REVIEW

Key aspects of analyzing microarray gene-expression data

James J Chen
US FDA,
Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, Jefferson, AR 72079, USA
Tel.: +1 870 543 7007;
Fax: +1 870 543 7662;
E-mail: jamesj.chen@fda.hhs.gov

One major challenge with the use of microarray technology is the analysis of massive amounts of gene-expression data for various applications. This review addresses the key aspects of the microarray gene-expression data analysis for the two most common objectives: class comparison and class prediction. Class comparison mainly aims to select which genes are differentially expressed across experimental conditions. Gene selection is separated into two steps: gene ranking and assigning a significance level. Class prediction uses expression profiling analysis to develop a prediction model for patient selection, diagnostic prediction or prognostic classification. Development of a prediction model involves two components: model building and performance assessment. It also describes two additional data analysis methods: gene-class testing and multiple ordering criteria.

The DNA microarray technology has been used increasingly in disease diagnosis, studying biological functions, identifying biomarkers and predicting clinical outcomes. One major challenge with the use of the microarray technology is the analysis of a massive amount of data with various sources of variability. A wide range of gene-expression data analysis methods have been proposed for various applications, ranging from the simple fold-change for identifying differentially expressed genes, to complex computational algorithms for tumor classification. Recently, the MicroArray Quality Control (MAQC) consortium suggested: 'Fold-change ranking plus a nonstringent p-value cutoff can be used as a baseline practice for generating more reproducible signature gene lists' [1]. Many researchers have questioned this approach [2,3]. This report gives an overview of the microarray gene-expression data analysis and discusses several key aspects of a successful data analysis. It will focus on the two common goals, class comparison and class prediction, and describe various data analysis methods, including gene selection, multiple testing, classification/prediction, gene-class testing (GCT) and multiple ordering criteria.

Microarray gene-expression data

Microarray is a device that measures transcription levels of hundreds or thousands of messenger (m)RNAs within a biological sample. Gene-expression levels are quantified from the image excited by a laser scanner. Signal intensities reflect the amount of transcript present for the gene in the mRNA sample. A microarray study is generally a comparative experiment, in which the relative expression levels are compared among the samples rather than the determination of absolute intensity measures of each sample. In two-color array experiments, the underlying principle is a competitive hybridization between two samples. The measured intensities reflect relative abundances of the two samples. In one-color experiments, the intensity is an absolute measure of gene expression; however, inferences are made regarding the expression levels for a gene in different samples but not regarding the level of expression of one gene in relation to other genes.

Experimental unit

In most microarray experiments, the experimental unit refers to an independent biological RNA source (sample) to which the treatments are applied [4,5]. The experimental unit may be represented by a tissue sample or a sample of cells from a cell culture. The array is not the experimental unit. However, in many studies, an array and an experimental unit have a one-to-one correspondence, so the array is not distinguishable from the experimental in such studies. Analysis of a microarray study involves an assessment of variation of biological samples among the experimental units.

A microarray experiment is a multistep process and each step is a potential source of variation. The sources of variation associated with the assay of mRNAs in each experimental unit are collectively referred to as technical variation. The variation from different RNA sources is referred to as biological variation. Replication is a key to the reliability of the data and accuracy of data

Keywords: class comparison, class prediction, cross validation, gene class testing, gene selection, multiple selection criteria, multiple testing, significance analysis

future medicine part of fsg

analysis. In parallel, there are two types of replication: technical replication and biological replication. Technical replication refers to replication in which the mRNA is from the same pool. Technical replication is used to minimize technical artifacts. In microarray quality control experiments, technical replication may be used to assess laboratory or platform reproducibility. The biological variation in measured gene expression comes from variation among the experimental units. It reflects the variability among the different biological samples used in the experiment. Different biological samples represent independent biological replicates to reflect the variability in the population of interest. The common goal of microarray studies is to make inference between different populations. Biological replication is used to allow the generalization of experimental results from sample to population. Statistical tests should be based on the biological variance and the sample size refers to the number of biological samples (experimental units).

Preprocessing

There are inherent characteristics of measured raw intensity data that can affect the data analysis. After data collection, many factors in generating intensity measurements need to be considered prior to data analysis. There are many experimental variables, such as differences in labeling, hybridization and detection. Intensity measurements should be adjusted to minimize systematic biases. This adjustment is referred to as normalization. Many normalization methods have been proposed [6–9]. Proper normalization strategies depend on the experimental design and the data collection process. In addition, the intensities contain background contribution that can be addressed in diverse ways. There are low signal intensities or inconsistent spots across arrays. Finally, some arrays may have multiple probes that measure the same gene; the intensities from these probes can be combined to generate a single expression level for the gene. Data preprocessing is the first step of data analysis. There are many options and methods for data filtering, local and regional backgrounds, multiple probes and normalization and transformation. Microarray platform manufacturers also have recommended data preprocessing protocols. Preprocessing attempts to correct technical biases; an overcorrection by introducing greater biases can result in different study conclusions [10,11]. Finally, data quality is a prerequisite for valid analysis and conclusions; preprocessing cannot salvage poor quality data.

