Elyn C. Tan
SUID: 4862645
e-mail address: eht@stanford.edu
Course: Biochem 218
Professor Doug Brutlag

A Critical Review of Statistical Methods for Differential Analysis of 2-sample Microarrays

The advent of microarray technology has made it possible to study the variation of expression for many genes simultaneously. The most common types of microarray experiments involve the comparison of gene expression across two or more kinds of tissue samples or of samples obtained under different experimental conditions. The analysis of gene expression data can be done from various levels of complexity: 1) at the level of single genes, where one seeks to determine whether a particular gene is differentially expressed under control and experimental conditions; 2) at the level of multiple genes, where attempts to classify genes into known classes (discriminant analysis/supervised learning) or to identify new or unknown classes (cluster analysis/unsupervised learning) are made through the analysis of common functionalities, interactions, regulation, etc.; and 3) at the systemic level, where the goal is to identify underlying gene and protein networks responsible for the patterns observed (1). This paper focuses on the first level of analysis.

One of the major goals of microarray data analysis is the identification of genes that are differentially expressed across two or more kinds of tissue samples or samples obtained under different experimental conditions. Gene expression patterns are thought to be different for various tissue types and for tissues at various stages of development and disease states. Genes that show differential expression between diseased tissue and normal tissue will allow for the identification of biomarkers for disease class predictions as well as the ability to fine-scale predictions of drug responses. Hence, a large number of statistical methods that will allow the researcher to systematically assess and measure the significance of any observed changes in gene expression levels have been proposed. In this paper, I will critically review some of the methods and modeling approaches. Clearly, the generation of experimental data is not enough; one must be able to sift through the large volume and substantial variation of the data and to extract biological information that is meaningful to the study. One of the major tasks of this paper is to compare and contrast the underlying assumptions, strengths, and weaknesses of each statistical method, with the recognition that different methods are suitable for different questions and types of analyses.

Expression Ratios and the fold-change approach

To illustrate the application of the various statistical methods discussed in this paper, I will take as my central example a hypothetical experiment involving the comparison of gene activity in a tumor versus a normal tissue. To investigate any changes in gene expression levels of some set of genes of interest, the scientist will need to position known DNA base sequences of each gene of interest into a pre-specified position on the microarray plates. Tumor cells will then be hybridized onto the plate, and these cells will generate messenger RNA's (mRNA) in proportion to the gene's actual activity in the cell. One can then visualize the expression levels of each gene by using a red dye to depict the effect the scientist is interested in and a green dye to measure any background activity and serve as a control. A differential expression ratio R/G can be derived for each spot. After hybridization, a number of preprocessing steps usually follow, such as dimension reduction, data normalization and data transformation to adjust

for any systemic variations or dye bias that may have occurred during the course of the experiment (2). Often, the logarithms of the expression ratios rather than the ratios themselves are used, because log ratios are easier to model and interpret. For example, taking the base2 logarithm of the expression ratios converts the multiplicative effect of the ratios into additive effects that allow us to more easily interpret the results. A gene that is upregulated by a factor of 2 has a $\log_2$(ratio) of 1, a gene downregulated by a factor of 2 has a $\log_2$(ratio) of -1, and a gene expressed at a constant level has a $\log_2$(ratio) of 0. One can also take the natural log or base10 log of the expression ratios, and the choice of which base to use depends on the researcher, as long as s/he is consistent (3).

Initially, measurements of differential expression were assessed simply by comparing the ratio of expression levels between the two conditions, a method known as the fold change approach. Genes with ratios above a fixed cut-off k (that is, those whose expression underwent a k-fold change) were said to be differentially expressed. However, this method has been proven to be unreliable because it fails to take into account measurement error (variance). For example, an excess of low-intensity genes may be mistakenly identified as differentially expressed because their fold-change values have a larger variance than the fold-change values of high intensity genes. The fold-change approach will also fail to find genes that show a highly reproducible but small difference in relative expression values. Methods that take variability into account are therefore preferred and have been found to be considerably more reliable than the fold-change approach (4).

