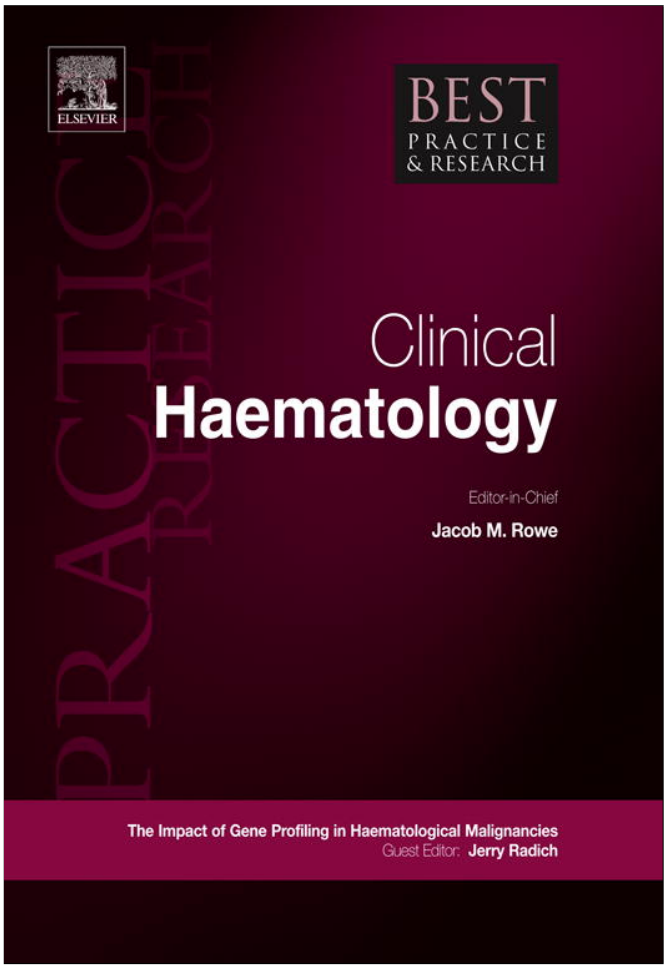

Best Practice & Research Clinical Haematology 22 (2009) 271–282

Contents lists available at ScienceDirect

Best Practice & Research Clinical Haematology

journal homepage: www.elsevier.com/locate/beha

8

Analysis of DNA microarray expression data

Richard Simon, D.Sc., Chief, Biometric Research Branch *

Biometric Research Branch, Division of Cancer Treatment & Diagnosis, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA

Keywords:
bioinformatics
biomarkers
gene expression signatures
microarray data analysis

DNA microarrays are powerful tools for studying biological mechanisms and for developing prognostic and predictive classifiers for identifying the patients who require treatment and are best candidates for specific treatments. Because microarrays produce so much data from each specimen, they offer great opportunities for discovery and great dangers or producing misleading claims. Microarray based studies require clear objectives for selecting cases and appropriate analysis methods. Effective analysis of microarray data, where the number of measured variables is orders of magnitude greater than the number of cases, requires specialized statistical methods which have recently been developed. Recent literature reviews indicate that serious problems of analysis exist a substantial proportion of publications. This manuscript attempts to provide a non-technical summary of the key principles of statistical design and analysis for studies that utilize microarray expression profiling.

Published by Elsevier Ltd.

Introduction

DNA microarray technology has found broad use in basic and translational cancer research. Our objective here is to provide a non-technical summary of the key principles of statistical design and analysis for studies that utilize microarray expression profiling. Because microarrays produce so much data from each specimen, they offer great opportunities for discovery and great dangers or producing misleading claims. Effective analysis of microarray data, where the number of measured variables is orders of magnitude greater than the number of cases, requires specialized statistical methods which have recently been developed. The literature review by Dupuy and Simon studies relating gene

* Tel.: +1 301 496 0975; Fax: +1 301 402 0560.
E-mail address: rsimon@nih.gov

1521-6926/$ – see front matter Published by Elsevier Ltd.
doi:10.1016/j.beha.2009.07.001

expression profiles to cancer outcome found serious problems of analysis in approximately 50% of publications [1]. Here we will attempt to provide a non-technical summary of the key principles of statistical design and analysis for studies that utilize microarray expression profiling and illustrate some of the important analysis principles using BRB-ArrayTools software [2].

Platform specific data pre-processing

DNA microarrays are assays for quantifying the types and amounts of mRNA transcripts present in a collection of cells. The chemical and physical mechanisms by which this quantification is accomplished varies widely among microarray platforms. In many cases the mRNA extracted from a sample of cells is reverse transcribed to fluorescently labeled complementary DNA (cDNA). The labeled cDNA is then placed on a solid surface on which strands of polynucleotide probes have been attached in specified positions. The labeled cDNA molecules hybridize to the probes to which they share sufficient sequence complementarity and the quantity of cDNA bound to each polynucleotide probe is quantified by illuminating the solid surface with laser light of a frequency tuned to the fluorescent label employed and measuring the intensity of fluorescence over each probe on the array. This intensity of fluorescence should be approximately proportional to the number of molecules of cDNA bound to the probe. Dual label microarrays often co-hybridize labeled transcripts from a specimen of interest with differently labeled transcripts from a reference source of RNA. For each probe represented on the array, a relative measure of abundance of the corresponding transcripts in the specimen of interest relative to the common reference source is obtained. This relative intensity is often expressed as a ratio or log ratio. Affymetrix GeneChips™ have oligonucleotide probes lithographically synthesized directly on the silicon surface of the array. Consistent probe geometry and sample circulation minimize probe specific inter-array variability and so single label protocols are generally used with GeneChips™.