Gene selection

A microarray experiment is conducted to study the changes in gene expressions under different experimental conditions of interest. The main objectives of most microarray studies can be classified broadly into the three applications: class comparison, class prediction and class discovery. Class comparison aims at selecting over- or under-expressed genes by comparing expression profiles between different experimental samples (e.g., with or without exposure to a specific drug or toxic compound). Class prediction aims to predict the class membership of a new sample from a gene-expression prediction function. For example, personalized medicine may apply classification algorithms to predict patient response to therapy for treatment assignment. Class discovery usually refers to identifying previously unknown sample subtypes or refining an existing sample class from the study of gene-expression profiles. In early microarray studies, class discovery used clustering methods, such as hierarchical clustering or k-means, to group genes or samples with similar expression patterns [12,13]. In clinical applications, the main objective is to develop a prediction model. For prediction purposes, the clustering analysis is not efficient as it does not use the class membership information.

A fundamental step in microarray data analysis is to select a subset of genes from the original gene set such that their expression levels are related to the experimental conditions, that is, to select a set of genes that shows differential expressions under different conditions. Gene selection can be separated into two steps. The first step is to calculate a discriminatory score that will rank the genes in order of evidence of differential expressions. The second step is to determine a threshold cutoff from the ranked scores to select the differentially expressed genes.

Gene ranking

The most common microarray experiment is to study changes in gene expression between two biological samples (treatments or tissues). Following image analysis, the intensity data are preprocessed and properly normalized. The data are typically transformed in $\log_2$ scale. Differential expression can be evaluated by comparing the difference between the mean expressions of the two samples. The difference, denoted by M, represents the fold-change between two samples. Owing to a wide range of magnitudes and variability among different genes in the array, the fold-changes cmputed for different genes are not directly

comparable as measures of evidence of differential expressions. The M statistic can be standardized by dividing scale parameters (e.g., standard deviation), denoted by T = M/s. The T-statistic has the general expression that represents the test statistics for two sample comparisons (Box 1). The T-statistic includes the fold-change M statistic as a special case by setting s = 1; and T becomes student's T-statistic when s is the sample standard deviation of the mean difference. Various T-statistic variants have been proposed in microarray data analysis [14–24]. Tusher and colleagues [16] proposed the significance analysis of microarrays (SAM) statistic, $M/(s_0+s)$, by adding a penalty s_0 to the sample standard deviation in the denominator to account for a very small standard deviation that results in a large T-value, where s_0 is computed from the data [101]. When sample size is small, the variance estimate s^2 may be imprecise; a number of shrinkage estimators have been proposed to improve reliability of variance estimates [18–23]. A shrinkage variance estimator is a weighted average of the within-gene variance and a pooled variance from all genes. These methods all seem to work well; they are particularly useful when the sample size is small. The T statistic is a measure for gene ranking but it does not provide a criterion to determine a cutoff.

Statistical significance testing is designed to provide a cutoff by computing the p-value from a statistic. The p-value is the probability of an outcome as extreme as or more extreme than the observed outcome, given no difference (null hypothesis is true). A small p-value indicates evidence of differential expressions, either over- or under-expression. Thus, the genes can be ranked according to the p-values. The p-value of a test statistic is computed either from the distribution of the test statistic or from re-sampling methods (Box 2). The T-statistic and the corresponding p-value may give slightly different rankings when the p-values are computed from the re-sampling method. In general, the p-value rankings computed from the T-statistic or various T-variants are similar, but the p-value ranking and the fold-change rank from the M-statistic are somewhat different [25].

Comparisons among several treatments or factorial designs are analyzed in linear model framework. Kerr and colleagues [26] proposed a single analysis of variance (ANOVA) model for an entire microarray experiment. Jin and colleagues [27] fitted separate ANOVA models for each gene. Tsai and colleagues [10] proposed a generalized ANOVA model and performed the analysis for each gene.

Box 1. Statistics for comparing two samples.

In a two sample comparison, most statistics used in microarray data analysis have the expression of the T-statistic, T = M/s, where M is mean difference (fold-change) and s is an estimate of standard deviation of M.

Setting s = 1, the T-statistic becomes the fold-change M-statistic. The fold-change was used in earlier years as a criterion to select differentially expressed genes because of very few replications. The use of fold-change for gene ranking is deficient in some aspects as it does not account for the variability of the expression levels among genes. It assumes all genes in the array have the same variance. When the number of replicates is small, genes with larger variances have a good chance of exhibiting larger fold-changes, even if they are not differentially expressed.

Setting s as sample standard deviation of M, T becomes the known student's T-statistic. Different sample standard deviation estimates have been used, such as $s_A + s_B$ [14] and $(s_A^2 + s_B^2)^{1/2}$ [15] by assuming the variances for the two samples are not necessarily equal. The T-statistic is a scaled M-statistic. The T-statistic across genes is comparable. However, because of a wide range of variations, genes with a very small standard deviation can have a good chance of giving a large T-value. This often happens to the nonexpression or very low expression genes and it can be alleviated by a filtering in data preprocessing.

