9-2007

A trend pattern assessment approach to microarray gene expression profiling data analysis

Kahai Cao
University of Nebraska at Omaha

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Javeed Iqbal
University of Nebraska Medical Center

John W.C. Chan
University of Nebraska Medical Center

A trend pattern assessment approach to microarray gene expression profiling data analysis

Kajia Cao [a,*], Qiuming Zhu [a], Javeed Iqbal [b], John W.C. Chan [b]

[a] *Department of Computer Science, University of Nebraska at Omaha, United States*
[b] *Department of Pathology and Microbiology, University of Nebraska Medical Center, United States*

Abstract

We study the problem of how to assess the reliability of a statistical measurement on data set containing unknown quantity of noises, inconsistencies, and outliers. A practical approach that analyzes the dynamical patterns (trends) of the statistical measurements through a sequential extreme-boundary-points (EBP) weed-out process is explored. We categorize the weed-out trend patterns (WOTP) and examine their relation to the reliability of the measurement. The approach is applied to the processes of extracting genes that are predictive to BCL2 translocations and to clinical survival outcomes of diffuse large B-cell lymphoma (DLBCL) from DNA Microarray gene expression profiling data sets. Fisher's Discriminate Criterion (FDC) is used as a statistical measurement in the processes. It is found that the weed-out trend analysis (WOTA) approach is effective for qualitatively assessing the statistics-based measurements in the experimentations conducted.

Keywords: Gene expression profiling; Microarray data analysis; Boundary points; Dynamical patterns; Trend evaluations; Fisher's discriminate criterion

1. Introduction

The accuracy and reliability of the statistics extracted from multidimensional datasets possibly contaminated with noises, uncertainties, outliers, and measurement errors is an important issue in pattern analysis and data mining researches (Arning et al., 1996; Knorr and Ng, 1999; Yu et al., 1999; Knorr et al., 2000; Breunig et al., 2000; Ramaswamy et al., 2000). Here, "accuracy" means that the statistical measurement fits to (or reveals) the underlying distribution of the dataset, and "reliability" means that the measurement is not significantly affected by small perturbations (adding, changing, or removing small percent of measurements) of samples to the dataset.

Most work on this subject has been conducted in the field of robust statistics previously (Huber, 1981). The methods usually make assumptions about the data distributions, the statistical distribution parameters, and the types or numbers of the irregularly distributed boundary points (Huber, 1981). Robust estimates are consistent estimates of the unknown parameters at the idealized model. Because of robustness, they will not drift too far away if the model is approximately true. However, the incongruence between the relatively small number of data samples collected in many practice problems and the high dimensions of the data set, such as the Microarray gene expression profiles, often makes the robust statistics models hard to be justified (the so-called curse of dimension). Moreover, the robust statistical measurements are easily biased and distorted by the uncertainty and inaccuracy of

Abbreviations: extreme-boundary-points, EBP; weed-out trend pattern, WOTP; weed-out trend analysis, WOTA; diffuse large B-cell lymphoma, DLBCL; Fisher's discriminate criterion, FDC.

* Corresponding author. Tel.: +1 402 554 2186; fax: +1 402 554 3400.
E-mail address: kcao@mail.unomaha.edu (K. Cao).

the sample values, the inexact categorization of the specimen, and irregularity of the sample distributions.

While robust statistics and error-cleaning techniques have been in the main streams of study in statistics, the assessment of the accuracy and reliability of statistical measurement has always been a popular problem of exploration (Yu et al., 1999; Maddala and Yin, 1997). There are many evaluation methods that have been suggested and applied in practice. The "t-test" and "p-value" computation are two most conventional, meanwhile computationally expensive, quantitative measurements (Dudoit et al., 2000; Sellke et al., 2001). Though the t-test and p-value approaches have high computational cost, no other quicker and more effective methods were seen yet.

Methods that try to clean up the dataset before applying the statistical measurement have been explored extensively. For example, many outlier detection methods and algorithms have been proposed and studied (Brown et al., 2000; Knorr et al., 2000; Yang and Zhang, 2002). The majority of these methods are based on certain discriminant approaches that calculate the distances (Euclidean distance, Mahalanobis distance, or some others) of the data points to their cluster means and eliminate the points that are $n\sigma$ (often $n = 3$) from the means, where σ is the standard deviation. The method has an essential fault because it bases its evaluation on the statistical measures (means and variances) that are the objects of the verification themselves. Moreover, it was pointed out by Huber that a two-step procedure (first clean the data by applying some rule for outlier rejection, then use classical estimation and testing procedures on the remainder) may be more difficult to work out than that of a straight robust procedure (Huber, 1981).

In this paper, we study an approach that evaluates the dynamical trends of the statistical measurements through a sequential extreme-boundary-point (EBP) weeding-out process. We found that an analysis of the trend of the statistical measurements via a focus on the changes of the EBPs of a dataset can provide a simple and quick assessment of the accuracy and reliability of the measurement to certain extend. Specifically, we analyze the perturbations of the statistical measurements and the trend patterns with respect to the variations of the boundary settings to assess the reliability of the measurement. We apply the Fisher's Discriminate Criterion to the processes of identifying genes that are predictive to BCL2 translocations and to clinical survival outcomes for diffuse large B-cell lymphoma from the DNA Microarray gene expression profiling datasets. The reliability of the FDC measurement on the data set is then evaluated using our weed-out trend analysis (WOTA) approach. Genes with higher reliability measurement according to the weed-out trend pattern analysis are extracted as the outcome predictors. The results are compared with gene set extracted using other methods. It is seen that the WOTA approach is effective in identifying a more accurate set of genes in the experiments conducted.

The paper is organized as follows. In Section 2, the sequential EBP weed-out and dynamical trend pattern analysis approach for qualitative evaluation of statistical measurement is described. Section 3 presents the application of the approach to Microarray data analysis using FDC measurement. The results of the experimentation on WOTA are presented and examined in Section 4. Section 5 gives conclusion remarks.

2. The WOTA approach

2.1. Measurement reliability $\rho_m(X)$

Let $\Psi_t(X)$ be the underlying (real) statistical distribution (the model) of a dataset X (where the subscript "t" means "true"). Assuming X' is a dataset that fits exactly to the $\Psi_t(X)$ and letting Ω be a set of data points within the dataset of X' which may or may not fit into the distribution $\Psi_t(X)$, we have

$$X = X' \cup \Omega \quad \text{or} \quad X = X' \cap \Omega,$$

where the cardinality $|\Omega| < |X'|$.

