Development of biomarker classifiers from high-dimensional data

Songjoon Baek, Chen-An Tsai and James J. Chen
Submitted: 14th January 2009; Received (in revised form): 21st February 2009

Abstract
Recent development of high-throughput technology has accelerated interest in the development of molecular biomarker classifiers for safety assessment, disease diagnostics and prognostics, and prediction of response for patient assignment. This article reviews and evaluates some important aspects and key issues in the development of biomarker classifiers. Development of a biomarker classifier for high-throughput data involves two components: (i) model building and (ii) performance assessment. This article focuses on feature selection in model building and cross validation for performance assessment. A ‘frequency’ approach to feature selection is presented and compared to the ‘conventional’ approach in terms of the predictive accuracy and stability of the selected feature set. The two approaches are compared based on four biomarker classifiers, each with a different feature selection method and well-known classification algorithm. In each of the four classifiers the feature predictor set selected by the frequency approach is more stable than the feature set selected by the conventional approach.

Keywords: class prediction; cross-validation; feature selection; frequency of selection; stable feature set

INTRODUCTION
Recent advances in biotechnology have accelerated research in the development of molecular biomarkers of exposure, toxicity, disease risk, disease status and response to therapy. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacological responses to therapeutic or other interventions [1]. Biomarker exploration has centered around identifying and profiling genomic, proteomic and metabolomic endpoints coupled with appropriate phenotypic, physiological, pharmacogenomic, or clinical or surrogate endpoints for various applications. A common goal of biomarker studies is to develop a biomarker classifier that can be used for safety assessment, disease diagnostics and prognostics and prediction of response for patient assignment [2, 3]. High-throughput genomic, proteomic and metabolomic data are characterized by a large number of variables with a relatively small number of samples. In a typical microarray gene expression data set, the number of genes is in the tens of thousands but the number of samples rarely exceeds a few hundred and often less than one hundred. In most studies, the majority of genes are irrelevant to the treatments (or phenotypes) of interest; these genes may confuse the analysis and make the classification unnecessarily complex. Selection of a subset of relevant biomarker predictors to enhance predictive performance becomes an integral part in the development of biomarker classifier.

The most important consideration for evaluation of a biomarker classifier is whether it can accurately predict new samples based on the selected feature predictors. Two methods are commonly used to assess performance of a classifier: the split-sample procedure and cross-validation procedure. In the

Corresponding author. Dr James J. Chen, HFT-20, Jefferson, AR 72079, USA. Tel: +1-870-543-7007; Fax: +1-870-543-7662; E-mail: jamesj.chen@fda.hhs.gov
Songjoon Baek is a postdoctoral Fellow at the National Center for Toxicological Research, U.S. Food and Drug Administration. His research is into prediction algorithms.
Chen-An Tsai is an Associate Professor at the China Medical University, Taiwan. His research interests include bioinformatics, biostatistics, statistical machine learning, and the development of statistical approaches for medical data.
James J. Chen is Mathematical Statistician, National Center for Toxicological Research (NCTR), U.S. Food and Drug Administration. He heads the biostatistics program at NCTR. His current research interests are statistical methods for biomarker identification, and statistical modeling for quantitative risk assessment.
split-sample procedure, the sample dataset is randomly split into two subsets: a training set for model building and a test set for performance assessment. Because data are often insufficient to split into two samples for model building and testing, the cross validation procedure is commonly used for performance assessment. Cross validation involves repeatedly splitting the data into a training set containing most of the samples and applying the prediction rule to the test set of the remaining samples to estimate the prediction accuracy rate. The prediction accuracy estimate is the average accuracy of the numerous training-test partitions. The estimate is, presumably, the prediction accuracy rate to be expected for the classifier developed using the whole dataset. Since cross validation only uses a fraction of samples, it may be expected that the estimated accuracy rate may depend on the number of fold partitions in the splitting between the training and test samples. For example, the accuracy estimate from a leave-one-out cross validation is expected to be higher than the estimate from a 2-fold cross validation, since the 2-fold cross validation uses only half of the samples while the leave-one-out uses almost the entire sample in building the prediction model. Cross validation is described in the next section.