Many variance function methods have been developed to improve the denominator of the T-statistic. Tusher and colleagues [16] adjusted the denominator by adding a penalty s_0 to the sample standard deviation, $s+s_0$, to account for a very small standard deviation. Tusher and colleagues [16] estimated s_0 by minimizing a coefficient of variation while Efron and colleagues [17] used a percentile of the distribution of sample standard deviations. Baldi and Long [18] replaced the variance estimate with a Bayesian estimator based on a hierarchical prior distribution. Lonnstedt and Speed [19] proposed an empirical Bayes approach that combines information across genes. Other variance function methods have also been proposed using a similar strategy of borrowing information across genes [20–23]. The adjusted variance for the denominator has the general expression: $(ws_1^2 + (1-w)s^2)^{1/2}$ where s_1^2 is a variance of M based on all genes in the arrays and $0 \le w \le 1$ is a weight depending on the sample size. This estimator, known as shrinkage estimator, is a weighted combination of gene-specific variance and variance from all genes. This estimator shrinks the individual variance toward their variance. The T-statistic has the weight w = 0. When the sample size is large, the T-statistic generally works well. However, when sample size is small, the estimate s can be imprecise, resulting in an unreliable test.

The T-statistic is a signal:noise ratio. In general, the mean difference M in the numerator can be replaced by a location parameter, such as median difference, and standard deviation s in the denominator is replaced by a scale parameter estimate, such as a shrinkage standard deviation.

Assigning a significance level

A p-value cutoff divides the genes into two sets. Ideally, the selected set would contain the differentially expressed genes and the nonselected set would contain the nondifferentially expressed genes. However, the selected gene set will have false positives; likewise, the nonselected genes will have false negatives. Owing to the variation of the biological data, it is not possible to have an optimal cutoff that simultaneously minimizes both false-positive and false-negative errors. The tradeoff between the two errors depends on the application. In class comparison, procedures that allow very few false positives may be appropriate when a small number of genes are selected to be validated by a follow-up confirmation. However, in the class prediction or class discovery setting, where the intent is to develop genomic profiles or classifiers, the omission of informative genes would have a much more serious consequence than the inclusion of noninformative genes. In such cases, procedures with fewer false negatives may be more desirable.

The false-positive probability or p-value is computed under testing a single gene. Since hundreds or thousand of genes are tested, determining a cutoff should be in terms of an overall false-positive error. The familywise error measure is commonly used in testing multiple end points when the number of tests is small. With a large number of genes involved in the comparisons, the false-discovery rate (FDR) error measure [28] is a more useful approach for determining a significance cutoff (Box 3). FDR is the probability of false selections among those selected genes. This approach allows the findings to be made, provided that the investigator is willing to accept a small fraction of false-positive findings. The FDR approach can be used in two different ways, either controlling FDR [28] or estimating FDR error [29,30]. For the desired FDR level, Benjamini and Hochberg [28] proposed a procedure to determine the cutoff so that FDR is controlled on average. On the other hand, Tsai and colleagues [30] proposed a procedure to estimate the conditional FDR for the desired selected number of genes.

The FDR approach emphasizes the false-positive error in determining the cutoff. As a result of small sample sizes, the FDR approach can result in a short significant list and a large false-negative error. Delongchamp and colleagues proposed a receiver-operating characteristic (ROC) approach to determine an optimal cutoff based on minimizing the total cost from making false-positive and false-negative errors [31]. In the ROC approach, the investigator is required to specify the ratio of the cost for making a false-positive error over the cost for making a false-negative error and an estimate of the number of nondifferentially expressed genes. Chen and colleagues illustrated an application of the ROC approach to determining a p-value cutoff [25]. The ROC approach is designed to reliably eliminate most of the undifferentially expressed genes from further consideration while essentially keeping all genes whose functions are potentially different in the biology under study.

Box 2. Computing p-value.

The p-value of a test statistic can be computed by two approaches: mathematical derivation or resampling method. In the mathematical derivation, the distribution or its approximation of a test statistic is theoretically derived. The p-value probability can be computed either by a numerical integration or by statistical computation methods. For example, the Student's t-distribution is mathematically derived under the assumption that the normalized intensity data are normally distributed. The probability of a T-distribution can be computed numerically. The shrinkage T-statistic, such as the random variance T-statistic [20] or the moderated T-statistic [22], has a T-distribution; their distributions were derived by the authors under the normality assumption. However, the distributions of many T-statistics do not have a closed mathematical expression; their probabilities cannot be computed directly.

The resampling method is a general approach to calculating p-values of any test statistics. Using a resampling method without replacement to compute p-value is referred to as the permutation test. The p-value of a permutation test is the cumulative sum of the probability of the observed outcome and the probability of all more extreme outcomes. The permutation test is based on the observed experimental outcomes. The p-value has the interpretation as the probability for the experimental outcome as observed or more extreme, under the experimental conditions. The permutation test does not require any assumption for the underlying normalized intensity distribution; it has been shown to be more powerful than the parametric approach when the sample size is at least five [56]. When the sample size is small, the number of possible permutations is limited. In this case, the permutation test becomes infeasible. For small sample size experiments, the shrinkage statistic designed [20–23] by borrowing information is preferable.

