BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3128–3136
doi:10.1093/bioinformatics/bth372

Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens

Glen A. Satten[1,*], Somnath Datta[2], Hercules Moura[1], Adrian R. Woolfitt[1], Maria da G. Carvalho[3], George M. Carlone[3], Barun K. De[3], Antonis Pavlopoulos[1] and John R. Barr[1]

[1]Division of Laboratory Science, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA 30341, USA, [2]Department of Statistics, University of Georgia, Athens, GA 30602, USA and [3]Division of Bacterial and Mycotic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

Received on February 13, 2004; revised on June 9, 2004; accepted on June 15, 2004
Advance Access publication June 23, 2004

ABSTRACT

Motivation: Application of mass spectrometry in proteomics is a breakthrough in high-throughput analyses. Early applications have focused on protein expression profiles to differentiate among various types of tissue samples (e.g. normal versus tumor). Here our goal is to use mass spectra to differentiate bacterial species using whole-organism samples. The raw spectra are similar to spectra of tissue samples, raising some of the same statistical issues (e.g. non-uniform baselines and higher noise associated with higher baseline), but are substantially noisier. As a result, new preprocessing procedures are required before these spectra can be used for statistical classification.

Results: In this study, we introduce novel preprocessing steps that can be used with any mass spectra. These comprise a standardization step and a denoising step. The noise level for each spectrum is determined using only data from that spectrum. Only spectral features that exceed a threshold defined by the noise level are subsequently used for classification. Using this approach, we trained the Random Forest program to classify 240 mass spectra into four bacterial types. The method resulted in zero prediction errors in the training samples and in two test datasets having 240 and 300 spectra, respectively.

Availability: Fortran code for standardization and denoising is available at the supplementary information website.

Contact: gsatten@cdc.gov

Supplementary information: http://www.stat.uga.edu/~datta/Massspec/supp.html

*To whom correspondence should be addressed.

INTRODUCTION

Mass spectrometry (MS) data are increasingly used to characterize proteins in biological specimens. The data consist of signal intensity at a large number of closely spaced mass to charge (m/z) ratios (see Thiele, 2003 for a general introduction to MS data). MS data have some similarities with microarray datasets; both represent outcomes of high-throughput designs and consequently feature high-dimensional multivariate data. An important difference is that microarray data are essentially unordered, while signal intensity at adjacent m/z ratios can be expected to be similar. This feature of MS data can be used to standardize and denoise mass spectra using data from a single spectrum.

Mass spectra have been proposed for use as biomarkers, e.g. for predicting cancer on the basis of the protein profile (Ball *et al*., 2002; Petricoin *et al*., 2002; Sorace and Zhan, 2003; Hawkins *et al*., 2003), in classification of two tissue samples (Wu *et al*., 2003; Purohit and Rocke, 2003), in the discovery of single nucleotide polymorphisms (SNPs) and mutations (Böcker, 2003), and for whole bacteria identification (reviewed by Fenselau and Demirev, 2001; Lay, 2001). Statistical models have been developed to match a given spectrum against a peptide database (Bafna and Edwards, 2001) and for identification of proteins (Nesvizhskii *et al*., 2003; Havilio *et al*., 2003). A number of papers have reported near perfect separation of two tissue samples on the basis of standard classifiers and using various sets of m/z ratios (including some in the so-called 'noise zone'). Some caution is required when interpreting such results, and differences in two samples may be of non-biological nature (e.g. arising purely due to artifacts resulting from the way the two samples were prepared) (Sorace and Zhan, 2003; Baggerly *et al*., 2004). We take the

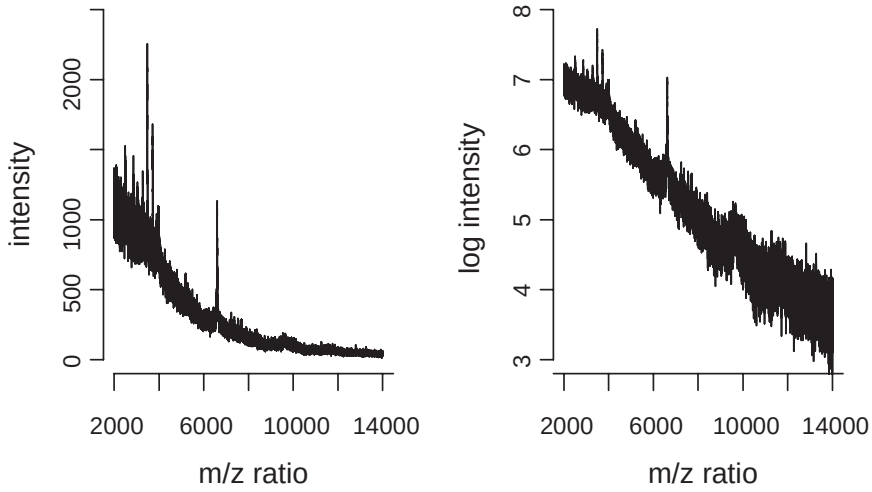

Fig. 1. A raw *B.anthracis* spectrum (left) and its (natural) log-transformed version (right).

viewpoint that all identification should be made using only features that well exceed a noise threshold, to ensure that the resulting classification algorithm has scientific validity. To this end, we provide an estimator of the noise scale that can be calculated separately for each spectrum.