The "pre-processing" steps of analysis of array data are somewhat platform specific. Because the lithographically synthesized probes in GeneChips™ are relatively short and not error-free, multiple probes are used for each transcript target and a summary measure of intensity per "probe set" is computed as a pre-processing step. Various methods of robust model based statistical estimation have been developed for such probe set summaries as reviewed by Irizarry et al. [3]. BRB-ArrayTools incorporates a platform-specific data importer for many popular platforms, including Affymetrix.cel files. It provides options for several of the probe-set summarization methods.

For a variety of technical reasons, the overall level of fluorescence intensity differs among arrays. The adjustment for such differences in overall intensity of single label arrays is called *normalization.* The simplest kind of normalization involves multiplicatively transforming all the intensities on each array by a factor so that all arrays have the same median probe intensity. More sophisticated methods such as quantile normalization essentially replace a probe intensity which is in the p'th intensity percentile on an array by the intensity of the p'th percentile of a selected reference array [4]. All such methods assume, however, that the variation across arrays of most probes is due to technical factors, not true biological effects. When this assumption is not appropriate, normalization should be based on probes for *housekeeping genes* considered uniformly expressed for the specimens under analysis. Normalization of data from dual label arrays is different than for single label arrays. For dual label arrays it is the log ratio of intensities that must be adjusted for inter-array technical variation in the relative intensity of the two labels. The simplest approach is to scale the ratios on each array so that the median log-ratio over the probes on the array is zero. Another commonly used approach lets the scale factor be intensity level dependent. Normalization methods are reviewed by Park et al. [5].

Pre-processing may also include "filtering" out probes with low intensity or minimal variation among the arrays being analyzed and thresholding intensity levels on dual label arrays to a lower limit of detection so that computed log-ratios are not extreme. Pre-processing should not, however, be based on differential expression among any phenotypes or classes as that may seriously bias subsequent analyses [6].

Objectives of microarray studies

Effective microarray experiments require careful planning based on clear objectives [7]. The objective drives the selection of specimens and the specification of an appropriate analysis strategy [7].

The large numbers of genes whose expressions can be measured in a single hybridization creates an even greater than usual need for careful planning of the methods of analysis so that biologically meaningful conclusions, rather than spurious associations are reported.

The objectives of many studies utilizing DNA microarrays can be categorized as either *gene discovery*, *class prediction*, or *class discovery*. Gene discovery, also called class comparison, focuses on determining which genes are differentially expressed among samples representative of pre-defined classes. The classes may represent different tissue types, diseased tissue or normal tissue of the same cell type, or the same tissue under different experimental conditions. The defining characteristic of gene finding/class comparison is that the classes are defined independently of the expression profiles. For example, Korn et al. [8] evaluated expression profiles from breast tumors pre and post chemotherapy to identify those genes whose expression was modified by treatment. Yang et al. [9]. studied gene expression changes in metastatic breast tumors pre and post Erlotinib treatment. Sotiriou et al. [10] evaluated genes whose expression was correlated with clinico-pathological characteristics of breast tumors. Desai et al. [11] evaluated genes differentially expressed among different transgenic mouse models of breast cancer. The phrase *gene discovery* is somewhat more general than *class comparison* as it can include finding the genes whose expression is correlated to a quantitative measurement or a survival time.

With class prediction the emphasis is on developing a computable function that can be used to predict which class a new specimen belongs to based on its expression profile. This usually requires finding which genes are informative for distinguishing the pre-defined classes, estimating the parameters of the mathematical function used, and estimating the accuracy of the predictor [7,12]. Class prediction is important for medical problems of diagnostic classification, prognostic prediction and treatment selection. For example, van't Veer [13] and van De Vijver [14] developed and evaluated predictors of which patients with primary breast cancer are at high risk for recurrence after local treatment alone. Ma et al. [15] developed such a predictor for patients with estrogen receptor positive primary breast cancer who received Tamoxifen monotherapy after local therapy. Ayers et al. [16] developed a predictor of complete pathologic response to neoadjuvant chemotherapy in patients with breast cancer. Jansen et al. [17] developed a predictor of response to Tamoxifen for patients with metastatic breast cancer.