This article reviews some key aspects and issues in the development of biomarker classifiers. The primary focus is on the feature selection since this is the most fundamental problem in the development of a prediction model from high dimensional data. Four biomarker classifiers, based on four feature selection methods in conjunction with four well-known classification algorithms, are used in the evaluation. The predictive accuracies of the four classifiers with feature selection and without feature selection are compared. A frequency approach for selection of more stable feature predictors is presented and compared with the ‘conventional’ approach.

DEVELOPMENT OF BIOMARKER CLASSIFIER

A biomarker classifier is a mathematical function that translates the biomarker values to a set of categories [4]. These categories correspond to predicted outcomes. Let $D = (t_1, t_2, \ldots, t_n)$ be the sample data set consisting of $n$ labeled observations. Each sample consists of two parts, $t_i = (x_i, y_i)$, where $x_i$ is a vector of predictors from an unknown space, and $y_i$ is the class label (e.g., 0 or 1) for the observations. The problem is to predict a future unlabelled observation $x_0$ by the prediction rule $p(x_0 | D) = y_0$ (0 or 1) built on the observed dataset $D$.

Development of a biomarker classifier involves two steps: (i) model building and (ii) performance assessment. Typically, the data are divided into a training set and a test set; the prediction rule (model) is developed on the training set and then is used to classify samples in the test set to assess its performance.

Model building

The model building step can be summarized into four components: (i) selecting feature predictors; (ii) selecting a classification algorithm; (iii) specifying the training parameters of the prediction model; and (iv) fitting the prediction model to the training samples.

Because of the large number of predictors involved, many predictors are often noisy in nature and irrelevant for prediction. The use of all predictors to build the prediction model can suppress or reduce the performance of a classifier. Selection of a subset of the most informative predictor variables, called feature selection, is commonly addressed in classification [5, 6]. Feature selection is the most important task in the model building step with high dimensional data. Depending on classification algorithms, there are two general approaches to feature selection: filters and wrappers. The filter approach removes out irrelevant predictors according to some pre-determined criterion such as a $p$-value. Classical classification algorithms, such as Fisher’s linear discriminant analysis and k-nearest neighbor, often use the filter approach to select relevant predictors prior to classification [7–10]. The filter approach is a stand-alone prior step, regardless of which classification algorithm will be used. The affects of the selected predictors on the performance of the algorithm are not taken into account. The wrapper approach finds a subset of predictors and evaluates its relevance while building the prediction model. The classification algorithms, such as stepwise logistic regression, classification trees [11, 12] and support vector machines with recursive feature elimination [13] use the wrapper approach.

The second task in the model building is the selection of a classification algorithm. A prediction model is a mathematical function constructed on
After selecting the algorithm, the prediction model needs to be specified before model fitting. Model specification means specifying all aspects of the model including feature selection criteria, specific functional form of the prediction model, approach and criterion for performance assessment, etc. [17]. For example, support vector machines can use either the linear kernel or radial basis function; in the k-nearest neighbor classification, k can be pre-specified or estimated from cross validation. For those classification algorithms using the filter approach, the feature selection method must be specified. Classification algorithms with wrapper approach also need to specify the feature selection criterion such as the number of predictors to be selected. In many classification algorithms, a cut-point to translate a quantitative predictive index into a class label must be specified. For example, the logistic regression typically uses 0.5 as a cut-point to separate two classes based on an equal misclassification cost.

After selecting the prediction algorithm to be used, the prediction model is fitted to the training dataset. The objective is to search for a prediction function that optimizes the class separation. In model fitting, the parameters of the prediction model are estimated to minimize the probability of misclassification error when equal misclassification costs are assumed.

