BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1636–1643
DOI: 10.1093/bioinformatics/btg210

Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data

Baolin Wu[1], Tom Abbott[2], David Fishman[5], Walter McMurray[2], Gil Mor[3], Kathryn Stone[2], David Ward[4], Kenneth Williams[2] and Hongyu Zhao[1,4,*]

[1]Department of Epidemiology and Public Health, [2]HHMI Biopolymer/Keck Laboratory, [3]Department of Obstetrics and Gynecology, [4]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA and [5]Department of OB/GYN, Northwestern University School of Medicine, Chicago, IL, USA

Received on December 18, 2002; revised on March 3, 2003; accepted on March 6, 2003

ABSTRACT

Motivation: Novel methods, both molecular and statistical, are urgently needed to take advantage of recent advances in biotechnology and the human genome project for disease diagnosis and prognosis. Mass spectrometry (MS) holds great promise for biomarker identification and genome-wide protein profiling. It has been demonstrated in the literature that biomarkers can be identified to distinguish normal individuals from cancer patients using MS data. Such progress is especially exciting for the detection of early-stage ovarian cancer patients. Although various statistical methods have been utilized to identify biomarkers from MS data, there has been no systematic comparison among these approaches in their relative ability to analyze MS data.

Results: We compare the performance of several classes of statistical methods for the classification of cancer based on MS spectra. These methods include: linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbor classifier, bagging and boosting classification trees, support vector machine, and random forest (RF). The methods are applied to ovarian cancer and control serum samples from the National Ovarian Cancer Early Detection Program clinic at Northwestern University Hospital. We found that RF outperforms other methods in the analysis of MS data.

Contact: hongyu.zhao@yale.edu

Supplementary information: http://bioinformatics.med.yale.edu/proteomics/BioSupp1.html

1 INTRODUCTION

In the past several years, microarray technology has attracted tremendous interest as it provides the potential ability to monitor the expression of an entire genome on a single chip. Thus, researchers can generate a ‘snap shot’ view of the expression level of thousands of genes simultaneously (Nature Genetics, 1999). However, microarray technology is inherently limited. It is directed at analyzing mRNA rather than the actual biological effector, which usually is the resulting protein molecule; it is blind to the array of protein post-translational modifications (e.g., phosphorylation) which often modulate protein function; levels of mRNA expression often correlate poorly with the actual *in vivo* protein concentration due to differential rates of mRNA translation and varying protein half-lives.

Proteins, which carry out and modulate the vast majority of chemical reactions which together constitute ‘life’, are really our targets of interest. Proteomics is an integral part of the process of understanding biological systems, pursuing drug discovery, and uncovering disease mechanisms. Because of their importance and very high level of variability and complexity, the analysis of protein expression and protein : protein interactions is as potentially exciting as it is a challenging task in life science research (Science, 2001). Comparative profiling of protein extracts from normal versus experimental cells and tissues enables us to potentially discover novel proteins that play important roles in disease pathology, response to stimuli, and developmental regulation. However, to conduct massively parallel analysis of thousands of proteins, over a large number of samples, in a reproducible manner so that logical decisions can be made based on qualitative and quantitative differences in protein content is an extremely challenging endeavor.

Mass spectrometry (MS) is increasingly being used for rapid identification and characterization of protein populations. The relative ease of operation of matrix assisted laser desorption ionization (MALDI) coupled with time-of-flight detection and its characteristic generation of (mostly) singly charged peptide and protein ions makes this MS platform the current method of choice for disease biomarker discovery. MALDI-MS uses

[*]To whom correspondence should be addressed.

a nitrogen UV laser (337 nm) to generate ions from high mass, non-volatile samples such as peptides and proteins. The key to this technique, which was discovered several years ago, is that in the presence of an energy absorbing matrix like α-cyano-4-hydroxy cinnamic acid (CHCA), large molecules like peptides ionize instead of decomposing. In this technique, purified or partially purified proteins are mixed with a crystal-forming matrix, placed on an inert metal target, and subjected to a pulsed laser beam to produce gas phase ions that traverse a field-free flight tube and then are separated according to their mass/charge ratio (m/z). The key is that the time is proportional to the square root of m/z. Since MALDI has the inherent advantage that most ions are singly charged, the mass of the analyte usually is equal to m/z and each analyte usually produces only a single ion type. By recording the time-of-flight, we can measure the mass of the peptide/protein ions. The resulting data format is very simple: paired mass/charge ratio versus intensities. (See http://info.med.yale.edu/wmkeck/prochem/procmald.htm for more information on MS.)