Li and Wong (5) introduced a more sophisticated fold-change approach to analyzing oligonucleotide array data. They first fit a model that accounts for random, array- and probe-specific noise, and then evaluated whether the 90% confidence interval for each gene's fold-change excludes 1.0. Unlike standard fold change approaches, this method incorporates available information about variability in the gene-expression measurements. However, because the error model is fitted to the entire data set, it can suffer when the data set is either too small or too heterogeneous. Other model-based methods designed for two-color arrays (6) also incorporate data-derived estimates of variation. However, before I move on to discuss the various statistical methods that have been proposed to measure differential expression, I will first address the problem of multiple testing and the evaluation of significance.

Significance and the problem of multiple testing

Regardless of the test statistic used, one needs to convert it to a p-value to determine its significance. Standard methods for computing p-values often employ the use of a statistical distribution table that lists the threshold value of the test statistic needed to determine significance. However, these tabulated values rely on the assumption that the data are sampled from normal populations with equal variances. Permutation tests, which are carried out by repeatedly shuffling the samples' class labels and computing t statistics for the genes in the shuffled data enables one to assess significance without assuming normality. Unfortunately, permutation tests are time- and effort-consuming; among its disadvantages are the complexity of its derivation and the requirement that the size of the dataset be large enough to allow for a sufficient number of distinct permutations to be obtained (1).

The issue of multiple testing is crucial in the analysis of microarrays as most microarray experiments often monitor the expression levels of thousands of genes, requiring that thousands of statistical tests to be computed. If we are to simply assume a standard p-value for every experiment, we run the risk of accumulating large numbers of false positive results. To illustrate, if one uses a p-value equal to 0.01 to monitor the expression levels of 5000 genes, one should expect a false positive error rate on the order of 50 genes, which is in most cases, an unacceptably high false positive rate. What follows is a brief discussion of the methods that have been proposed to address the issue of multiple testing.

Family-wise error-rate control

One approach to multiple testing is to control the family-wise error rate (FWER), which is the overall probability that at least one gene is incorrectly identified as differentially expressed over a number of statistical tests. One way to control for FWER is to increase the stringency applied to each individual test. This can be done by performing a Bonferroni correction, where the desired significance level is divided by the total number of tests conducted. Unfortunately, standard Bonferroni corrections assume independence of the different tests and an acceptable FWER could be achieved for microarray data only if the Bonferroni threshold is set at a very stringent level (7). Such stringency often results in the non-identification of any genes as differentially expressed. A step-down correction method was designed by Westfall and Young (8), and this method allows for dependence between the different tests, but can still be overly restrictive in some cases. Permutation-based one-step correction procedures have also been proposed as alternatives to the Bonferonni correction. The latter tests have been found to perform better compared to the standard Bonferroni, although a disadvantage is their computational complexity relative to the Bonferroni procedure.

False-discovery-rate control

For microarray studies that focus on finding sets of predictive genes, an alternative approach to multiple testing considers the false discovery rate (FDR), which is the probability that a given gene identified as differentially expressed is a false positive. The FDR is typically computed after a list of differentially expressed genes has been generated (9). Unlike a significance level, which is determined before looking at the data, FDR is a post-data measure of confidence. It uses information available in the data to estimate the proportion of false positive results that have occurred. A simple method for bounding the FDR is proposed by Benjamini and Hochberg (9). Benjamini and Hochberg's method assumes independent tests and sets an upper bound for the FDR by a step-up or step-down procedure applied to individual P values. In this method, the calculated P values of each independent test are ordered from P(1) being the most significant to P(n) being the least significant. The analyst can then formulate a rule R that will specify when a null hypothesis is rejected. For example, R could be set as: "Reject H_i if P_i is among the smallest 1% of the P-values and $P_i \leq 0.001$". Benjamini and Hochberg were then able to prove that the FDR of R is the expected proportion of rejected H_i that were actually true. They identified an algorithm that allows the specification of a preset value α which serves as the upper bound of the FDR of R where FDR (R α) ≤ α. The analyst can therefore use R as a measure of significance and be assured that the FDR from using R will be less than or equal to the preset value α derived from Bejamini and Hochberg's algorithm.

Another method is the positive false-discovery rate (pFDR) proposed by Storey (10). It multiplies the FDR by a factor Π_0, which is the estimated proportion of non-differentially expressed genes among all the studied genes. It has been found that the FDR criteria or its variants allow for a higher false positive rate than FWER procedures, and can therefore be a valuable alternative when more stringent analyses fail to identify potential leads.

