Using DNA Microarrays For Diagnostic and Prognostic Prediction

Richard Simon

Summary

DNA microarrays are a potentially powerful technology for improving diagnostic classification, treatment selection, and prognostic assessment. There are, however, many potential pitfalls in the use of microarrays that result in false leads and erroneous conclusions. Effective use of this technology requires new levels of inter-disciplinary collaboration with statistical and computational scientists. This paper provides a review of the key features to be observed in developing diagnostic and prognostic classification systems based on gene expression profiling. It also attempts to outline some of the steps needed to develop initial microarray research findings into classification systems suitable for broad clinical application.

Expert Rev. Mol. Diagn. 3(5), 587-595 (2003)

Keywords: Microarray, gene expression profiling, diagnostic classification, prognostic classification, pharmacogenomics

Richard Simon, D.Sc.
Chief, Biometric Research Branch
Division of Cancer Treatment & Diagnosis
National Cancer Institute
9000 Rockville Pike
MSC #7434
Bethesda MD 20892
Tel (301)496-0975
Fax (301)402-0560
rsimon@nih.gov

1. Introduction

DNA microarray experiments require planning. Planning is driven by experimental objectives. Good DNA microarray experiments have clear objectives. Microarray studies are usually not based on gene specific mechanistic hypotheses, but clear objectives ensure that they are focused investigations with a high likelihood of successfully answering important biomedical questions. One type of objective commonly encountered in DNA microarray experiments is identification of genes differentially expressed among pre-defined phenotypic classes of samples. We will refer to this as the *class comparison* objective. *Class prediction* is a related, but distinct, objective that is often relevant for medical studies. Class prediction involves development of a classification function that can accurately predict the biologic group, diagnostic category or prognostic stage of a patient based on an expression profile of tissue from that patient. With class comparison or class prediction, the phenotype classes are defined in advance independently of the gene expression data.

There are many published examples of class prediction. Hedenfalk et al. developed a predictor of whether a breast tumor came from an individual with a germline BRCA1 or BRCA2 mutation based on the gene expression profile of the tumor [1]. Golub et al. developed a predictor of whether an acute leukemia was of lymphoblastic or myloblastic type [2]. Wang et al. developed a predictor of whether a melanoma would respond to IL2 based treatment based on the gene expression profile [3].

There is a related class of prognostic prediction problems where the objective is to predict patient outcome based on the gene expression data e.g. Shipp et al. [4] and Rosenwald et al. [5]. This is similar to class prediction, although outcome may be measured continuously. Many of the methodologic issues pertinent to class prediction are also important for prognostic prediction. For simplicity of exposition, however, I will focus attention on the problem of class prediction for most of the paper and make some additional comments at the end about prognostic prediction.

Class discovery involves finding groupings of the samples that are relatively homogeneous with regard to gene expression, or finding grouping of the genes that appear to be co-expressed. In class discovery the classes are not pre-defined. An example of class discovery was the study by Bittner *et al.* examining expression profiles for advanced melanomas [6]. Alizadeh *et al.* [7] also performed class discovery in examining the expression profiles of patients with diffuse large B cell lymphoma. Often the purpose of class discovery is to discover clues to heterogeneity of disease pathogenesis.

DNA microarray technology is sufficiently mature to support the development of powerful diagnostic and prognostic classification. Because it is a genome-wide technology, it is tremendously powerful. Many laboratories are not well prepared to use the technology effectively, however, as it requires sophistication in study planning and data analysis beyond most previous single gene/protein assays. Our objective here is to highlight important aspects of the process of developing and evaluating gene expression based class predictors in order to facilitate the effective use of DNA microarray technology for clinical applications. Most of the methods recommended in this paper are available in the BRB-ArrayTools software developed by Simon and Peng-Lam [8] and available for non-commercial purposes without charge from the National Cancer Institute. We also discuss clinical translation of microarray based classification methods.

2. Inappropriateness Of Cluster Analysis For Class Prediction

Cluster analysis is widely used for all types of studies although it is often not effective for class comparison or class prediction. Cluster analysis refers to an extensive set of methods of partitioning samples into groups based on the pair-wise distances of their expression profiles. Cluster analysis is considered an *unsupervised* method because phenotype class information is not utilized. Cluster analysis is generally based on *global* gene expression. The pair-wise distance measures between expression-profiles are generally computed with regard to all genes represented on the array, or all genes which are well measured with sufficiently high signals. Since the genes that distinguish the particular classes of interest may be few in number relative to the full set of genes, the pair-wise distances used in cluster analysis may not reflect the influence of these relevant genes. Consequently, the clusters obtained may not be closely related to the phenotype classes of interest. Cluster analysis does not provide statistically valid quantitative information about which genes are differentially expressed between the classes. Investigators often use simple average fold change measures or visual inspection of a cluster-image display to provide information about differentially expressed genes, but these approaches do not account for variability in expression across samples nor do they account for multiple comparison issues. One of the most common errors in the analysis of DNA microarray data is use of cluster analysis and simple fold change statistics for problems of class comparison and prediction.