MS data sets are increasingly used for protein profiling in cancer research. An important goal of this endeavor is the ability to predict cancer on the basis of peptide/protein intensities. The identification of phenotypic expression patterns that correlate strongly to a defined pathological condition might well represent a significant step towards early detection and/or the development of novel therapies in which these molecules might serve as clinical targets (Fung *et al.*, 2000).

As MS is increasingly used for protein profiling, significant challenges have arisen with regard to analyzing the data sets. These include peak identification and alignment, MS spectrum normalization, and data set visualization, among others. These pre-processing steps are arguably critical and we are currently evaluating them carefully. The final and most important step is the classification of disease status based on MS data. Recent publications on cancer classification using MS data sets have mainly focused on identifying biomarkers in serum to distinguish between cancer and normal samples. Basically these studies first use approximate criteria to select a subset of variables acquired on a 'training set' of samples, statistical methods are then applied to this subset to identify the most important variables as biomarkers. Finally, the performance of these biomarkers is determined based on their ability to classify samples. The statistical methods used to select biomarkers include T-statistics (Guoan *et al.*, 2002), classification methods such as trees (Bao-Ling *et al.*, 2002), genetic algorithms and self-organizing-maps (SOM) (Petricoin *et al.*, 2002), and artificial neural network (Ball *et al.*, 2002). There is a rapidly growing literature on the use of MS to identify peptide and protein biomarkers. Virtually all of these reports are based on MS data obtained via surface enhanced laser desorption ionization (SELDI) and the use of only a single methodology to identify the biomarkers. Here we utilize MALDI MS to obtain the data set and then use it to compare the performance of several well-known classification methods which can be reasonably easily adapted to the analysis of MS data sets. It is our ongoing endeavor to evaluate their relative performance in analyzing MS data sets. Although it is a very important aspect of biomarker discovery, this report does not cover various options for biomarker selections. Rather, we will compare two criteria for biomarker selection and then use selected biomarkers to compare several classification methods. We review below some classification methods and assess their performance on an ovarian cancer MALDI-MS data set obtained by the Keck Laboratory at Yale University (http://info.med.yale.edu/wmkeck/) as described in Section 4. Figure 1 shows the MALDI-MS data from one cancer and one normal serum sample.

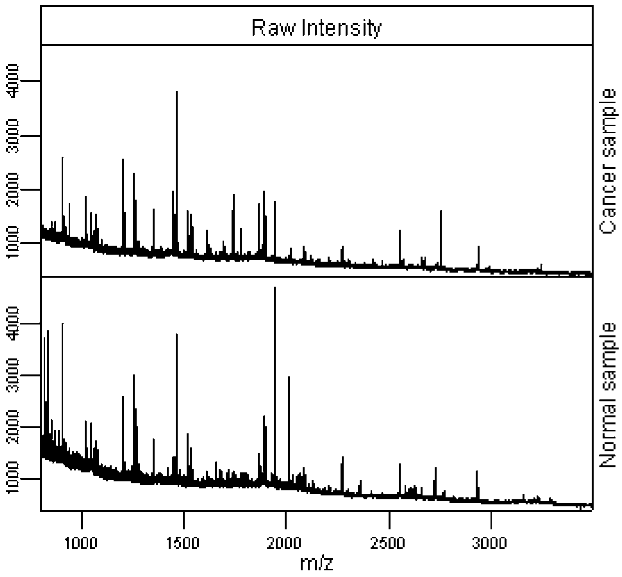

Fig. 1. MALDI-MS sample plots.

2 MATERIALS AND METHODS

2.1 Discriminant methods

For our following discussion, we can summarize our MS data set for n samples in a $p \times (n + 1)$ matrix: $(mz, \mathbf{X}) = (mz, X_1, \ldots, X_n)$ where p is the number of m/z ratios observed, mz is a column vector denoting the measured m/z ratios, and the X_i are the corresponding intensities for the ith sample. We also have a vector $\mathbf{Y} = (y_1, \ldots, y_n)$ to denote the sample cancer status. Our goal is to predict y_i based on the intensity profile $X_i' = (x_{1i}, x_{2i}, \ldots, x_{pi})$. For our ovarian cancer data set, there are two classes, cancer or normal, and the class labels y_i can be defined as 1 or 2, respectively. Generally we can have g classes, and the goal of statistical analysis is to use the class information to reveal the structures of the data. A predictor or classifier partitions the space $\mathbf{X}$ of protein intensity profiles into two disjoint subsets, A_1 and A_2, such that for a sample with intensity profile $X = (x_1, \ldots, x_p) \in A_j$ the predicted class is j.

Classifiers are built from observations with known classes, which comprise the learning set (*LS*) $L = \{(X_1, y_1), \ldots, (X_{n_L}, y_{n_L})\}$. Classifiers can then be applied to a test set (*TS*) $T = \{X_1, \ldots, X_{n_T}\}$, to predict the class for each observation. If the true classes y are known, they can be compared with the predicted classes to estimate the error rate of the classifiers.