An ideal statistical measurement is a function $\Psi_i(X)$, for $i = 1, 2, \ldots$, such that $\lim_{\Omega \to \Phi} \Psi_i(X) = \Psi_t(X)$, where Φ denotes a null set (the subscript "i" means the ith ideal measurement on X). Based on our assumption, we will then have $\Psi_i(X') = \Psi_t(X)$.

Let $\rho(\cdot)$ be a reliability measurement function. A sample reliability of dataset X is defined as

$$\rho_s(X) = \vartheta[\Psi_i(X), \Psi_t(X)]$$

where ϑ is a distance operator applied to $\Psi_i(X)$ and $\Psi_t(X)$. Obviously, $\rho_s(X)$ is dependent on the dataset Ω within X.

Let $m(X)$ be a statistical measurement performed on the dataset X, and simultaneously, $\Psi_m(X)$ be the distribution obtained from applying the $m(X)$, a measurement reliability of the $m(X)$ can be defined as

$$\rho_m(X) = \vartheta[\Psi_m(X), \Psi_t(X)]$$

The $\rho_m(X)$ is related to both the statistical measurement $m(X)$, and the sample reliability $\rho_s(X)$, which relates $\rho_m(X)$ indirectly to the existence of set Ω. That is, there are two main factors affecting the reliability of a statistical measurement, namely, (1) the fault of the measurement technique itself, and (2) the fault of the sample data set.

In this paper we focus on the measurement reliability factor sourced on sample data set and try to assess the measurement reliability through rectifications of the sample reliability, i.e., the influence of set Ω. Note that the definition of "reliability" is different from the definition of "robustness." As it was defined in (Huber, 1981), "robustness" signifies insensitivity to deviations from the assumptions, e.g., about randomness, independence, and distribution models, etc. The "reliability" signifies how the statistical measurement is insensitive to deviations of the sample distributions (due to the influences of outliers

and noises), i.e., the quality of the dataset. Correspondingly, we deal with the problem of how the reliability of a statistical measurement can be evaluated with respect to the sample reliability, that is, in terms of how the set Ω is deviating from the true model defined on X'.

It is noted that the sample reliability of the dataset is associated with a number of different factors. Some of the factors that affect the sample reliability include:

(i) Sampling noise which is often inevitable. However, we want to (1) assess how much and how serious the noise affects the measurement; and (2) constrain or limit the effect of noises to the measurement.
(ii) Data acquisition errors that include, for example, the mislabeling of data category (labeling mistake) and wrong placement or inclusion of irrelevant data points.
(iii) Outliers which are a serious kind of uncertainty in statistical measurement. An outlier could be a legitimate data point – reflecting the essential nature of the problem domain. There is no absolutely effective way to detect and remove outliers from a dataset without a known model of the dataset.

In general, it is hard to find one computational method that can attenuate all these factors and result in a reliable statistical measurement. Therefore, our intension of research described in this paper is not to find a method that gives a reliable statistical measurement, but to assess the reliability of a statistical measurement by applying an analytical approach on the measurement results. Our study is conducted on the external view rather than the internal mechanism, of the statistical measurement.

2.2. Extreme-boundary-points (EBPs)

Let $\mu(X)$ be the sample mean computed on the dataset X, an EBP is defined as the data point x_i such that

$$x_i \in X$$

and

$$\forall j, \quad (x_j \in X) \wedge (j \neq i) \Rightarrow [d(x_i, \mu) \geqslant d(x_j, \mu)]$$

That is, an EBP is a sample point in dataset X such that its distance to the mean μ of X is maximum. It is also true that the EBPs are often the noises, outliers, or sampling errors. That is, they are more likely belonging to the members in the perturbation set Ω, though it is impossible to make a firm claim of this in many situations.

While the data points come with randomness, the EBP usually plays a much more significant role in deviating the statistic parameters of the data set. The proper treatment of these boundary points can improve the accuracy of statistics. Meanwhile, an analysis and assessment of the effect of these EBPs provide means to assess the reliability (and accuracy) of the statistics derived from the data set externally.

The effect of EBPs on the sample mean μ and sample variance σ^2, two fundamental statistical measurements, of dataset $X = [x_1, x_2, \ldots, x_n]$, can be quantitatively evaluated. Let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Let x_k be an EBP that is d_k distance away from the mean μ that is, $\|x_k - \mu\| = d_k$. The sample mean, μ', and sample variance, $\sigma^{2'}$, for the dataset $X' = X - [x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n]$, which is the dataset X with the removal of x_k, can be calculated as: (taking $x_k = \mu + d_k$)

$$\begin{aligned}
\mu' &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i - x_k\right) = \frac{1}{n-1}\sum_{i=1}^{n} x_i - \frac{1}{n-1}x_k \\
&= \frac{1}{n-1}\sum_{i=1}^{n} x_i - \frac{1}{n-1}(\mu + d_k) \\
&= \frac{1}{n-1}n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n-1}(\mu + d_k) \\
&= \frac{n}{n-1}\mu - \frac{1}{n-1}(\mu + d_k) = \mu - \frac{1}{n-1}d_k \\
\sigma^{2\prime} &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \mu)^2 - (x_k - \mu)^2\right) \\
&= \frac{1}{n-1}\sum(x_i - \mu)^2 - \frac{1}{n-1}(x_k - \mu)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{n-1}(\mu + d_k - \mu)^2 \\
&= \frac{1}{n-1}n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right) - \frac{1}{n-1}d_k^2 \\
&= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}d_k^2 = \sigma^2 - \frac{1}{n-1}[d_k^2 - \sigma^2]
\end{aligned}$$

That is, the variations of the mean and variance with respect to the removal of an EBP x_k are the $(n-1)$th factor of d_k and $d_k^2 - \sigma^2$, respectively. However, these quantities neither reveal the sample reliability of the dataset X, nor the reliability of the statistical measurements μ and σ^2 on the dataset X directly. It tells that the number of data points, n, is an important factor to the magnitude of variations of the measurement with respect to the removal of certain data points from the set.

2.3. EBP weed-out trend patterns (WOTP)

To see how a statistical measurement is less variant (i.e, more reliable) with respect to the presence of a perturbation set Ω in dataset X, we adopt a method that sequentially weeds out some EBPs from the dataset, and assess the trend pattern of the resulting statistical measurements. The method is based on the principle that if a statistical model assumption accurately reflects the true distribution parameters of the data set, and the size of the data set is

reasonably big, then the elimination of one or a few (a small percentage) data point should not significantly alter the overall value of the measurement.