Here we consider the use of matrix-assisted laser desorption ionization/time-of-flight (MALDI-TOF) mass spectra to classify whole-organism bacterial samples. Currently, identification of bacterial samples may take several days. Rapid and unequivocal identification of whole organism bacterial samples such as anthrax spores and other biowarfare agents is highly desirable. Unlike other proteomic applications in which complex samples are prefractionated to select a particular group of compounds of interest, mass spectra obtained from whole organisms can be quite complex and are often very noisy. Additionally, MALDI-TOF analyses tend to be less reproducible from sample to sample and from spectrum to spectrum when compared to other mass spectrometry techniques. The ionized peptides and proteins tend to be large; for this reason, we scan 1000–14 000 m/z region that we have found contains the most significant peaks in whole-bacterial specimens. However, in this region mass spectrometer resolution is lower than the more-commonly used 500–3500 m/z region, and there is little advantage in using a reflectron (Thiele, 2003) to increase resolution. Finally, it was not advantageous to include an internal mass calibration standard to correct small discrepancies in the mass values of each spectrum in these datasets. This was because the automated software that rejects or accepts candidate spectra (Thiele, 2003) too often selected spectra where the internal standard was the only peak. As a result, our automatically acquired spectra are noisier and more variable than typical MS data. Raw mass spectra generally require preprocessing before any further analysis. This is especially true of our data, given the high noise level. It is desirable to have a preprocessing algorithm that uses only data from a single spectrum, so that data in the training set and future test samples are handled in an equivalent manner.

Figure 1 shows a typical spectrum from our dataset. Common characteristics of our bacterial mass spectra are sharp peaks that tower above the baseline noise level; a non-uniform baseline; and higher noise level associated with higher baseline values. The heteroscedasticity as a function of m/z ratio has led most researchers to analyze log-transformed spectra. However, on the log scale our spectra were sufficiently noisy that the peaks no longer tower above the noise level; after log transformation there is even some difficulty in deciding what is signal and what is noise.

The purpose of this paper is 2-fold. First, we present a novel preprocessing method which we have found useful in analyzing noisy mass spectra; these tools may in fact be useful in analyzing any mass spectra. They consist of a standardization step and a denoising step. After denoising, only spectral features that exceed a spectrum-specific noise threshold are retained. We then present a data analysis in which 240 bacterial specimens are categorized into four bacterial types using the Random Forest (RF) algorithm of Breiman (1999, 2001). The ensuing classification algorithm is tested on two additional datasets of bacterial spectra. The goal of this analysis is to show that our standardization and denoising algorithms retain sufficient information to allow categorization of bacterial spectra. In a recent study, Wu *et al*. (2003) declared RF to be superior to several other classifiers for classifying two tissue types using mass spectra. We also find the performance of RF coupled with the preprocessing methods to be extremely good in classifying the four bacterial types we considered. We also consider two related questions about recognition of different strains of the same bacterial species. Specifically, we consider whether the RF algorithm, when trained on one strain of a bacterial species, can recognize another closely

related strain; and whether the information in our standardized and denoised spectra allows the RF algorithm to differentiate between two related strains of the same bacterial species when this information is included in the training set.

SYSTEMS AND METHODS

Bacterial specimens

Our dataset consisted of mass spectra of the following bacterial isolates: *Escherichia coli* (ATCC25922), *Streptococcus pyogenes* (SS1662), *Bacillus anthracis* (ATCC4229; Pasteur strain) and two strains of *Streptococcus pneumoniae* (denoted as 18C-A and 18C-B). While we expect *B.anthracis*, *E.coli*, *S.pyogenes* and *S.pneumoniae* will produce different mass spectra, the two *S.pneumoniae* isolates may be expected to be closely related as they are both of serotype 18C, differing in that one attaches better to tissue culture cells.