Methods for Differential Analysis

Most methods that have been proposed to assess differential analysis are based on using the two-sample t-test or a minor variation of the t-statistic, but they differ in how to associate a. statistical significance level (p) to the corresponding summary statistic. As mentioned in the previous section, differences in how a significance level is assigned could lead to possibly large differences in the numbers of genes detected and the number of false-positives and false negatives. For analysts to choose between different statistical methods, it is important that they understand the various modeling assumptions underlying each method, particularly in relation to how each method determines the corresponding significance level or p-value associated with the test statistic.

The t-Test

A straightforward method is the traditional t-test. Suppose that Y_{ij} is the expression level of gene *i* in array *j*. *i* can take on the values (1, 2, 3, …, n) depending on the number of genes one is interested in. Values of *j* can equal 1, …, J1 where J1 is the sample size or number of repetitions under one condition, for example, the control (normal) condition, and *j* = J1+1, . . ., J2 is the sample size or number of repetitions under a different experimental (tumor) condition. A general statistical function is

$$Y_{ij} = a_i + b_i x_j + \varepsilon_{ij}$$

where $x_j = 0$ for array *j* where $1 < j < J1$ from the normal group, and $x_n=1$ for array *j* from the tumor group. a_i and $(a_i + b_i)$ are therefore the mean expression levels of gene *i* under the control and experimental conditions, respectively. To determine if a particular gene is differentially expressed, one must test the null hypothesis:

H_0: $a_i + b_i = a_i$ (or $b_i = 0$) against the alternative H_1: $b_i \neq 0$.

There are several versions of the t-test, depending on whether the sample size is large and whether it is reasonable to assume that the gene expression levels have an equal variance under the two conditions. To test whether gene *i* is differentially expressed under the two conditions, we can take the sample means Y_{i1} and Y_{i2} where

$$\bar{Y}_{i(1)} = \frac{\sum_{j=1}^{J1} Y_{ij}}{J1}, \qquad \bar{Y}_{i(2)} = \frac{\sum_{j=J1+1}^{J1+J2} Y_{ij}}{J2} \qquad \textbf{Eq. (1)}$$

and their corresponding variances s^2_{i1} and s^2_{i2}. in order to get the t-statistic:

$$t_i = \frac{\bar{Y}_{i(1)} - \bar{Y}_{i(2)}}{\sqrt{s^2_{i(1)}/J1 + s^2_{i(2)}/J2}} \qquad \textbf{Eq. (2)}$$

The resulting t-statistic can be used to determine which genes are significantly differentially expressed given a particular p-value. The p-value is usually calculated based on the distribution of the test statistic under the null hypothesis (also referred to as the null distribution of the test statistic) which may be specified or estimated via different modeling assumptions. Under the normality assumption for Y_{ij}, t_i approximately has a t-distribution with degrees of freedom d_i = J1 + J2 – 2 under a standard t-test. A problem with the standard t-test is that it assumes that gene expression levels have equal variances under the two conditions (eg, tumor and normal cells). Because the sample sizes J1 and J2 are often small in microarray experiments, there is evidence to support unequal variances, making the standard t-test not an ideal method (11). A simple t-test specifically designed to handle the possibility of having unequal variances is the Welch t-test. The Welch t-test, like the standard t-test, also requires the assumption that Y_{ij} is normally distributed but allows for unequal variances under the two conditions. Under this assumption, the distribution of t_i can be approximated with degrees of freedom d_i equal to

$$d_i = \frac{(s^2_{i(1)}/J1 + s^2_{i(2)}/J2)^2}{(s^2_{i(1)}/J1)^2/(J1-1) + (s^2_{i(2)}/J2)^2/(J2-1)}. \qquad \textbf{Eq. (3)}$$

When t exceeds a certain threshold depending on the confidence level selected, the two populations are considered to be different. The Welch t-test has been found to have a relatively good performance compared with other alternative t-tests, such as the standard t-test. (12). A problem with the standard- and Welch t-test is that they often have low power because of the small sample size. In addition, the variances estimated from each gene are not stable; for example, if the estimated variance for one gene is small, the t value can, simply by chance, be large even when the corresponding fold change is small. The fundamental problems of the t-statistic as defined are the normality assumptions imposed a priori and that it is subject to large fluctuations given small changes in the error variance SE_i, the square root of which gives the denominator of the t-statistic as shown in Eq. 1 (13).