We denote a classifier built from a learning set L by $C(\cdot, L)$; the predicted class for an observation x is $C(x, L)$. Below we briefly review several well-known discrimination methods which are compared in our study. Most of the methods discussed below have also been compared in the context of using microarray data to distinguish various cancer types (Dudoit *et al*., 2002). General references on the topic of discriminant analysis include Mardia *et al*. (1979), McLachlan (1992), and Ripley (1996).

Linear discriminant analysis (*LDA*), *quadratic discriminant analysis* (*QDA*). LDA (linear discriminant analysis) was first described by Fisher (1936). It seeks a linear combination xa of the sample intensity $X = (x_1, \ldots, x_p)$ which has a maximal ratio of the separation of the class means to the within-class variance, that is, maximizing the ratio $a^{\mathrm{T}}Ba/a^{\mathrm{T}}Wa$, where W denotes the within-class covariance matrix, i.e. the covariance matrix of the variables centered on the class mean, and B denotes the between-classes covariance matrix. These two matrices can be calculated as follows. Let M be the $n \times p$ matrix of class means, and G be the $n \times g$ matrix of class indicator variables (so $g_{ij} = 1 \iff$ case i is assigned to class j). Let $\bar{x}$ be the means of the variables over the whole sample, then the sample covariance matrices are

$$W = \frac{(X - GM)^{\mathrm{T}}(X - GM)}{n - g}$$

$$B = \frac{(GM - 1\bar{x})^{\mathrm{T}}(GM - 1\bar{x})}{g - 1}$$

Different denominators have been used in covariance matrices. Here we follow the notation in Venables and Ripley (2002). The criterion used in LDA is very intuitive. LDA is a non-parametric method that is also a special form of a maximum likelihood discriminant rule for multivariate normal class densities with the same covariance matrix. An alternative approach to discrimination is via probability models. Let π_c denote the prior probabilities of the classes, and $p(x|c)$ the densities of distributions of the observations for class c. Then the posterior distribution of the classes after observing x is

$$p(c|x) = \frac{\pi_c p(x|c)}{p(x)} \propto \pi_c p(x|c)$$

The allocation rule which makes the smallest expected number of errors chooses the class with maximal $p(c|x)$; this is known as the *Bayes* rule. Now suppose the distribution for class c is multivariate normal with mean μ_c and covariance Σ_c. Then the *Bayes* rule minimizes

$$\begin{aligned} Q_c &= -2\log(p(x|c)) - 2\log(\pi_c) \\ &= (x - \mu_c)\Sigma_c^{-1}(x - \mu_c)^{\mathrm{T}} + \log(|\Sigma_c|) - 2\log(\pi_c) \end{aligned}$$

The difference between the Q_c for two classes is a quadratic function of x, so the method is known as QDA and the boundaries of the decision regions are quadratic surfaces in the x space. LDA is a special case of QDA where classes have common covariance matrix.

k-Nearest neighbor (*KNN*). KNN classifiers are based on finding the k nearest examples in some reference set, and taking a majority vote among the classes of these k examples, or, equivalently, estimating the posterior probability $p(c|x)$ by the proportions of the classes among the k examples. We can measure 'nearest' by Euclidean distance or by one minus correlation. Here we consider using $k = 1, 3$ under Euclidean distance.

Bagging, boosting classification trees. Constructing classification trees may be seen as a type of variable selection. Possible interactions between variables are handled automatically, and so is monotonic transformation of the variables. These issues are reduced to which variables to divide on, and how to achieve the split in building a classification tree. Specifically we construct trees by recursive splits of subsets of the samples into two child subsets, starting with all the samples. Each terminal node is assigned a class label and the resulting partition corresponds to a final classifier. There are several forms of trees. Here we use the *CART*—classification and regression trees. For a detailed technical discussion of *CART*, see Breiman *et al*. (1983).

Aggregating classifiers could dramatically improve predictive accuracy (Breiman, 1996, 1998). In classification, the multiple classifiers are aggregated by majority votes, i.e. the final class is the one predicted by the majority of the predictors. Breiman (1998) studied the bias and variance properties of the aggregated predictors. The key is the possible instability of the prediction method, i.e. whether small changes in the learning set result in large changes in the predictor. CART is an unstable classifier that can benefit from aggregation. Here we aggregate trees which are grown until they perfectly fit the data. The simplest form of bagging is using bootstrap to produce pseudo-replicates. In our study, we aggregated 50 bootstrap samples to produce a pool of classification trees. This algorithm works in the following way (suppose our sample is S having n samples):