The WOTP approach concentrates on distinguishing four basic pattern types (in terms of the shapes and slopes of the curve) with respect to the measurement variations under the varying EBP weed-out conditions (the number of EBPs weeded out). These four basic pattern types are defined in our research as follows:

(1) Steady patterns – The variations of the measurements over all weed-out points are all within a certain range that are relatively small.
(2) Monotony patterns – The variations of the measurement are either positive or negative over all weed-out cases.
(3) Conic (convex or concave) patterns – The variations of the measurement show an up-and-down (or a down-and-up) trend, such as a conic section, over the process.
(4) Oscillation (wave) patterns – The variations of the measurement show more than two up-and-down (or down-and-up) trends over the process.

The reason that we categorize these patterns into different types is that these mathematically disciplinary curves can help us to better understand the weed-out trends and their effects to statistical measurement reliability. It is also possible to define a trend coefficient κ which can be used as a threshold for further categorization of the above four basic pattern types. More specifically, a coefficient κ can be applied to measure the swinging variations along the sequence of data points that forming the patterns. The difference between two consequent measurement values under the varying weed-out conditions, for example, from weeding out i EBPs to $i+1$ EBPs is considered in this measurement, as presented in Section 3.3.

With the use of the κ, the four pattern types of above can be further divided into 11 WOTPs for a qualitative assessment of the reliability of the corresponding statistical measurement. These WOTPs are:

A. Under κ Steady pattern – There presents as a straight line that can be drawn in parallel to the number of weed-out EBP axis such that no any variation of the measurements is κ distance away from this line.
B. Under κ Monotonic rising pattern – If the variation of the measurement over each case of weed-out test is less than or equal to κ, and the κ is always positive in each case, we call the pattern "Under κ monotonic rising".
C. Under κ Monotonic dropping pattern – If the variation of the measurement over each case of weed-out test is less than or equal to κ, and the κ is always negative in each case, the pattern is "Under κ monotonically dropping".
D. Under κ Conic-Valley pattern – The variations of the measurement show a down-and-up trend but no variation has a magnitude greater than κ in any of the variations.
E. Under κ Conic-Mountain pattern – The variations of the measurement show an up-and-down trend but no variation has a magnitude greater than κ in any of the variations.
F. Under κ Oscillation pattern – The variations of the measurement show more than two up-and-down (or down-and-up) trends but no single variation is greater than κ.
G. Over κ Monotonic rising pattern – If the variation of the measurements in some cases of weed-out tests exceed the κ, and the κ is always positive in each case, we call the pattern "Over κ monotonic rising".
H. Over κ Monotonic dropping pattern – If the variations of the measurement in some cases of weed-out tests exceed the κ, and the κ is always negative in each case, the pattern is "Over κ monotonically dropping".
I. Over κ Conic-Valley pattern – The variations of the measurement show a down-and-up trend with some variations having magnitude greater than κ.
J. Over κ Conic-Mountain pattern – The variations of the measurement show an up-and-down trend with some variations having magnitude greater than κ.
K. Over κ Oscillation pattern – The variations of the measurement show more than two up-and-down (or down-and-up) trends with some variations having magnitude greater than κ.

The above pattern categories give a qualitative indication of the reliability of an associated statistical measurement in the order from A to K. A score can be assigned to each of these pattern categories, from A to K, with A having the highest score and K the lowest. That is, a statistical measurement showing a WOTP of category A is considered to be a most reliable measurement. The reliability of the measurement decreases when the resulting WOTP category falls down the list from pattern A towards the pattern K. Note that in many cases, only the four basic WOTPs are needed to a rough assessment of the reliability of the statistical measurement. The more detailed categorization of the WOTP can be considered as a way to provide a more quantitative assessment of the statistical measurement.

Some examples of these patterns can be seen in Fig. 2. In our research, algorithms are developed for analyzing the EBP weed-out trend, extracting the WOTPs, and recognizing the pattern categories. The statistical measurements of target data sets are then assessed in terms of these patterns and the pattern parameters.

2.4. WOTA algorithm

Let $m(X)$ be a statistical measurement function applied to a dataset $X=[x_1,x_2,\ldots,x_n]^k$. Let ${}^kX$ denote the dataset

that has k extreme-boundary-points weeded out, that is $|^kX| = n - k$. The WOTA algorithm we use in our research can be described as the following.

Algorithm – WOTA

// This algorithm extracts a WOTP from applying $m(X)$ on dataset X multiple times, and classifies the WOTP into one of the 11 categories.//

Inputs:

X: Sample data set;

d: Maximum number of EBPs to be weeded out;

Outputs:

$WOTP[\,]$: An array that holds the WOTP values of the $m(X)$ measurements;

CX: A score according to the WOTP category on dataset X.

Uses:

$m(X)$: A statistical measurement function applied to X;

$EBP(X)$: A function that finds the EBP of X;

$CX(P)$: A function that calculates according to the WOTP patterns with the trend coefficient κ.

Process:

1. $WOTP[0] \quad m(X)$,
2. For ($k = 1$ to d) do
   - 2.1 Compute the mean μ_k of data set $^{(k-1)}X$; // When $k = 1$, $^{k-1}X = X$.
   - 2.2 $EBP_k \quad EBP(^{k-1}X)$; // find the EBP of ^{k-1}X and assign it to EBP_k
   - 2.3 $^kX \quad ^{k-1}X - EBP_k$; // ^kX is a dataset from ^{k-1}X with EBP_k removed.
   - 2.4 $WOTP[k] \quad m(X - k)$;
3. $WOTP[\,] \quad WOTP[\,]$ with an application of moving average computation;
4. $CX \quad CX(WOTP[\,])$.

The WOTA algorithm of the above is only intended to give an objective assessment of the reliability of the measurement externally, with respect to the given dataset by assigning the WOTP to one of the 11 categories. As pointed out before, it is by no means intended to improve the efficiency, accuracy, or robustness of the statistical measurement internally. However, it can be used to complement to the statistical measurements for selecting a better data analysis process. We applied the method to a number of Microarray data analyses tasks, to help make decisions on the selection of genes that are statistically significant to the given criteria. The problems, experiments, and results are presented in the next two sections.

