

## Regular Article

---

# Biomarkers that discriminate multiple myeloma patients with or without skeletal involvement detected using SELDI-TOF mass spectrometry and statistical and machine learning tools

Sudeepa Bhattacharyya<sup>a</sup>, Joshua Epstein<sup>b</sup> and Larry J. Suva<sup>a,\*</sup>

<sup>a</sup>*Department of Orthopaedic Surgery, Center for Orthopaedic Research, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA*

<sup>b</sup>*Myeloma Institute for Research and Treatment, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA*

**Abstract.** Multiple Myeloma (MM) is a severely debilitating neoplastic disease of B cell origin, with the primary source of morbidity and mortality associated with unrestrained bone destruction. Surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) was used to screen for potential biomarkers indicative of skeletal involvement in patients with MM. Serum samples from 48 MM patients, 24 with more than three bone lesions and 24 with no evidence of bone lesions were fractionated and analyzed in duplicate using copper ion loaded immobilized metal affinity SELDI chip arrays. The spectra obtained were compiled, normalized, and mass peaks with mass-to-charge ratios ( $m/z$ ) between 2000 and 20,000 Da identified. Peak information from all fractions was combined together and analyzed using univariate statistics, as well as a linear, partial least squares discriminant analysis (PLS-DA), and a non-linear, random forest (RF), classification algorithm. The PLS-DA model resulted in prediction accuracy between 96–100%, while the RF model was able to achieve a specificity and sensitivity of 87.5% each. Both models as well as multiple comparison adjusted univariate analysis identified a set of four peaks that were the most discriminating between the two groups of patients and hold promise as potential biomarkers for future diagnostic and/or therapeutic purposes.

**Keywords:** SELDI, surface-enhanced laser desorption/ionization, Multiple Myeloma, biomarkers

**Abbreviations:** SELDI-TOF: surface-enhanced laser desorption/ionization-time-of-flight; MM: Multiple Myeloma; PLS-DA: partial least squares-discriminant analysis; RF: random forest; ROC: receiver operating curve.

## 1. Introduction

Multiple myeloma (MM) a B-cell neoplasia characterized by the clonal expansion of plasma cells in the bone marrow encompasses approximately 1% of all hematologic malignancies within the United States [1].

---

\*Corresponding author: Larry J. Suva, Ph.D., Department of Orthopedic Surgery, Center for Orthopedic Research, University of Arkansas for Medical Sciences, 4301 W. Markham St, #644, Little Rock, AR 72205, USA. Tel.: +1 501 526 6110; Fax: +1 501 686 8987; E-mail: SuvaLarryJ@uams.edu.

MM is the most malignant of the plasma cell dyscrasias, which include the precursor stages of monoclonal gammopathy of unknown significance (MGUS) and indolent or smoldering myeloma [2,3]. Recent clinical and experimental observations suggest that the development of lytic bone disease may drive the progression of MM [4]. Lytic bone disease is usually absent in MGUS, is very limited and asymptomatic when present in indolent myeloma, but is frequently associated with the advanced stages of the disease [5]. The osteolysis associated with late stage MM is perhaps one of the most debilitating manifestations of the disease. The symptoms often include hypercalcemia, pathological fractures with acute and chronic pain, reduced mobility and the inability to fully participate in normal day-to-day activities [2,3,6,7]. In radiographic skeletal scans, myeloma bone disease appears as circular “punched-out” areas in involved hematopoietic bone marrow sites, as diffused osteopenia, or their combination [5].

The bone destruction that is associated with MM progression results from the activation of osteoclasts (bone resorbing cells) and the suppression of osteoblast (bone forming cells) activity in the myelomatous bone marrow [4,8,9]. Clinically, changes in bone turnover rates (measured by increased osteoblastic and osteoclastic activity) precede the progression of MGUS to overt myeloma by as long as 3 years [10]. Although several bone-resorbing cytokines including IL-1 $\alpha$ , IL-1 $\beta$ , TNF- $\alpha$  and IL-6 have been found to be associated with increases in myeloma proliferation and osteoclast activity [11,12], definitive diagnostic tests and early detection strategies remain elusive. Thus, the identification of changes in protein biomarkers in myeloma patient serum may be the first indication of disease progression, and enable the early detection of MM at its presymptomatic stage.

The advent of SELDI-TOF-MS has provided the means for analyzing a broad array of proteins of different physical properties directly in patient samples [13, 14]. SELDI-TOF MS was selected as the diagnostic tool of choice, based on the impressive accuracy of the technology when applied to the diagnosis of a variety of cancers, including ovarian [15,16], prostate [17], breast [18,19] and pancreas [20,21]. This technology, combined with bioinformatics and/or statistical data analysis, has proven especially useful in diagnosing diseases where available tests are either too invasive or perhaps limited by poor diagnostic accuracy, such as in the progression of MM.

In the study described here, serum from MM patients with and without skeletal complications was profiled

to identify protein patterns indicative of bone disease status. Using univariate statistical analysis as well as linear (PLS-DA) and non-linear (RF) classification algorithms we were able to generate a diagnostic fingerprint that holds great promise as a potential serum biomarker profile for the diagnosis and treatment of MM progression.

## 2. Materials and method

### 2.1. Sample collection and preparation

In all cases, blood was collected using a University of Arkansas for Medical Sciences (UAMS) IRB approved protocol from patients in the Myeloma Institute for Research and Therapy (MIRT) at UAMS. Serum samples were obtained from 24 MM patients with no radiographic evidence of bone metastasis and 24 multiple myeloma patients with greater than three radiographically identified bone lesions, based on a review of patient records. Serum was obtained and stored in small aliquots, at or below  $-80^{\circ}\text{C}$  until processing. No samples underwent more than two freeze-thaw cycles before SELDI-TOF MS analysis [22].