Bacterial cells of *E.coli*, *S.pyogenes* and the two *S.pneumoniae* isolates were obtained after being grown overnight in trypticase soy/sheep blood agar plates under 5% CO_2 at 37°C. Vegetative cells of *B.anthracis* were cultured anaerobically in sheep blood agar plates for 24 h, and were then transferred to sporulation medium plates and incubated for 5 days at 37°C to induce spore formation. The organisms were harvested and washed in cold Tris–sucrose buffer (0.01 M Tris, 0.025 M sucrose, pH 7.0), gamma-irradiated (2.1 Mrad for 3 h at 4°C) and were confirmed to be non-viable by culture. Samples were stored at −80°C until MALDI-TOF MS analysis. All samples were stored under the same conditions of media, temperature and atmosphere.

Mass spectrometry

The bacterial isolates were analyzed by MALDI-TOF MS. Whole cells were mixed with matrix and the suspensions were deposited onto a stainless steel plate. MALDI-TOF MS analyses were performed using intact cells. The MALDI matrix consisted of a 10 mg/ml solution of 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid) (Sigma-Aldrich, St. Louis, MO) in 50% acetonitrile and Milli-Q grade water containing 0.1% trifluoroacetic acid. Mass spectra were acquired using an Applied Biosystems Voyager DE-STR MALDI-TOF mass spectrometer (Framingham, MA) equipped with a nitrogen laser (337 nm, 3 ns pulse width). Analyses were performed in linear-delayed extraction positive ion mode at accelerating voltages of 20 or 25 kV; extraction delay times varied from 280 to 320 ns depending on the chosen mass range. The instrument was mass calibrated before sample analysis. After initial manual laser intensity optimization and baseline data acquisition, spectra were acquired in automatic control mode. Samples were plated on 100-well stainless steel MALDI targets, and all spectra were generated the same day the samples were plated. The mass spectrometer was programmed to examine signals from a maximum of 30 randomly positioned non-overlapping spots in each sample well. The spectra from the first 10 of these spots that had detectable signals were summed into a final profile mass spectrum. This procedure was repeated four times per well, i.e. four profile mass spectra were obtained for each well. Increasing the number of spectra per profile spectrum increases profile spectrum quality by decreasing the variability between profile spectra, but results in slower data acquisition. With our choice of 10 spectra per profile and 4 profile spectra per well, 6 h were required to analyze each plate. The profile mass spectra were exported in text format, using Microsoft Visual Basic for Applications macros within the mass spectrometer data analysis software (Data Explorer™), in preparation for further statistical analysis.

Standardization

A standardization step is necessary to account for the non-uniform baseline and variability in maximum intensity across spectra. The literature on statistical analysis of proteomics data is limited. As discussed in the introduction, log-transformation appears inappropriate for these data. We propose a novel standardization procedure, in which each spectrum is standardized using only information from that spectrum.

Let x_i denote the m/z ratios at which intensity y_i is measured, and adopt the convention that the values x_i are in ascending order. We standardize the spectra by replacing each intensity $y(x_i)$ with

$$y_i^* = \frac{y_i - Q_{0.5}(x_i)}{Q_{0.75}(x_i) - Q_{0.25}(x_i)},$$

where $Q_\alpha(x)$ is an estimate of the α-th quantile of spectral intensities at m/z ratio x. That is, we center the spectra using a (local) estimate of the median spectral intensity, and divide by a (local) estimate of the interquartile range. We chose interquartile range over the SD as a measure of scale as it is less likely to be sensitive to outlying values (peak intensities).

For any $0 < \alpha < 1$, the local α-th quantile of the intensity at the i-th m/z ratio x_i is estimated to be a weighted average of $Q_\alpha^+(x_i)$ and $Q_\alpha^-(x_i)$, where

$$Q_\alpha^+(x_i) = \min_y\{W_h(x_i, y) \geq \alpha\},$$

$$Q_\alpha^-(x_i) = \max_y\{W_h(x_i, y) \leq \alpha\},$$

and

$$W_h(x_i, y) = \frac{\sum_{j=\max(1,i-h)}^{\min(n,i+h)} I\{y_j \leq y\}\left\{1 - \frac{(i-j)^2}{h^2}\right\}}{\sum_{j=\max(1,i-h)}^{\min(n,i+h)}\left\{1 - \frac{(i-j)^2}{h^2}\right\}},$$

where $I\{\mathcal{C}\} = 1$ if $\mathcal{C}$ is true and 0 otherwise, and where h is a user selectable width defining a neighborhood of x_i. In our analyses, we chose $h = 500$. Note that our spectra comprise