3. Components of Class Prediction

Most class predictors do not use all of the genes. One step in developing a class predictor is determining which genes to include in the predictor. This is generally called *feature selection*.

Feature selection is particularly important in microarray studies because the number of variables that are informative for distinguishing the classes of interest may be very small relative to the total number of genes represented on the array. The influence of the genes that actually distinguish the classes may be lost among the variation of the other genes unless we select the informative genes to be utilized by the class predictor.

The second main component of a class predictor is specification of the mathematical function that will provide a predicted class label for any given expression vector x. There are many kinds of predictor functions such as diagonal linear discriminant analysis, logistic regression [9], nearest neighbor predictors [10], support vector machines [11], decision trees [12], and neural networks [13].

Most classifiers predict the class $\hat{c}$ of a specimen based on a vector *x* of gene expression levels (log ratios or log signal values) and a vector of parameters *b*. For example the model can be written as $\hat{c} = f(x,b)$ where the function f corresponds to the type of model as described in the previous paragraph and in some cases may have a complex form with non-linear terms. The parameters in the *b* vector often represent weights assigned to the predictive variables included in the model. The parameters be assigned values before the predictor can provide specific predictions. These parameters are in many ways equivalent to the regression coefficients of ordinary linear regression. The machine learning literature calls the process of specifying the parameters "learning the data" but it is equivalent to fitting the parameters of a non-linear regression model. Even neural network models are essentially non-linear regression models, although they are often represented as something more exotic [14].

After selecting the kind of class predictor to be used, the predictor is fitted to a set of data. Before the model parameters can be determined, the genes must be selected. There is usually at least one parameter to specify for each genes included in the model. If the gene expression values are utilized in a non-linear manner, then there will be more than one parameter per gene. For example, if the model tries to represent the interaction of genes in determining class, then the number of parameters may exceed the square of the number of genes included in the model. For some kinds of predictors there is a cut-point that must be specified for translating a quantitative predictive index into a predicted class label (eg 0 or 1) for binary class prediction problems. Completely specifying the predictor means specifying all of these aspects of the predictor, the type of predictor, the genes included and the values of all parameters.

4. Estimating Accuracy of a Class Predictor

It is important to estimate the accuracy of class prediction for future samples for which the phenotype class is unknown? Knowing that there are highly statistically significant genes that are differentially expressed between the classes is not enough. We want to know how accurately we can predict which class a new sample is in. For a future sample, we will apply a fully specified predictor developed using the data available today. If we are to emulate the future predictive setting in developing our estimate of predictive accuracy, we must set aside some of our samples and make them completely inaccessible until we have a fully specified predictor that has been developed from scratch without utilizing those set aside samples.

To properly estimate the accuracy of a predictor for future samples, the current set of samples must be partitioned into a training set and a separate test set [15]. The test set emulates the set of future samples for which class labels are to be predicted. Consequently the test samples cannot be used in any way for the development of the prediction model. This means that the test samples cannot be used for estimating the parameters of the model and they cannot be used for selecting the genes to be used in the model. This later point is often overlooked.

The most straightforward method of estimating the accuracy of future prediction is the *split-sample* method of partitioning the set of samples into a training set and a test set as described in the previous paragraph. Rosenwald et al. [5] used this approach successfully in their international study of prognostic prediction for large 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 fully specified predictor, 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.

Cross-validation is an accepted alternative to the split sample method of estimating prediction accuracy when it is applied properly. There are several forms of cross-validation. Here we will describe *leave-one-out cross-validation (LOOCV)*. LOOCV starts like split-sample cross validation in forming a training set of samples and a test set. With LOOCV, however, the test set consists of only a single sample; the rest of the samples are placed in the training set. Cross-validation is similar to the split sample method in that the single sample in the test set is placed aside and not utilized *in any way* in the development of the class prediction model. Using only the training set, the informative genes are selected and the parameters of the model are fit to the data. Let us call M_1 the model developed with sample 1 in the test set. When this model is fully developed, it is used to predict the class of sample 1. This prediction is made using the expression profile of sample 1, but obviously without using knowledge of the true class of sample 1. Symbolically, if $\underline{x}_1$ denotes the complete expression profile of sample 1, then we apply model M_1 to $\underline{x}_1$ to obtain a predicted class $\hat{c}_1$. This predicted class is compared to the true class label c_1 of sample 1. If they disagree, then the prediction is in error. Comparing the predicted class $\hat{c}_1$ to the true class c_1 for sample 1 is an unbiased comparison because the expression profile for sample 1 was not used in any way for developing the model M_1 on which the prediction $\hat{c}_1$ was based.

Then a new training set - test set partition is created. This time sample 2 is placed in the test set and all of the other samples, including sample 1, are placed in the training set. A new model is constructed from scratch using the samples in the new training set. Call this model M_2 . Model M_2 will generally not contain the same genes as model M_1. Although the same algorithm for gene selection and parameter estimation is used, since model M_2 is constructed from scratch on the new training set, it will in general not contain exactly the same gene set as M_1. After creating M_2, it is applied to the expression profile $\underline{x}_2$ of the sample in the new test set to obtain a predicted class $\hat{c}_2$. If this predicted class does not agree with the true class label c_2 of the second sample, then the prediction is in error.