### 2.2. Patient demographics

Serum specimens were analyzed from forty-eight age- and sex-matched archived MM patient samples (24 were with  $\geq 3$  bone lesions (bone disease) (average age  $55.2 \pm 9.7$  years) and 24 without apparent bone lesions (no bone disease) (average age  $55.4 \pm 8.9$  years), diagnosed with MM between February 1998 and December 2001. Blood samples were obtained from all patients during their regular clinic attendance. The bone disease group comprised 9 females and 15 males, and the no bone disease group comprised 6 females and 18 males.

All patients were treated with a number of agents, according to the status of their MM. In addition, all MM patients with evidence of bone disease were treated with the standard anti-catabolic bisphosphonate therapy (pamidronate) used at our Institution. Treatments in the no bone disease group varied due to disease status and included Coumadin, or dexamethasone or procrit. No patients in this group required or were treated with bisphosphonate.

Table 1  
Highest ranking 1.5 fold peaks with significant multiple comparison adjusted p values (<0.1)<sup>a</sup>

1.5 fold peaks (mass in KDa)	Fraction	Median Fold change (up bone lesion)	Raw p-values (WRS) <sup>b</sup>	Adjusted p-values <sup>c</sup>	Significant in PLS-DA <sup>d</sup>	RF <sup>e</sup>
<b>5.802</b>	<b>F6</b>	<b>-1.6</b>	<b>0.00014</b>	<b>0.00308</b>	✓	✓
<b>6.639</b>	<b>F3 +4</b>	<b>-2.7</b>	<b>0.00020</b>	<b>0.00308</b>	✓	✓
<b>6.443</b>	<b>F3 +4</b>	<b>-2.4</b>	<b>0.00153</b>	<b>0.01555</b>	✓	✓
<b>2.019</b>	<b>F2</b>	<b>+2.0</b>	<b>0.00201</b>	<b>0.01555</b>	✓	✓
<b>10.472</b>	<b>F1</b>	<b>-2.1</b>	<b>0.00593</b>	<b>0.03677</b>	✓	✓
14.898	F1	+1.5	0.01334	0.06894	✓	✓
11.705	F6	-1.6	0.01753	0.07763	✓	✓
6.512	F1	-1.8	0.02586	0.09078	✓	✓
5.094	F1	-2.2	0.02753	0.09078	✓	✓
11.744	F3 + 4	+1.6	0.02928	0.09078		

a. The peaks in bold were significant at the level of  $p < 0.05$  after multiple comparison adjustments.

b. Wilcoxon rank sum test with t-approx.

c. Multiple comparison adjustments via false discovery rate.

d&e. Among the top ten peaks in partial least squares-discriminant analysis and random forest classification model respectively.

### 2.3. Serum processing and fractionation

To increase the sensitivity of peak detection and alleviate signal suppression effects on low-abundance proteins from preponderant species such as albumin, serum samples were fractionated into six fractions on the basis of their isoelectric point [21]. Serum samples were loaded into each well of a 96-well filter plate pre-filled with an anion exchange sorbent (Serum Fractionation kit, CIPHERGEN Biosystems, CA) and eluted in a stepwise pH gradient using a BIOMEK 2000 (Beckman Coulter, Fullerton, CA) liquid-handling robot according to the manufacturer's protocol. The six fractions obtained in this stepwise fashion, designated F1 through F6, contained flow-through plus proteins eluted with buffers of pH 9, pH 7, pH 5, pH 4, pH 3 and organic solvent, respectively. Each serum sample was diluted approximately 10 fold during fractionation in 50 mM Tris-HCl with the pH adjusted for the different fractions and containing 0.1% nonionic detergent.

### 2.4. Protein chip SELDI TOF-MS analysis

Three different chip chemistries {metal binding IMAC3 (present name: IMAC30), strong anion exchange SAX2 (present name: Q10) and weak cation exchange WCX2 (present name: CM10), CIPHERGEN Biosystems} were initially evaluated in a pilot study to determine which type provided the best spectral profiles in terms of peak number and resolution (data not shown). The IMAC3 metal binding chip consistently captured the most peaks in the majority of the fractions and was selected for analysis. Each fraction was individually loaded on to the IMAC3 chip arrays, except for

fractions 3 and 4 which were combined before loading, as fewer peaks were observed in these fractions during the preliminary study. The serum samples from each fraction were diluted 1:5 fold in phosphate-buffered saline (PBS) and applied to the wells of a 96-well bio processor containing 8-spot IMAC3 chips (CIPHERGEN) previously activated with 100 mM CuSO<sub>4</sub>, as described by the manufacturer. The bio processor was then sealed and incubated with the samples for an hour with vigorous agitation on a Micromix 5 platform shaker. Excess sera was discarded, and the chips washed three times with PBS and twice with deionized water before being removed from the bio processor, and air dried for 20 minutes. A saturated solution of sinapinic acid in 50% acetonitrile, 0.5% trifluoroacetic acid (0.5  $\mu$ l) was then applied to each spot of the protein chip arrays. Each spot surface was allowed to dry for 10 minutes before another application of 0.5  $\mu$ l of the sinapinic acid solution. All sample handling procedures were carried out using the BIOMEK 2000 robotic system, minimizing errors due to technician intervention.

ProteinChips were placed in the Protein Biological System II C mass spectrometer reader (CIPHERGEN) and the time-of-flight spectra generated by averaging 156 laser shots collected in the positive mode at a laser intensity of 180; detector sensitivity of 8, and focus lag time of 782 ns. All data acquisition parameters were optimized to detect peaks in the range of 2–20 kDa, as this range contains the majority of the resolved protein/peptide peaks. Mass accuracy was calibrated using the All-in-one peptide and All-in-one protein molecular weight standards (CIPHERGEN). Each chip generated included a randomly assigned control sample (pooled serum from normal healthy individuals) in order to assess inter assay and inter spot variability.