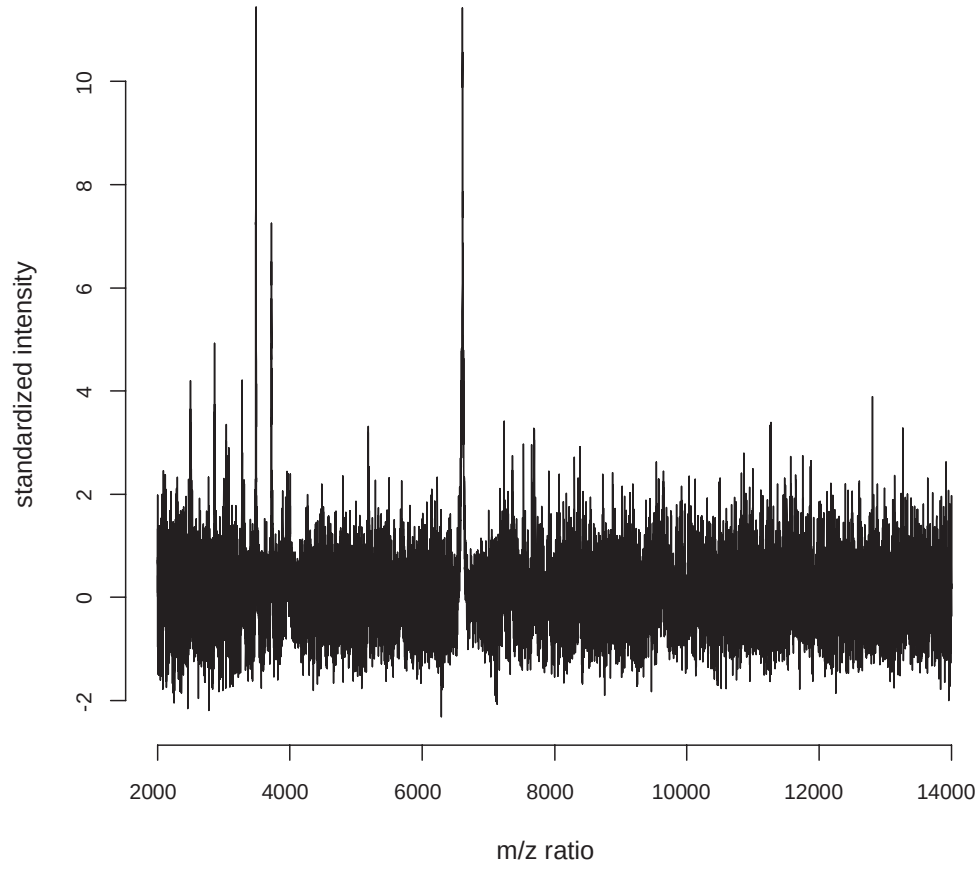

Fig. 2. The *B.anthracis* spectrum of Figure 1 after standardization.

intensity values at 18 190 m/z ratios (i.e. 18 190 m/z-intensity pairs). The function $W_h(x_i, y)$ is the proportion of weights $\{1-(i-j)^2/h^2\}$ that correspond to intensities y_j that are less than or equal to y. Note that the denominator of $W_h(x_i, y)$ is never zero because it is the sum of non-negative terms and the summand corresponding to $j = i$ is 1. If $Q_\alpha^-(x_i) = Q_\alpha^+(x_i)$ then take $Q_\alpha(x_i) = Q_\alpha^+(x_i)$. When $Q_\alpha^-(x_i) < Q_\alpha^+(x_i)$, let $\alpha^\pm = W_h[x_i, Q_\alpha^\pm(x_i)]$ and take $Q_\alpha(x_i) = [(\alpha^+ - \alpha)/(\alpha^+ - \alpha^-)]Q_\alpha^-(x_i) + [(\alpha - \alpha^-)/(\alpha^+ - \alpha^-)]Q_\alpha^+(x_i)$. This choice for Q_α was proposed by Ducharme *et al.* (1995).

In Figure 2, we illustrate the result of this standardization applied to the *B.anthracis* spectrum of Figure 1. The resulting spectrum has a flat baseline and tall peaks; moreover, the noise intensity is quite constant across the spectrum. Finally, because the standardized spectrum is a ratio of intensities, standardized spectra can be compared directly. Note that for our application, standardization using an estimate of the noise is preferable to the usual standardization algorithm [standardization so that the tallest peak has unit intensity, see e.g. Banks and Petricoin (2003)] because the peak spectral intensities for the same bacterial specimen may vary from spectrum to spectrum, depending on how many bacterial fragments of a certain size are generated in the laser desorption step. Further, unlike the usual standardization algorithm, our standardization is not affected if the tallest peak is attenuated due to signal saturation, since the scale of each spectrum is determined by the low-intensity noise level.

Denoising

Although standardized spectra have a common scale and are fairly homoscedastic, they are still a mixture of noise and signal. It is desirable to denoise a spectrum before using it for classification. Even if noisy spectra allow efficient or even perfect classification, denoising ensures that the features used for classification correspond to real peaks. This increases confidence in the scientific validity of the classification procedure (see, e.g. Sorace and Zhan, 2003). Again, we seek a method that uses only the information in a single spectrum. Note from Figure 2 that the standardized spectrum $y^*(x)$ can be negative; in fact the median of the $y^*(x)$ values is typically zero. While it may be difficult to separate noise from signal using those standardized intensities that are positive, the negative standardized intensities presumably represent pure noise. This is evidenced in Figure 3, where we show the q–q plot of the standardized y^* values corresponding to the *B.anthracis* spectrum of Figures 1 and 2.