Class comparison

Class comparison involves comparing expression profiles from different exposure conditions or different tissue types. The main interests are in identifying genes whose expression levels are altered by treatments and/or to establish expression profiles between classes. In this application, gene selection often targets a limited number of candidate differentially expressed genes for confirmation and further study. The FDR approach is appropriate. The differentially expressed genes so selected are referred to as statistically significant. However, if an observed change is small with a very small standard deviation, then the

Box 3. Multiple testing.

The significance level of a p-value is defined under a single gene test. Since hundreds or thousands of p-values are calculated, simply using the p-value to determine a cutoff without adjusting for the multiplicity testing effect will increase the chance of false positives. For instance, in a study of 10,000 genes, there could be 100 apparent 'significant' expression changes found using a p-value cutoff of 0.01, when in truth none are differentially expressed. The family-wise error rate (FWE) and false-discovery rate (FDR) are two commonly used false-positive error measures in the analysis of multiple end points (genes).

The FWE approach is commonly used in testing multiple clinical points. An FWE-controlled procedure guarantees that the probability of one or more false positives is not greater than a predetermined level, regardless of how many genes are tested. In large-scale microarray experiments, the FWE approach could present a problem, since this analysis tends to screen out all but very few genes that show extreme differential expressions.

The FDR error measure considers the expected proportion of false-positives among the selected genes [22]. Essentially, FDR considers the probability of false selections among those selected genes. The FDR approach allows the investigator to select the potential differentially expressed genes, while accepting a small fraction of false findings. In the FDR approach, the genes are ranked according to the p-values: $p_{(1)} \leq \ldots \leq p_{(r)} \leq \ldots \leq p_{(m)}$.The notation $p_{(i)}$ represents the i-th ordered p-value and m is the number of genes in the analysis. For example, $p_{(1)}$ is the smallest and $p_{(2)}$ is the second smallest. For the desired FDR level, a FDR-controlled procedure [28] is to find the largest $p_{(r)}$, such that $(m * p_{(r)}/r) \leq FDR$. Those r genes with p-values less than or equal to $p_{(r)}$ are selected as differentially expressed genes. The FDR approach can be used differently. For the desired selecting number of genes, Tsai and colleagues [29] proposed a procedure to estimate the conditional FDR. In either use, if r is the number of genes selected, then an approximate FDR estimate is $(m * p_{(r)}/r)$.

For the m genes in the array, denote m_0 and m_1 as the numbers of truly nondifferentially and differentially expressed genes, respectively. If r genes are selected and the probability of false-positive error for the last ranked gene is $p_{(r)}$, then the expected number of false positives for selecting r genes is $m_0 * p_{(r)}$. Given r selected genes, the expected number of true negatives is $m_0 - m_0 * p_{(r)}$; thus, the expected number of false negatives is $(m-r)-(m_0-m_0*p_{(r)})$. Let C_{FP} be the cost for making a false-positive error and C_{FN} be the cost for making a false-negative error. The total expected cost for selecting the r (top-ranked) genes is the sum of the false-positive cost and the false-negative cost. That is, $COST_{total} = m_0\ p_{(r)}\ C_{FP} + [(m-r) - (m_0-m_0*p_{(r)})] * C_{FN}$. The optimal cut-off that minimizes the total expected cost can be estimated numerically. The receiver-operating characteristic approach requires the prior knowledge of m_0 and m_1. Hsueh and colleagues described several methods to estimate m_0 [57].

change can be identified as statistically significant, even if it may not be significant 'biologically'. In a two sample comparison, it is common to use the p-value to decide a cutoff and then focus on the genes that pass the cutoff with a higher fold-change comparison. This analysis can be plotted by the so-called volcano plot [27]. The approach of using the p-value cutoff as the primary criterion and followed by the fold-change provides the control of false-positive error and, in the meantime, preserves the desired biological significance.

Recently, there appears to be interest in promoting reproducibility of selected genes as a desirable objective [1]. The MAQC Consortium [1] suggested a fold-change cutoff with a nonstringent p-value cutoff to improve reproducibility. Reproducibility of a gene list is not the same as reproducibility of expression measurements and it does not imply accuracy, sensitivity (the probability of selecting truly differentially expressed genes) or specificity (the probability of not-selecting truly nondifferentially expressed genes). The MAQC Consortium [1] used the percentage of overlapping genes as the measure of reproducibility in the evaluation of a gene-selection procedure. There are several problems with using the percentage of overlapping genes as a measure for a gene-selection criterion. The percentage of overlapping genes can increase or decrease irregularly as a cutoff changes; it will be 100% reproducible if all genes are selected, regardless of how many genes are truly differentially expressed. The use of a particular fold-change cutoff assumes that, for any given gene that is biologically significant, its biologically meaningful change between two experimental conditions is judged by this same fold-change cutoff. When there are no treatment effects, the fold-change cutoff with a nonstringent p-value cutoff could have a false discovery rate of 100%. Fold-change cutoff is widely considered inadequate [25,32,33]. The statistical (p-value) approach is much more than a way of gene ranking; it provides a measure to estimate the false-positive error probability for a decision. The approach of using the p-value cutoff as the primary criterion and followed by the fold-change or a pathway analysis provides the control of false-positive error and, in the meantime, preserves the desired biological significance.