Algorithm 1, Bagging

```
(1) Sample with replacement to form $N$ bootstrap samples
    $\{B_1, \ldots, B_N\}$.
(2) Use $B_k$ to construct Tree classifier $T_k$, and predict $S$
    using $T_k$.
(3) Final prediction is un-weighted average.
```

Boosting was first proposed by Freund and Schapire (1997), and it was also called arcing and studied by Breiman (1998). The basic idea is to adaptively resample the original data so that the weights are increased for those most frequently misclassified samples. The final prediction is based on weighted or un-weighted voting. It is conjectured that boosting is a special form of RF (Breiman, 2001) (see below). Here we consider two special forms of *Boosting*: *arc-x4* and *arc-fs*, following the descriptions of Breiman (1998).

Algorithm 2, Arc-fs details

(1) At first step, initialize $p_i^{(1)} = 1/n$. Let $P^{(1)} = \{p_1^{(1)}, \ldots, p_n^{(1)}\}$.

(2) At kth step, using the current probabilities $P^{(k)}$, sample with replacement from sample S to get the training set S_k and construct tree classifier T_k using S_k.

(3) Run S down T_k and let $d(i) = 1$ if ith case is classified incorrectly, otherwise zero.

(4) Define $\epsilon_k = \Sigma_i p_i^{(k)}, \beta_k = (1 - \epsilon_k)/\epsilon_k$, update $k + 1$ step probabilities by

$$p_i^{(k+1)} = \frac{p_i^{(k)} \beta_k^{d(i)}}{\Sigma_j p_j^{(k)} \beta_k^{d(j)}}$$

If $\epsilon_k = 0, \epsilon_k \geq \frac{1}{2}$, re-initialize $p_i^{(k+1)} = 1/n$.

(5) After K steps, $\{T_1, \ldots, T_K\}$ are combined using weighted voting with T_k having weight $\log(\beta_k)$.

Algorithm 3, Arc-x4 details

(1) Same as **Arc-fs**

(2) Same as **Arc-fs**

(3) Run S down tree classifier T_k and let $m(i)$ be the number of misclassifications the ith case by $\{T_1, \ldots, T_k\}$.

(4) Update $k + 1$ step probabilities defined by $p_i^{(k+1)} = p_i^{(k)}(1 + m(i)^4)/\Sigma_j p_j^{(k)}(1 + m(j)^4)$

(5) After K steps, $\{T_1, \ldots, T_K\}$ are combined by un-weighted voting.

Support vector machine (*SVM*). The observed m/z ratio for the ith subject X_i can be thought of as a point in $\mathbb{R}^p$. An intuitive binary classifier would be to construct a hyperplane separating cancer subjects from normal subjects in this $\mathbb{R}^p$ space. But for most problems, there is no hyperplane which can successfully separate different classes. The idea of SVM is to map the data into a higher dimension space and separate them there. For technical details, please see Vapnik (1998) and Burges (1998). The algorithm used here is described at http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html

Random forest (*RF*). RF (Breiman, 2001) combines two powerful ideas in machine learning techniques: bagging and random feature selection. Bagging, as described above, stands for bootstrap aggregating, which uses resampling to produce pseudo-replicates to improve predictive accuracy. By using random feature selections, we can significantly improve our predictive accuracy. Here we use the RF program from (Breiman, 2001), and it works as follows.

Algorithm 4, RF

(1) Sample with replacement to form N bootstrap samples $\{B_1, \ldots, B_N\}$.

(2) Use each sample B_k to construct a Tree classifier T_k to predict those samples that are not in B_k (called *out-of-bag* samples). These predictions are called *out-of-bag* estimators.

(3) Before using T_k to predict *out-of-bag* samples, if we randomly permute the value for one variable for these *out-of-bag* samples, intuitively the prediction error is going to increase. And the amount of increase will reflect the importance of this variable.

(4) When constructing T_k, at each node splitting we first randomly select m variables, then we choose one best split from these m variables.

(5) Final prediction is the average of *out-of-bag* estimators over all Bootstrap samples.

3 DATA SET AND PRE-PROCESSING

3.1 Data set

We have obtained ovarian cancer and control serum samples from the National Ovarian Cancer Early Detection Program at Northwestern University Hospital. The Keck Laboratory then subjected these samples to automated desalting and MALDI-MS on a Micromass M@LDI-R instrument http://www.micromass.co.uk as described generally at http://info.med.yale.edu/wmkeck/prochem/biomarker.htm. This data set consists of MS spectra that extend from 800 to 3500 Da and that were obtained on serum samples from 47 patients with ovarian cancer and 44 normal patients. Based on our evaluation, two of the normal spectra are of poor quality and are excluded in our analyses. Figure 2 shows the overall case and control median log intensities based on 89 samples.