## 2.5. Data processing

Acquisition and preprocessing of all spectral data was performed using Ciphergen ProteinChip software version 3.1. All peaks were baseline corrected and peak intensities normalized to the total ion current of  $m/z$  between 2000–20,000 Da as described previously [21]. All spectra were normalized on a fraction by fraction basis and individually examined to exclude spectra that showed ion-current saturation or a lack of peaks. Spectra with normalization factors below 0.5 or above 2.5 were recollected at slightly lower (170) or higher (190) laser intensities so that all normalization factors were between 0.5 and 2.0. Biomarker Wizard software (Ciphergen) was used to compile spectra and detect peaks that were consistently present across a minimum of 10% of the spectra with a signal-to-noise ratio of  $\geq 2.0$ . Selected peaks were clustered using a second-pass peak selection within a 0.3% mass window. Sample statistics were performed on spectra from each of the fractions separately, by groups of profiles (MM without bone lesions vs. MM with  $\geq 3$  bone lesions). Peak intensities were considered statistically significantly different at  $p$ -value  $\leq 0.05$ .

## 2.6. Normalization across fractions and Intra-sample correlation assessment

All peak intensities from all fractions were compiled, transformed to their base-2 logarithms, and then centered and scaled on a peak-by-peak basis to means of zero and standard deviations of one (Fig. 1).

Samples giving rise to paired spectra were used to assess intra-sample correlations by determining the Pearson correlation among pairs for each peak. The Pearson correlation coefficients had a median value of 0.89 with the interquartile range lying between 0.80 and 0.94. Even though samples were applied robotically and in a randomized fashion on the protein chip surfaces, high median intra-sample correlations were observed due to high spot-to-spot reproducibility of the SELDI system. Hence, the paired spectra from each patient were averaged together on a peak-by-peak basis for subsequent data analysis.

## 2.7. Feature selection and Univariate analysis

For each peak, the median patient-averaged intensity was calculated for the bone lesion (bone) and no bone lesion (no bone) groups. The difference in group medians was reported as a ratio, the fold change. Peaks were

first pre-selected for all subsequent analysis using the biological criterion of having a  $> 1.5$  fold change in the median peak intensity level between the groups. These peaks, referred to as 1.5 fold peaks henceforth, were assessed for statistical significance via Wilcoxon rank-sum test with t-approximation. Multiple-comparison adjustment of  $p$ -values was via false discovery rate and the Step-down Permutation procedure of Westfall and Young [23] using 100,000 random permutations of class labels. A peak with a multiple-comparison-adjusted  $p$ -value  $< 0.05$  was considered statistically significantly different.

All statistical analyses were performed using SAS version 9.0 (SAS institute) and S-plus version 6.2 (Insightful Corporation) statistical software.

## 2.8. Partial least squares discriminant analysis

PLS-DA attempts to find variance in the set of predictor variables (X-data) that correlates with variance in the response variables (Y-data). PLS was developed as an econometric technique [24] but has been used as a useful tool in classifying microarray [25,26] and SELDI data [27].

In the MM sample set, the centered and scaled 1.5-fold peaks were used as the predictor variables while the Y-data set was created by indicating the classes (bone lesion vs. no bone lesion). The significant PLS-components were determined by leave-one-out cross-validation. The model was also validated by external validation on an independent data set that was created by setting aside 10 randomly selected specimens from the sample set, 5 from each group. The remaining 38 samples were used as a training set to build the model which was then applied to the independent test set and the classification accuracy was recorded. The variable influence on projection (VIP) [28] and PLS regression coefficients were used to determine the peaks that were most important in driving the separation between the classes.

## 2.9. Random forest classifier

The random forest algorithm [29] (available freely from <http://www.stat.berkeley.edu/users/breiman/RandomForests/>) was conducted by bagging a classification tree coupled with random feature selection. Bagging was performed by re-sampling with replacement 2500 bootstrap subsets from the MM data set containing 70% of the data from each group. No transformation, centering or scaling was performed on the peak

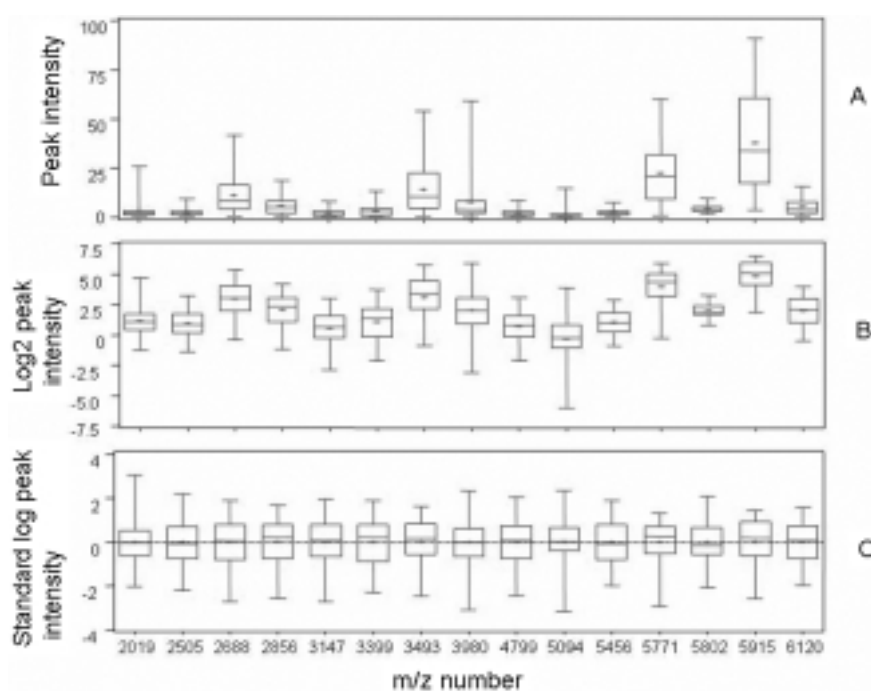


Fig. 1. Peak intensities from different fractions were log transformed, centered to means of zero and scaled to standard deviation of one. A) Raw peak intensities (prior to transformation and standardization) of a fraction of the peaks with  $m/z$  between 2–6.2 KDa. B) The same peaks after  $\log_2$  transformation. C) After centering and scaling, the same peaks showed increased uniformity in their distribution.