Class discovery is different than gene finding or class prediction because it does not involve any pre-defined classes. Instead, it involves grouping together of specimens based on similarity of their expression profiles with regard to the genes represented on the array. *Cluster analysis* algorithms are used for generating the groups. Cluster analysis algorithms are called "unsupervised" because the grouping is not driven by any phenotype external to the expression profiles, such as tissue type, stage, grade or response to treatment. The objective of clustering expression profiles of tumors is to determine new disease classifications. For example, Perou [18] characterized expression profiles of primary breast tumors into four patterns which they called basal-like, luminal-like, Erb-B2+, and normal-like. Cluster analysis is an exploratory analysis method, however, and even random expression profiles can be clustered. It is generally difficult to evaluate the meaningfulness of a set of clusters except by comparing them with regard to existing phenotypes [19]. Cluster analysis is over-utilized, however, and in many cases it provides misleading results [1]. Most cancer studies involving microarray expression profiling really have class comparison or class prediction objectives. For such studies, "supervised methods" are usually more effective [7].

Study design

Clear objectives are essential for the design of effective microarray studies. The objectives indicate the kinds of samples that should be included and the number of such samples. The statistical power for identifying differentially expressed genes or for developing classifiers is generally determined by the number of *biological replicates* in each class. These are distinguished from *technical replicates* which are just repeat assays of the same RNA samples. For most commercial microarray platforms, random technical variation is small relative to biological variation and so there is little value in obtaining technical replicates of RNA extractions [20]. Systemic variation over time remains a problem, however, for some platforms and so it is important to perform the assays in a manner that does not confound

phenotype classes with assay performance. For example, in comparing expression of p53 mutant cell lines to p53 wild type cell lines, one should avoid assaying all the mutants with one set of reagents on one week and the wild type cell lines with a different set of reagents on another week. If a large number of samples are to be assayed, the phenotype classes should be intermixed in the group assayed at each time. Pooling samples is generally not advantageous [21]. Although the number of arrays may be reduced by pooling, the number of samples needed may be substantially increased. When dual-label arrays are used, there are additional design issues to be addressed; e.g. whether to use a common reference RNA or to pair the samples from different classes for co-hybridization on each array. Dobbin et al. provide a thorough discussion of this issue including ways of avoiding the need for performing dye-swap technical replicates [22]. Dobbin and Simon provide formulas and graphs for determining the number of experimental/biological replicates needed for class comparison problems [23] or for developing a predictive classifier [24].

Gene finding (class comparison)

In the earliest microarray studies, investigators performed class comparison by examining fold change differences for each gene between a microarray of a single specimen from one class and a specimen of the other class. This is not really meaningful, however, because the comparison may reflect sample differences or assay differences, rather than class differences. Using replicate arrays for measuring expression for one sample from each of two classes does not help much. Such *technical replicates* do not satisfy the crucial need for studying multiple tumors of each type. Individual microarrays of independent biological replicates from each phenotype class of interest is generally needed, not assay replicates of the same RNA specimen or microarrays of pooled biological replicates [7,20].

Often the genes are ranked with regard to their degree of differential expression among the classes, and a cut-point determined on the ranking in order to control the number of false positive claims. If there are two classes, the absolute value of a standard t statistic could be used for the ranking. The t statistic is the difference in the class specific means of log expression divided by an estimate of the standard error of the difference. If there are few samples per class, however, the ranking will be unstable because the estimates of within-class variation, made separately for each gene, will be too imprecise. Improved methods based on t statistics which borrow variance information among genes are recommended if there are less than 10 samples per class. [25,26] These methods are called regularized t-tests, random variance t-tests or empirical Bayes t-tests. They are also applicable for cases with more than two classes. They are based on the assumption that the within class variances for different genes come from the same distribution, but not that they are equal.

Although standard statistical methods like regularized t-tests are often used for creating the ranking, the novel aspect of this analysis that must be taken into account is that there are generally tens of thousands of genes analyzed. Hence, more stringent standards of statistical significance for claiming differential expression must be used. If, for example, there are 10,000 genes represented on the array, then in comparing expression for samples from two classes, one would expect 500 false positive claims of statistical significance at the traditional 5% significance level ($0.05 \times 10{,}000$). This is not acceptable. By using a stringent threshold of significance the number of false positive findings can be limited; a threshold of $p < .001$ results in 1 false positive gene per 1000 genes analyzed on average.

In comparing gene expression profiles among classes, biostatisticians today generally prefer reporting the "false discovery rate" (FDR) for the comparison as a whole rather than the statistical significance level for individual comparisons [27]. The false discovery rate is the proportion of false positives among the genes claimed to be differentially expressed among the classes. For example, suppose one claims a gene to be differentially expressed if the univariate significance level is less than 0.001. Then for 10,000 genes analyzed the expected number of false positives is about 10 (since most of the 10,000 of genes are not expected to be differentially expressed). If there are 40 genes for which the univariate significance level is less than 0.001, then the false discovery rate is about 10/40 or 25%; that is, one in four of the reported genes are likely to be false positives.

There are powerful multivariate methods for comparing expression profiles between classes and identifying differentially expressed genes in a manner that controls the FDR and takes into account the

correlation among the genes [28,29]. The multivariate test of Korn et al. is similar in spirit to the very popular SAM method of Tusher et al. The former permits somewhat greater control over the statistical confidence with which the proportion of false discoveries is limited to the target level. Both the multivariate permutation test and SAM are available in BRB-ArrayTools. For this software SAM was re-programmed in FORTRAN and so it is much faster than other available versions.