The process described in the previous paragraph is repeated n times where n is the number of biologically independent samples. Each time it is applied, a different sample is used to form the single-sample test set. During the n steps, n different models are created and each one is used to predict the class of the omitted sample. The prediction errors are totaled and that is the leave-one-out cross-validated estimate of the prediction error. With two classes, one can use a similar approach to obtain cross-validated estimates of the sensitivity, specificity, and ROC curve [16].

If we use all of the data to select genes and construct a model, there is no independent data left to validly estimate prediction error. A commonly used but completely invalid estimate is called the *re-substitution* estimate [17]. You use all the samples to develop a model M. Then you predict the class of each sample i using it's expression profile $\underline{x}_i$; $\hat{c}_i = M\left(\underline{x}_i\right)$. The predicted class labels are compared to the true class labels and the errors are totaled.

Simon et al. [17] performed a simulation to examine the bias in estimated error rates for class prediction. In a simulated data set, twenty expression profiles of 6000 genes were randomly generated from the same distribution. Ten profiles were arbitrarily assigned to "Class 1" and the other ten to "Class 2", creating an artificial separation of the profiles into two classes. Since no true underlying difference exists between the two classes class prediction will perform no better than a random guess for future biologically independent samples. Hence, the estimated error rates for simulated data sets should be centered around 0.5 (i.e, ten misclassifications out of twenty).

Figure 1 shows the observed number of misclassifications resulting from each level of cross-validation for 2000 simulated data sets. It is well-known that the re-substitution estimate of error is biased for small data sets and the simulation confirms this, with an astounding 98.2 % of the simulated data sets resulting in zero misclassifications even though no true underlying difference exists between the two groups. Moreover, the maximum number of misclassified profiles using the re-substitution method was only one.

Two types of leave-one-out cross-validation were studied. In one approach the differentially expressed genes to be used in the class predictor were selected using all of the data before starting the cross-validation process. This is partial cross-validation. With proper cross-validation, the gene selection is re-done for each leave-one-out training set.

Figure 1 shows that partial cross-validation is nearly as problematic as no cross-validation. Cross-validating the prediction rule after selection of differentially expressed genes from the full data set does little to correct the bias of the re-substitution estimator: 90.2 % of simulated data sets still result in zero misclassifications. It is not until gene selection is also subjected to cross-validation that we observe results in line with our expectation: the median number of misclassified profiles jumps to eleven, although the range is large (0 to 20).

The simulation results underscore the importance of cross-validating all steps of predictor construction in estimating the error rate. A study of breast cancer also illustrates the point: van 't Veer *et al.* [18] predicted clinical outcome of patients with axillary node-negative breast cancer (metastatic disease within 5 years versus disease-free at 5 years) from gene expression profiles, first using the re-substitution method and then using a fully cross-validated approach. The investigators controlled the number of misclassified recurrent cases (i.e., the sensitivity of the test) in both situations, so here we focus attention on the difference in estimated error rates for the disease-free cases. The improperly cross-validated method and the properly cross-validation result in estimated error rates of 27% (12 out of 44) and 41% (18 out of 44), respectively. The improperly cross-validated method results in a seriously biased under-estimate of the error rate, probably largely due to over-fitting the predictor to the specific data set. While van ′t Veer *et al.* report both estimates of the error rate, the properly cross-validated estimate was reported only in the supplemental results section on the website and the invalid estimate received more attention. Another example of this occurred in a study where classification trees were built from gene expression data to classify specimens as normal colon or colon cancer [19]. The authors used a procedure that only cross-validated steps that occurred *after* selection of genes for inclusion in the predictor from the full data set. As our simulation shows, not subjecting gene selection to cross-validation can result in a large bias. Other examples are described by Ambroise and McLachlan [20].

5. Class Prediction Algorithms

5.1 Feature selection

The most commonly used approach to feature selection is to identify the genes that are differentially expressed among the classes when considered individually. For example, if there are two classes, one can compute a t-test or a Mann-Whitney test for each gene. The log-ratios or log-signals are generally used as the basis of the statistical significance tests. The genes that are differentially expressed at a specified significance level are selected for inclusion in the class predictor. The stringency of the significance level 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.

Several authors have developed methods to identify optimal sets of genes which together provide good discrimination of the classes [21], [22], [23], [24]. These algorithms are generally very computationally intensive, some requiring a large cluster of parallel computers. Unfortunately, it is not clear whether the increased computational effort of these methods is warranted. In some cases, the claims made do not appear to be based on properly cross-validated calculations; all of the data being used to select the genes and cross-validation used only for fitting the parameters of the model. Thorough studies comparing the performance of such methods to the simpler univariate methods are needed.