Class prediction

Class prediction aims at developing a prediction model that accurately predicts the class membership of a new sample from the available gene-expression data set. The prediction models can

be used to discriminate between different biologic phenotypes or to predict the diagnostic category, prognostic stage of a patient or treatment response. Development of a prediction model involves two components: model building and performance assessment. Typically, the data are divided into a training set and a test set; the prediction model is developed on the training set and is then used to classify samples in the test set to assess its predictive accuracy (Box 4).

Model building

In most microarray studies, most genes in the arrays are not differentially expressed; these genes are irrelevant to the prediction. The use of all genes can suppress or reduce the performance of a prediction algorithm. Selection of discriminatory (feature) genes is critical to the accuracy of prediction. Depending on classification algorithms, two general approaches are used for feature selection: filters and wrappers. The filter approach filters out irrelevant genes according to some predetermined criterion, such as p-value cutoff. Alternatively, genes can be selected based on the individual predictive accuracy by performing gene-by-gene prediction using, for example, the simple logistic regression. Classification algorithms, such as the Fisher's linear discriminant analysis, *k*-nearest-neighbor and support vector machines [34–37], use the filter approach without involving gene selection. Prediction models for cancer classification mainly apply the filter approach, where the gene set used in building the prediction model are preselected. The wrapper approach finds a subset of genes and evaluates its relevance while building the prediction model. The classification algorithms, such as stepwise logistic regression, classification tree [38,39] and support vector machines with recursive feature elimination [40], used the wrapper approach.

For a selected classification algorithm with appropriate gene selection method, the prediction model is fit to the training data set.

Performance assessment

In the development of a prediction model, the most important question is the ability of the model to predict a future sample. To ensure an unbiased assessment of accuracy, the prediction model is developed in one data set; the prediction model is applied to another data set to estimate the predictive accuracy. The most straightforward method of estimating the accuracy is the split-sample validation in which one portion of the data is held out (test dataset) while the classifier is being developed on the remaining data (training dataset) and then is tested on the test dataset [41]. In practice, data are often insufficient to split into a training set and a test set for validation. Instead, cross-validation is used to evaluate the performance of a prediction model. Cross-validation involves repeatedly splitting the data into a training set containing most of samples and applying the prediction rule to the test set made up of the remaining samples. The predictive accuracy rates are estimated from the test data [42].

In cross-validation, the test data must be completely independent of the training data from which the prediction model is built. The cross-validation needs to repeat all steps of model development, including gene selection and model construction, within each stage. In using the filter approach, gene selection must be conducted in the training set to avoid selection bias [43,44].

Future perspective

Normalization

Normalization (transformation) has been an active research area in microarray data analysis since microarray technology was introduced [45]. The primary goal of normalization is to properly eliminate or minimize the systematic variation, such as microarray construction in probe printing, sample preparation procedures, hybridization and washing procedures, detecting method and so on. As technology has improved with better experimental design and control of sources of variability, it has recently been recognized that, except for technician factors, the biological variation is the main source of variability. The recent MAQC project used different manufacturer-recommended normalization procedures for each platform; the normalized intensity data within a platform are highly consistent in two distinct biological samples. Tsai and colleagues showed that an array-by-array Lowes normalization procedure can have a large impact on the result of data analysis [10]. As the technology becomes mature, the normalization factors may be incorporated in the data analysis as covariates [46].

Gene-class testing

In class comparison, after selecting the list of significant genes, a GCT or over-representation analysis often follows to determine whether any gene class (e.g., a pathway) is over-represented in the significant list compared with the whole list. The Fisher's exact test is typically used to assess the significance for an over-representation class [47]. This approach has several short comings [48–50]. First, the division

Box 4. Class prediction.

Class prediction is used to predict the class membership of a new sample using a classification algorithm. As a result of insufficient data for performance assessment, in classification, the original sample data set is typically divided into two subsets: a training set and a test set. The classification algorithm is built from the training samples and then its prediction rule is applied to the test samples. Class prediction generally consists of two components: building of a classification model, including determining a prediction algorithm, selecting the predictor set and fitting the prediction model to training data and assessment of the performance of the prediction model.

Classification involving a large number of predictors presents a challenge to the development of accurate classifiers. Traditional classification algorithms, such as the logistic regression and Fisher's linear discriminant analysis [27], require the number of predictors to be less than the number of samples. In microarray experiments, a large number of genes are measured. Most genes are not differentially expressed: they are noisy and not useful for prediction. The use of all genes can suppress or reduce the performance of a classification algorithm. Selection of a subset of predictors to improve predictability, known as feature selection, has been an important issue in the development of prediction systems in data mining.

The classification algorithms, such as the logistic regression and Fisher's linear discriminant classifiers, form the prediction rule based on a preselected set of predictors without involving selection of predictors. Some algorithms, such as the Classification Tree [38] and support vector machines with recursive feature elimination [40], have incorporated feature selection into model building. These algorithms find a subset of predictors and evaluate its relevance for the classification; their classification rules are built from an optimal predictor subset. In feature selection, having the selection of predictor variable independent of the classification algorithm is referred to as the filter approach; having the selection of predictor variables incorporated while building the prediction model in the training phase is referred to as the wrapper approach.