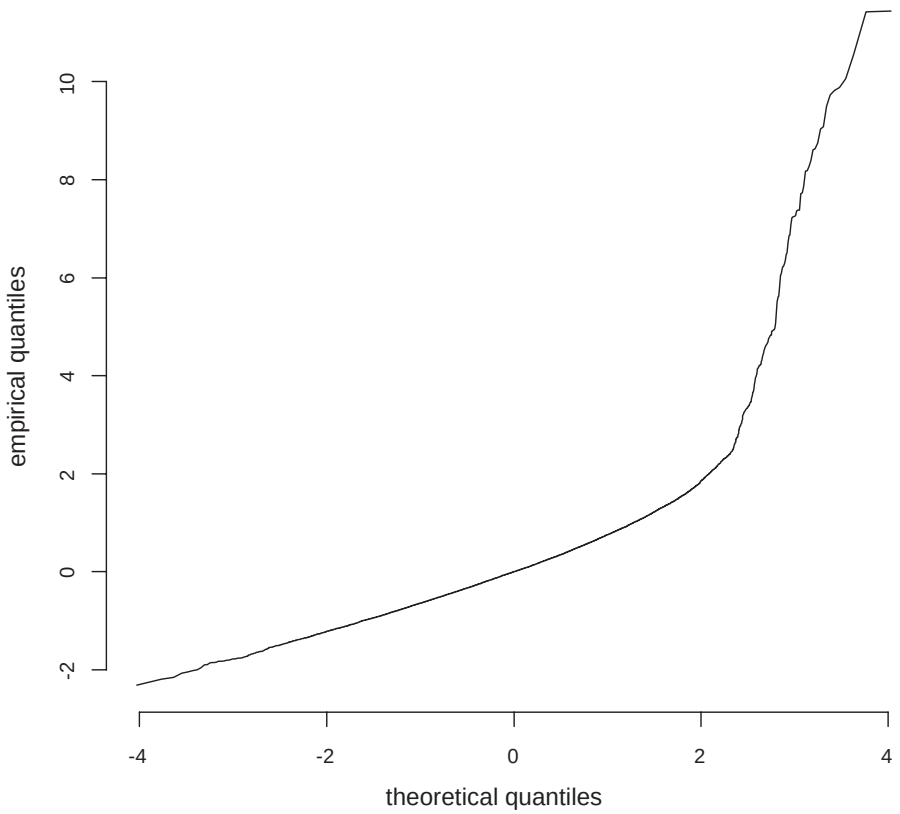

Fig. 3. A standard normal q–q plot of the standardized spectrum values for the *B.anthracis* spectrum of Figures 1 and 2.

The q–q plot of Figure 3 compares the empirical quantiles of the distribution of standardized intensities to the quantiles of the standard normal distribution. A linear q–q plot is evidence that data are normally distributed. The linearity of the q–q plot for negative values (and small positive values) shows that the standardized spectra $y^*(x)$ consists of a signal part $s(x) \geq 0$ and a noise part $\epsilon(x)$ that is approximately normally distributed with zero mean and SD σ. As a result, the sample root mean square $\hat{\sigma}$ of the negative values of $y^*(x)$ can serve as an estimator of σ.

Having estimated the noise scale for a given spectrum, we propose the following hard thresholding criterion to denoise the standardized spectrum:

$$\tilde{y}(x) = \begin{cases} y^*(x) & \text{if } y^*(x) \geq 6\hat{\sigma} \\ 0 & \text{otherwise.} \end{cases}$$

Use of $6\hat{\sigma}$ as a cutoff to eliminate normally distributed noise is a very stringent criterion. For certain applications, a lower threshold such as $5\hat{\sigma}$ may be more appropriate. It is also possible to use a soft thresholding rule such as

$$\tilde{y}(x) = \max\{0, y^*(x) - 6\hat{\sigma}\}.$$

We use hard thresholding to preserve the usual relationship between peak height and concentration. Standardization and denoising of a single spectrum takes ∼5 s using a Compaq Evo W4000 (2.8 GHz Pentium IV processor).

Random Forest classifier

The RF algorithm of Breiman (1999, 2001) is arguably the best statistical classifier currently available. In a recent study, Wu *et al*. (2003) establishes its superiority in classifying two groups of MS samples over other classification methods including the support vector machine for MS data in the 500–3500 m/z range. RF is based on classification trees (Breiman *et al*., 1984). However instead of constructing one tree, it constructs multiple trees via resampling (bagging); also at each node of each tree RF uses only spectral intensities at a random selection of m/z ratios as potential classification variables. The final class membership is taken to be the class that is most frequently predicted by the classification trees.