**Performance assessment**

The most important consideration in classification is the ability to predict future samples accurately. The current samples are used in two ways: (i) to develop the prediction model in the modeling building phase and (ii) to assess performance of the prediction model when it is applied to predicting future samples. The bias and variance are two common measurements to assess the performance assessment of a classifier. There two quantities depend on the prediction model and sample size. An important issue in the performance assessment is to provide unbiased estimates of error (or accuracy) rates of the classifier [18, 19]. The main principle to obtain unbiased estimates is that the samples used in the model building should be completely independent of the samples for performance assessment. Typically, the samples are partitioned into a training set and a separate test set. The training set is used for model development, and the test set is used for performance assessment. The test samples emulate the future samples for which class labels are to be predicted from the prediction model developed in the training phase. Cross validation is often used for performance assessment because it can generate a large number of different training and test sample partitions.

Cross validation involves repeatedly splitting the data into training and test datasets. The prediction accuracy estimate is the average accuracy of the numerous training-test partitions. Specifically, in V-fold cross-validation, the entire data set is divided into V subsets of roughly equal size, and the classification analysis is iterated V times. Each time, the prediction rule is trained on (V−1) subsets together and then applied to the remaining subset as the test dataset. (For example, a 10-fold cross-validation first creates 10 partitions (subsets) from the whole dataset. A prediction model is developed using nine subsets for training and then the model is applied to the remaining one for testing. The analysis repeats 10 times for each of the 10 partitions as the test dataset.) After completion of all V subsets (i.e., all samples are classified), the sensitivity, specificity and accuracy rates of the prediction model
are computed across all \( V \) subsets. The entire process can be repeated \( K \) times with different partitions of \( V \) subsets. The averages over the \( K \) repetitions are calculated. Note that when \( V \) equals the total sample size, the \( V \)-fold cross validation is known as the leave-one-out cross validation. In leave-one-out cross validation, one repetition (\( K = 1 \)) is done since repetition is not possible. It should be noted that cross validation needs to repeat all steps of model development, including feature selection and model construction within each stage. In using the filter approach, gene selection must be conducted in the training set to avoid selection bias [18, 19].

The cross validation estimate is presumably the accuracy to be expected to predict future samples for the classifier developed using the whole dataset. Since cross validation only uses a fraction of samples, its estimates may depend on the number of folds used. As discussed, the accuracy estimate from the leave-one-out cross validation may be expected to be higher than the estimate from the 2-fold cross validation; also, the leave-one-out estimate may better reflect the accuracy rate estimated from the whole dataset since it uses almost the entire dataset in training the classifier.

The four biomarker classifiers are applied to a publicly available colon tumor dataset [20] to illustrate the feature selection and \( V \)-fold cross validation: the support vector machine with recursive feature elimination (SVM-RFE) [13, 21], a random forest algorithm with mean decrease in accuracy (RF-MDA) [22], a diagonal linear discriminant analysis with between–within (DLDA-BW) [14], and a naïve Bayes with \( t \)-statistic (NB-T). A brief description of each of the four classifiers is given in the Supplementary Data.

### Example

**Colon cancer data.** The colon tumor data set [20] consists of 22 normal and 40 colon tumor tissue samples with more than 6500 genes profiled. This data set contains the expression levels of the 2000 genes. The data set is available at http://sdmc.lit.org.sg/GEDatasets/Datasets.html#ColonTumor.

Table 1 shows the accuracy with standard deviation of the four classifiers for leave-one-out, 10-, 5- and 2-fold cross validation from 100 repetitions. The upper panel shows the accuracy rates using all 2000 genes without feature selection and the lower panel shows the accuracy rates with a feature selection of 50 genes.