intensities prior to input into the RF algorithm. A Tree-classifier was constructed on each of these subsets and predictions were recorded on the remaining 30% of the data (called out-of-bag samples). When constructing the decision trees, 5 random inputs from the 1.5 fold eligible predictors were attempted at each split. Final prediction from the RF classifier is the out-of-bag estimator averaged over all the 2500 bootstrap samples. The amount of increase in the prediction error when the value of a splitter for the out-of-bag samples was randomly perturbed was used to score the importance of the splitters in constructing the tree-classifiers.

### 3. Results

#### 3.1. The compiled dataset and univariate analysis

Forty-eight age- and sex-matched archived MM patient serum samples (24 were with  $\geq 3$  bone lesions (bone disease) (average age  $55.2 \pm 9.7$  years) and 24 without apparent bone lesions (no bone disease) (average age  $55.4 \pm 8.9$  years) diagnosed with MM between February 1998 and December 2001 were assayed by SELDI-TOF MS were selected from the UAMS Myelo-

ma Institute for Research and Therapy (MIRT) tissue bank. The bone disease group comprised 9 women and 15 men, whereas the no bone lesion group contained 6 women and 18 men.

After ion-current normalization on a fraction by fraction basis, 94, 95, 97, 96 and 93 spectra were compiled from fractions 1, 2, 3 + 4, 5 and 6, respectively. A total of 168 peaks were resolved from all the fractions in the 2–20 KDa mass range.

A total of 31 peaks showed median fold change  $> 1.5$  between the two groups of which 5 peaks were found to be significant with multiple comparison adjusted  $p$ -values  $< 0.05$  (Table 1). Four of these peaks ( $m/z$  5.802 kDa from F6, 6.639 kDa and 6.443 kDa from F3 and 10.472 kDa from F1) were found to be higher in sera from patients without bone lesions (1.56 to 2.68 fold) while one peak (2.019 from F2) was higher in patients with bone lesions (1.92 fold). Representative protein spectra from different bone disease versus no bone disease samples, showing the univariately significant peaks are shown in Fig. 2).

#### 3.2. Partial least squares discriminant analysis

The objective with PLS-DA is to find models that allow the maximum separation among classes of ob-

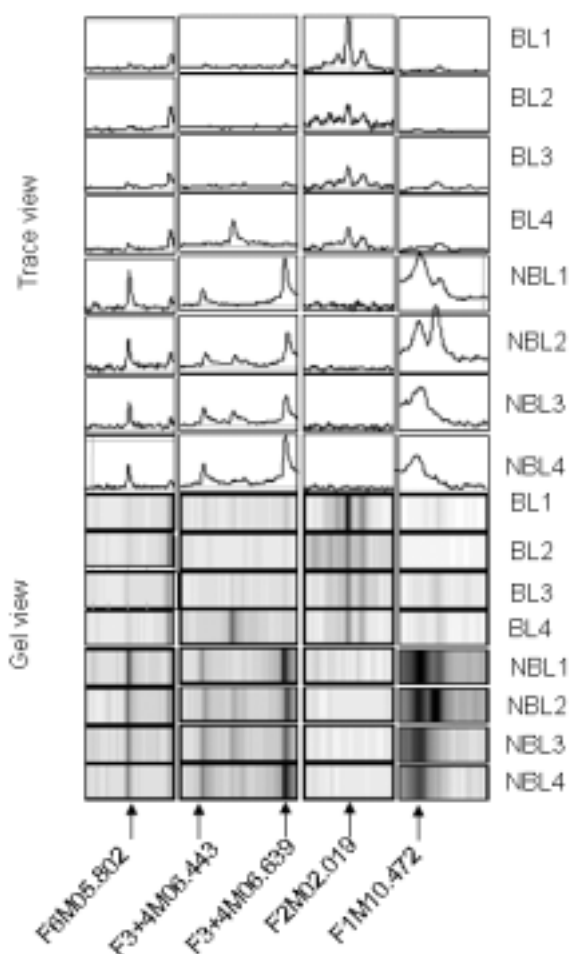


Fig. 2. Differences in peak intensities in the normalized spectra of bone lesion (BL) versus no bone lesion (NBL) groups. The top eight spectra depict the differences in expression levels in TraceView mode while the lower eight depict the difference in GelView mode for the univariately significant peaks of  $m/z$  5.802 (F6), 6.443 & 6.639 (F3 + 4), 2.019 (F2) and 10.472 (F1), indicated by the arrows.

jects in high dimensional datasets [30]. Unlike principal component analysis where only the X-scores are chosen to explain as much of the predictor variation as possible, in PLS the X- and Y-scores are selected such that the relationship between successive pairs of scores is as strong as possible. Thus, PLS attempts to extract the latent factors that account for as much of the manifest predictor variation as possible while modeling the responses well [30].