Some investigators have used linear combinations of gene expression values as predictors [9] [25]. *Principal components* are the orthogonal linear combinations of the genes showing the greatest variability among the cases. The principal components are sometimes referred to as singular values [26]. Using principal components as predictive features provides a vast reduction in the dimension of the expression data, but has two serious limitations. One is that the principal components are not necessarily good predictors. The second problem is that measuring the principal components requires measuring expression of all the genes. The method of gene shaving attempts to provide linear combinations with properties similar to the principal components that does not require measuring all of the genes [27].

5.2 Class Prediction Algorithm

Many algorithms have been used effectively with DNA microarray data for class prediction. Dudoit et al. [12] compared several algorithms using publicly available data sets. The algorithms compared included nearest neighbor classification and several variants of linear discriminant analysis and classification trees. A linear discriminant is a function

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

where x_i denotes the log-ratio or log-signal for the i'th gene, w_i is the weight given to that gene, and the summation is over the set F of features (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 or greater than d respectively.

Several kinds of class predictors used in the literature have the form shown in (1). They differ with regard to how the weights are determined. The oldest form of linear discriminant is Fisher's linear discriminant [13]. 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. indicated that Fisher linear discriminant analysis did not perform well unless the number of selected genes was small relative to the number of samples; otherwise 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 case of Fisher linear discriminant analysis in which the correlation among genes is ignored [12]. By ignoring such correlations, one avoids having to estimate many parameters, and obtains a method which performs better when the number of samples is small. Golub's weighted voting method [2] and the compound covariate predictor of Radmacher et al.[15] are similar to diagonal linear discriminant analysis and tend to perform very 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 use a predictor of the form of equation (1). The weights are determined by optimizing an error rate criterion, however, instead of a least-squares criterion as in linear discriminant analysis [11]. 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 [28].

Khan et al.[25] reported accurate class prediction among small, round blue cell tumors of childhood using an artificial neural network. The inputs to the ANN were the first ten principal components of the genes; that is, the 10 orthogonal linear combinations of the genes that accounted for most of the variability in gene expression among samples. Their neural network used a linear transfer function with no hidden layer and hence it was a linear *perceptron* classifier of the form of equation (1). Most true artificial neural networks have a hidden layer of nodes, use a non-linear transfer functions and individual features as inputs. Such a "real" neural network would likely not perform as well as the linear model of Khan et al. because of the number of parameters to be estimated would be too large for the available number of samples.

In the study of Dudoit et al. [12], 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 feature set F of genes selected to be useful 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 of features F. 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. A variant of nearest neighbor classification is k-nearest neighbor classification. For example with 3-nearest neighbor classification, you find the three samples in the training set which are closest to the sample $\underline{y}$. The class which is most represented among these three samples is the predicted class for $\underline{y}$.

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 diagonal linear discriminant analysis or nearest neighbor classification. Ben-Dor et al. [28] also compared several methods on several public datasets and found that nearest neighbor classification generally performed as well or better than more complex methods.

6. Clinical Applicability

Many of the publications on microarray based diagnostic classification are proof-of-principle studies for which established diagnostic methods already exist. For example, distinguishing tumors of different organs or that originate in very distinct cell types is usually a relatively easy diagnostic problem by traditional histopathologic means. It is also usually an easy problem for microarray class prediction because there may be hundreds of differentially expressed genes. In some cases it may take very few samples of each class in order to develop an accurate class predictor..

Using gene expression profiles to distinguish among disease states that appear similar histologically or predicting response to treatment of patients with the same stage of a given type of disease are generally more difficult tasks requiring more samples. There are an increasing number of examples of success for clinically relevant problems. For example, Pomeroy et al. were able to use expression profiles to distinguish among several types of embryonal tumors of the central nervous system and to predict clinical outcome of children with medulloblastomas. Shipp et al. [4] and Rosenwald et al. [5] developed gene expression based predictors of long-term survival for patients with large B-cell lymphoma who received combination chemotherapy. Vasselli et al. [29] developed a gene expression based prognostic index for patients with kidney cancer, and Wang et al. [3] developed a gene expression based predictor of complete response to IL2 based therapy for patients with advanced melanoma. van 't Veer et al. have published a gene expression based class predictor of long term disease free survival for stage I patients with breast carcinoma [18] [30] and similar studies have been reported in a wide range of diseases.

6.1 Prognostic Markers

In oncology, in spite of the extensive literature on cancer prognostic markers, relatively few have been adopted into clinical practice. There are often conflicting literature reports on prognostic markers and the process of development of a prognostic marker is much less clearly defined than is the process of drug development [31]. Most of the problems that have hindered the study and development of prognostic markers exist for DNA microarray based expression profiles as markers. For example, there are multiple platforms and protocols for measuring expression profiles, most studies do not evaluate either inter-laboratory assay reproducibility or intra-laboratory reproducibility on multiple samples of the same tissue specimen.

Some of the problems that exist in the prognostic marker literature derive from the non-prospective nature of most marker studies. Clinical drug trials are generally prospective, with patient selection criteria, primary endpoint, hypotheses and analysis plan specified in advance in a written protocol. The consumers of clinical trial reports have been educated to be skeptical of *data dredging* to find something “statistically significant” to report in clinical trials. They are skeptical of analyses with multiple endpoints or multiple subsets, knowing that the chances of erroneous conclusions increase rapidly once one leaves the context of a focused single hypothesis clinical trial.