3. Application of WOTA to FDC for microarray data analysis

In the following we present the application of the WOTA approach to Fisher's Discriminate Criterion (FDC) for Microarray gene expression profiling data analysis. FDC (Fisher, 1936) is a well-known parametric method for identifying data attributes and their projections that are most likely to be separable among different classes. It has been popularly used in recent years for identifying genes predictive to certain biological phenomena from DNA Microarray gene expression profiling datasets that are differentially expressed (Brown et al., 2000). The approach described in this paper is an attempt to find out how to attenuate the effects of measurement uncertainties.

3.1. On the microarray data analysis

DNA Microarray as a rapidly developing technique in biology and biomedicine provides an effective means for monitoring the expression levels of thousands of genes simultaneously (Granjeaud et al., 1999; Alizadeh and Staudt, 2000; Saluz et al., 2002). DNA Microarrays are used to identify a molecular predictor of a specific translocation and survival outcome after chemotherapy for diffuse large B-cell lymphoma (DLBCL) (Rosenwald et al., 2002; Shipp et al., 2002; Iqbal et al., 2004).

There were many ways discussed in literatures for identifying genes that are indicative to certain diseases or health disorders (Zhu et al., 2004). Most methods focused on the scoring of genes for relevance detections. General approaches include (1) parametric methods, such as the principal component analysis (PCA) (Oja, 1992), independent component analysis (ICA), and separation correlation metric (SCM), or known as the Fisher's discrimination criterion (FDC) (Fisher, 1936); and (2) non-parametric methods, such as the threshold number of misclassification (TNoM) (Ben-Dor et al., 2000), projection pursuit regression (PPR) (Friedman and Tukey, 1974), support vector machines (SVM) (Brown et al., 2000), neural networks, expectation maximization (EM), etc.

However, there are number of issues that compromise the accuracy and reliability of the statistical measurements and analysis of the Microarray gene expression profiling data. These issues include: (1) the existence of noise, outlier, and uncertainties in the sample data, (2) the imbalance of the available number of cases in each of the disease categories, (3) the incongruence of the number of cases (data points) versus the number of genes (the data dimensions) to be analyzed, and (4) the inaccuracy and uncertainty of the clinical/pathologic diagnostic characterization of the cases.

The parametric methods make use of a set of statistical metrics derived from the gene expression profiling dataset under the assumption of certain statistical models. Statistical methods are usually reliable and accurate in large data set analysis. However, the incongruence between the relatively small number of data samples collected in current practice and the large dimensions of the genes profiled often makes the statistical models not justifiable. Moreover, the statistical measurements are easily biased and distorted by the outliers resulted from noise corruptions

taking place in the data acquisition processes. Examination of our experimentation dataset showed that those outliers often value at a magnitude of 4–10 times away from the normal values. These values severely deviate from the major statistical parameters (the mean and variance values) on the dataset that has a count of 20–40 samples.

The non-parametric methods do not rely on the assumptions of the statistical models and parameters. Rather, they work toward the objectives (such as discrimination or prediction) by applying certain non-statistical metrics or protocols directly on the individual data samples. The methods would be advantageous at constraining and attenuating the effects of outliers. However, the diversity of measurement metrics and the uncertainty (which includes the imprecision and incompleteness) of the individual data samples often make them hard to get a consistent result in different experimentations.

It is noted that different algorithms/evaluations often result in different set of genes in the gene expression profile analyses. It is therefore desirable to establish a way of assessment of the reliability of the measurements for the varying approaches, so that the merit of the resulting gene sets can be better assured.

3.2. FDC for DLBCL data analysis

3.2.1. The FDC measurement

The FDC can be expressed as the following. Let ω_1 and ω_2 be the labels of two different sample classes (e.g., surviving vs. fatal cases like in the DLBCL data set of our study). The method is aimed at maximizing a criterion

$$J(\underline{W}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3.1}$$

where μ_i, $i = 1, 2$, is the mean vector of the projection of the data samples of classes ω_i in a direction $\underline{W}$, respectively. That is

$$\mu_i = \frac{1}{n_i} \sum_{\underline{X} \varepsilon w_i} \underline{W}^T \underline{X} \tag{3.2}$$

where $\underline{X} = [x_1, x_2, \ldots, x_n]$ represents the gene expression vector (expression values of an individual gene over all sampling cases). The n_i is the number of data samples in class ω_i. The σ_i^2, $i = 1, 2$ is the scatter (or variance) matrix for the projected samples of class ω_i in direction $\underline{W}$, respectively.

$$\sigma_i^2 = \frac{1}{n_i} \sum_{x \in \omega_i} (\underline{W}^T \underline{X} - \mu_i)^2 \tag{3.3}$$

When limiting the projection vector $\underline{W}$ to the form of [1 0 ...], [0 1 ...], ... (i.e., axes of the Euclidean coordinates), the criterion represents a measurement of individual genes according to its mean and variance parameters with respect to the original class designations. Let FDC_k denotes such a measurement on gene g_k, i.e, on vector $x_k = [x_{k1}, x_{k2}, \ldots]$, the criterion can then be expressed as

$$\mathrm{FDC}_k = \frac{(\mu_{k1} - \mu_{k2})^2}{\left(\sum_{j=1}^{n_1} (x_{kj}^{(1)} - \mu_{k1})^2\right) \Big/ n_1 + \left(\sum_{j=1}^{n_2} (x_{kj}^{(2)} - \mu_{k2})^2\right) \Big/ n_2} \tag{3.4}$$

where the $x_{kj}^{(1)}$ and $x_{kj}^{(2)}$ are gene expression values corresponding to classes ω_1 and ω_2 respectively. The separability of the genes with respect to the tumor outcome classes thus can be ranked by these FDC_k values. We denote the FDC_k as the FDC measurement of the gene numbered k in this paper.

3.2.2. The measurement reliability of FDC

It was mentioned by Zhu et al. that the FDC measurement alone does not provide an overall good indication of the genes in relation to the clinic outcomes (Zhu et al., 2004). According to the studies of Yang and Zhang (2002), FDC is not an absolute criterion for yielding accurate classifications. They pointed out that the method should be combined with other statistical or non-statistical correlation analyses to diminish some parametric side effects. In Zhu et al.'s paper, a set of methods called Cross-projection (CP) and Discrete Partition (DP) are proposed to fuse with the FDC in order to diminish the side effects of the outliers on FDC (Zhu et al., 2004). The algorithmic fusion approach provides an overall better measurement on the gene analysis in the experiments.