Most of the methods used for finding genes whose expression is correlated with a phenotype can be used with categorical phenotypes, quantitative phenotypes or survival time phenotypes. The measure of correlation used for each gene varies depending on the type of phenotype of interest. For categorical phenotypes the multivariate methods such as SAM and the multivariate test of Korn et al. are based on computing regularized t-tests for each gene. For survival time phenotypes p values from univariate proportional hazards regression analyses can be used.

Cluster analysis is often used in a potentially misleading way for identifying differentially expressed genes. Investigators may generate a gene list using an inadequately stringent univariate significance level of 0.05 or 0.01. The samples are then clustered with regard to the expression profiles for the selected genes. The fact that the samples from the classes are separated in this cluster analysis is taken as validation that the genes are really differentially expressed. This supervised form of cluster analysis is invalid. If one generates expression profiles for two classes using random numbers with no real difference between the two classes, there will be about 500 false positives per 10,000 genes. If one clusters the randomly generated samples with regard to those selected genes that were found significant at the 0.05 level, the samples will be separated [1].

Finding differentially expressed sets of genes

One problem encountered in the analysis of gene expression data is biologically interpreting lists of genes identified as differentially expressed among compared classes. Although many software packages provide biological annotations for the genes found differentially expressed, a more recent approach compares the classes with regard to the expression of pre-defined biologically meaningful gene sets. There are several advantages to this approach in addition to ease of interpretation. Since the number of gene sets tested is generally much less than the number of genes represented on the array, the magnitude of the multiple comparison problem is reduced. Also, expression patterns of genes in a gene set can reinforce each other and do not have to be individually significant at a very stringent level as required for the post-hoc annotation methods.

There is a large variety of such Gene Set Enhancement methods [30,31] which provide a score for summary differential expression for each gene set. BRB-ArrayTools provides four such methods [32] and a wide variety of pre-defined gene sets including Gene Ontology groups, genes in annotated metabolic or signaling pathways, genes on the same chromosome arm, genes that are targets of the same transcription factor, genes containing the same protein domain and genes on the same experimentally determined signature of response to pathway activation or silencing.

Time-course expression data

Time course data is also used for gene finding although there are really no pre-defined classes or phenotypes. Typically gene expression is measured at intervals following an experimental intervention and the objective is to identify genes whose expression is changing with time. For example, BRB-ArrayTools fits a quadratic function to the time course of individual genes and tests whether the linear and quadratic coefficients are both zero. Those statistical significance levels are used to control the false discovery rate using the method of Benjamini and Hochberg [27]. Those genes are clustered to sort them into sets showing similar patterns over time and a heat map is provided. The time-course analysis tool also provides for identifying genes whose variation with time differs based on some other phenotype such as treatment. A line plot of average gene expression over time for each phenotype group is provided for each gene with a significant interaction between time course and phenotype. Specialized software for supervised time course analysis is also provided by Storey [33] and by Leek [34].

Predictive classifier development

A class predictor, or *predictive classifi*er, is a computable function which can be used to predict a class from an expression profile. Predictive classifiers can be of considerable importance for guiding treatment selection in medicine although several levels of validation are needed before such classifiers are "ready for prime time" [35]. In developing a predictive classifier the emphasis should be on predictive accuracy, sensitivity, specificity, and positive and negative predictive values, not on controlling the false discovery rate, goodness of fit to the data, or the statistical significance of regression coefficients.

Development of a predictive classifier requires specification of (i) the mathematical/computational function used to relate an expression profile to a class; (ii) the genes whose expression levels are utilized in the prediction; and (iii) the parameters; e.g. the weights placed on expression levels of individual genes and cut-points used in the prediction [7,36]. It is often not recognized that a predictive classifier is not just a set of genes, it is a completely specified computable function that can be used to classify individual patients whose expression profiles are determined.

The development of a predictive classifier is similar to the development of a statistical regression function, except that the former predicts class identifier rather than a continuous value. Statistical regression models are generally built using data in which the number of cases (n) is large relative to the number of candidate variables (p). In the development of class predictors using gene expression data, however, the number of candidate predictors is generally orders of magnitude greater than the number of cases. This has two important implications. One is that only simple class prediction functions should be considered because functions with too many degrees of freedom will over-fit the data and predict poorly for independent samples [37]. The second important implication is that the data used for evaluating the class predictor must be distinct from the data used for developing it. It is almost always possible to develop a class predictor even on completely random data which will fit the training data almost perfectly but be completely useless for prediction with independent data [6].

A wide variety of algorithms have been used effectively with DNA microarray data for class prediction. Many predictive classifiers are based on linear discriminants of the form

$$l(\underline{x}) = \sum_{i \in G} w_i x_i \tag{1}$$

where x_i denotes the log-ratio or log-intensity for the i'th gene, w_i is the weight given to that gene, and the summation is over the set G of genes selected for inclusion in the class predictor. For a two-class problem, there is a threshold value d, and a sample with expression profile defined by a vector $\underline{x}$ of values is predicted to be in class 1 or class 2 depending on whether $l(\underline{x})$ as computed from equation (1) is less than the threshold d or greater than d respectively.