The PLS-DA model generated here has three significant PLS components determined by leave one out cross-validation. The number of factors chosen usually minimizes the predicted residual sum of squares (PRESS), which is a measure of the predictability of the model [31]. Usually PRESS is re-expressed as  $Q^2$ ,

the cross-validated  $R^2$  and is calculated as  $1 - \text{PRESS}/\text{SS}$  where SS is the sum of squares of the response corrected for the mean. In PLS-DA analysis both  $R^2$  and  $Q^2$  are important parameters to evaluate the predictive power of the model being investigated [28,31]. The final 3-component model had  $R^2(Y) = 0.79$  and  $Q^2(Y) = 0.59$ . Thus, three of the PLS-DA components alone were able to explain 79% of the response variance. These data suggest that there is information in the spectra that correlates extremely well with the two group differences. The separation of the training observations in the three PLS components are shown in Fig. 3. As well, the predictive ability of the model expressed as  $Q^2$  is also very good. Generally, an accumulated predicted variation share larger than 0.5 is regarded as good [28].

External validation performed on the model by applying it on a separate holdout test set of 10 randomly selected samples showed a prediction accuracy of 100% i.e., all 10 out of 10 samples were correctly classified by the model (Table 2). External validation by leave-one-out cross validation on the entire data set produced a prediction accuracy of 96%. The predicted classification was determined using the simple-rule that if the predicted class membership of the bone lesion group was greater than 0.5 then classify the sample as bone-lesion class. In order to determine which peak variables contributed most in driving the separation between the two classes, the variable influence on projection (VIP) and PLS regression coefficients were analyzed. While VIP parameters point to the variables that contribute most in explaining both X- and Y-data [28,32], the coefficients indicate which X-variables contribute in modeling the Y-variables structure. The VIP ranking of the top ten peaks with high PLS coefficients ( $>0.1$ ) and VIP parameters ( $>0.8$ ) that contributed most in the PLS-DA model are shown (Fig. 4). The highest contributors are  $m/z$  11.705 and 5.802 from F6, 6.639 and 6.443 from F3, 14.898 from F1 and 2.019 from F2.

### 3.3. Random forest classification

The random forest algorithm [29] uses an ensemble of classification trees that can achieve both low-bias and low-variance by averaging over a large number of low-bias, high-variance but low-correlation trees [33, 34]. The algorithm operates as follows. First from a training set of  $n$  molecules, bootstrap samples of the same size are drawn randomly, with replacement. In the process, some molecules are left out while others are repeated in the sample. The left out molecules

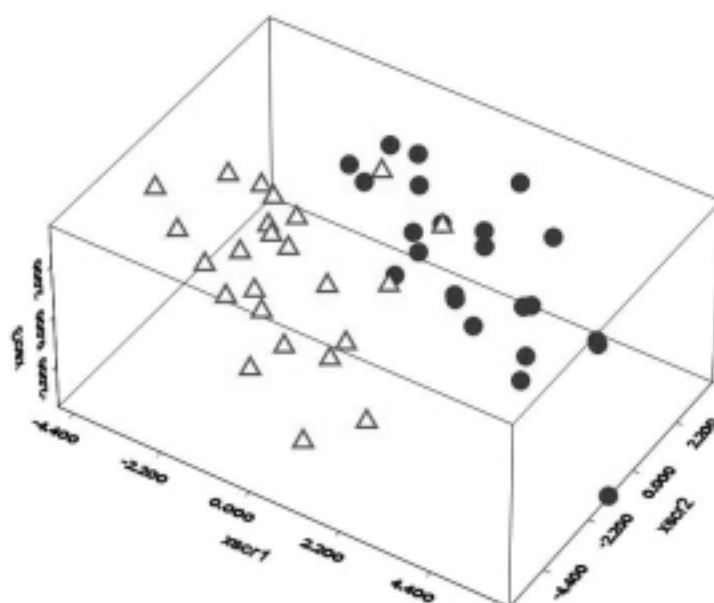


Fig. 3. PLS score plots based on the peak intensities identified in the bone lesion and the no bone lesion groups of MM patients. Triangles represent no-bone lesion group; solid circles the bone lesion group. A clear separation between the two classes is observed.

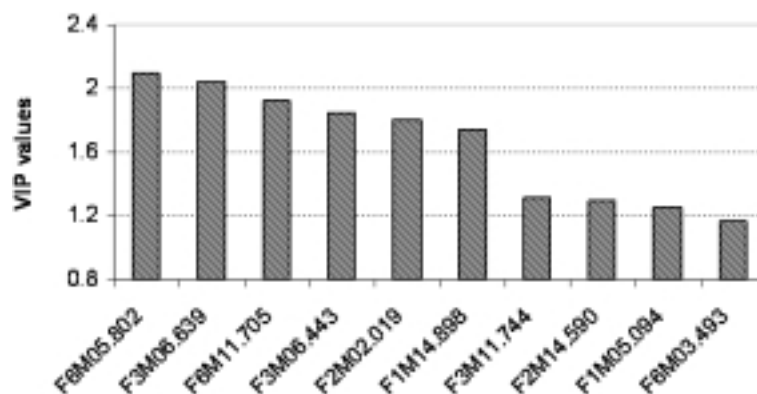


Fig. 4. The VIP ranking of the ten peaks with the highest PLS coefficients ( $>0.1$ ) and VIP parameters ( $>0.8$ ) that contributed most in the PLS-DA model.

constitute the 'out-of-bag' sample. For each bootstrap sample a classification tree is grown. At each node the best split is chosen from among a randomly selected subset of predictor variables (rather than all). The trees are grown to the maximum size till no further splits are possible and not pruned back. Since the 'out-of-bag' samples have not been used in the tree construction, they are used to estimate the ensemble prediction performance. RF prediction for a molecule is computed by averaging the tree predictions over trees for which the given molecule was 'out-of-bag'. The RF method correctly classified 21 out of 24 MM samples with bone lesion (87.5% sensitivity) and also correctly classified