The situation with respect to the FDC measurement can be illustrated by the following examples. The dataset we use here comes from "The Leukemia/ Lymphoma Molecular Profiling Project" (LLMPP) (it is represented in Section 4.1). The data set includes 7399 genes and 240 cases, but only 65 cases are studied here. We take a look at the gene #2967 and #6641, and give their respective FDC measurements in terms of the ranks of the measurements over a total of 7399 genes in Table 1. It is seen that the gene #2967 has a relative high rank (=17) in its original FDC measurement, while gene #6641 has a relative low rank (=147) in its original FDC measurement. However, the ranks change significantly when some EBPs are weeded out of the data set. In the table, the columns of FDC-1, FDC-2, and FDC-3 indicate the FDC measurement with 1, 2, and 3 EBPs taken out. The overall expression values of these two genes are shown in Fig. 1. It is obvious that the original FDC measurements do not properly reflect the statistical characteristics of the gene expressions, that is, the measurements are not reliable.

Table 1
FDC measurements of gene #2967 and #6641 in different test cases

Gene #	FDC rank	FDC-1 rank	FDC-2 rank	FDC-3 rank
2967	17	118	665	1322
6641	147	19	4	10

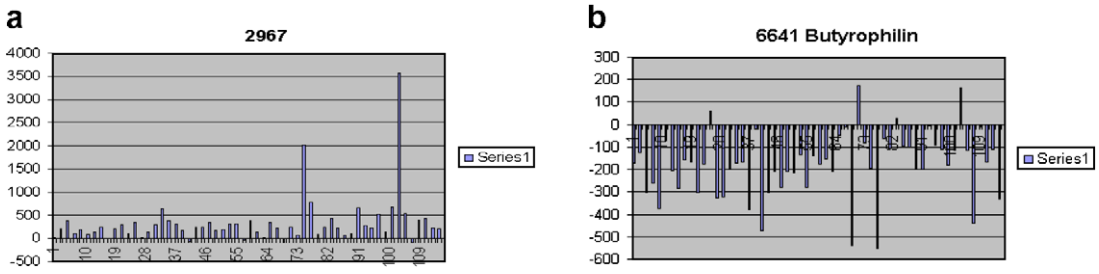

Fig. 1. Overall gene expression values of (a) Gene #2967 and (b) Gene #6641.

3.3. Applying WOTA to FDC measurement – algorithm

To find that how FDC is sensitive to the unreliable values in the dataset and to extract genes that are both meaningful and reliable to the objectives, we experimented on the dataset applying WOTP. The process of the computation is described as follows.

1. Preprocessing the Microarray gene expression profiling dataset.
2. For each gene g_k
   Compute FDC value by using formula (3.4), where $x_{kj}^{(1)}$ and $x_{kj}^{(2)}$ are gene expression values corresponding to *Group* 0 and *Group* 1 respectively; and μ_{k1}, μ_{k2} can be calculated by using formula (3.2)
   Calculate the Euclidean distance between x_{kj} and μ_k by

$$d(x_{ki}, \mu_k) = (x_{ki} - \mu_k)^2 \tag{3.5}$$

   Sort the distance in a *descending* order, choose first ε (ε is a small number of the whole samples) cases, i.e., the ε cases with farthest distances to the mean μ_k.
   Re-compute the FDC values on the data set with the removal of the ε cases of EBPs, one at a time, by using the formula (3.4), where μ_{k1}, μ_{k2} are the new means after the possible elimination of extreme values. Note that ε FDC values are obtained on eliminating $0, 1, 2, 3, \ldots, \varepsilon$ EBPs respectively. These FDC values are recorded as elements of WOTP and kept in an array WOTP[ ].
   Following the steps 2.1–2.4, we obtain the WOTP for every gene g_k.
3. Sorting the genes according to the highest FDC values obtained in step 2, and then choose the first N genes with both highest FDC values and WOTA score.
4. Take the cross-over fusion on the lists of the ε FDC measurement results and select a set of genes that have high ranks on combination of the results.

In our experiment, we choose $\varepsilon = 15$, because 15 cases is the number about 20% of the total case of number 65, which is close to a quarter of the dataset such that we can have plenty of FDCs to look for WOTPs. Some typical WOTPs of the gene sets are shows in Fig. 2a–k below. A moving average of the FDC measurements is also show in the figures. The κ for categorize the measurements is selected through try- and error experimentations. In the results presented in Fig. 2, we had the value set to 0.1 which, we believe, gives us a reasonable categorization of the trend patterns. Other κ may also be selected, which will lead to certain variations of the assessment results.

4. Experiment results and analysis

4.1. BCL2 translocation correlative gene extraction

The data set used in our experiment was derived from "The Leukemia/Lymphoma Molecular Profiling Project" (LLMPP) which was also used by Iqbal et al. (2004). Among the 240 cases of DLBCL measured in gene expression profiling, 129 cases were studied for the presence of a specific translocation involving a gene called BCL2. Iqbal et al. mapped the *BCL2* translocation data into the gene expression defined subgroups of DLBCL. From a specific subgroup (GCB) that was positive (+) for *BCL2* translocation was combined into *Group* 1 as *BCL2* translocation positive cases. The negative (−) cases formed *Group* 0. *Group* 1 contains 29 cases, whereas *Group* 0 has 36 cases. These 65 cases have gene expression profiling data with 7399 clones for each case. These cases of DLBCL have gene expression profiles determined by complementary DNA (cDNA) Microarray technology (Rosenwald et al., 2002). All the values in the data set are based on the value of the expression ratio R/G (*Cy3*/*Cy5*) i.e., tumor sample to reference standard. According to the suggestion of Yang et al. for avoiding the data normality assumption (Dudoit et al., 2000), we preprocessed the data set by Box-Cox transformation and zero mean normalization (z-score). The method transforms the response $y \to t_\lambda(y)$ where the family of transformations indexed by λ is

$$t_\lambda(y) = \begin{cases} \frac{(y^\lambda - 1)}{\lambda} & \text{when } \lambda \neq 0 \\ \log(y) & \text{when } \lambda = 0 \end{cases}$$

For fixed $y > 0$, $t_\lambda(y)$ is continuous in λ. The λ is chosen by using maximum likelihood. Here, we choose to use $t_\lambda(y) = \frac{R}{G}$, and $\lambda = 0$. After the transformation, the distribution of the data set is approximately normal distributed (Dudoit et al., 2000).