For each bootstrap replicate, about one-third of the original data are not included in that simple random sample with replacement, and these 'out-of-bag' samples are used to assess the prediction error of the procedure. Besides reporting the estimated prediction errors, RF also produces a list of how important the spectral intensity at each m/z ratio is in the classification process, as measured by the change in prediction error when values of the spectral intensity at a given m/z ratio are randomized, compared with their proper assignment.

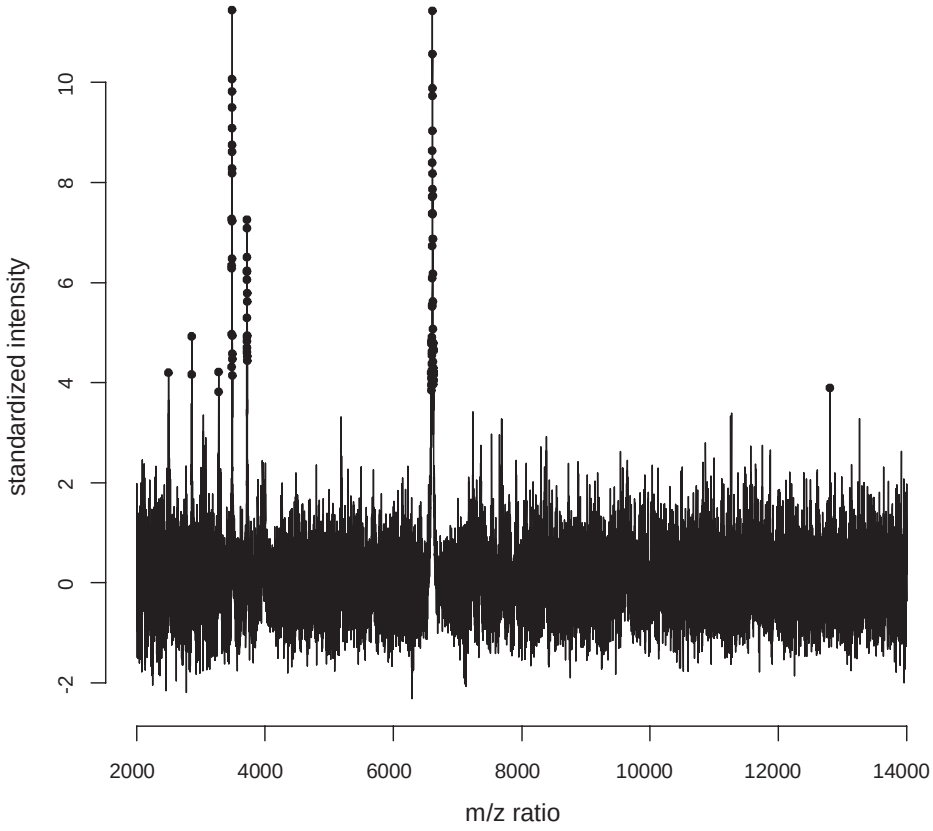

Fig. 4. Signals selected by denoising the standardized *B.anthracis* spectrum using hard thresholding and a cutoff of 6σ.

Training and test datasets

A training set of 60 spectra each from *B.anthracis*, *E.coli*, *S.pneumoniae* 18C-A and *S.pyogenes* was created using the procedures described above. Two test sets were created, the first having 60 spectra each from *B.anthracis*, *E.coli*, *S.pneumoniae* 18C-A and *S.pyogenes* and the second having 60 spectra each from *B.anthracis*, *E.coli*, *S.pneumoniae* 18C-A, *S.pneumoniae* 18C-B and *S.pyogenes*. For each dataset, we used 15 wells per bacterial strain and took four spectra from each well. All samples for a given dataset were run on the same plate, but separate plates were used for the training and each test dataset, and each set was run on a different day. A latin square design was used to avoid possible systematic effects corresponding to well position on the plate.

RESULTS

Training set data

All 240 spectra were standardized and denoised as described above. For illustration, we exhibit the results of standardization and denoising filters applied to the *B.anthracis* spectrum featured in Figure 1. The dots on the standardized spectrum in Figure 4 denote the signal values retained after noise removal. Overall, 3075 m/z ratios were retained after standardization and denoising (i.e. had non-zero values in least one of the 240 spectra). We set RF to grow 500 trees using $55 \approx \sqrt{3075}$ (randomly selected) input variables at each node, the value suggested by Breiman (2001). The estimated classification error rate was 0%, and the RF algorithm successfully classified all of the bacterial specimens in the training set. Figure 5 plots the importance measures (heights) of all the variables (m/z ratios) used by the RF classifier. This plot can be used to determine which peaks should be investigated further by sequencing.