Linear discriminant classifiers differ with regard to how the weights are determined. The oldest form of linear discriminant is Fisher's linear discriminant. The weights are selected so that the mean value of $l(\underline{x})$ in class 1 is maximally different from the mean value of $l(\underline{x})$ in class 2. The squared difference in means divided by the pooled estimate of the within-class variance of $l(\underline{x})$ was the specific measure used by Fisher. To compute these weights, one must estimate the correlation between all pairs of genes that were selected in the feature selection step. The study by Dudoit et al. [38]. indicated that Fisher's linear discriminant analysis did not perform well unless the number of selected genes was small relative to the number of samples. The reason is that in other cases there are too many correlations to estimate and the method tends to be un-stable and over-fit the data.

Diagonal linear discriminant analysis is a special form of Fisher linear discriminant analysis in which the correlation among genes is ignored. By ignoring such correlations, one avoids having to estimate many parameters, and obtains a method which generally performs better when the number of samples is small. Golub's weighted voting method [39] and the Compound Covariate Predictor of Radmacher et al. [36] are similar to diagonal linear discriminant analysis and tend to perform well when the number of samples is small. They compute the weights based on the univariate prediction strength of individual genes and ignore correlations among the genes.

Support vector machines are very popular in the machine learning literature. Although they sound very exotic, linear kernel support vector machines do class prediction using a predictor of the form of equation (1). The weights are determined by optimizing a mis-classification rate criterion with

a penalty for large weights which tends to prevent over-fitting [40]. Although there are more complex forms of support vector machines, they appear to be inferior to linear kernel SVM's for class prediction with large numbers of genes [41].

In the study of Dudoit et al. [37,38], the simplest methods, diagonal linear discriminant analysis, and nearest neighbor classification, performed as well or better than the more complex methods. Nearest neighbor classification is based on a set G of genes selected to be informative for discriminating the classes and a *distance function* $d(\underline{x}, \underline{y})$ which measures the distance between the expression profiles $\underline{x}$ and $\underline{y}$ of two samples. The distance function utilizes only the genes in the selected set G. To classify a sample with expression profile $\underline{y}$, compute $d(\underline{x}, \underline{y})$ for each sample $\underline{x}$ in the training set. The predicted class of $\underline{y}$ is the class of the sample in the training set which is closest to $\underline{y}$ with regard to the distance function d. The distance function is usually either the standard Euclidean distance or one minus the correlation between the expression profiles $\underline{x}$ and $\underline{y}$. A variant of nearest neighbor classification is *nearest centroid classification* in which the new expression profile $\underline{y}$ is compared to the mean expression profile for the training samples of each class. Those mean expression profiles are called centroids. The *shrunken centroid* method of Tibshirani et al. is a popular and effective form of nearest centroid classification which incorporates both automatic selection of the gene set G and adjustment of class specific centroids to account for the bias of gene selection [42].

Dudoit et al. also studied some more complex methods such as classification trees and aggregated classification trees. These methods did not appear to perform any better than the simpler methods. Ben-Dor et al. [41] also compared several methods on several public datasets and found that nearest neighbor classification generally performed as well or better than more complex methods.

The models described above are generally applied with the gene set G specified by identifying the genes that are differentially expressed among the classes when considered individually. For example, if there are two classes, one can compute a regularized t-test for each gene. The log-ratios or log-intensities are generally used as the basis of the statistical significance tests. The genes that are significantly differentially expressed at a specified significance level are selected for inclusion in the class predictor. The stringency of the significance level used controls the number of genes that are included in the model. If one wants a class predictor based on a small number of genes, the threshold significance level is made very small. Issues of multiple testing or false positives are not really relevant, however, because the objective is just to select features for inclusion in the model and the threshold significance level is just a "tuning parameter". Some methods do not use p values at all but merely select the k most differentially expressed genes, and specify k arbitrarily or by optimizing using cross-validation.

Several authors have developed methods to identify optimal sets of genes which together provide good discrimination of the classes [43–46]. Some of these kinds of algorithms are very computationally intensive. Several independent evaluations have, however, indicated that the increased computational effort of these methods is not warranted [47,48].

Predictive classifier validation

Analytical validation

At least three levels of validation should be distinguished. First is analytical validation of an assay for measuring the classifier. Analytical validation has traditionally meant that the assay accurately measures what it claims to measure. This presumes, however, the existence of a *gold standard* way of measuring the true value. For gene expression based predictive classifiers, there is usually no gold standard. Whereas an RT-PCR assay might be accepted as a gold standard for measuring gene expression of an RNA sample, this ignores questions of the representativeness of the RNA sample for the target tissue prior to biopsy. For assays in which there is no gold standard, analytical validation generally means reproducibility and robustness. Sometimes robustness of the assay is distinguished from robustness of tissue handling.

Careful development of an analytically validated assay is important for all later steps of validation. Dobbin et al. reported that in order to ensure good inter-laboratory reproducibility in using the Affymetrix GeneChip system, a pilot study and development of a common protocol was necessary [49].