<table>
<thead>
<tr>
<th>Genes</th>
<th>CV method</th>
<th>SVM-RFE (%)</th>
<th>RF-MDA (%)</th>
<th>DLDA-BW (%)</th>
<th>NB-T (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 genes</td>
<td>LOO</td>
<td>85.48</td>
<td>82.26</td>
<td>64.52</td>
<td>61.29</td>
</tr>
<tr>
<td>10-fold</td>
<td>83.72 (2.55)</td>
<td>81.47 (2.46)</td>
<td>66.32 (2.79)</td>
<td>64.58 (3.38)</td>
<td></td>
</tr>
<tr>
<td>5-fold</td>
<td>83.31 (2.76)</td>
<td>79.50 (3.49)</td>
<td>66.58 (3.63)</td>
<td>64.23 (4.01)</td>
<td></td>
</tr>
<tr>
<td>2-fold</td>
<td>81.16 (4.31)</td>
<td>74.11 (6.29)</td>
<td>65.31 (8.10)</td>
<td>63.74 (6.81)</td>
<td></td>
</tr>
<tr>
<td>50 genes</td>
<td>LOO</td>
<td>83.87</td>
<td>85.48</td>
<td>88.71</td>
<td>75.81</td>
</tr>
<tr>
<td>10-fold</td>
<td>83.92 (2.33)</td>
<td>83.95 (2.59)</td>
<td>86.21 (2.84)</td>
<td>74.35 (1.38)</td>
<td></td>
</tr>
<tr>
<td>5-fold</td>
<td>83.29 (2.80)</td>
<td>83.55 (2.95)</td>
<td>82.82 (2.78)</td>
<td>74.03 (2.21)</td>
<td></td>
</tr>
<tr>
<td>2-fold</td>
<td>80.90 (4.60)</td>
<td>78.19 (5.37)</td>
<td>78.06 (4.82)</td>
<td>73.63 (3.11)</td>
<td></td>
</tr>
</tbody>
</table>

CROSS VALIDATION ACCURACY AND SELECTION OF STABLE FEATURE VARIABLES

In cross validation, different training samples often result in dissimilar feature sets from diverse data splits. The selected features are presumed to be most relevant to the sample distinction. Michiels et al. [23] showed that the list of genes identified as
predictors of cancer prognosis was highly unstable; the selected genes were highly dependent on the training samples. This section evaluates the performance of using cross validation to estimate predictive accuracy to predict future samples, and presents a frequency approach for selection of more stable feature predictors.

Several researchers have selected the feature variables based on the frequency of selections from the cross validation [24, 25]. The frequency of selections is a measure of the likelihood to be selected for the predictor. The most frequently selected predictors should be the most relevant and most stable predictor variables. The selected feature set from this approach would be different from the feature set selected from the whole dataset to be used for predicting future samples. Most importantly, the predictive model built from the feature set that composed of the most frequently selected predictors is not the model tested in the cross-validation. We compared the frequency approach with the ‘conventional’ approach by splitting the colon dataset into two subsets described below.

The data were randomly split, stratified by the tissue type, into two sets: set A and set B. Each set had 31 samples. The prediction accuracy rates were computed using set A as the training set to predict set B, and vice versa. Specifically, in the conventional approach, the prediction model consisting of 50 top-ranked predictors selected from the full training set (say, set A) is applied to its test set (set B) to estimate the prediction accuracy. In the frequency approach, the V-fold cross validation was conducted on set A. The classifier was trained on the (V – 1) subsets (in set A) with a selection of 50 top-ranked genes and then applied to the remaining subset (in set A). After completion of all V subsets, the prediction model was built from the 50 most frequently selected genes among the V cross validation. The prediction model was then applied to the test set (set B) to estimate the prediction accuracy. The procedure was repeated 100 times for different splits of set A and set B. Table 2 shows the averages of accuracy for the predictions of A to B (A → B) and B to A (B → A) over the 100 repetitions.

These results are consistent with the results based on the original full dataset shown in Table 1. The SVM-RFE appears to have the highest accuracy followed by RF-MDA and DLDA-BW. The NB-T has the lowest accuracy. Note that the conventional approach in Table 2 is one-half of the 2-fold cross validation in Table 1, while the 2-fold cross validation in Table 1 is the averaged accuracy from A to B and B to A. The two accuracy estimates in Tables 1 and 2 are close for all four classifiers. Table 2 shows that the predictive accuracy based on the most frequently selected predictors appears to be comparable with the accuracy estimate that would be obtained using the whole dataset from 2-fold cross validation (Table 1) or the conventional approach (Table 2).