Prognostic marker studies are generally performed with no written protocol, no eligibility criteria, no primary endpoint or hypotheses and no defined analysis plan. The analysis is performed on tissue for previously treated patients for whom enough information and enough tissue is available. Consequently, the analysis is often much less structured, the patient population more heterogeneous, and there are many subset analyses.

Because of the number of genes available for analysis analysis, microarray data can be a veritable fountain of false findings unless statistical care is applied. This is one of the reasons why it is important that the authors of microarray based publications with clinical implications should make their raw data available for independent analysis by others. Independent analysis is not sufficient, however, for establishing clinical applicability of microarray findings.

The initial study should be conducted with attention to the statistical principles described in previous sections. For example, if the study has not been correctly internally cross-validated, then the claims are not likely to be worth confirming or attempting to translate to clinical applicability. Assuming that the initial study is performed properly, however, it might be considered a phase II study, and the next step should be to conduct a phase III study that is focused on testing the specific classifier developed by the initial study [31]. We will assume that the outcome of the initial study is a microarray based classifier that claims to distinguish patients who have good outcome following some specified type of treatment. The phase III study should be conducted with a written protocol. One of the components of the protocol is to define patient selection criteria so that the study is focused on a medically meaningful population. Often in prognostic marker studies the patient population is very broadly defined and the fact that a marker is "prognostic" may not have therapeutic relevance. The marker may be prognostic because it is correlated with disease stage or some other know prognostic marker. Broad populations are also often heterogeneously treated and so finding that a marker is prognostic in such a population may be difficult to interpret. Prognostic markers that do not have therapeutic implications are rarely used. Consequently, it is important to focus the phase III study on a patient population which is medically meaningful from a therapeutic point of view. The population should be adequately diagnosed and staged using conventional procedures and be relatively homogeneous with regard to diagnosis, stage and treatment.

The phase III trial should be designed to test the classifier developed in the previous study. The classifier should be fully specified in the protocol including the genes included, the mathematical form of the classifier, parameter values and cut-off thresholds for distinguishing the classes or prognostic groups.

The phase III study should attempt to perform the microarray assay in a manner as similar as possible to the way it would be performed broadly outside of a research setting if the diagnostic classifier were adopted. Consequently, careful thought is required in determining whether the same platform should be used for the phase III trial as for the phase II trial. If the platform is changed, then clearly some intermediate steps will be needed to translate the classification algorithm from use on the phase II platform to the platform used in the phase III trial.

Even if there is not a change in platform, intermediate steps may be required to prepare the classifier for use with multiple laboratories performing the array profiling. The phase II trial may have had all of the microarrays performed at a single location by a research laboratory and it may be advisable to conduct the phase III trial in a manner more similar to the way it would be performed if the classifier were adopted for national use.

Generally this will mean that several laboratories will be conducting the microarray assays. Consequently, the protocol for the phase III study should specify procedures to be used for conducting the assay. It is also useful to conduct pilot studies of inter-laboratory reproducibility by shipping pieces of tissue to different laboratories. This study can also address tissue heterogeneity issues. Unless inter-laboratory reproducibility is sufficiently high, it is not advisable to proceed with the phase III trial.

If the classifier was developed using a dual-label array platform, then use of the classifier in other laboratories requires that they use the same common reference RNA as was used for the initial study. Since different batches of the common reference will be utilized for classifying subsequent patients, calibration studies will generally be required to ensure that the expression profile of the common reference does not change and to adjust the classifier for small changes.

One important design issue for the phase III trial is whether the study will be performed with prospective accrual of patients or retrospectively based on frozen tissue. Prospective accrual is desirable for many reasons. One can never be sure that the patients for whom one has adequate preserved tissue are representative of the population of patients presenting for treatment. It is difficult to assure that a retrospective cohort was adequately staged and treated, and the data available may be incomplete. It is also difficult to assess whether a diagnostic procedure is practical unless it is studied in the real-time context of presenting patients who need to be evaluated and treated. Prospective accrual is also important for evaluating the diagnostic classifier in the context of real-time tissue handling. Because microarrays are RNA assays, tissue handling issues are of great importance.

The objective of the phase III trial is to test the hypothesis that the classifier can separate the uniformly staged and treated patients into groups of differing outcome. Consequently, the treatment should not generally be changed based on the microarray classifier. In some cases with other kinds of markers, phase III trials are designed to determine whether a risk based treatment assignment strategy can improve patient outcome. That is a more complex and study. For example, the classifier may divide patients into a predicted good prognosis group G, and a poor prognosis group P under standard therapy S. The clinical trial may be structured to assign standard treatment S to patients in G, but use a new experimental treatment E for patients in P. Such a trial would be difficult to interpret, however, because there is no concurrent control group for evaluating the prognosis based treatment assignment. One could randomize patients in P to either receive standard treatment S or experimental treatment E. This would not provide a test of the classifier. The new treatment E might have been superior to the standard S both for good prognosis and poor prognosis patients and the prognostic classifier might not be needed. By randomizing both good prognosis and poor prognosis patients to treatments E or S and comparing treatments within the two subsets separately one can determine whether risk based treatment assignment is useful, but this would require a very large study.