In classifying the risk of recurrence for patients with node negative and estrogen receptor positive breast cancer receiving Tamoxifen treatment, the investigators utilized DNA microarray gene expression profiling to identify the informative genes, but then transferred to an RT-PCR platform based on primers for use with paraffin embedded formalin fixed tissue. They performed detailed studies on sources of variation of the assay in order to assure reproducibility of results [50].

Clinical validation/correlation

Most reports describing the development of a predictive classifier based on gene expression do not address analytical validation, they address *clinical correlation* sometimes referred to as *clinical validation*. For predictive classifiers developed from microarray gene expression data, it is essential to separate the cases used for developing the classifier from the cases used for evaluating the classifier. With traditional regression modeling where the number of candidate variables is much less than the number of cases, separation of training and validation cases is often not practiced. Failure to observe this key separation principle with microarray based classifiers, however, results in enormous bias in the resulting estimate of predictive accuracy [6].

The most straightforward way of ensuring separation is the *split-sample* method of partitioning the set of samples into a training set and a test set. Rosenwald et al. [51] used this approach successfully in their international study of prognostic prediction for large B cell lymphoma. They used two thirds of their samples as a training set. Multiple kinds of predictors were studied on the training set. When the collaborators of that study agreed on a single fully specified prediction model, they accessed the test set for the first time. On the test set there was no adjustment of the model or fitting of parameters. They merely used the samples in the test set to evaluate the predictions of the model that was completely specified using only the training data.

The split-sample method is often used with so few samples in the test set, however, that the validation is almost meaningless. One can evaluate the adequacy of the size of the test set by computing the statistical significance of the classification error rate on the test set or by computing a confidence interval for the test set error rate.

Michiels et al. [52] suggested that multiple training-test partitions be used, rather than just one. The split sample approach is most useful, however, when one does not have a well defined algorithm for developing the classifier. When there is a single training set-test set partition, one can perform numerous exploratory analyses on the training set and utilize biological information about the genes to develop a classifier and then test that classifier on the test set. With multiple training-test partitions however, that type of flexible approach to model development cannot be used. If one has an algorithm for classifier development, it is generally better to use one of the cross-validation or bootstrap resampling approaches to estimating error rate (see below) because the split sample approach does not provide as efficient a use of the available data [53].

Cross-validation is an alternative to the split sample method of estimating prediction accuracy [36] while preserving the key separation principle. With leave-one-out cross-validation, one omits one case and develops a predictive classifier on the remaining cases $n - 1$. That classifier is used to classify the omitted case and one records whether the prediction was correct or not. Then a different case is omitted, the one omitted the first time is included, and a new classifier is developed from scratch on the new training set of n-1 cases. That classifier is then used to classify the case omitted and one records whether the prediction was correct or not. This continues leaving each case out, one at a time, and the total number of mis-classifications determined. Molinaro et al. describe and evaluate many variants of cross-validation and bootstrap re-sampling for classification problems where the number of candidate predictors vastly exceeds the number of cases [53].

The cross-validated prediction error is an estimate of the prediction error associated with application of the algorithm for model building to the entire dataset. The model building process must be repeated from scratch for each loop of the cross-validation and so the process must be completely algorithmic. In particular, the gene selection must be repeated for each loop of the cross-validation. Simon et al. [6] showed that if you use the full dataset to select genes, and then cross-validate only the fitting of the prediction model for those genes, you obtain a highly biased estimate of prediction accuracy. Their results underscore the importance of cross-validating all steps of predictor construction

in estimating the error rate. Failure to do this is one of the most common and most serious errors made in using cross-validation [1].

It can also be useful to compute the statistical significance of the cross-validated estimate of classification error. This determines the probability of obtaining a cross-validated classification error as small as actually achieved if there were no relationship between the expression data and class identifiers. A flexible method for computing this statistical significance was described by Radmacher et al. [36]. This method of computing statistical significance of cross-validated error rate for a wide variety of classifier functions is implemented in the BRB-ArrayTools software [2].

Medical utility

The third level of validation of a predictive classifier is determining whether the classifier has medical utility. A classifier generally has medical utility only if it enables physicians to make better treatment decisions. Many classifiers are developed using a convenience sample of specimens not selected for purposes of addressing a question of medical decision making. Consequently they often include a heterogeneous group of patients who have received a variety of treatments [54]. For example, many prognostic factor studies in breast cancer include node negative and node positive ER negative and ER positive patients, those who received cytotoxic chemotherapy and those who received Tamoxifen alone. Showing that a new classifier is prognostic for such a mixed group generally has little apparent therapeutic value and such classifiers are rarely used [55]. It doesn't make the classifier therapeutically relevant to show in a multivariate analysis that the new classifier is more statistically significant than standard prognostic variables. Unless the cases represent the participants of a carefully selected clinical trial, there may be an insurmountable gap between clinical correlation and medical utility.