The distribution of the frequency of the selected genes provides a measure of stability of a feature selection method. The frequency distributions of selections of 50 genes over the 100 repetitions were evaluated for the two approaches. The frequencies of the selections from the two predictions (A → B) and (B → A) were combined. Thus, there were a total of 200 selections with 50 genes in each selection. If all 2000 predictors have no discriminability, then the expected frequency of selections in each repetition is (200 x 50)/2000 = 5 for each gene. The

<table>
<thead>
<tr>
<th>CV method</th>
<th>Direction</th>
<th>SVM-RFE (%)</th>
<th>RF-MDA (%)</th>
<th>DLDA-BW (%)</th>
<th>NB-T (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convention (W/O CV)</td>
<td>A → B</td>
<td>82.26 (7.27)</td>
<td>77.48 (8.25)</td>
<td>77.51 (8.99)</td>
<td>72.87 (6.60)</td>
</tr>
<tr>
<td></td>
<td>B → A</td>
<td>79.74 (6.78)</td>
<td>78.45 (8.43)</td>
<td>78.55 (9.29)</td>
<td>73.11 (6.28)</td>
</tr>
<tr>
<td>Frequency LOO CV</td>
<td>A → B</td>
<td>82.74 (6.88)</td>
<td>78.03 (8.80)</td>
<td>78.61 (8.82)</td>
<td>72.45 (6.67)</td>
</tr>
<tr>
<td></td>
<td>B → A</td>
<td>80.26 (6.86)</td>
<td>77.92 (8.78)</td>
<td>79.63 (8.95)</td>
<td>71.53 (6.59)</td>
</tr>
<tr>
<td>Frequency 10-fold CV</td>
<td>A → B</td>
<td>82.93 (6.63)</td>
<td>74.75 (8.36)</td>
<td>78.39 (10.1)</td>
<td>72.06 (6.92)</td>
</tr>
<tr>
<td></td>
<td>B → A</td>
<td>80.87 (6.48)</td>
<td>76.97 (8.18)</td>
<td>79.08 (8.99)</td>
<td>71.63 (6.81)</td>
</tr>
<tr>
<td>Frequency 5-fold CV</td>
<td>A → B</td>
<td>83.16 (5.95)</td>
<td>76.76 (8.07)</td>
<td>76.97 (9.99)</td>
<td>73.61 (7.08)</td>
</tr>
<tr>
<td></td>
<td>B → A</td>
<td>81.51 (6.16)</td>
<td>77.72 (7.99)</td>
<td>77.81 (10.0)</td>
<td>74.42 (7.22)</td>
</tr>
<tr>
<td>Frequency 2-fold CV</td>
<td>A → B</td>
<td>84.13 (5.77)</td>
<td>75.06 (9.38)</td>
<td>79.48 (8.25)</td>
<td>73.32 (8.40)</td>
</tr>
<tr>
<td></td>
<td>B → A</td>
<td>81.90 (6.36)</td>
<td>75.16 (9.37)</td>
<td>80.07 (8.24)</td>
<td>72.88 (8.16)</td>
</tr>
</tbody>
</table>

The data are randomly split into two sets, A and B. The prediction accuracy is computed using set A as the training set to predict set B, and vice versa, over the 100 repetitions.
maximum number of selections for a discriminatory gene is 200. Table 3 shows the cumulative frequencies (in step of 10) of the 50 most selected genes from the two approaches. In both approaches, NB-T shows the most consistency. However, NB-T has the poorest predictive accuracy (Table 2). On the other hand, SVM-RFE shows the least consistency, but it has the highest predictive accuracy. Within each algorithm, the frequency approach shows better consistency than the conventional approach. Since the primary determinant of the utility of a classifier is its ability to predict future samples as accurately as possible, a feature selection method should have a greater than or comparable predictive accuracy. Given the colon dataset and using the frequency approach, SVM-RFE has the highest accuracy (Table 2) and the most stable predictor set (Table 3), while NB-T has the least accuracy. Also, for all classifiers, the frequency approach produces more stable feature sets than the conventional approach while maintaining comparable accuracy. It should be noted that for a different dataset, a different classification algorithm may be the most accurate. Perhaps, using the frequency approach will provide the most stable predictor set.