3.2 Pre-processing

Due to the noisy nature of the data set, pre-processing is an important step in the analysis of MS data. The raw intensities

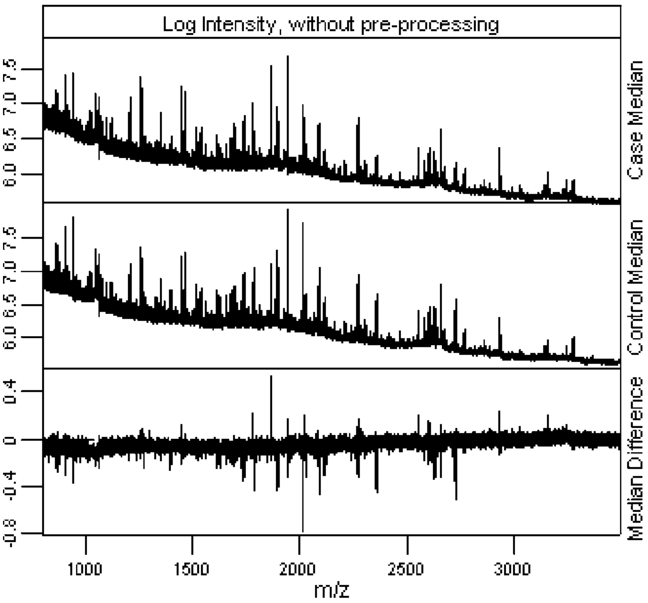

Fig. 2. Median log intensity for 89 samples.

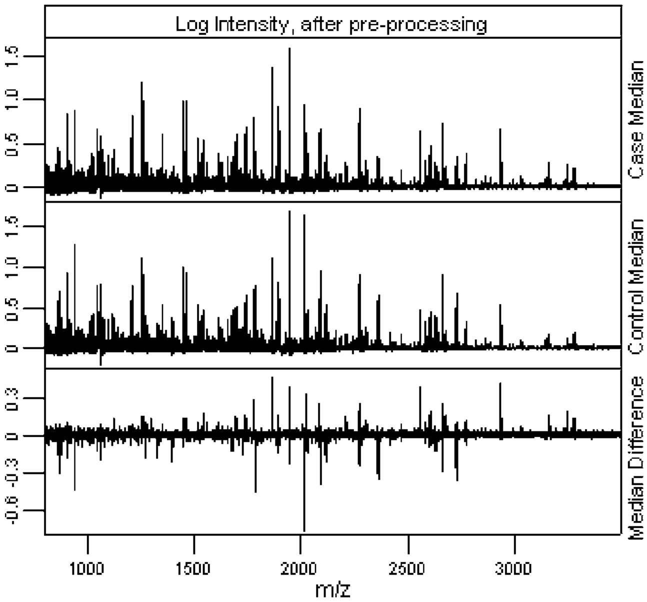

Fig. 3. Median log intensity after pre-processing.

have a wide dynamic range. Taking the log of the intensities decreases the magnitude and variation within this range. Before we submit the data set to our classifiers, we have to carry out some pre-processing (e.g. background subtraction to remove the effect of chemical and electronic noise, peak identification, etc.) which will be described in a subsequent publication. Figure 3 shows the median intensities after all pre-processing.

3.3 Peptide/protein marker selection

For some discriminant analysis methods discussed above, the number of features that can be handled has to be smaller than the number of observations, e.g. the LDA and QDA methods. Therefore, we cannot use all the intensity values from an MS data set for these classification methods. Instead, we have to identify certain m/z ratios as inputs to these methods, and it is apparent that this feature selection step is critical in the analysis of MS data and comparison of various methods. To make the comparison as valid as possible, we feed the same set of m/z ratios to all classification methods and compare their performance on our data. This practice will likely penalize those methods that can utilize as many features as possible in classifications. In our analysis, we use two methods to select m/z ratios used in classification analysis. For the first method, we rank the variables, i.e. m/z ratios, based on normalized difference between two groups (cancer group and normal group), which is the T-statistic, and then we select variables based on the absolute values of the t-statistics. In our study, we evaluate the effects of selecting 15 and 25 markers. In order to evaluate the effects of LDA and QDA, we must verify that there is a sufficient number of samples. So there is a practical limit on the number of markers that we can use. For the second method of choosing variables in classification analysis, we use the by-product of the RF program. The RF program outputs a variable importance measure. This measure is derived from assessing the decrease in prediction accuracy after random permutation of each variable in the feature set. The idea is that if we randomly permute the observed values of an important variable, this will result in substantially decreasing our ability to classify each individual in the sample set. In our analysis, we also select 15 and 25 markers from a customized RF algorithm which will be described in a subsequent publication. We also compare marker selection based on RF and the normalized difference between groups.

