criterion does not miss many genes that are biologically significant.

The results of the queries are shown in Figure 4(a) for the leukemia dataset and Figure 4(b) for the multicancer dataset. In the figures, the percentages of PubMed hits for each criterion among additional genes and missing genes are plotted in blue and red bars respectively. We seek a criterion that has the largest positive difference between percentages of hits on additional genes and on missing genes; i.e., it gains more than it loses. Ideally, the best criterion is the one with the maximum percentage of hits on additional genes and minimum percentage of hits on missing genes.

For the relatively simple leukemia dataset, ADS turns out to be the ideal best criterion, and it is followed closely by ICE. Both of these criteria yield high TPRs in the simulated results. Despite the large number of selected genes, WRS has the worst performance with the smallest positive difference between percentages of hits on additional genes and on missing genes. Thus, more potentially useful genes are selected by ADS and ICE than other criteria, especially WRS. The fact that all criteria give positive differences shows that they are all better than WTS in this dataset. For the more complex multicancer dataset, MDS has the best performance followed by FCS. Both of these criteria give low FPR in the simulated results. Higher TPR criteria, such as ADS and ICE, do not perform well because they give relatively large percentages of hits on the missing genes. In this dataset, WRS has more hits on the missing genes than the additional genes. Thus, not only does it have the worst performance among other criteria considered, it has a poorer performance than WTS.

To further investigate the additional genes selected by ADS or MDS on the leukemia dataset or multicancer dataset, we

Table 2(a). Selected differentially expressed genes for the leukemia dataset, and multi-cancer dataset.

Gene	ADS *p*-value	WTS *p*-value	Note
HLA-C	0.025	0.912	A large HLA association study in leukemia (17), (18) was carried out on the International Bone Marrow Transplant Registry data, which consist of 1,834 patients with ALL, AML, and chronic myelogenous leukemia (CML) treated between 1969 and 1985. These studies showed that HLA-Cw3 and -Cw4 are both susceptibility markers for all of the three major types of leukemia.
PRTN3	0.042	0.147	Dengler et al. (19) investigated the expression of PRTN3 in samples of bone marrow from healthy individuals and patients with different types of leukemia by using immunocytochemical staining and flow cytometric quantitation. The results indicated that PRTN3 may be differentially expressed between AML and ALL.
IL6 *	0.071/0.074	0.222/0.220	IL-6 is a pro-inflammatory and immunosuppressive cytokine. Significant levels of IL-6 were found to be expressed in peripheral blood or bone marrow (BM) cells from AML patients, but not on ALL patients (20).

Table 2(b). Selected differentially expressed genes for the multicancer dataset.

Gene	MDS *p*-value	WTS *p*-value	Note
TIAM1	0.044	0.375	TAIM1 binds to c-MYC (21), which is important for the control of cell growth and apoptotic cell death. Overexpression of TIAM1 inhibited the c-Myc apoptotic activity (22). Previous studies also show that TIAM1 is related to metastasis of breast tumor cell (23) and colorectal carcinoma (24).
BCL2A1	0.069	0.251	BCL2A1 encodes a member of the BCL-2 protein family which has been shown to retard apoptosis in various cell lines and mutations of it may lead to cancers (25). High expression of BCL2A1 also contributes to the apoptosis resistant phenotype in B-cell chronic lymphocytic leukemia (26).

** Note: two clones of the same gene are identified.*

manually examined all the hits and looked through the references to decide if the genes are actually meaningful. Some of these meaningful genes and their known functions are listed in Table 2. Some of the additional genes selected by ADS on the leukemia dataset are related to tumors or leukemia in general, and some of them have been identified previously in the literature as differentially expressed between AML and ALL samples. Similarly, some of the additional genes selected by MDS on the multicancer dataset are related to apoptosis, growth-control, and cell-cycle control. These genes have been found to participate in a tumorigenesis pathway. The adjusted *p*-values obtained from ADS/MDS and WTS for all of these genes are also listed in the table. Many of them have adjusted *p*-values larger than 0.20 under WTS. Thus, they would not be selected by WTS even at higher *p*-value thresholds (0.10–0.15). The conclusion from this analysis is that ADS and MDS were able to select several biologically significant genes that were missed by the traditional WTS. Most of these genes have changes in higher moments of the distributions of their expressions. The distributions of expression levels for two of these genes, HLA-C and TIAM1, are shown in Figure 5. The distributions of HLA-C are similar to the distributions of the exemplary Gene 1 in Figure 1. The distribution of TIAM1 under normal samples has a higher variance than the distribution under cancerous samples. These explain why HLA-C and TIAM1 are not preferred by WTS.