21 out of 24 MM samples with no bone lesions (87.5% specificity) (Table 3). The area under the receiver operating characteristic (ROC) curve for the 'out-of-bag' cases was 0.91. It is generally accepted that the area under the ROC curve  $>90\%$  is satisfactory in diagnosis. The RF analysis used 25 out of the 31 predictors to construct the predictive models. The predictors were ranked based on an importance measure between 0 and 1. A major importance measure indicates that random permutation of that peak variable causes the samples to be misclassified more often and hence that peak variable is important. Six peaks had an importance score  $>0.35$  (Fig. 5). Interestingly, five of these six peaks

Table 2  
PLS-DA prediction results on the separate test set

Patient-ID	Pred. Classification (Class: Bone Lesion)	Pred. Class	Actual Class
Pt06	0.690773	BL <sup>a</sup>	BL
Pt25	0.596842	BL	BL
Pt34	0.642156	BL	BL
Pt45	0.788041	BL	BL
Pt53	0.722192	BL	BL
Pt17	0.093782	NBL <sup>b</sup>	NBL
Pt28	0.228806	NBL	NBL
Pt44	0.100527	NBL	NBL
Pt54	0.200266	NBL	NBL
Pt66	0.258923	NBL	NBL

<sup>a</sup>. myeloma patients with bone lesions.

<sup>b</sup>. myeloma patients with no bone lesions.

Table 3  
'Out of bag' prediction results by RF classification

	Total cases	Correct	Percent
sensitivity	24 (BL)	21	87.5
specificity	24 (NBL)	21	87.5

were also identified as top contributors in the PLS-DA model. The specific peaks are m/z 5.802 from F6, 6.639 and 6.443 from F3, 2.019 from F2 and 14.898 from F1.

#### 4. Discussion

A major challenge confronting the management of MM is a way to positively impact patient survival through the diagnosis of disease progression. Proteomic signature analysis accurately distinguished serum samples from patients with MM bone disease and no bone involvement. The development of our technique depended not only on proteomic technology (SELDI), but has also involved critical post-acquisition analysis of obtained spectra. Currently, the choice of spectral analysis algorithms by various investigators relies on personal preference, although these choices have profound effects on the diagnostic accuracy [34,35,33,37].

Due to the multifactorial nature of MM bone disease [5], it is highly likely that a combination of multiple markers will be necessary to diagnose the disease with high specificity and sensitivity. To search for these diagnostic biomarkers serum samples from 48 MM patients with and without bone disease were examined in this study. Each serum sample was fractionated to increase the peak resolution in SELDI spectra and normalized peak information from each of the fractions merged into an input data matrix that included 48 samples and 168 feature (predictor) vectors. Such small datasets are often a reality in biomedical research, since

obtaining large number of serum samples for many diseases can be difficult and expensive [38].

Our approach to the analysis of this small, but important MM patient sample set was novel in two ways. First, we adopted a non-statistical criterion of predictor variable selection based on a >1.5 fold change in the median peak intensity level between the groups. This was because our objective was not only prediction, but also to identify a small set of proteins with good predictive performance that could subsequently be used for diagnostic and/or therapeutic purposes in the progression of MM. We hypothesized that this pre-selection criterion is biologically relevant for both research and diagnostic purposes. In addition, we determined that this criterion did not adversely affect the prediction outcome of the study. In fact, when all 168 predictors were used as input variables against 48 samples, the sensitivity and specificity predicted by the RF algorithm decreased to 83.3% and 75% respectively, presumably, due to noise in the data.

Variable selection derived in an iterative fashion from the importance measures based on RF itself was performed and the error cost determined from the models built at each iterative step. The misclassification error rate did not show improvement using any of these models. In PLS-DA there was a small (2%) increase in the cumulative fit to the Y-data, when all 168 predictors were used. However, 8 out of the first 11 important peaks based on the VIP scores were preserved in both models. Thus, although rule-based peak selection may miss peaks of very low amplitude this does not appear to be a significant problem, presumably because of the large number of discriminating peaks.

Second, to increase the robustness of our analytical approach a linear (PLS-DA) as well as a nonlinear (RF) classification algorithm was selected. Linear classification techniques like PLS-DA make the assumption of the existence of a linear relation between the predictor variables which may not be the case always in spectral data. Both these unrelated algorithms, that have been shown to perform extremely well on high dimensional data, were able to obtain high prediction accuracy on the MM patient data set, as well as point to the same 4 peak variables as the top predictors of the respective classification models. It is worth noting that this set of 4 variables were also found to be statistically significant based on multiple comparison adjusted WRS test with t-approximation.

Multivariate data-analysis tools such as PLS-DA are sensitive to pre-processing steps like scaling, centering and where appropriate, transformation [30]. In or-



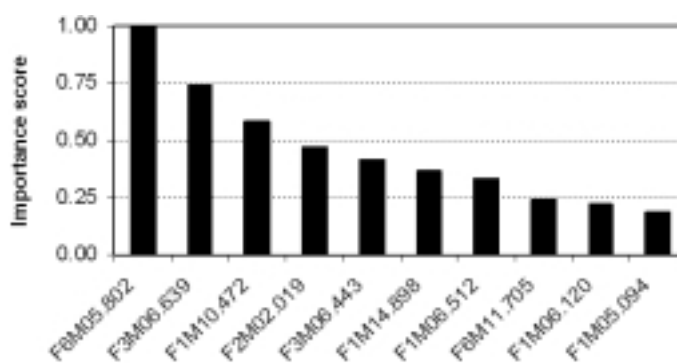


Fig. 5. Variable importance score of the highest ranking 10 peaks from RF. The first six peaks had an importance score  $>0.35$ .

der to ensure that efficient biological comparisons can be made between the samples based on the peak intensity information compiled from six different serum fractions that were collected on separate days, log<sub>2</sub>-transformation, centering and scaling were applied to the input data set for the PLS-DA algorithm. This significantly increased the class separation along the first three PLS components. This preprocessing step, however, was not performed for the RF algorithm which is not sensitive to any kind of rescaling, transformation or modification.