Variations of the t-test

Modifications of the t-test have been proposed to address the difficulty in estimating the error variance which is subject to erratic fluctuations in various sample sizes. For simplicity, in this section, I will assume that we have a series of *n* replicate arrays and are interested in knowing whether a particular gene is differentially expressed. We let M be the log differential expression ratio $\log_2$ R/G for each spot and $\bar{M}$ be the mean log ratio of the expression levels of the gene in question. A simplified version of Eq. 2 of the t-statistic is therefore

$$t = \frac{\bar{M}}{s/\sqrt{n}} \qquad \textbf{Eq. (4)}$$

where s is the standard deviation of the log differential expression ratio across the replicates for a particular gene.

Lonnstedt and Speed (14) adopted a parametric empirical Bayes approach whereby they produced a B-statistic that serves as an estimate of the log posterior odd

ratio of differential expression versus non-differential expression. The B-statistic could be seen as a variant of a penalized t-statistic where

$$t = \frac{\bar{M}}{\sqrt{(a + s^2)/n}} \qquad \textbf{Eq. (5)}$$

with *a* equals the penalty estimated from the mean and standard deviation of the sample variances s^2. The B-statistic is essentially the logarithm of a ratio of probabilities where the numerator is the probability that the gene is differentially expressed and the denominator is the probability that the gene is not differentially expressed. Both probabilities are estimated using the entire data and are called posterior probabilities, thus the reference that the B-statistic is a logarithm of the posterior odds of differential expression. An advantage of the B-statistic is that it allows for gene-specific variances while also combining information across many genes, making it a more stable estimate than the ordinary t-statistic. However, a shortcoming of the B-statistic is that it is subject to the validity of several parametric assumptions regarding the data.

Tusher, Tibshirani, and Chu (7) have proposed the significance analysis of microarrays' (SAM) version of the t-tests which uses penalized t-statistics of the form

$$t = \frac{\bar{M}}{(a + s)/\sqrt{n}} \qquad \textbf{Eq. (6)}$$

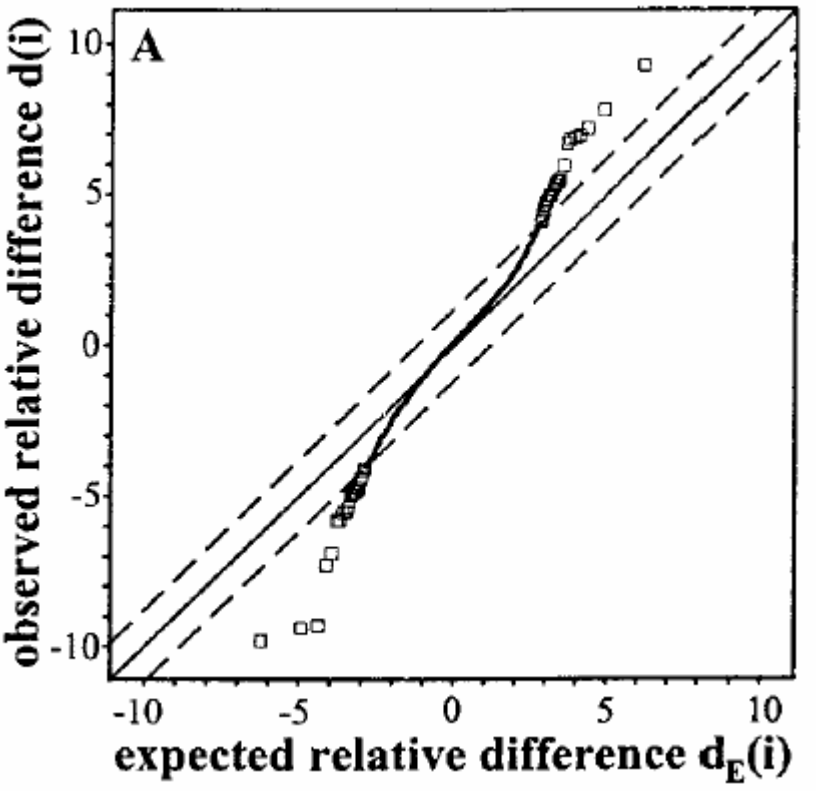
where the penalty *a* is a small positive constant. Tusher *et al*.'s model used an *a* that minimizes the coefficient of variation of the absolute t-values