Applying the WOTA method to the DLBCL data set, and evaluating the trend patterns of the FDC values through weeding out 15 EBPs, we obtained a set of 35 genes that are most reliably correlated with the BCL2

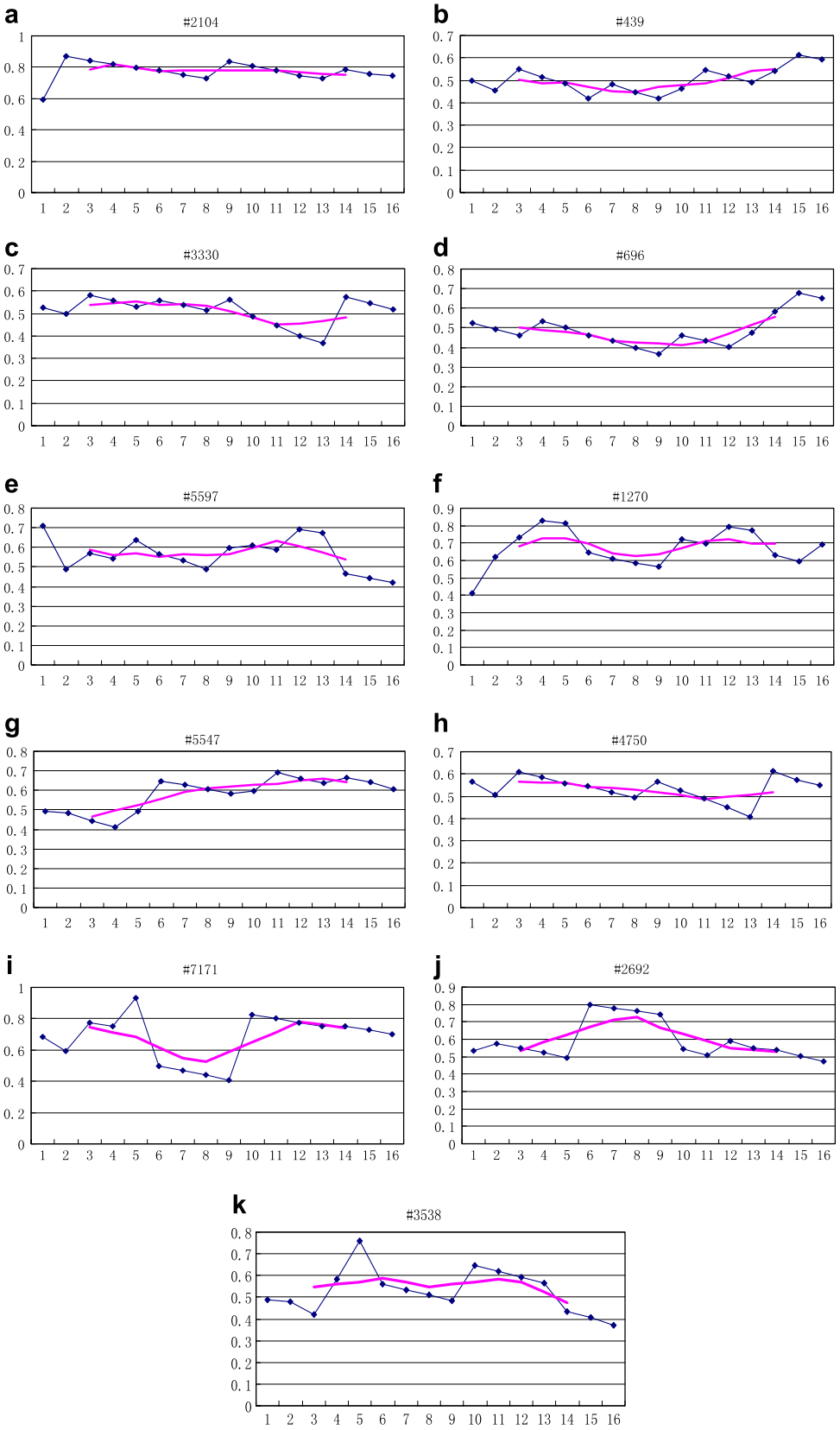

Fig. 2. Selected examples of WOTP in FDC measurements of Microarray DLBCL gene expression dataset: (a) example of WOTP category A, (b) example of WOTP category B, (c) example of WOTP category C, (d) example of WOTP category D, (e) example of WOTP category E, (f) example of WOTP category F, (g) example of WOTP category G, (h) example of WOTP category H, (i) example of WOTP category I, (j) example of WOTP category J and (k) example of WOTP category K.

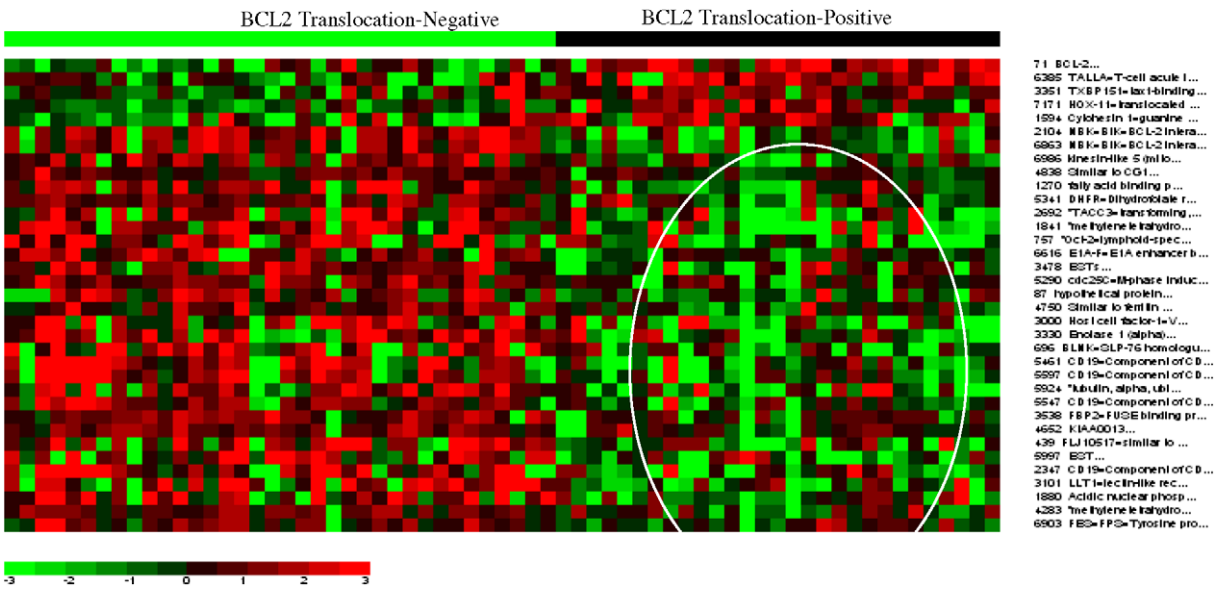

Fig. 3. Expression profile of genes extracted using FDC measurement WOTA.