These high dimensional data sets require a combination of feature selection methods and robust classifiers that can handle complex data and effectively recognize hidden patterns. We used two such classifiers, PLS-DA and RF that have been successfully applied in high dimensional microarray data [25,26,39] as well as SELDI data [27,34,40]. Wu et al. has recently shown that the RF algorithm outperformed other methods like linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbor (k-NN), bagging and boosting classification trees in classifying SELDI generated ovarian cancer data [40]. RF uses bagging (bootstrap aggregation) to combine unstable learners and random variable selection for tree building. Each tree is left unpruned to obtain low-bias trees, while bagging and random variable selection results in low-correlation of the individual trees. PLS-DA is an excellent classifier that is resilient to noise in the data and performs dimensional reduction and discrimination among the groups in a simultaneous fashion [30]. Thus, PLS-DA produces more optimal models compared to Principal Component Analysis-Discriminant analysis.

The RF algorithm produced an 'out-of-bag' sensitivity of 87.5% as well as a specificity of 87.5% on the MM data set. 'Out-of-bag' or OOB testing [34,41, 42] is very similar to a cross validation that is repeated

many times with each replication starting with a random reordering of the data. In small data sets which lack a separate test set OOB results have been shown to be fully reliable and produce prediction accuracies nearly identical to that obtained from separate test sets. The three-component PLS-DA model generated with leave-one-out cross-validation was able to explain almost 80% of the response variance in the data. External validation performed on a separate test set produced a prediction accuracy of 100% while leave-one-out cross validation on the training data produced a prediction accuracy of 96%.

The four top-ranked peaks (of m/z 5.802 from F6, 6.639 and 6.443 from F3, and 2.019 from F2) derived from t-tests, multiple comparison adjusted significance testing, PLS-DA and the RF models were identified to be of high value in the diagnosis of skeletal involvement in MM. It is particularly important to sequence and identify these protein biomarkers. These protein peaks had a  $>1.5$  fold difference in expression level between the two groups of MM patients. Once the individual biomarkers are identified and sequenced, a more conventional and cost-effective method detection should result in less patient-to-patient variability and improve diagnosis. The identification of these biomarkers would also help in the understanding of the pathophysiology of MM. The identification of these specific biomarkers and the subsequent development of a diagnostic assay are ongoing.

This diagnostic method needs to be validated using additional patient samples, especially since the accuracy of our test for identifying MM bone disease compares favorably with that reported using the same technology to diagnose ovarian cancer (sensitivity 100%, specificity 95%) [15], breast cancer (sensitivity 93%, specificity 91%) [18,19], pancreatic cancer (100% sensitivity and 93.5% specificity) [20,21] and prostate cancer (100% specificity and sensitivity) [17].

We adopted a novel approach to data analysis by using a number of different but complementary techniques. The power of our diagnostic method is many-fold: first, similar results were obtained using unrelated analytical techniques, indicating the robustness of the selected proteomic approach; second, the majority opinion (based on multiple analyses) improves predictive accuracy, since algorithms tend to misassign different patients. For the few individual errors, collective opinion should either improve the performance of the worst algorithm or leave it the same [43].

A major advantage of the proteomic biomarker signatures approach described here is that it does not assume differences in the spectral peaks obtained from different patient samples. From a bone disease diagnosis standpoint, there is no need to identify further the nature of new peaks associated with bone disease (positive biomarkers), nor to explain why peaks only present in MM patients with no bone involvement disappear in patients with bone disease (negative biomarkers). Nevertheless, biomarkers may be useful for improving our understanding of the pathophysiology of MM bone involvement.

In summary, this report describes the accurate diagnosis of bone involvement in MM using proteomic signature analysis. The same approach could equally be used to improve the early diagnosis of the disease, such as MGUS and eventually other bone diseases such as osteoporosis.

## Acknowledgements

The authors wish to thank Eric Siegel, for critical review of the manuscript and outstanding bioinformatics advice and guidance. These studies were supported by an Arkansas Breast Cancer Research fund grant to LJS and NIH CA55819-06 to JE.