The SAM version of the t-test differs from the one proposed by Lonnstedt and Speed in that the penalty is applied to the sample standard deviation *s*, rather than to the sample variance s^2. SAM identifies genes with statistically significant changes in expression by assimilating information from a set of gene-specific t-tests (7). Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. With this modification, genes with small fold changes will not be selected as significant. To "increase" the sample sizes, Tusher *et al.* computed relative differences from permutations of each of their hybridizations. They were then able to calculate the expected relative difference $d_{\varepsilon i}$ which is defined as the sum of the largest relative differences of each permutation over the total number of permutations. Essentially, the expected relative difference equals to the average relative difference over all the permutations. Potentially significant genes with differential expression were identified by plotting the observed relative difference of the gene d_i vs the expected relative difference $d_{\varepsilon i}$. The resulting "SAM plot" will show that most genes are located along the line $d_i = d_{\varepsilon i}$. Genes displaced from the line $d_i = d_{\varepsilon i}$ by a distance greater than some threshold value Δ were then considered 'significant'. They initially set Δ=1.2 as their baseline estimate, but had also calculated the difference between the number of genes assessed as significant and the number of false positives for decreasing Δ's. Essentially, SAM derives the P value for each gene from permutations of the available experimental data

To identify the false discovery rate, horizontal cutoffs $\text{cut}_{up}(\Delta)$ and $\text{cut}_{low}(\Delta)$ were defined such that $\text{cut}_{up}(\Delta)$ is the smallest d_i among genes that are significantly induced and $\text{cut}_{low}(\Delta)$ is the least negative d_i among genes that are significantly

repressed. The number of falsely significant genes corresponding to each permutation was computed by counting the number of genes that exceeded the horizontal cutoffs for induced and repressed genes. By defining the cutoffs according to the information presented by the data itself, SAM allows for asymmetric cutoffs for induced and repressed genes. Such flexibility offers advantages not available in standard t-tests which impose symmetric horizontal cutoffs for both induced and repressed genes, disregarding the possibility that induced and repressed genes may behave differently under certain experimental conditions.

An example of a SAM plot is shown in Figure 1 taken from Tusher et al.'s 2001 article (7). The figure depicts the scatter plot of the observed relative difference di versus the expected relative difference $d_{\varepsilon i.}$ The solid line represents the group of genes where $di = d_{\varepsilon i.,}$ and the dotted lines are drawn at a distance Δ from the solid line. Points indicated by squares in the plot represent potentially significant genes.



Regression Modeling Approach

Thomas, Olson, Tapscott, and Zhao (16) proposed a regression modeling approach where the constants a_i and b_i from Eq. 1

$$Y_{ij} = a_i + b_i x_n + \varepsilon_{ij}$$

are derived using a weighted least squares method, and the variance of b_i was estimated using the robust variance estimator. Their test statistic is

$$t_i = \hat{b}_i \Big/ \sqrt{\mathrm{Var}(\hat{b}_j)}, \qquad \textbf{Eq. (7)}$$

$$\hat{b}_i = \bar{Y}_{i(1)} - \bar{Y}_{i(2)} \quad \text{and} \quad \mathrm{Var}(\hat{b}_i) = \frac{s^2_{i(1)}}{J_1}\frac{J_1 - 1}{J_1} + \frac{s^2_{i(2)}}{J_2}\frac{J_2 - 1}{J_2} \qquad \textbf{Eq. (8)}$$

The t_i statistic from Thomas' model has a similar form to the ordinary t-statistic with the difference being in how the variances are estimated. While the ordinary t-statistic uses the unbiased sample variances, the maximum likelihood estimator of the variance under an assumption of normality for Y is used in Thomas' model. An advantage of Thomas' model is that it works even if the random errors ε_{ij} have different variances for different genes *j*, or even different variances for the same gene *j* under the two conditions. A shortcoming of Thomas' model is that in order to use the maximum likelihood variance estimator, one needs to have large sample sizes J_1 and J_2 in order to achieve the same power as an ordinary least squares estimate. This model is also unlikely to work well in experiments with small sample sizes as small sizes are likely to result in biased estimators.