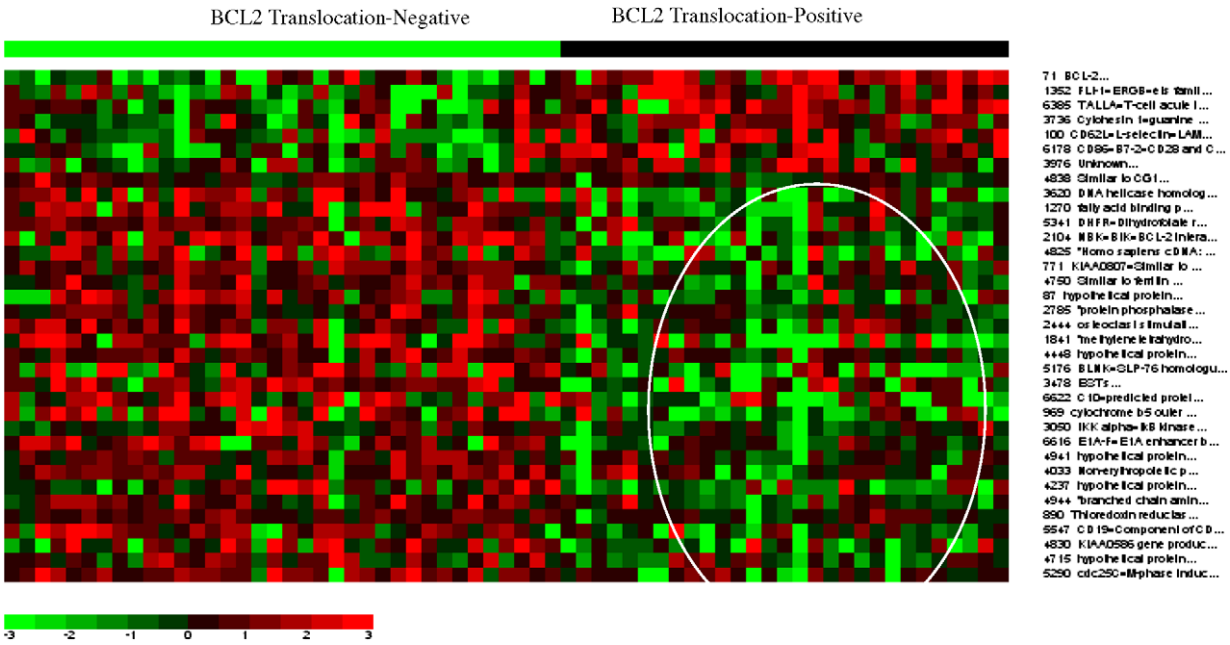

Fig. 4. Expression profile of genes having highest FDC values without WOTA.

translocation. Fig. 3 gives the visual pattern of the selected genes, with the $t(14;18)$ negative cases on the left and $t(14;18)$ positive cases on the right. To give a comparison we also show in Fig. 4 the visual pattern of 35 genes that have high FDC values without going through the WOTA processes.

From both of the figures, we can clearly see that selected genes are well divided into two groups, one performs high (red[1]) to the BCL2 translocation positive cases while the other group performs high to the BCL2 translocation negative cases. the result gene sets in two experiments are different, 37% of them are overlapped. This means that these 37% of the genes are more reliable than the others in terms of the FDC measurement with respect to out WOTA. In another study, a more quantitative result was obtained, as described in the section below.

4.2. Results on clinical outcome prediction

Two gene expression profiling studies of DLBCL for identifying genes predictive of clinical outcomes have been

[1] For interpretation of color in figures, the reader is referred to the Web version of this article.

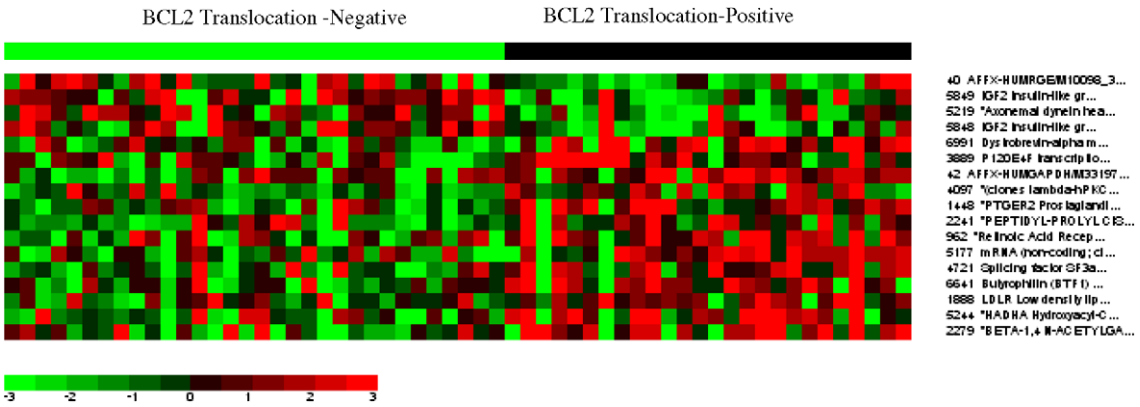

Fig. 5. Genes indicative for survival outcome extracted by applying FDC.