The 2-fold cross validation shows the best consistency (stability) in selection of feature predictors for all four classifiers. The consistency decreases as the fold partition increases. Both the sample sizes and correlations among predictors can have effects on the consistency of the selection of feature predictors. In general when the training sample size is large, such as LOO, and/or the correlations among the predictors are large, then the selected predictors become less consistent (instable). Because of heterogeneity among samples, there may be a number of discriminatory genes. Also, genes correlated with these discriminatory genes can also separate samples. This also explains instability of the feature set in microarray classifiers. Furthermore, both t-statistic and BW ratio perform univariate gene-by-gene analysis; the ranking of variable importance is less affected by the partition size and correlations in the univariate analysis.

**DISCUSSION**

Classification is an area of scientific endeavor that involves the use of statistical learning techniques to
develop algorithms for classifying unknown samples through supervised training on samples of known class. Classification has received renewed attention in recent years due to advancements in biotechnology that facilitate the use of high-dimensional genomics information in biomedical decision making for better diagnosis and treatment of disease. The U.S. Food and Drug Administration envisions clinical profiling to identify patients most likely to benefit from particular drugs and patients most likely to experience adverse reactions [26, 27]. Ideally, patient treatment should be based on an individual’s disease characteristics and risk factors. Large inter-individual differences highlight the need to develop predictive biomarkers for selecting the right treatment for the right patient. Clinical pharmacogenomic profiling of individual patients is a promising source of biomarker classifier which will enable assignment of drug therapies on a scientifically sound predictive basis to individual patients and replace one-size-fits-all practice.

Classification is a commonly used method for determining relationships between genes/proteins or gene-clusters/protein–complexes for identifying biological functions or predicting specific biological outcomes (or diseases). Given the nature of the high dimensionality, feature selection is essential in order to enhance relationships among the underlying structures or to improve prediction accuracy. Selection of a subset of predictors is often used in attempts to identify a minimum number of non-redundant predictors that are useful. Ideally, these predictors should contain independent information for prediction in order to minimize the number of leads or to provide a simple prognostic/diagnostic tool. Typically there are millions of subsets so an exhaustive search is computationally prohibitive. The set of predictors selected may vary substantially among classifiers or even among different iterations in cross validation, although the models predict about equally well. It is not feasible to come up with a general procedure to determine the optimal predictor set combined with a classification algorithm that gets the best accuracy. The optimal predictor set may depend on a classification algorithm and can vary from data to data.

Recently, Baek et al. [28] investigated several feature selection methods using the filter approach. Wrapper feature selection typically is a built-in a procedure for specific classification algorithms, such as the SVM-RFE and RF-MDF in this article. However, their investigation included several tree-based wrapper methods and the (SVM-)RFE wrapper method in combination with the DLDA, NB, kNN and Shrunken Centroids [29] classification algorithms in the evaluation. Baek et al. [28] also showed that feature selection improved prediction accuracy for some classification algorithms, and the ‘optimal’ number of predictors varied with the fold partition, feature selection method and classification algorithm. Their results showed that the simple $t$-statistic and BW ratio tended to perform the best overall for two datasets analyzed. This also re-confirms that the optimal predictor set can vary from data to data.

The accuracy rate is estimated by applying the prediction model based on the current samples that presumably emulate the population of future samples for which the class labels are to be predicted. If the samples in the present study do not adequately represent the future samples, then the estimates of prediction accuracy can be biased. For example, if the sample size in the present study is inadequate to reflect the variability that might be seen for subsequent samples, then the estimated prediction accuracy is likely too optimistic. Because of heterogeneity in patient populations and complexity of the disease and genetic and genomic factors, ideally, it is best to have a sufficiently large collection of data composed of patients representing the patients for which the classifier will be applied in the future. This article focuses on the use of cross validation for performance assessment in the development of classifiers. However, the use of an independent test set is a gold standard in clinical and preclinical validation of biomarker classifiers.

Sensitivity and specificity are the measures of a classifier performance estimated based on the number of positive and the number of negative samples from the current sample data. In practice, we are more interested in the probability that the sample is positive given a positive prediction, positive predictive value and the probability that the sample is negative given a negative, positive predictive value. These two measures depend on the underlying prevalence of the outcome categories. These two measures are the most directly useful indicators and can be directly linked to the specific clinical purpose of the prediction.