Nonparametric tests

The Wilcoxon rank sum and Kruskall-Wallis tests have also been used as alternatives to the t-test in two-sample comparisons of microarray data (13). Because they are nonparametric, they avoid the possibly questionable parametric assumptions used in the t-test. However, nonparametric tests have the disadvantage of requiring that the two samples have distribution functions with the same shape, with the only difference being their location parameters. In addition, the use of nonparametric tests comes at the price of a lowered statistical power. Thomas et al (16) demonstrated that the application of the Wilcoxon test to Golub's (17) leukemia data resulted in the absence of any findings of significant differentially expressed genes. In general, if samples are normally distributed and a nonparametric test is used, one would need larger sample sizes relative to the size required in parametric test in order to detect statistical significance. For these reasons, nonparametric tests are not ideal unless substantial nonnormality is believed to exist.

A Bayesian Framework

The Bayesian approach uses the Bayesian theorem

$$\mathrm{P(Hypothesis|Data)} = \frac{\mathrm{P(D|H)P(H)}}{\mathrm{P(D)}} \qquad \textbf{Eq. (9)}$$

which enables the propagation of information conditioned under various background information or assumptions. In particular, the theorem allows the microarray analyst to compute the posterior probability of any hypothesis or model H in light of the probability P(D|H) [data likelihood] and P(H) [the prior probability of the hypothesis H] given any background information the scientist may have available (13).

One of the pervading problems of differential analysis is the difficulty in obtaining accurate estimates of the standard deviation of individual genes based on only a few measurements. To supplement the weak empirical estimates of single-gene variances

across a small number of replicates, a more robust estimate of variance is obtained by pooling genes with similar expression levels. The Bayesian approach incorporates the observation that a reciprocal relationship exists between variance and gene expression levels, and that genes expressed in similar levels exhibit similar variance.

Baldi and Long's regularized t-test

Baldi and Long (18) developed a Bayesian statistical framework that regularizes the t-test in order to account for small sample sizes/ number of replications. The regularized t test combines information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator for a gene-specific t test. The regularized t-statistic has the form:

$$t = \frac{\overline{M}}{\sqrt{\frac{v_0 \cdot s_0^2 + (n-1) \cdot s^2}{v_0 + n - 2}}} \qquad \textbf{Eq. (10)}$$

where v_0 is a tunable parameter that determines the relative contributions of gene-specific and global variances and n is the number of replicate measurements for each condition. They first assumed that the expression-level measurements of a gene in a given situation have a *roughly* Gaussian distribution and that each observation is independent of the others. Other distributions could also be used and still retain the general Bayesian framework that they proposed. They then calculate the likelihood of the data D conditioned on the background information (based on their distribution assumption) that the mean and standard deviations of the expression-level of the gene follow a normal distribution. Because a Bayesian approach requires the introduction of a prior distribution $P(\mu, \sigma^2)$, Baldi and Long assumed that the prior and the posterior have the same function form, and therefore used a conjugate prior α for their model. Use of the conjugate prior allowed them to apply the Bayes theorem to get the posterior distribution. An advantage of using a conjugate prior is that it is convenient and it allows for the possibility that μ, σ^2 are not independent. From the above assumptions, they were able to derive the distribution of the posterior, $P(\mu, \sigma^2|D, \alpha)$, which combines all the information from the prior and the data D. Using this distribution, they were able to find the mean, degrees of freedom, and sum of squares of the posterior, all of whose formulas are listed in their paper. In this discussion, I will focus only on how the posterior mean μ_n was derived and how it can be used in assessing whether a gene is differentially expressed.

Before they are able to derive the mean μ_n of the posterior, Baldi and Long had to specify a prior mean μ_0, which they assumed to be equal to the sample mean, m. The posterior mean was then found to be the convex weighted average of the sample mean and the prior mean. The resulting posterior distribution $P(\mu, \sigma^2|D, \alpha)$, contains all the relevant information about all possible values of μ and σ^2. One can then get single point estimates of μ and σ^2 of the control and treatment group and derive various information such as computing for a more robust estimation of the variance or to find the probability $P(\mu_c = \mu_t|D, \alpha_t, \alpha_c)$. The regularized t-test approach has been implemented in a web-based program called CyberT . The default point estimate used in CyberT is the mean of the posterior estimate derived from the posterior distribution where:

$$\mu = m \quad \text{and} \quad \sigma^2 = \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2} \qquad \textbf{Eq. (11)}$$

The need for regularization of the variance used in the t-test is therefore achieved through the assumption of the conjugate prior. This allows for an estimate of the empirical variance, which is modulated by the tunable parameter v_0 that takes into account information obtained from the background variance.