3.4 Study design

Here we want to compare the performance of the classifiers discussed above based on their prediction error rate. Since a test data set was not available, cross-validation within the original data set was utilized to provide a nearly unbiased estimate of the prediction error rate. Breiman and Spector (1992) demonstrated that leave-one-out cross-validation has high variance if the prediction rule is unstable, because the leave-one-out training sets are too similar to the full data set. 5-fold or 10-fold cross-validation displayed lower variance. Efron and Tibershirani (1997) proposed a 0.632+ bootstrap method, which is a bootstrap smoothing version of cross-validation and has less variation. We applied both methods to LDA, QDA and NN classifiers. We ran 100 cycles of 10-fold cross-validation and 0.632+ bootstrap error rate estimation. In the 0.632+ rule, we used 100 bootstrap samples. In estimating the 0.632+ error of QDA, bootstrap samples often caused the covariance matrix to be singular if we used 25 markers,

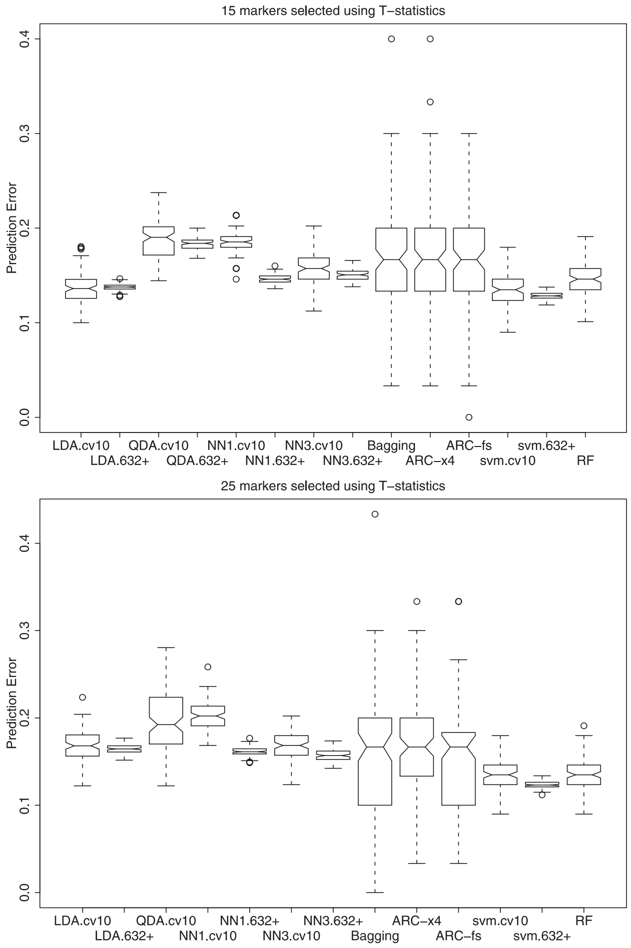

Fig. 4. Error summary for T-statistics marker selection.

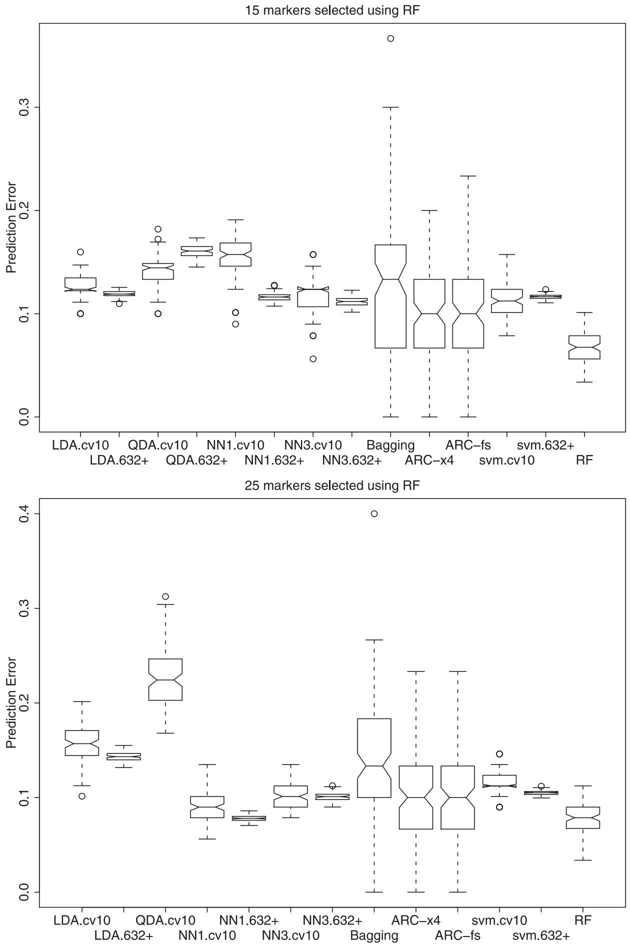

Fig. 5. Error summary for RF marker selection.

so we calculated the 0.632+ error rate of QDA only for 15 markers.