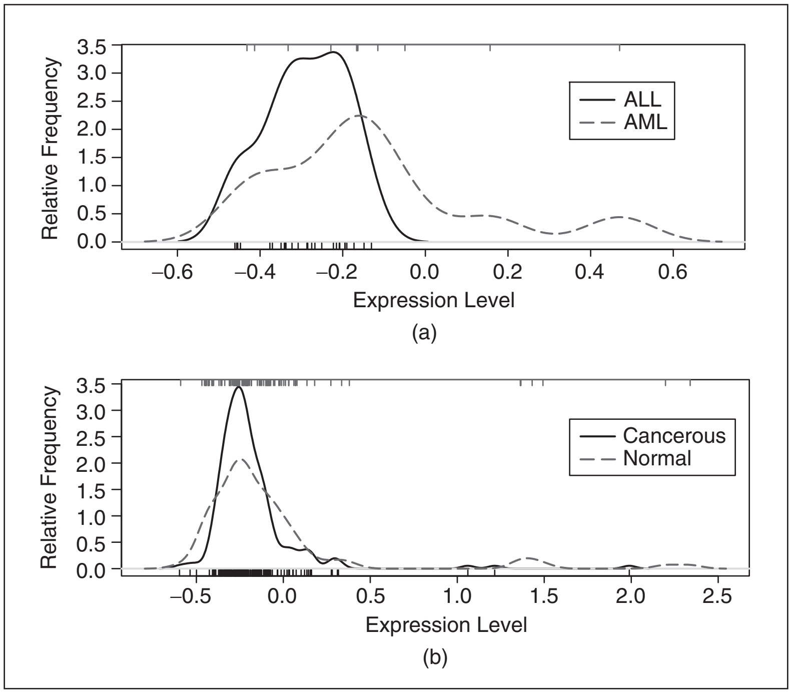

Fig. 5. Estimated distributions of the expression levels of two genes selected by ADS and MDS. (a) Estimated distributions of HLA-C selected by ADS from the leukemia dataset. (b) Estimated distributions of TIAM1 selected by MDS from the multicancer dataset.

The same analysis was also performed on the missing genes of ADS and MDS. For the leukemia dataset, only two of the 45 missing genes have more than one hit, and only one of them is found to be related to AML. None of the 45 genes were found to be previously reported as differentially expressed on AML and ALL samples. For the multicancer dataset, a few of the missing genes of MDS are related to cancer. However, all *p*-values for these genes are between 0.10 and 0.13 under MDS. They could be selected by MDS at a slightly increased threshold. Furthermore, the *p*-values of most missing genes on both datasets are close to 0.10 under WTS. It appears that most of the genes missed by ADS or MDS are not important.

Summary and Conclusions

One of the major concerns in detecting changes in higher moments is these changes may be due to outliers or process errors that are not biologically significant. For example, a larger variance observed in the expression levels may simply due to the larger variation in the data collecting process. Several outliers, which exhibit some extreme expression levels than the rest of the samples, may also increase the variance or skewness of the expression levels significantly. So it is very important to reduce the effect of outliers and process errors by proper experimental designs [27], such as technical replicates and biological replicates, before high sensitivity criterion, such as ADS, can be applied.