Efron et al.'s Empirical Bayesian Approach

Efron, Tibshirani, Storey, and Tusher (15) developed an EB approach that starts with the assumption that a gene $gene_i$ is either differentially expressed or not differentially expressed. Let Y be the derived test statistic, where Y can be either a standard t-statistic, a Wilcoxon rank-sum estimate, d_i which is a t-statistic with an added constant *a* as described in the section on t-test modifications, or any other summary test statistic.

The following standard notations are used:
p_0 is the prior probability that a gene is not differentially expressed .
$f_o(y)$ is the prior density function of Y_i if $gene_i$ is not differentially expressed.
$1 - p_0$ or p_1 is the prior probability that a gene is differentially expressed.
f_1 is the prior density function of Y_i for a differentially expressed $gene_i$. One can then derive the mixture density function f(y) where:

$$f(y) = p_0\text{-}f_0(y) + p_1 f_1(y) \qquad \textbf{Eq. (12)}$$

A direct application of Bayes' theorem yields the posterior probabilities:

$$p_0(y) = [\, p_0 f_0(y) \,] / f(y) \qquad \textbf{Eq. (13)}$$

which is the probability that $gene_i$ is not differentially expressed given Y_i=y and

$$p_1(y) = 1 - [p_0 f_0(y) \,] / f(y) \qquad \textbf{Eq. (14)}$$

which is the probability that $gene_i$ is differentially expressed given Y_i=y.

Unlike Baldi and Long who specified the prior distributions required by a full Bayesian analysis, Efron et al. derived the posterior probabilities $p_0(y)$ and $p_1(y)$ empirically by using the massively parallel structure of microarray data. In their initial analysis, they used data from Hedenfalk's 2001 microarray experiment which compares the gene activity differences in BRCA1 tumors versus BRCA2 tumors. In this experiment, tumors from 22 women were analyzed, 7 with BRCA1 mutation, 8 with BRCA2 mutation, and 7 that had neither of the two mutations. Gene expression levels of 3,226 genes were analyzed. They then derived the Wilcoxon rank sum statistic Y_i for each $gene_i$ and plotted the relative expression differences of each $gene_i$ for BRCA2 versus BRCA1 tumors. This plot allowed them to estimate $f_0(y)$ as well as derive the distribution f(y) empirically by using a Poisson regression fit to the Y values. Details of the derivations can be found in Efron *et al*.'s article (15).

Using the most conservative estimate of the prior null probability p_0=1 (which minimizes the probability of assigning a gene to be differentially expressed), Efron *et al.* were able to show that

$$p_1(y) = 1 - [p_0 f_0(y) \,] / f(y) \geq 0.90 \qquad \textbf{Eq. (15)}$$

Using Efron *et al.*'s approach, the analyst can derive the posterior probability of activity differences for each gene without having to run 3,226 separate t- or Wilcoxon-tests to identify which of them can be confidently labeled as differentially expressed.

Mixed Modeling Approaches

Artifacts and flaws in experimental design, such as channel-specific variabilities and confounding of treatment effect with dyes have led to the rise of mixed models to account for limitations in microarray technologies and their applications (19). Their general form allows for the testing of effects of any identifiable source of variability. Unfortunately, this strength of a mixed model approach is also its weakness: the mixed-model approach has been criticized on the grounds that it attempts to estimate too many parameters from sparse data. However, certain adaptations that can be applied to minimize potential problems of a mixed-model approach have been described.

MDSS Algorithm

The Maximum Difference Subset (MDSS) algorithm developed by Weiler, Patel, and Bhattacharya (19), combines cluster analysis, classical statistical tests and machine learning in a way that incorporates classification accuracy into the criterion for finding differentially expressed genes. The MDSS approach consists of the following steps:

1) Perform a classical statistical test (e.g., a t-test) for each gene in the two sample groups (e.g., normal vs. tumor).
2) Rank each gene in descending order according to the magnitude of the measure (e.g. t-statistic for a t-test) and find the largest threshold value of the measure that succeeds in discriminating between the two groups. For example, find the largest significance level that succeeds in discriminating between the tumor and normal group. This gene set is called the 'initial MDSS'.
3) Remove individual samples from the total set and store a list of genes that are significant beyond the threshold value. Continue to remove individual samples and create new lists of genes with significance values greater than the threshold value. Each new list created after the removal of one sample results in an additional individual MDSS list.
4) Identify the genes that are common to all the individual MDSSs. This gene set is assumed to comprise of genes that are differentially expressed between the two sample groups and is called the 'overall MDSS'. This set also passes the critierion of predictive utility and may also be used for class classification and prediction.
5) Use a clustering algorithm to verify if the overall MDSS returns the correct classification. Adjust the threshold value in step 2 if the set fails to pass the test and do the test again.