Most classification algorithms are designed to work only with two class (binary) problems. A standard way to deal with multi-class problems is
to consider them as a collection of binary subproblems, and then combine their solutions. In this context, two approaches are most commonly employed: the one-versus-all (OVA) and the one-versus-one (OVO) [30]. Scholkopf and Smola [31] indicated that there was probably no clear winner for multi-class problem. Usually, the performance of feature selection methods was data-dependent due to computational demands of many tuning parameters. Furthermore, Li et al. [32] showed that multi-class problem was much more complicated than the binary one for gene expression datasets in a comparative study. The difficulties are due to the large number of genes relative to the number of samples. As a result, the classification accuracy degrades very rapidly as the number of classes increases. The dataset with a large number of classes shows a lower accuracy value regardless of the feature selection and classification methods.

In this article, we use the microarray colon data for illustration. Microarray gene expression profiles can provide more information than classic morphology and provide an alternative to morphology-based tumor classification systems. Several studies have shown the fundamental power of microarrays to analyze gene expression in colon, breast, and other tumors, and these studies have demonstrated the potential utility of expression profiling for classifying tumors [20, 33, 34]. However, microarray data produce an excessive amount of data which presents challenges of selecting a subset of informative genes to build a classifier with either clear biological interpretation or some implication in the molecular mechanism of the tumorigenesis. Many algorithms have been developed to classify disease types based on the expression of selected genes, and significant gains have been made in the accuracy of disease classification [35, 36]. Many studies have shown that improved performance can be achieved when using a selected subset of features, as opposed to using all available data [37–39]. Increases in accuracy achieved through the selection of predictive features can complement and enhance the performance of classification algorithms, as well as improve the understanding of disease classes by identifying a small set of biologically relevant features [34].

In microarray experiments, biomarkers often refer to differentially expressed genes. The differentially expressed genes are typically identified using a univariate statistical significance test, such as t-test, by controlling the false positive error. Differentially expressed genes can be used to develop biomarker classifiers (filter approach). Also, differentially expressed genes are potential biomarkers for specific endpoints. These potential biomarkers, individually or as a set, may be further developed to be probable biomarkers or valid biomarkers for clinical use [40]. On the other hand, a biomarker classifier consists of a set of discriminatory genes described by a specified prediction model. The individual genes in the predictor set are not biomarker, but they may be developed as biomarkers for specific endpoints. In general, there is no theoretical estimation of the optimal number of selected predictors for a given specific classification algorithm on a particular application. In some applications, simply identifying exactly the optimal predictor set may not be as important as developing a classifier that performs the clinical prediction accurately. It is preferable to have all the selected predictors be informative and the established classifier be biologically interpretable. When two procedures have similar prediction accuracy, the procedure which gives a more stable gene set is more desirable (Table 2). In this article, we propose a frequency approach for feature selection to obtain a more stable predictor set while maintaining comparable accuracy.

**Key Points**

- A biomarker classifier is developed to classify a risk factor or predict a response of a new sample from the available data, including and not limited to genomics, proteomics and metabolomics.
- Development of a biomarker classifier involves two phases: (i) model building and (ii) performance assessment. The available data are partitioned into a training set and a separate test set. The training set is used for model development, and the test set is used for performance assessment.
- Feature selection is used in an attempt to identify the most predictive subset of feature variables among all possible subsets. Feature selection is an essential part of the development of classifiers from high dimensional data.
- There are two general approaches to feature selection: filters and wrappers. The filter approach selects the feature variables in a prior step before building the prediction model. The wrapper approach selects feature variables while building the prediction model.
- The primary determinant of the utility of a classifier is its ability to predict future samples as accurately as possible, a feature selection method should have a greater than or comparable predictive accuracy as compared to the alternatives.
SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The views presented in this article are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration. The authors would like to thank three reviewers for providing valuable comments and suggestions that improved this article considerably.

FUNDING

China Medical University of Taiwan, ROC (grant CMU 96-127, in part).

References