We have presented and demonstrated the operation of two new criteria, ADS and the MDS, for identifying differentially expressed genes. These two criteria were compared with several commonly used criteria, namely WTS, WRS, FCS, and ICE. Experiments with simulated data show ADS to be more powerful than the WTS. When high-sensitivity screening is required, ADS appears to be preferable to WTS. When an FPR similar to WTS is desired, MDS should be used. The popular Wilcoxon rank sum is a more conservative approach that should be employed when the lowest FPR is desired, even at the expense of lower TPRs. ICE is a less desirable criterion because it does not perform well for data generated by the normal model. FCS gave results similar to those of WTS. Evaluation of these algorithms using real biological datasets showed that ADS and MDS flagged several biologically significant genes that were missed by WTS, besides selecting most of the genes that are also selected by WTS.

Lit-Hsin Loo received his Ph.D. in electrical and computer engineering from Drexel University, Philadelphia, Pennsylvania, in 2004. His dissertation was on identifying differentially expressed genes from microarray experiments. After his graduation, he became a postdoctoral research fellow at the Bauer Center for Genomics Research at Harvard University, where he worked on image segmentation and feature extraction algorithms for high-throughput microscopy imaging. In 2005, he moved to Dallas, Texas, to join the Green Comprehensive Center for Computational and Systems Biology at the University of Texas Southwestern Medical Center. His current research concentrates on studying signaling pathway and characterizing responses of individual cells to external perturbations.

Samuel Roberts was employed at GlaxoSmithKline following his D.Phil. in artificial intelligence and machine learning. He worked as a statistician, specializing in the application of multivariate statistics to the prediction of the toxicity of drug candidates from high-dimensional datasets deriving from microarray, mass spectrometry, and NMR spectroscopy. He is now a senior application engineer at The MathWorks in the United Kingdom, focusing on supporting customers in the pharmaceutical, biotechnology, and life science industries in the use of Matlab and other MathWorks tools.

Leonid Hrebien received a B.S. in electrical engineering, an M.S. in biomedical engineering, and a Ph.D. from Drexel University, Philadelphia, Pennsylvania, in 1972, 1975, and 1980, respectively. He is an associate professor of electrical and computer engineering at Drexel University, a Senior Member of IEEE, and a Fellow of the Aerospace Medical Association. His research interests are in the areas of biomedical systems; the study and mitigation of acceleration effects on cardiovascular and cerebrovascular functions; and the analysis, modeling, and estimation of large arrays of complex and noisy biological signals and data. The goal of this work is to develop efficient and robust analysis and screening techniques to aid in studying efficacy and toxicity of pharmaceutical compounds and the effects of intoxicants and stress on human physiology.

Moshe Kam was educated at Tel Aviv University (B.S, 1977) and Drexel University (M.Sc 1985; Ph.D. 1987). Currently he is the Robert Quinn Professor of Electrical and Computer Engineering at Drexel University, director of Drexel University's National Security Agency Center of Excellence in Information Assurance Education, and technical coordinator of the Department of Defense-sponsored project ACIN (Applied Communications and Information Networking). Dr. Kam's professional interests are in system theory, detection and estimation, information assurance, robotics, navigation, and control. Within these areas he has worked on architectures for decision fusion and cooperative control and applied them to detection problems in large-scale databases, distributed control systems, and mobile networks. Kam has received the C.H. MacDonald award for the Outstanding Young Electrical Engineering Educator, an NSF Presidential Young Investigator Award, and an IEEE Third Millennium Medal. Since 2003 he has served on the IEEE Board of Directors as director of IEEE Region 2 (2003–2004) and vice president for educational activities (2005–2006).

Address for Correspondence: Moshe Kam, Dept. of Electrical and Computer Engineering, Drexel University, 3141 Chestnut St., Philadelphia, PA 19104 USA. Phone: 215 895 6920. Fax: 215 895 1695. E-mail: kam@minerva.ece.drexel.edu

References

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[2] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci.,* vol. 98, no. 26, pp. 15149–15154, 2001.

[4] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Machine Learning*, New Brunswick, NJ, 1994, pp. 121–129.

[5] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, 2004.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[7] S. Dudoit, J.P. Shaffer, and J.C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statistical Sci*., vol. 18, no. 1, pp. 71–103, 2003.