For bagging and boosting, we used a random split to partition the observed data into a training set (59 samples) and a testing set (30 samples) to estimate error rate. We repeated this 100 times, and the RF prediction error estimate is based on out-of-bag estimation, which we believe is reasonably accurate. To assess the error rate variation, we repeated the whole procedure 100 times, with each error estimate based on 100 trees.

These calculations were carried out for selected markers using RF and the normalized difference between groups.

4 RESULTS

4.1 Prediction error rates

We use boxplots to summarize the error rates. Figure 4 summarizes the errors for using T-statistics to select markers and Figure 5 summarizes the errors for using RF to select markers. In these plots, the postfix ‘cv10’ means estimating error using 10-fold cross-validation, ‘0.632+’ means estimating error using 0.632+ bootstrap method, NN1 for k-nearest neighbor with $k = 1$, NN3 for k-nearest neighbor with $k = 3$.

First, for the estimates of error rates based on different methods (cross-validation or 0.632+ rule), we can see that the 0.632+ rule provides a more stable estimate of the error rate than 10-fold cross-validation for LDA, QDA, kNN, and SVM classifiers. The error results for bagging and boosting trees are highly variable. Although the variance of the error estimates for RF is not as small as those based on the 0.632+ rule, it is certainly quite comparable and much less than those based on bagging and boosting.

As for error rate, RF consistently performs well among all the scenarios considered. When a total of 15 markers selected through t-statistics are used, SVM has the lowest error rate among all classifiers, whereas the error rate based on LDA is the second lowest one among all the classifiers. The error rate based on RF closely follows the top two methods. As the number of markers selected increases from 15 to 25, the relative advantage of LDA over RF no longer holds. SVM has the lowest error rate and RF has close performance. In

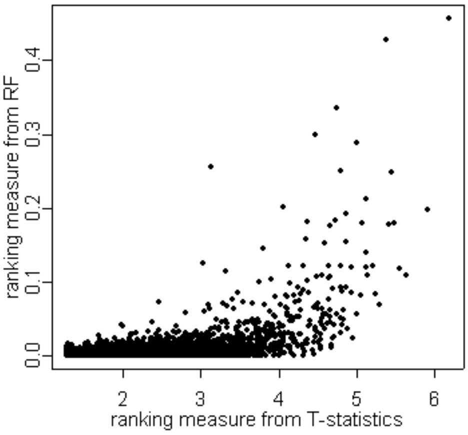

Fig. 6. Variable ranking comparison.

addition, error rates based on RF have consistent low variation, which suggests that the error rate from RF is very reliable.

When the variables selected are derived from importance measures based on RF, it is not surprising that RF outperforms all other methods. Based on these sets of variables, the relative performance of LDA stays the same. But the QDA becomes worse when 25 markers are used, which is due to the unstable estimate of the covariance matrix in QDA. Because all variables, instead of a pre-selected subset of variables, can be utilized in RF as well as bagging and boosting methods considered here, the ability to incorporate many more variables to build classifiers represents a distinct advantage over the methods that are limited by the number of variables that can be considered in an analysis. As a result, the error rate based on RF using the variables selected by RF has lower prediction error rate than the minimum error rate achieved using variables selected through T-statistics.

4.2 Choice of predictor variables

LDA and QDA are not stable using a large number of variables. In using bootstrap to estimate the error rate, the covariance matrix for the bootstrap samples was often singular.

4.3 Variables identified from T-statistics and RF

Here we compare the variable selection based on T-statistics and RF program. Figure 6 plots the ranking measures of selected peaks based on T-statistics and the importance measures from RF. We can see that both measures will be able to capture a common set of variables, i.e. the variables corresponding to the points in the upper tight region of this figure. However, there do exist discrepancies between these two measures, resulting in different performance of various classifiers based on the selected variables.

5 DISCUSSION

In this report, we have compared results obtained with several well-known classification methods to distinguish ovarian cancer patients from normal individuals based on MS data obtained on serum samples. Overall, we have found that the RF approach both leads to an overall lower misclassification rate as well as to a more stable assessment of classification errors. Therefore, our preliminary analyses suggest that RF and methods similar in nature to RF may be more useful than other methods to classify samples based on MS data. This conclusion has been confirmed by applying these classification methods to a completely different data set on autism which yielded similar results (data not shown). Compared to LDA and QDA methods, RF has the advantage of not requiring the number of variables used to be less than the number of subjects in the study, which is a clear advantage for the analysis of MS data as the number of m/z versus intensity data points is very large. In addition, RF is able to handle interactions among variables. Although many methods have been compared in this report, there also are some additional methods, e.g. neural networks, that we have not yet compared. This is an ongoing endeavor, and we are in the process of evaluating these other methods as well.