One could use a trial design to randomize all patients to either receive standard treatment S or to receive a prognosis group based treatment assignment. The latter might be for all patients in G to receive S and for all patients in P to receive E. Such a trial is properly controlled, but it has other defects. It would require a huge sample size because the good risk patients in both randomization groups receive the same treatment. Consequently the differences observed between the two groups will be limited. The other main defect is that this trial is really a trial of treatments S versus E for the poor risk patients. If experimental treatment E is not more effective than S for poor risk patients, then even a huge randomized trial will be negative. If E is better than S for poor risk patients, then the classifier is useful because it was used to identify the poor risk patients. A defect of the design, however, is that the evaluation of the classifier is intertwined with the evaluation of the new treatment E. On the other hand, this design tests whether the classifier is clinically useful for treatment selection.

Often, it may be better to confirm the effectiveness of the classifier for distinguishing risk groups without tying its evaluation to some specific new treatment. If the classifier can effectively classify patients into prognosis groups, then it may be useful to a range of different investigators who are studying a variety of new treatments. Consequently the simpler design of treating patients uniformly independent of the value of the classifier may be preferable. Such a study can be performed most rapidly using frozen tissue when available for an appropriate cohort of patients. In some cases it may be effective to first conduct a confirmatory trial based on archived tissue, but otherwise structured as a phase III trial as described here. If this trial confirms the effectiveness of the classifier, then a prospective trial could be considered.

6.2 Pharmacogenomic Markers

Two important developments in therapeutics are the use of molecularly targeted drugs and the growing recognition that many common diseases are molecularly and genetically heterogeneous. Although one hears debate about whether pharmaceutical companies want to develop drugs focused on molecularly defined subsets of disease populations, there are compelling advantages to doing so. If one has an a classifier for identifying a subset of patients likely to benefit from a given treatment, then a clinical trials focused on the subset become enormously more efficient than a trial that includes the full range of patients. In the context of treatments with side effects (most drugs), and treatments paid for by third parties (most drugs), it will become increasingly more difficult to develop drugs in the traditional manner. Molecularly targeted drugs will be developed on targeted populations, and the effective ones will have a greater benefit to side-effect ratio and benefit to cost ratio than many current drugs in non-targeted application.

Microarray classifiers distinguishing responders from non-responders can be developed during phase II clinical trials of a new treatment. The classifiers should be developed using the statistical practices described in previous sections of this paper, because the pace of drug development will dictate that the classifier be used in selecting patients for the randomized phase III clinical trials of the new treatment. It will be useful to have phase II experience with patients treated with the standard treatment S as well as those treated with the new treatment E so that a classifier can be developed to identify patients who are predicted to be more responsive to the new treatment E than to standard treatment S.

This approach is outlined in Figure 2 for a phase III superiority trial comparing the new treatment E to a standard treatment S with pre-screening of patients using the classifier developed in phase 2 trials. Only patients predicted to be responsive to treatment E are randomized. The treatment could potentially also be evaluated by a separate clinical trial in patients not predicted to be responsive to E using the classifier but this could be a difficult trial to get patients and physicians to participate in.

Having the new treatment E developed in the context of a targeted population obviously introduces complexities as well as potential benefits. If the targeted phase III trial is successful, a widely applicable assay is needed for the delivery of the new treatment broadly. This is always a challenge, and is certainly no simpler for an RNA based assay such as a microarray based classifier. The developmental steps for studying inter-laboratory reproducibility and intra-specimen reproducibility as described in section 6.1 are no less important in the context of targeted drug development.

References

1. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, al. e: Gene expression profiles of hereditary breast cancer. *New England Journal of Medicine* 344:539-548, 2001

2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531-537, 1999

3. Wang E, Miller LD, Ohnmacht GA, Mocellin S, Perez-Diez A, Petersen D, Zhao Y, Simon R, Powell JI, Asaki E, Alexander HR, Durray PH, Herlyn M, Restifo HP, Liu ET, Rosenberg SA, Marincola FM: Prospective molecular profiling of melanoma metastases suggests classifiers of immune responsiveness. *Cancer Research* 62:3581-3586, 2002

4. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, al. e: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8:68-74, 2002

5. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, al. e: The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 346:1937-1947, 2002

6. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, V.Sondak, Hayward N, Trent J: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406:536-540, 2000

7. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, al. e: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503-511, 2000

8. Simon R, Peng-Lam A: BRB-ArrayTools Users Guide, Version 3.0. *Technical Report, Biometric Research Branch, National Cancer Institute, http://linus.nci.nih.gov/brb*, 2003

9. West M, Blanchette C, Dressman H: Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Science* 98:11462-11467, 2001

10. Pomeroy SL, Tamayo P: Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature* 415:436-442, 2002

11. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science* 98:15149-15154, 2001