Classifiers are sometimes described as either *prognostic* or *predictive*. Prognostic factors provide information about the prognosis of a group of untreated or homogeneously treated patients. Predictive factors provide information about response or benefit from a specific treatment. Either kind of classifier derives medical utility, however, if it enables better treatment decisions. For example, OncotypeDx was developed as a prognostic classifier for patients with node negative, estrogen receptor positive breast cancer receiving Tamoxifen [50]. It's medical utility is based on identifying a subset of such patients whose long term disease-free survival is sufficiently good that cytotoxic chemotherapy might not be warranted. OncotypeDx is currently being tested in a very large prospective clinical trial. It can be more difficult to establish medical utility of a predictive classifier. For example, several major studies have reported gene expression classifiers for outcome following chemotherapy for diffuse large B cell lymphoma [51,56]. Withholding a potentially curative therapy from a patient based on an imperfect test where no established alternative therapeutic options are available is, however, difficult.

Simon et al. [57–59] have discussed prospective clinical trial designs for co-development of new drugs and companion diagnostics and these will not be reviewed here. Establishing the medical utility of a classifier for use of an established treatment can be more difficult. Medical utility depends on a variety of factors including other treatments available, availability of more easily measured predictive factors and practice standards [12]. It may be more difficult to conduct prospective trials that involve withholding widely used treatments.

In general, establishing medical utility requires demonstrating that a clinically meaningful measure of patient benefit is improved based on using the new classifier compared to not using the classifier. The direct approach would involve randomizing patients to treatment determination based on practice standards or based on the genomic classifier. The genomic classifier has clinical utility if treatment outcome is improved overall for the group randomized to classifier based treatment assignment. The genomic classifier also has clinical utility if outcome is the same for the two randomized groups but the patients randomized to classifier determined treatment have reduced adverse events, inconvenience or cost. This kind of prospective clinical trial design is very inefficient and rarely practical. It generally requires an enormous sample size because many or most patients in both randomization group receive the same treatment [60,61].

An alternative design is to measure the classifier on all eligible patients and determine before randomization whether recommended treatment assignment would differ between conventional

medical guidelines and the classifier based strategy. Then, the only patients randomized are those for whom the two strategies result in different treatment assignments. This approach entails the cost of measuring the classifier on all patients, but results in a much smaller clinical trial than that described above. This is the approach used for the design of the MINDACT clinical trial to prospectively evaluate the medical utility of a 70 gene expression signature for breast cancer [62]. By measuring the classifier on all eligible patients one can also compare practice standard treatment to classifier guided treatment separately for the patient subsets in which the assignments are discordant and that can improve the usefulness of the study. Nevertheless, such prospective randomized studies generally require very large numbers of patients and may require long follow-up times until results are available.

In some cases it may be possible to utilize archived tumor samples from patients treated in a randomized clinical to reasonably simulate the analysis that would have been performed in a prospective trial. This is a viable strategy only when archived specimens are available for a large proportion of the patients in an appropriately designed previously conducted randomized trial. Concern about whether the patients for whom samples are available are representative of the whole is minimized when adequate archived specimens are available on almost all patients, but even for prospective randomized trials there is always concern about whether randomized patients are representative of the entire population of patients. The retrospective strategy is not credible unless the plan for the retrospective analysis is completely specified in writing prior to performing assays on the archived specimens. The classifier must be completely determined by data external to the clinical trial used for retrospective analysis [63]. Because retrospective classification of archived specimens will not accurately reflect the challenges of tissue handling and assay performance encountered prospectively in a time frame that enables real-world treatment selection, it is important to separately establish the analytical validity of the assay for use with archived tissue.

Summary

DNA microarrays provide great opportunity for discovery and development of predictive oncology but also great opportunity for developing false claims. The review of the literature of use of DNA microarrays in studies of cancer outcome by Dupuy and Simon indicated that about 50 percent of studies contained at least one major flaw in the analysis serious enough to raise questions about the claims. Dupuy and Simon developed guidelines for the analysis of DNA microarray data in conjunction with outcomes of cancer patients, illustrated by a list of Do's and Don'ts [1]. BRB-ArrayTools software is a resource for improving the analysis of microarray expression data that can be useful for both biomedical investigators and statisticians. There are currently about 9000 registered users of this software in over 65 countries. It is freely available for non-commercial purposes from the National Cancer Institute at http://linus.nci.nih.gov/brb.

Conflict of interest statement

None declared.

References

*[1] Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 2007;99:147–57.
[2] Simon R, Lam A, Li MC, et al. Analysis of gene expression data using BRB-ArrayTools. Cancer Inform 2007;2:11–7.
[3] Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics 2006;22(7):789–94.
[4] Bolstad BM, Irizarry RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185–93.
[5] Park T, Yi SG, Kang SH, et al. Evaluation of normalization methods for microarray data. BMC Bioinformatics 2003;4:33.
[6] Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–8.
[7] Simon R, Korn EL, McShane LM, et al. Design and analysis of DNA microarray investigations. New York: Springer Verlag; 2003.
[8] Korn EL, McShane LM, Troendle JF, et al. Identifying pre-post chemotherapy differences in gene expression in breast tumors: a statistical method appropriate for this aim. Br J Cancer 2002;86:1093–6.