Test set performance and discrimination of bacterial strains

We used the RF model obtained using the training set to classify the bacterial spectra from the two test sets after they were standardized and denoised as described above. Only intensities at the 3075 m/z ratios identified in the training set were used in this classification. In both test sets, all bacterial spectra were correctly classified. In the second set it should be noted that the 60 *S.pneumoniae* 18C-B specimens were all identified as *S.pneumoniae*, even though only *S.pneumoniae* 18C-A isolates were used in constructing the training set. In Tables 1 and 2 we show the classification results for the test datasets.

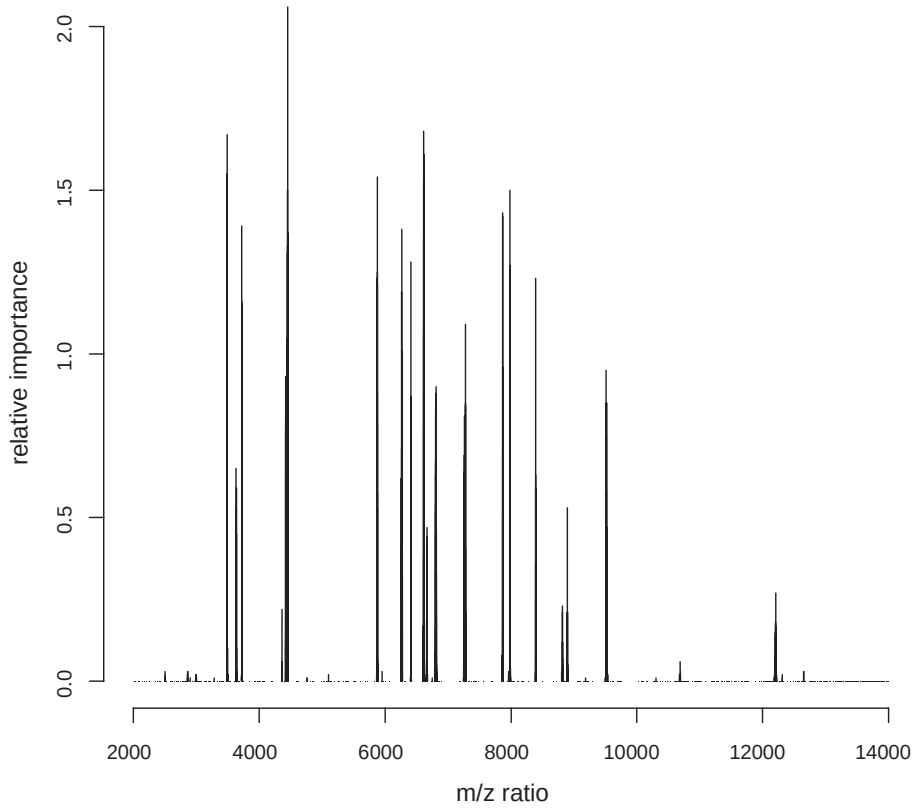

Fig. 5. Relative importance of spectral intensities as a function of m/z ratio. The value plotted on the y-axis represent the relative importance of the intensity at each m/z ratio in classifying bacterial samples using the RF classifier.

Table 1. The results of classification of the first test dataset by RF

Predicted class	True class			
	B.anthracis	*E.coli*	*S.pneumoniae* 18C-A	*S.pyogenes*
B.anthracis	60	0	0	0
E.coli	0	60	0	0
S. pneumoniae 18C-A	0	0	60	0
S.pyogenes	0	0	0	60

To determine whether there was sufficient information in the standardized and denoised *S.pneumoniae* spectra to distinguish between bacterial subtypes, we used the Random Forest program on the second test set, treating it as a training set with five categories. Using intensities at the 3226 m/z ratio values for which at least one spectrum in the second test set had a non-zero intensity after denoising, we fit the RF program using $58 \approx \sqrt{3226}$ variables per node. As with the original training set, all spectra were correctly classified and the estimated error rate was 0%. To validate this result, we reserved 15 spectra from each of the five bacterial strains as a 'test set' and fit the RF to the remaining 45 spectra. Using intensities at the 2889 m/z ratio values identified in the 225 'training set' spectra and 55 variables per node; as shown in Table 3, all 75 'test set' spectra were in fact correctly classified.

DISCUSSION

MALDI-TOF MS has the potential to substantially reduce the time required to classify whole-organism bacterial specimens. However, the spectra are substantially noisier than those of tissue samples. We have developed novel methods for standardizing and denoising whole-organism mass spectra, and have shown that the processed spectra can be used for classification with high reliability. In our data, even closely related bacterial strains could be resolved with high confidence. Using only spectral features that exceed a noise threshold for classification increases confidence in the resulting algorithm.