12. Dudoit S, Fridlyand J, Speed TP: Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97:77-87, 2002

13. Ripley BD: *Pattern recognition and neural networks*. Cambridge U.K., Cambridge University Press, 1996

14. Faraggi D, Simon R: A neural network model for survival data. *Statistics in Medicine* 14:73-82, 1995

15. Radmacher MD, McShane LM, Simon R: A paradigm for class prediction using gene expression profiles. *Journal of Computational Biololgy* 9:505-512, 2002

16. Swets J: Measuring the accuracy of diagnostic systems. *Science* 240:1285-1293, 1988

17. Simon R, Radmacher MD, Dobbin K, McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 95:14-18, 2003

18. Veer LJvt, Dai H, Vijver MJvd, He YD, Hart AA, Mao M, al. e: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530-536, 2002

19. Zhang H, Yu CY, Singer B, Xiong M: Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Science* 98:6730-6735, 2001

20. Ambroise C, McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Science* 99:6562-6566, 2002

21. Bo TH, Jonassen I: New feature subset selection procedures for classification of expression profiles. *Genome Biology* 3:0017.0011-0017.0011, 2002

22. Ooi CH, Tan P: Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 19:37-44, 2003

23. Deutsch JM: Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics* 19:45-54, 2003

24. Kim S, Dougherty ER, Barrera J, Chen Y, Bittner ML, Trent JM: Strong feature sets from small samples. *Journal of Computational Biology* 9:127-146, 2002

25. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, al. e: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7:673-679, 2001

26. Alter O, Brown PO, Botstein D: Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Science* 97:10101-10106, 2000

27. Hastie R, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1:0003.0001-0003.0021, 2000

28. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: Tissue classification with gene expression profiles. *Journal of Computational Biololgy* 7:559-584, 2000

29. Vassilli J, JH JHS, Iyengar SR, Maranchie J, Riss J, Worrell R, Torres-Cabala C, Tabios R, Mariotti A, Steraman R, Merino M, Walther MM, Simon R, RD RDK, Linehan WM: Predicting survival in patients with metastatic kidney cancer by gene expression profiling in the primary tumor. *Proceedings of the National Academy of Science* In Press, 2003

30. Vijver MJv, He YD, Veer LJvt, Dai H, Hart AAM, Voskuil DW, al. e: A gene expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 347:1999-2009, 2002

31. Simon R, Altman DG: Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 69:979-985, 1994

Figure 1

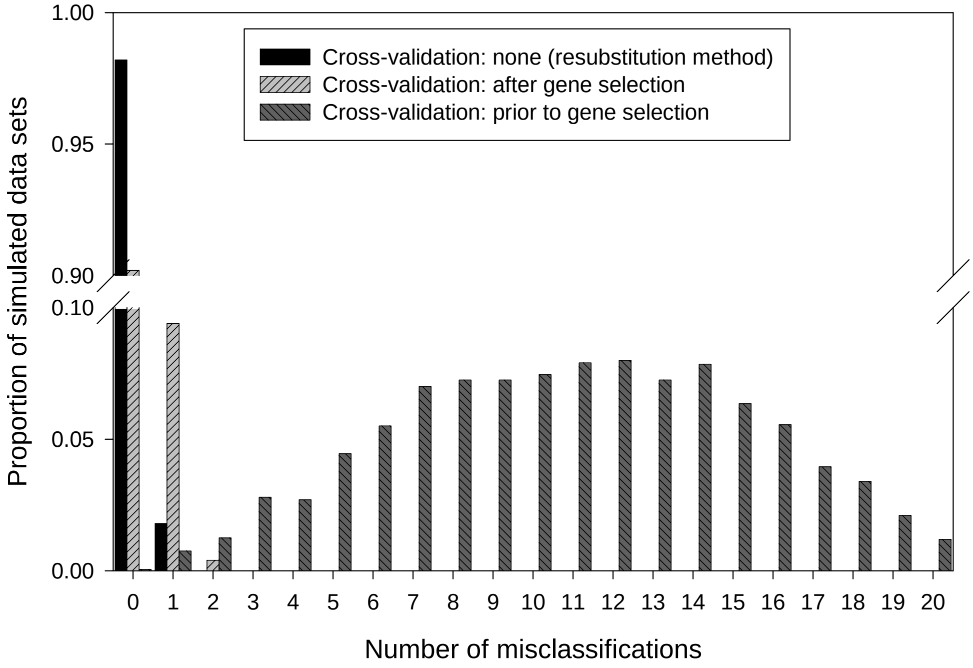
Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Number of misclassifications