Assessment of the accuracy of the prediction model is a critical step in the development of a prediction model. One problem in fitting a prediction model to microarray data is over-fitting the data. Microarray gene-expression data are characterized by the number of genes far exceeding the number of samples. The predicted model can fit the original data well, but may predict poorly for new data. An unbiased assessment of the accuracy of the prediction model is important. In the development and validation of a class prediction model it is best to have a sufficiently large collection of data. In practice, data are often insufficient to set aside a test set for validation. Cross validation is used to evaluate the performance of a prediction model.

In a V-fold cross-validation, the entire data set is divided into V subsets of roughly equal size and the classification analysis iterates V times. Each time, the prediction rule is trained on (V-1) subsets together and then applied to the remaining subset as the test data set. After completion of all V subsets (all samples are classified), the sensitivity, specificity and concordance rates of the prediction rule are computed across all V subsets. The entire process may be repeated b times with different partitions of V subsets. The averages over the b trials are calculated. The simplest cross-validation is known as 'leave-one-out',' where V is the total sample size.

of genes into differential and nondifferential expression groups is arbitrary; the genes in the nondifferential expression list are discarded, regardless of their p-values. Second, the Fisher's exact test approach simply counts the number of genes in the list; the order of genes is not taken into consideration. Third, the correlation structure of genes is not taken into consideration.

GCT is a statistical approach to determine whether some functionally predefined classes of genes are differentially expressed. A gene class refers to a group of genes with related functions or a set of genes grouped together based on biologically relevant information, such as a metabolic pathway, protein complex or gene ontology (GO) category. Several GCT procedures have been proposed [48–50]. The procedure uses a global test statistic to compute the p-value of each gene class and the classes are ranked accordingly. A typical GCT approach can be summarized as follows:

- All genes are ranked by computing a test statistic (or p-value) that measures the association between the expressions and the treatment conditions;
- For each gene class or functional category, a class score is calculated as a summary measure of the class;
- Resampling methods are used to generate the null distribution of the class score for each gene class;
- Statistical significance is assessed by comparing the observed functional score with the percentile of the null distribution of the gene class.

By considering the distribution of the entire set of genes, this approach is more powerful and interpretable. The success of a GCT requires development of multivariate statistics and computational algorithms for testing hundreds of variables and a well-characterized gene class mapped to microarray probes.

Multivariate ordering criteria

A microarray experiment can generate different gene lists by different filters, normalizations or analysis methods for different study objectives. In some studies, it may be interested in selecting a subset of genes from the multiple gene lists. For example, the fold-change and p-value are two

commonly known criteria to select differentially expressed genes under two experimental conditions. These two selection criteria often result in incompatible selected gene sets. Also, consider a mouse experiment to study differences in gene expression among the three p53 genotypes: wild-type (+/+), knock-out (-/-) and heterozygous (+/-). In class comparison, a statistical analysis typically consists of a comparison among the three genotypes. An important follow-up analysis is the comparisons between the knock-out and wild-type mouse and between the heterozygous and wild-type mouse. The Dunnett's test is frequently used to generate the differentially expressed gene lists for the two comparisons. Often, it is also interested in the genes that show differences in both comparisons. Chen and colleagues recently proposed layer ranking algorithms to provide a single preference gene list from multiple gene lists generated by different ranking criteria [51].

Development of prediction models

The US FDA envisions clinical pharmacogenomic profiling to identify patients most likely to benefit from particular drugs and patients most likely to experience adverse reactions. The goal is to change medical practice from a population-based approach to an individualized approach. Personalized medicine uses the available data on each individual patient for assignment of more effective therapies, as well as better diagnosis and earlier interventions that might prevent or delay disease. A main objective of pharmacogenomic profiling is to identify a subset of genes to develop a genomic composite biomarker (GCB). A GCB may help in determining how the benefits and adverse effects of a drug vary among a target population of patients based on the genomic features of patients' germline and diseased tissue. By identifying groups of patients with a high probability of benefiting from therapeutics and avoiding serious adverse events, the therapeutic index of a drug can be substantially increased.

A GCB-based predictive model requires high accuracy, since the consequence of misclassification may result in suboptimal treatment or an incorrect risk profile. In class prediction, much research has focused on improving the predictive accuracy; less work has been done on exploring individual variables in respect to disease characteristics and treatment response. An analysis by Michiels and colleagues [52] showed that the list of genes identified as predictors of cancer prognosis was highly unstable. The selected gene set strongly depended on the selected patients in the training set. Class prediction will be more focused on the selection of the relevant genomic variables and clinical variables that can improve predictability. A GCB is a classifier that consists of a set of genes described by a prediction model with a specified threshold cutoff. Different prediction models or different threshold cutoffs will result in different sensitivity and specificity. A predictive model built from the GCB based on the disease phenotype and patient genotype should have better predictive accuracy and provide better guidance on treatment assessment. As a result of pre-conditions of each individual and population variability, the response to treatment may vary among different patients. In addition, there may be a probability for nonresponses and different costs associated with different outcomes. A decision analysis is needed to account for the probabilities of efficacy of treatment assignments.