[8] O.G. Troyanskaya, M.E. Garber, P.O. Brown, D. Botstein, and R.B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, no. 11, pp. 1454–1461, 2002.

[9] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci*., vol. 98, no. 9, pp. 5116–5121, 2001.

[10] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[11] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, pp. 546–554, 2002.

[12] R. Bijlani, Y. Cheng, D.A. Pearce, A.I. Brooks, and M. Ogihara, "Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED)," *Bioinformatics*, vol. 19, no. 1, pp. 62–70, 2003.

[13] A. Reiner, D. Yekutieli, and Y. Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures," *Bioinformatics,* vol. 19, no. 3, pp. 368–375, 2003.

[14] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statistical Soc*. B (Statistical Methodol.), vol. 7, pp. 289–300, 1995.

[15] S. Siegel and N.J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*. 2nd Ed. New York: McGraw-Hill, 1988.

[16] G.W. Snedecor and W.G. Cochran, *Statistical Methods*. 8th Ed. Ames, IA: Iowa State Univ. Press, 1989.

[17] M.M. Bortin, J. D'Amaro, F.H. Bach, A.A. Rimm, and J.J. van Rood, "HLA associations with leukemia," *Blood*, vol. 70, pp. 227–232, 1987.

[18] J. D'Amaro, F.H. Bach, J.J. van Rood, A.A. Rimm, and M.M. Bortin, "HLA C associations with acute leukemia," *Lancet*, vol. 2, pp. 1176–1178, 1984.

[19] R. Dengler, U. Munstermann, S. al-Batran, I. Hausner, S. Faderl, C. Nerl, and B. Emmerich, "Immunocytochemical and flow cytometric detection of proteinase 3 (myeloblastin) in normal and leukaemic myeloid cells," *Br. J. Haematol*., vol. 89, no. 2, pp. 250–257, 1995.

[20] K. Inoue, H. Sugiyama, H. Ogawa, T. Yamagami, T. Azuma, Y. Oka, H. Miwa, K. Kita, A. Hiraoka, and T. Masaoka, "Expression of the interleukin-6 (IL-6), IL-6 receptor, and gp130 genes in acute leukemia," *Blood*, vol. 84, no. 8, pp. 2672–2680, 1994.

[21] S. Pelengaris, M. Khan, and G. Evan, "c-MYC: More than just a matter of life and death," *Nature Reviews Cancer*, vol. 2, no. 10, pp. 764–776, 2002.

[22] Y. Otsuki, M. Tanaka, T. Kamo, C. Kitanaka, Y. Kuchino, and H. Sugimura, "Guanine nucleotide exchange factor, Tiam1, directly binds to c-Myc and interferes with c-Myc-mediated apoptosis in rat-1 fibroblasts," *J. Biological Chem*., vol. 278, no. 7, pp. 5132–5140, 2003.

[23] L.Y. Bourguignon, H. Zhu, L. Shao, and Y.W. Chen, "Ankyrin-Tiam1 interaction promotes Rac1 signaling and metastatic breast tumor cell invasion and migration," *J. Cell Biol.*, vol. 150, no. 1, pp. 177–191, 2000.

[24] L. Liu, D.H. Wu, and Y.Q. Ding, "Tiam1 gene expression and its significance in colorectal carcinoma," *World J. Gastroenterol*., vol. 11, no. 5, pp. 705–707, 2005.

[25] J.M. Adams and S. Cory, "The Bcl-2 protein family: arbiters of cell survival," *Science*, vol. 281, no. 5381, pp. 1322–1326, 1998.

[26] A.A. Morales, A. Olsson, F. Celsing, A. Osterborg, M. Jondal, and L.M. Osorio, "High expression of bfl-1 contributes to the apoptosis resistant phenotype in B-cell chronic lymphocytic leukemia," *Int. J. Cancer*, vol. 113, no. 5, pp. 730–737, 2005.

[27] Y.H. Yang and T. Speed, "Design issues for cDNA microarray experiments," *Nature Reviews Genetics*, vol. 3, pp. 579–588, 2002.