[9] Yang SX, Simon RM, Tan AR, et al. Gene expression patterns and profile changes pre- and post-erlotinib treatment in patients with metastatic breast cancer. Clin Cancer Res 2005;11:6226–32.
[10] Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population based study. Proc Natl Acad Sci U S A 2003;100(18):10393–8.
[11] Desai KV, Xiao N, Wang W, et al. Initiating oncogenic event determines gene-expression patterns of human breast cancer models. Proc Natl Acad Sci USA 2002;99:6967–72.
*[12] Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Cancer Inst 2005;97:866–7.
[13] van'tVeer LJ, Dai H, Vijver MJvd, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415:530–6.
[14] van-de-Vijver MJ, He YD, Veer LJvt, et al. A gene expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347(25):1999–2009.
[15] Ma XJ, Wang Z, Ryan PD, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:1–10.
[16] Ayers M, Symmans WF, Stec J, et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 2005;22(12): 2284–93.
[17] Jansen MPHM, Foekens JA, Staveren ILv, et al. Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. J Clin Oncol 2005;23(4):732–40.
[18] Perou CM, Serlie T, Eisen MB, et al. Molecular portraits of human breast tumors. Nature 2000;406:747–52.
[19] McShane LM, Radmacher MD, Freidlin B, et al. Methods of assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002;18:1462–9.
[20] Klebanov L, Yakovlev A. Is there an alternative to increasing the sample size in microarray studies? Bioinformation 2007; 1(10):429–31.
*[21] Shih JH, Michalowska AM, Dobbin K, et al. Effects of pooling mRNA in microarray class comparison. Bioinformatics 2004; 20:3318–25.
*[22] Dobbin K, Shih J, Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J Natl Cancer Inst 2003;95:1362–9.
*[23] Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005;6:27–38.
*[24] Dobbin K, Simon R. Sample size planning for developing classifiers using high dimensional DNA expression data. Biostatistics 2007;8:101–17.
[25] Wright GW, Simon R. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 2003;19:2448–55.
[26] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t test and statistical inferences of gene changes. Bioinformatics 2001;17:509–19.
[27] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 1995;57:289–300.
[28] Korn E, Troendle JF, McShane LM, et al. Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plan Inference 2004;124:379–98.
[29] Korn EL, Li MC, McShane LM, et al. An investigation of SAM and the multivariate permutation test for controlling the false discovery proportion. Statistics in Medicine 2007;26:4428–40.
*[30] Subramanian A, Tamayo P, Mootha VK. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–50.
[31] Tian L, Greenberg SA, Kong SW, et al. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A 2005;102(38):13544–9.
[32] Xu X, Zhao Y, Simon R. Gene sets expression comparison in BRB-ArrayTools. Bioinformatics 2008;24:137–9.
[33] Storey JD, Xiao W, Leek JT, et al. Significance analysis of time course microarray experiments. Proc Natl Acad Sci U S A 2005;102:12837–42.
[34] Leek JT, Monsen E, Dabney AR, et al. EDGE: extraction and analysis of differential gene expression. Bioinformatics 2006; 22:507–8.
[35] Simon R. When is a genomic classifier ready for prime time? Nat Clin Pract Oncol 2004;1(1):2–3.
*[36] Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–12.
[37] Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed T, editor. Statistical analysis of gene expression microarray data. London, New York, Washington D.C.: Boca Raton; Chapman & Hall/CRC; 2003.
*[38] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.
[39] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[40] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149–54.
[41] Ben-Dor A, Bruhn L, Friedman N, et al. Tissue classification with gene expression profiles. J Comput Biol 2000;7:559–84.
[42] Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567–72.
[43] Bo TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol 2002; 3(4):0017.1–0017.11.
[44] Ooi CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 2003;19:37–44.

[45] Deutsch JM. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics 2003;19: 45–54.
[46] Kim S, Dougherty ER, Barrera J, et al. Strong feature sets from small samples. J Comput Biol 2002;9:127–46.
[47] Lai C, Reinders MJT, Veer LJvt, et al. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 2006;7:235.
[48] Lecocke M, Hess K. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. Cancer Inform 2006;2:313–27.
[49] Dobbin K, Beer DG, Meyerson M, et al. Inter-laboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res 2005;11:565–72.
[50] Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
[51] 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 2002;346:1937–47.
[52] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple validation strategy. Lancet 2005;365:488–92.
*[53] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21(15):3301–7.
[54] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–85.
[55] Bast RC, Ravdin P, Hayes DF, et al. 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. J Clin Oncol 2001;19:1865–78.
[56] Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8:68–74.
*[57] Simon R. A roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23: 7332–41.
[58] Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 2005; 10:6759–63.
[59] Simon R. Using genomics in clinical trial design. Clin Cancer Res 2008;14:5984–93.
[60] Sargent DJ, Conley BA, Allegra C. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005;23(9):2020–7.
[61] Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol 2004;15: 1731–7.
[62] Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 2006;3(10):540–51.
[63] Simon R, Wang SJ. Use of genomic signatures in therapeutics development. Pharmacogenomics J 2006;6:166–73.