Development of a prediction model involves quality assessment, missing data treatment, background subtraction, normalization, gene selection, incorporation with clinical variables, data partitions for performance assessment and so on. The integration of these steps for prediction of future samples is challenging. Currently, a community-wide MAQC-II project has been working to characterize approaches and standard operating procedure for the development and validation of prediction models [102].

Finally, there are two aspects in the development of pharmacogenomic classifiers: optimizing treatment selection for individual patients and using predictive classifiers in conjunction with the development of new drugs in clinical design. For treatment selection, the clinical utility is whether the classifier predicts more accurately than the standard classification system. For experimental therapy in drug development, the clinical utility is to demonstrate effectiveness of the drug in a population identified by the classifier as being more likely to benefit [53–55]. The genomic technologies available today are sufficient to develop pharmacogenomic biomarker classifiers to more effectively assign patients to the proper treatment.

Disclaimer

The views presented in this paper are those of the author and do not necessarily represent those of the USA FDA. J Chen declares that there are no conflicts of interest.

Executive summary

- The experimental unit refers to an independent biological sample to which the treatments are applied. Statistical tests for inference between different populations should be based on the biological sample variance.
- Selection of differentially expressed genes (gene selection) can be separated into two steps: gene ranking and assigning a significance level.
- The statistical (p-value) approach is much more than a gene ranking; it provides a measure to estimate the false-positive error probability for a cutoff decision.
- In performance assessment, the test data must be completely independent of the training data from which the prediction model is built; the cross-validation needs to repeat all steps of model development.
- A gene class testing is a statistical approach to determine whether some functionally predefined classes of genes are differentially expressed under different experimental conditions.
- Layer ranking algorithms provide a preference gene list from multiple gene lists generated by different ranking criteria.
- Class prediction involves data quality assessment, missing data treatment, normalization, gene selection, performance assessment and so on. The integration of these steps for prediction of future samples is challenging.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. MicroArray Quality Control Consortium: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotech.* 24, 1151–1169 (2006).
2. Perket JM: Six things you won't find in the MAQC. *Scientist* 20(11), 68–72 (2006).
3. Klebanov L, Qiu X, Welle S, Yakovlev A: Statistical methods and microarray data. *Nat. Biotech.* 25, 25–26 (2007).
4. Chen JJ, Delongchamp RR, Tsai CA *et al.*: Analysis of variance components in gene expression data. *Bioinformatics* 20, 1436–1446 (2004).
   - Background and methods for estimating biological, within- and between-array and other sources of variation and their relative contributions to the overall variation.
5. Chen JJ, Chen CH: Microarray gene expression. In: *Encyclopedia of Biopharmaceutical Statistics (2nd Edition).* Chow S (Ed.). Marcel Dekker, Inc., NY, USA, 599–613 (2003).
6. Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7, 819–837 (2000).
   - First paper to present a general statistical framework for microarray data analysis.
7. Yang YW, Dudoit S, Luu P, Speed TP: Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids. Res.* 30, e15 (2002).
   - Introduces the scatter plot smoother Lowes fit for normalization.
8. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193 (2002).
9. Irizarry RA, Wu Z, Jaffee HA: Comparison of Affymetrix gene-chip expression measures. *Bioinformatics* 21, 1–7 (2005).
10. Tsai CA, Hsueh HM, Chen JJ: A generalized additive model for microarray gene expression data analysis. *J. Biopharm. Stat.* 14, 553–573 (2004).
    - Generalizes an analysis of variance model that incorporates Lowes normalization for microarray data analysis.
11. Qui LX, Kerr, KF, and Contributing members of the Toxicogenomics Research Consortium: Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.* 32, 5471–5479 (2004).
12. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95, 14863–14868 (1998).
13. Alizadeh AA, Eisen MB, Davis RE *et al.*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511 (2000).
14. Golub T, Slonim D, Tamayo P *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–536 (1999).
15. Dudoit S, Yang YH, Callow MJ, Speed TP: Statistical methods for identifying differential expressed gene in replicated cDNA microarray experiments. *Statistica Sinica* 12, 111–139 (2002).
16. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci USA* 98, 5116–5121 (2001).
    - Introduces the Significance Analysis of Microarrays test.
17. Efron B, Tibshirani R, Storey JD, Tusher V: Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 96, 1151–1160 (2001).
18. Baldi P, Long AD: A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519 (2001).
19. Lonnstedt I, Speed TP: Replicated microarray data. *Statistica Sinica* 12, 31–46 (2002).
20. Wright GW, Simon RM: A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19(18), 2448–2455 (2003).
    - Provides a shrinkage estimator. The procedure is available at BRB-Array Tools [103].
21. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK: Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* 19, 1945–1951 (2003).
22. Smyth GK: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3(1), A3 (2004).
23. Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA: Improved statistical tests for differential gene expression by shrinking