Figure 2

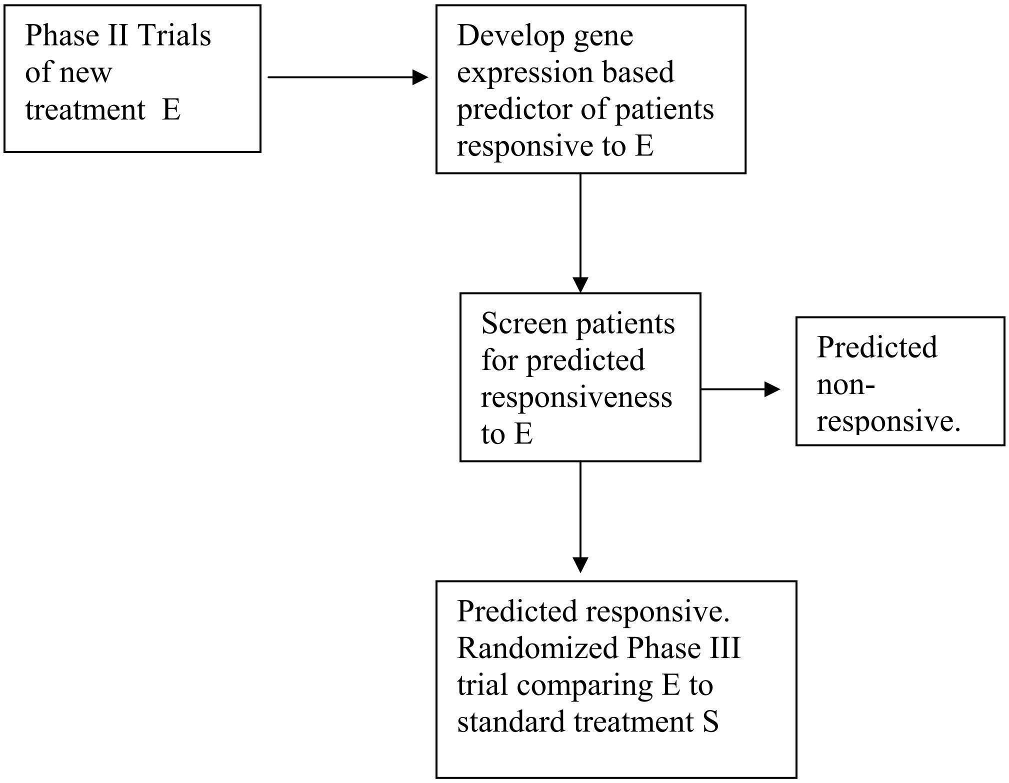
Phase II Trials of new treatment E
Develop gene expression based predictor of patients responsive to E
Screen patients for predicted responsiveness to E
Predicted non-responsive.
Predicted responsive. Randomized Phase III trial comparing E to standard treatment S

Figure Legends

Figure 1. Effect of various levels of cross-validation on the estimated error rate of a class predictor derived from 2000 simulated datasets. Class labels were arbitrarily assigned to the specimens within each dataset, so poor classification accuracy is expected. Class prediction was performed on each dataset as described in [17], varying the level of leave-one-out cross-validation. Vertical bars indicate the proportion of simulated datasets resulting in a given number of misclassifications. Figure previously appeared in Simon et al. [17].

Figure 2. Block diagram of strategy for utilizing gene expression profiling in clinical development of a new treatment (E). Gene expression based predictor of patients likely to respond to E is developed during Phase II studies. Only patients predicted to be responsive to E are entered into randomized Phase III trials comparing E to standard treatment S.

Key Issues

- DNA microarray technology is sufficiently mature for the development of medically important diagnostic and prognostic classification systems

- Microarray based classification systems are likely to be used only if they are therapeutically relevant
    - Therapeutic options are available
    - Patients included in microarray studies constitute a uniformly staged therapeutically meaningful group
- Microarray based response predictors developed during phase II treatment trials can serve for the selection of patients for highly efficient phase III trials

- DNA microarray data is often improperly analyzed, using cluster analysis where more powerful supervised methods are preferable

- Because the number of candidate genes that can be used in predictive models is orders of magnitude larger than the number of cases in most microarray studies:
    - Statistical methods must be properly used to avoid misleading claims.
    - Simple predictive models generally perform best for microarray data. The hype associated with complex models is often mis-guided.

- Invalid claims are often made for microarray based classification systems and classification algorithms because of a failure to properly cross-validate predictions
    - Cross-validation based on a pre-selected set of predictive genes is invalid

- Confirmatory phase III trials of fully specified gene expression based classification systems are needed before such systems are adopted into medical practice. The structure of such phase III trials is discussed.

Five Year Views

- DNA microarrays will be a standard laboratory tool for investigating biological mechanisms, discovering new disease taxonomies and developing therapeutically relevant diagnostic tests. .

- DNA microarrays will be an integral tool for clinical therapeutics in many disease areas. In clinical drug development, transcription profiling of specimens from responding and non-responding participants in phase II studies of a drug will enable response predictors to be developed. Such predictors will be used to select patients for phase III trials thereby increasing the efficiency of clinical development, and improving the therapeutic index of new drugs. This process will avoid many of the delays and uncertainties of developing assays for patient selection when biological mechanisms and pathways are incompletely understood.

- Microarray based predictive assays accepted into clinical practice will be those that are therapeutically relevant and biologically plausible.

- Elucidation of the biologic basis of disease and development of effective therapeutics will increasingly require inter-disciplinary teams of biologists and computational/statistical scientists working together in new organizational structures not currently found in industry, government or academia.