An advantage of the MDSS algorithm is that it learns first at which statistical threshold a particular gene set may have, eliminating the arbitrariness associated with setting a threshold of statistical significance, say, of alpha = 0.05. The MDSS approach also minimizes the effect of the normality assumption inherent in t-tests.

Conclusion

The statistics literature on microarray data analysis is quite recent and the search for more powerful statistical methods remains an area of active research. This paper reviewed some of the statistical methods that have been proposed to address the problems that are quite unique to the nature of microarray experimentation and data analysis. The advent of microarray technologies brought with it the power to retrieve large amounts of information at a relatively short period of time. It also brought forth the need to find reliable methods that could sift through the enormous amount of information retrieved to

gain a better understanding of the data and the problem at hand. The methods discussed in this paper could be seen as initial attempts that address the problem of multiple testing. The methods allow the analyst to assign statistical significance to differentially expressed genes without requiring that the restrictive assumptions of standard statistical methods be imposed. Ultimately however, statistical methods like the ones discussed that focus on the analysis of microarray data alone will most likely be insufficient. Methods that allow for the integration of microarray data with other sources of information will very likely need to be developed. For example, information from other clinical, patient, and experimental records could be combined with data from microarray experiments. Continued efforts to increase the power and reliability of statistical methods as well as to integrate knowledge from different sources will be necessary to harness the full potential of microarray technologies.

References:

(1) Kooperberg C, Sipione S, LeBlanc M, Strand A, Cattaneo E, and Olson J. (2002) Evaluating tests statistics to select interesting genes in microarray experiments. *Human Molecular Genetics*, **11(19),** 2223-2232.

(2) Quackenbush J. (2002) Microarray data normalization and Transformation. *Nature Genetics Supplement* **(32)**,496-501.

(3) Chen Y, Kamat V, Dougherty E, Bittner M, Meltzer P, and Trent J.(2002) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*, **18(9),** 1207-1215.

(4) Slonim, DK (2002). From patterns to pathways:gene expression data analysis comes of age. *Nature Genetics Supplement* **(32)** , 502-508.

(5) Li, C and Hung Wong, W. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*. **2**, research0032.

(6) Chen Y, Dougherty, ER and Bittner, M (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics*, **2**, 364-374.

(7) Tusher, V.G., Tibshirani, R. and Chu, G.(2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–512.

(8) Westfall, P.H. & Young, S.S. *Resampling-Based Multiple Testing*, 340 (John Wiley Sons, New York, 1993).

(9) Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical andpowerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289–300.

(10) Storey, J.D. (2002a). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B* **64**: 479-498.

(11) Newton MA, Kendziorshi CM, Richmond CS, Blattner FR, and Tsui, KW (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol*.,**8**, 37-52.

(12) Best, DJ and Rayner JC (1987) Welch's approximate solution for the Belurens-Fisher problem. *Technometrics* **29**, 205-210.

(13) Dudoit S, Yang YH, Callow MJ, and Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111-139.
(14) Lönnstedt, I. and Speed, T. P. (2002) Replicated microarray data. *Statistica Sinica* **12**, 31-46.
(15) Efron B., Tibshirani, R., Storey J. D., and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96, 1151-1160.
(16) Thomas JG, Olson JM, Tapscott SJ, and Zhao LP (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res*., **11**, 1227-1236.
(17) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.*Science*, **285**, 531-537.
(18) Baldi P, Long A (2001) A Bayesian framework for the analysis of gene expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17(6),** 509-519.
(19) Weiler JL, Patel S, Bhattacharya S. (2003) A Classification-Based Machine Learning Approach for the Analysis of Genome-Wide Expression Data. *Genome Research* **13**:503–512