variance components estimates. *Biostatistics* 6, 59–75 (2005).
24. Qin LX, Kerr KF: Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.* 32, 5471–5479 (2004).
25. Chen JJ, Wang SJ, Tsai CA, Lin CJ: Selection of differentially expressed genes in microarray data analysis. *Pharmacogenomics J.* (In Press) (2007).
26. Wolfinger RD, Gibson G, Wolfinger ED *et al.*: Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8, 625–637 (2001).
27. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* 29, 389–395 (2001).
28. Benjamini Y, Hochberg Y: Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* 57, 289–300 (1995).
• Introduces the concept of false discovery rate (FDR) for multiple testing.
29. Tsai CA, Hsueh HM, Chen JJ: Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* 59, 1071–1081 (2003).
• Describes several FDR error measures and illustrates the power of a FDR procedure improved by incorporating an estimate of the number of nondifferentially expressed genes.
30. Storey JD: A direct approach to false discovery rates. *J. Royal Stat. Soc. B* 64, 479–498 (2002).
31. Delongchamp RR, Bowyer JF, Chen JJ, Kodell RL: Multiple testing strategy for analyzing cDNA array data on gene expression. *Biometrics* 60, 774–782 (2004).
32. Miller RA, Galecki A, Shmookler-Reis RJ: Interpretation, design, and analysis of gene array expression experiments. J. *Gerontol. A* 56, B52–B57 (2001).
33. Smyth GK, Yang YH, Speed TP: Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.* 224, 111–136 (2003).
34. Hastie T, Tibshirani RT, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, NY, USA (2001).
35. Vapnik VN: *Statistical Learning Theory.* Wiley, NY, USA (1998).
36. Tsai CA, Lee TC, Ho IC, Yang UC, Chen CH, Chen, JJ: Gene selection for multi-class clustering and prediction. *Math. Biosci.* 193, 79–100 (2005).
37. Furey TS, Christianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: Support vector machine classification andvalidation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914 (2000).
38. Brieman L, Friedman JH, Olshen RA, Stone CJ, Steinberg D, Colla P: *CART: Classification and Regression Trees.* CRC Press LLC, CA, USA (1995).
39. Moon H, Ahn H, Kodell RL, Lin CJ, Baek S, Chen JJ: Classification methods for the development of genomic signatures from high-dimensional data. *Genome Biol.* 7, R121 (2006).
40. Guyon I, Weston J, Barnhill S, Vapnik V: Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422 (2002).
41. Rosenwald A, Wright G, Chan WC *et al.*: The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 346, 1937–1947 (2002).
42. Tsai CA, Chen CH, Lee TC, Ho IC, Yang UC, Chen JJ: Gene selection for sample classifications in microarray experiments. *DNA Cell Biol.* 23, 607–614 (2004).
• First paper to illustrate examples of cross-validation using a filter approach for class prediction.
43. Ambroise C, McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA* 99, 6562–6566 (2002).
• First paper to illustrate that cross-validation should extend to gene selection in estimating classification error rate.
44. Dupuy A, Simon R: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl Cancer Inst.* 99, 147–157 (2007).
•• Provides guidelines for statistical analysis of microarray data and presents a checklist of 'Dos and Don'ts'.
45. Chen Y, Dougherty ER, Bittner ML: Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2(4), 364–374 (1997).
46. Parrish RS, Delongchamp RR: Normalization of microarray data. In: *DNA Microarrays and Related Genomic Techniques: Design, Analysis, and Interpretation of Experiments.* Allison DB, Page GP, Beasley TM, Edwards JW (Eds), Chapman & Hall/CRC, FL, USA, 9–28 (2006).
47. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: Global functional profiling of gene expression. *Genomics* 81, 98–104 (2003).
48. Mootha VK, Lindgren CM, Eriksson K. *et al.*: PGC-1-α-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. *Nat. Genet.* 34, 267–273 (2003).
• First paper to propose a formal statistic for gene class testing.
49. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E: Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.* 29, 1213–1222 (2004).
50. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA* 102, 13544–13549 (2005).
51. Chen JJ, Tseng SL, Tsai CA, Chen CH. Gene selection with multiple ordering criteria. *BMC Bioinformatics* 8, 74 (2007).
52. Michiels S, Koscielny S, Hill C: Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365, 488–492 (2005).
53. Simon R: Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *J. Natl Cancer Inst.* 98, 1169–1171 (2006).
54. Simon R: Road map for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.* 23(29), 7332–7341 (2005).
55. Simon R: Validation of pharmacogenomic biomarker classifier for treatment selection. *Dis. Markers* 21, 1–8 (2005).
56. Tsai CA, Chen YJ, Chen JJ: Testing for differentially expressed genes with microarray data. *Nucleic Acids Res.* 31, e52 (2003).
57. Hsueh H, Chen JJ, Kodell RL: Comparison of methods for estimating number of true null hypothesis in multiplicity testing. *J. Biopharm. Stat.* 13, 675–689 (2003).

Websites

101. Significance Analysis of Microarrays www-stat.stanford.edu/~tibs/SAM/
102. US FDA: MicroArray Quality Control project on the evaluation of analysis protocols for deoxyribonucleic acid microarray data. www.fda.gov/nctr/Science/centers/toxicoinformatics/maqc/
103. National Cancer Institute Biometric Research Branch http://linus.nci.nih.gov/brb