In our analysis, we used hard thresholding and quantitative intensity values to classify bacterial specimens. We repeated all RF analyses reported in this paper using soft thresholding to determine the effect of thresholding type on classification success. We found the classification ability and error rates of RF were identical using either type of thresholding. Finally, we wondered if the success of RF in classifying our 240 specimens was solely due to the large number of explanatory variables. To assess this possibility, we generated 240 random 'spectra' consisting of 18 190 'intensities' that

Table 2. The results of classification for the 2nd test dataset using the RF classifier based on the four-class training set

Predicted class	True class				
	E.coli	*S.pneumoniae* 18C-A	*S.pneumoniae* 18C-B	*S.pyogenes*	*B.anthracis*
E.coli	60	0	0	0	0
S.pneumoniae 18C-A	0	60	60	0	0
S.pyogenes	0	0	0	60	0
B.anthracis	0	0	0	0	60

Table 3. The results of classification by RF for the 75 'test set' spectra of the second test data, treating the remaining 225 spectra of the 2nd test dataset as a training set with five classes

Predicted class	True class				
	E.coli	*S.pneumoniae* 18C-A	*S.pneumoniae* 18C-B	*S.pyogenes*	*B.anthracis*
E.coli	15	0	0	0	0
S.pneumoniae 18C-A	0	15	0	0	0
S.pneumoniae 18C-B	0	0	15	0	0
S.pyogenes	0	0	0	15	0
B.anthracis	0	0	0	0	15

were independently and identically distributed random variables having a normal distribution with mean 0 and variance 1. These 'spectra' were then randomly assigned to four 'groups' (subject to the constraint that each group has 60 spectra). We found that in strong distinction with our bacterial specimens, RF was unsuccessful in assigning group membership, achieving an estimated prediction error rate of 73% (close to the 75% error rate that would correspond to random classification into four groups with equal probability).

Procedures that enable rapid and reliable identification of bacterial specimens are an important step for bioterrorism preparedness. This report represents an initial step in this direction. Ultimately, we hope to develop methods to identify a large number of closely related bacterial strains using the approaches described here.

ACKNOWLEDGMENTS

We thank Dr Richard Faklam for providing the *S.pyogenes* strain used in this study and Dr Sandra Romero-Steiner for the two *S.pneumoniae* strains. S.D. was partially supported by the Centers for Disease Control and Prevention.

REFERENCES

Bafna,V. and Edwards,N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17**, S13–S21.

Baggerly,K.A., Morris,J.S. and Coombes,K.R. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**, 777–785.

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.

Banks,D. and Petricoin,E. (2003) Finding cancer signals in mass spectrometry data. *Chance*, **16**, 8–12.

Böcker,S. (2003) SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry. *Bioinformatics*, **19**, 44-53.

Breiman,L. (1999) Random forests-random features. *Technical Report 567*, Department of Statistics, University of California, Berkeley, CA.

Breiman,L. (2001) Random forests. *Technical Report 567*, Department of Statistics, University of California, Berkeley, CA. http://oz.berkeley.edu/users/breiman/randomforest2001.pdf

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Ducharme,G.R., Gannoun,A., Guertin,M.-C. and Jéquier,J.-C. (1995) Reference values obtained by kernel-based estimation of quantile regressions. *Biometrics*, **51**, 1105–1116.

Fenselau,C. and Demirev,P.A. (2001) Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrom. Rev.*, **20**, 57–171.

Hawkins,D.M., Wolfinger,R.D., Liu,L. and Young,S.S. (2003) Exploring blood spectra for signs of ovarian cancer. *Chance*, **16**, 19–23.

Havilio,M., Haddad,Y. and Smilansky,Z. (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.*, **75**, 435–444.

Lay,J.O.,Jr (2001) MALDI-TOF mass spectrometry of bacteria. *Mass Spectrom. Rev.*, **20**, 172–194.

Lu,B. and Chen,T. (2003) A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics*, **19**, ii113–ii121.

Nesvizhskii,A.I., Keller,A., Kolker,E. and Aebersold,R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4658.

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

Purohit,P. and Rocke,D.M. (2003) Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics*, **3**, 1699–1703.

Sorace,J.M. and Zhan,M. (2003) A data review and re-assesment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, **4**, 24.

Thiele,H. (2003) Mass spectrometry and bioinformatics in proteomics. *Chance*, **16**, 29–36, 51.

Wu,B., Abbott,T., Fishman,D., McMurray,W., Mor,G., Stone,K., Ward,D., Williams,K. and Zhao,H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.