Table 2
Genes extracted for DLBCL outcome prediction applying FDC with WOTA

Index	Gene#	Description	WOTP category
1	6991	Dystrobrevin-alpha mRNA	F – under κ Oscillation pattern
2	3889	P120E4F transcription factor mRNA	E – under κ Conic pattern
3	4097	(clones lambda-hPKC-beta[15,802]) protein kinase C-beta-1 (PRKCB1) mRNA	B – under κ Monotony pattern
4	5849	IGF2 Insulin-like growth factor 2 (somatomedin A)	F – under κ Oscillation pattern
5	2241	Peptidyl-prolyl CIS-TRANS isomerase mitochondrial precursor	E – under κ Conic pattern
6	5177	mRNA (non-coding; clone h2A)	E – under κ Conic pattern
7	4721	Splicing factor SF3a120	F – under κ Conic pattern
8	6641	Butyrophilin (BTF1) mRNA	E – under κ Conic pattern
9	2279	BETA-1,4 *N*-acetylgalac to saminyl transferase	B – under κ Monotony pattern

reported (Shipp et al., 2002; Rosenwald et al., 2002). Rosenwald et al. identified four functional groups of genes that are predictive of survival and one of the groups consists of genes that divide the tumor into distinct biologic subtypes (Rosenwald et al., 2002). Shipp et al. applied supervised learning method on an entire expression profiling dataset and identified 13 individual genes that are highly predictive to the survival outcomes (Shipp et al., 2002). We conducted the WOTA experiments by applying the FDC measurement on the same dataset which contains 58 DLBCL samples used by Shipp's group (www.genome.wi.mit.edu/MPR/lymphoma). First, a total of 17 individual genes were identified by applying the FDC measurement (Fig. 5). Then, the WOTP patterns of these genes were further studied, which leads to the selection of 9 genes in the list that show relatively more acceptable patterns under weed-out situations. Table 2 shows these genes along with their WOTP category notations.

To further evaluate the quality of the result we obtained, a simple linear discrimination as well as a quadratic discrimination process (Zhu et al., 2004), is applied to the 9 genes with respect to the original dataset. Table 3 gives the number of correctly identified clinical cases, versus the result reported in (Zhu et al., 2004; Shipp et al., 2002). The result shows an improvement to the previous approach: the total number of predicted genes is 46 in linear classifier which is better than result of Shipp's gene list, while 50 predicted genes in quadratic classifier which is visibly improved than Zhu's previous results. The survival prediction results show that not only the nine genes explored in the WOTA can give a more reliable prediction, but also is in a reduction to the number of genes needed to examine for determining the survivability.

Table 3
Comparison of classification results on different gene sets for survival prediction in DLBCL Microarray gene expression profiling

		Linear classifier			Quadratic classifier		
		Survival	Fatal	Total	Survival	Fatal	Total
1	The result in (Zhu et al., 2004)	27	19	46	30	18	48
2	With Shipp's gene (Zhu et al., 2004)	26	19	45	26	19	45
3	New result using our 9 genes	26	20	46	28	22	50

5. Conclusion

We studied the problem of how to assess the reliability of a statistical measurement on a dataset contaminated with noises, uncertainties, and outliers. The EBP based WOTA approach is a practical way for gaining a qualitative assessment of the measurements and providing useful hints to the selection and acceptance of the measurement under given circumstances. Particularly, we applied the

approach to the FDC measurements for selections of indicative genes in Microarray DLBCL gene expression profiling data analysis. The approach resulted in the extraction of genes that are more meaningful than the results obtained without the reliability assessment in both experiments. We must point out that the WOTA is not an approach of improving the FDC measurement internally. It is an approach for improving the results obtained from applying the FDC measurement by providing an additional selection process that is external to the FDC measurement. In this sense, the WOTA is a complementary process to a statistical measurement for improving the results derived from the measurements. We hope this assessment approach can be applied to more statistical evaluation processes and data analysis problems as we continue our research work.

Acknowledgements

The first author of this paper would like to thank Dr. Steve From of the Mathematics Department at the University of Nebraska at Omaha for the help he provided on some statistics concepts presented in this paper. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions that helped us greatly to improve the presentation of this paper.

References

Alizadeh, A.A., Staudt, L.M., 2000. Genomic-scale gene expression profiling of normal and malignant immune cells. Curr. Opin. Immunol. 12 (2), 219–225.

Arning, A., Agrawal, R., Raghavan, P., 1996. A linear method for deviation detection in large database. In: Proc. 2nd Internat. Conf. on Knowledge Discovery in Databases and Data Mining.

Ben-Dor, A., Fridman, N., Yakhini, Z., 2000. Scoring genes for relevance. Technical Report AGL-2000-13, Agilent Labs, Agilent Technologies.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: Identifying density-based local outliers. In: Proc. ACM SIGMOD 2000 Internat. Conf. on Management of Data, Dalles, TX.

Brown, M. et al., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA 97, 262–267.

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P., 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578, August 2000.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Reproduced from the Annals of Eugenics 7, 179–188.

Friedman, J.H., Tukey, J.W., 1974. A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Computers c-23 (9), 881–890.

Granjeaud, S., Bertucci, F., et al., 1999. Expression profiling: DNA arrays in many guises. Bioassays 21 (9), 781–790.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Iqbal, J., Sanger, W.G., Cao, K., Zhu, Q., et al., 2004. BCL2 translocation defines a unique tumor subset within the germinal center B-cell-like diffuse large B-cell lymphoma. Amer. J. Pathol. 165 (1), 159–166.

Knorr, E.M., Ng, R.T., 1999. Finding Intensional knowledge of distance-based outliers. In: Proc. 25th VLDB Conf., Edinburgh.

Knorr, E.M., Ng, R.T., Tucakov, V., 2000. Distance-based outliers: Algorithms and applications. VLDB J.: Very Large Databases 8 (3–4), 237–253.

Maddala, G.S., Yin, Y., 1997. Outliers, unit roots and robust estimation of nonstationary time series. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 15. Elsevier, pp. 237–266.

Oja, E., 1992. Principal components, minor components, and linear neural networks. Neural Networks 5, 927–935.

Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. In: Proc. ACM SIGMOD Conf.

Rosenwald, A., Wright, G., Chan, W.C., et al., 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New Engl. J. Med. 346 (25), 1937–1947.

Saluz, H.P., Iqbal, J., et al., 2002. Fundamentals of DNA-chip/array technology for comparative gene-expression analysis. Curr. Sci. 83 (7), 829–833.

Sellke, T., Bayarri, M.J., Berger, J., 2001. Calibration of *P*-values for testing precise null hypotheses. Amer. Statist. (55), 62–71.

Shipp, M.A. et al., 2002. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Med. 8 (1), 68–74.

Yang, J., Zhang, D., 2002. What's wrong with Fisher criterion? Pattern Recognition 35 (11), 2665–2668.

Yu, D., Sheikholeslami, G., Zhang, A., 1999. Findout: Finding outliers in very large datasets. Technical Report, 99-03.

Zhu, Q., Cui, H., Cao, K., Chan, J., 2004. Algorithmic fusion of gene expression profiling for diffuse large B-cell lymphoma outcome prediction. IEEE Trans. Inform. Technol. BioMed. 8, 79–88.