The pre-processing of MS outputs is a very critical step in the overall analysis of MS data set. Peak identification, spectrum alignment, as well as normalization undoubtedly all affect the performance of classifications. Because the focus of this report is on comparing various classifiers, and we believe it is likely that the relative performance of these classifiers will not be *differentially* affected by pre-processing the data, we have not discussed in detail the specific steps we have taken to pre-process MS data. The effects of pre-processing on classification analysis and biomarker identification will be reported elsewhere.

In this report, we use T-statistics to pre-select a set of variables as inputs for various classifiers. There are some limitations to this approach as it does not take into account interactions among variables, and more importantly, it is not stable when sample sizes are relatively small. We could consider a more robust form of variation estimation, and utilize the global variation to improve the variance estimates. Variance shrinkage is a very good strategy to improve the estimation of variance (Long *et al*., 2001). As our current sample size is relatively small, we are considering a more robust approach to estimation of variations. Although NN, RF and other tree-based methods are able to analyze many variables, we still believe variable selection is a critical issue. For example, the m/z ratios corresponding to background levels should not be considered in classification analysis, and keeping these background noises in the data will likely reduce the performance of any classifier. Therefore, in addition to data pre-processing, variable selection for classification analysis may represent another challenge for MS data analysis.

ACKNOWLEDGEMENTS

We thank reviewers for their constructive comments. This research is supported by NIH grant NHLBI N01-HV-28186, R01 GM59507, RR015837, NCI-EDRN, NCI U01 CA-98-028, and DOE grant DE-FG02-02ER63462.

REFERENCES

(1999). The chipping forecast. *Supplement to Nature Genetics*, **21**.

(2001). Proteomics. *Science*, **294**, 2074–2085.

Ball,G., Mian,S., Holding,F., Allibone,R.O., Lowe,J., Ali,S., Li,G., McCardle,S., Ellis,I.O., Creaser,C. and Rees,R.C. (2002). An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers. *Bioinformatics*, **18**, 395–404.

Bao-Ling,A., Yinsheng,Q., John,W.D., Michael,D.W., Mary,A.C., Lisa,H.C., John,O.S., Paul,F.S., Yutaka,Y., Ziding,F. and George,L.W.J. (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Breiman,L. (1996). Bagging predictors. *Mach. Learning*, **24**, 123–140.

Breiman,L. (1998). Arcing classiers. *Ann. Statistics*, **26**, 801–824.

Breiman,L. (2001). Randomforest. *Technical Report, Stat. Dept. UCB*.

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C. (1983). *Classification and Regression Trees*, Chapman & Hall.

Breiman,L. and Spector,P. (1992). Submodel selection and evaluation in regression: the x-random case. *Int. Stat. Rev.*, **60**, 291–319.

Burges,C.J.C. (1998, June). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**(457), 77–87.

Efron,B. and Tibershirani,R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.

Fisher,R.A. (1936). The use of multiple measurements in taxonomic problems. *An. Eugenics*, **7**, 179–188.

Freund,Y. and Schapire,R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

Fung,E.T., Wright,G.L.J. and Dalmasso,E.A. (2000). Proteomic strategies for biomarker identification: progress and challenges. *Curr. Opin. Mol. Therap.*, **2**, 643–650.

Guoan,C., Tarek,G.G., Chiang-Ching,H., Dafydd,G.T., Kerby,A.S., Jeremy,M.G.T., Sharon,L.R.K., David,E.M., Thomas,J.G., Mark,D.I. *et al.* (2002). Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors. *Clin. Cancer Res.*, **8**, 2298–2305.

Long,A.D., Mangalam,H., Chan,B.Y.P., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001). Genome expression profiling in *Escherichia coli* k12: improved statistical inference from DNA microarray data using analysis of variance and a bayesian statistical framework. *J. Biol. Chem.*, **276**, 19937–19944.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979). *Multivariate Analysis*. Academic Press, Inc., San Diego.

McLachlan,G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Petricoin,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simine,C., Fishman,D.A., Kohn,E.C. and Liotta,L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, **359**, 572–577.

Ripley,B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, New York.

Vapnik,V. (1998). *Statistical Learning Theory*. Wiley, New York.

Venables,W.N. and Ripley,B.D. (2002). *Modern Applied Statistics with S*, 4th edn. Springer, Berlin.