## References

- [1] G.J. Bublely and L.E. Schnipper, Multiple Myeloma, in: *Clinical Oncology, American Cancer Society Textbook*, G.P. Murphy, W. Lawrence, Jr., R.E. Lenhard, Jr., eds, American Cancer Society, Washington D.C., 1995, pp. 470–485.
- [2] M.M. Oken, Multiple Myeloma: prognosis and standard treatment, *Cancer Invest* **15** (1997), 57–64.
- [3] B.G.M. Durie and S.E. Salmon, A clinical staging system for multiple myeloma, *Cancer* **36** (1975), 842–854.
- [4] E. Tian, F. Zhan, R. Walker, E. Rasmussen et al., The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma, *N. Engl. J. Med* **349**(26) (2003), 2483–2494.
- [5] B. Barlogie, J. Shaughnessy, G. Tricot, J. Jacobson et al., Treatment of multiple myeloma, *Blood* **103**(1) (2004), 20–32.
- [6] G.D. Roodman, Biology of Osteoclast Activation in Cancer, *J. Clin. Oncol* **19**(15) (2001), 3562–3571.
- [7] G.R. Mundy, Myeloma bone disease, *Eur. J. Cancer* **34**(2) (1998), 246–251.
- [8] R. Bataille, D. Chappard and B. Klein, Mechanisms of bone lesions in multiple myeloma, *Hematol. Oncol. Clin. North Am.* **6**(2) (1992), 285–295.
- [9] T. Taube, M.N. Beneton, E.V. McCloskey, S. Rogers et al., Abnormal bone remodelling in patients with myelomatosis and normal biochemical indices of bone resorption, *Eur. J. Haematol.* **49**(4) (1992), 192–198.
- [10] R. Bataille, D. Chappard, C. Marcelli, P. Dessauw et al., Recruitment of new osteoblasts and osteoclasts is the earliest critical event in the pathogenesis of human multiple myeloma, *J. Clin. Invest.* **88**(1) (1991), 62–66.
- [11] B. Klein, Cytokine, cytokine receptors, transduction signals, and oncogenes in human multiple myeloma, *Semin. Hematol.* **32**(1) (1995), 4–19.
- [12] B. Barlogie, S. Jagannath, J. Epstein, N. Munshi et al., Biology and therapy of multiple myeloma in 1996, *Semin. Hematol.* **34** (1997), 67–72.
- [13] N. Tang, P. Tornatore and S.R. Weinberger, Current developments in SELDI affinity technology, *Mass Spectrom. Rev.* **23**(1) (2004), 34–44.
- [14] M. Merchant and S.R. Weinberger, Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry, *Electrophoresis* **21**(6) (2000), 1164–1177.
- [15] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine et al., Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* **359**(9306) (2002), 572–577.
- [16] K.R. Kozak, M.W. Amneus, S.M. Pusey, F. Su et al., Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: Potential use in diagnosis and prognosis, *Proc. Natl. Acad. Sci. USA.* **100**(21) (2003), 12343–12348.
- [17] Y. Qu, B.L. Adam, Y. Yasui, M.D. Ward et al., Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, *Clin. Chem.* **48**(10) (2002), 1835–1843.
- [18] J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang and D.W. Chan, Proteomics and Bioinformatics approaches for identification of serum biomarkers to detect breast cancer, *Clin. Chem.* **48**(8) (2002), 1296–1304.
- [19] C.P. Paweletz, B. Trock, M. Pennanen, T. Tsangaris et al., Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer, *Dis. Markers.* **17**(4) (2001), 301–307.
- [20] J. Koopmann, Z. Zhang, N. White, J. Rosenzweig et al., Serum Diagnosis of Pancreatic Adenocarcinoma Using Surface-Enhanced Laser Desorption and Ionization Mass Spectrometry, *Clin. Cancer Res.* **10**(3) (2004), 860–868.
- [21] S. Bhattacharyya, E.R. Siegel, G.M. Petersen, S.T. Chari et al., Diagnosis of pancreatic cancer using serum proteomic profiling, *Neoplasia* **6**(5) (2004), 674–686.
- [22] W.E. Grizzle, O.J. Semmes, W. Bigbee, L. Zhu et al., The Need for Review and Understanding of SELDI/MALDI: Mass Spectroscopy Data Prior to Analysis, *Cancer Informatics* **1**(1) (2005), 86–97.
- [23] P. H. Westfall and S.S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, Inc., New York 1993, NY.

- [24] H. Wold, Estimation of principal components and related models by iterative least squares, in: *Multivariate Analysis*, P.R. Krishnaiah, ed., Academic Press, New York, 1966, pp. 391–420.
- [25] D.V. Nguyen and D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18**(1) (2002), 39–50.
- [26] D.V. Nguyen and D.M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics* **18**(9) (2002), 1216–1226.
- [27] M. Lundquist, M.B. Caspersen, P. Wikstrom and M. Forsman, Discrimination of *Francisella tularensis* subspecies using surface enhanced laser desorption ionization mass spectrometry and multivariate data analysis, *FEMS Microbiol. Lett.* **243**(1) (2005), 303–310.
- [28] L. Eriksson, E. Johansson, N. Kettanah-Wold and S. Wold, *Introduction to multi and megavariate data analysis using projection methods (PCA and PLS)*, Umetrics AB, Umeå, 1999, 490.
- [29] L. Breiman, Random Forests, *Mach. Learning.* **45** (2001), 5–32.
- [30] P. Geladi and B. Kowalski, Partial least squares regression: A tutorial, *Anal. Chim. Acta* **35** (1986), 1–17.
- [31] K.R. Lee, X. Lin, D.C. Park and S. Eslava, Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method, *Proteomics* **3**(9) (2003), 1680–1686.
- [32] G. Musumarra, V. Barresi, D.F. Condorelli and S. Scire, A Bioinformatic Approach to the Identification of Candidate Genes for the Development of New Cancer Diagnostics, *Biol. Chem.* **384**(2) (2003), 321–327.
- [33] H. Hong, Y. Dragan et al., Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS), *BMC Bioinformatics* **6**(2) (2005), S5.
- [34] G. Izmirlan, The Applications of Bioinformatics in Cancer Detection, *Ann. NY Acad. Sci.* **1020** (2004), 154–174.
- [35] J.D. Wulfskuhle, L.A. Liotta and E.F. Petricoin, Proteomic applications for the early detection of cancer, *Nat. Rev. Can.* **3**(4) (2003), 267–275.
- [36] E.F. Petricoin and L.A. Liotta, Mass Spectrometry-based Diagnostics: The Upcoming Revolution in Disease Detection, *Clin. Chem.* **49**(4) (2003), 533–534.
- [37] N. Jeffries, Algorithms for alignment of mass spectrometry proteomic data, *Bioinformatics* **21**(14) (2005), 3066–3073.
- [38] H. Bensmail, B. Aruna, J. Semmes and A. Haoudi, Functional clustering algorithm for high dimensional proteomics data, *Biomed. Biotechnol.* **2** (2005), 80–86.
- [39] E.C. Gunther, D.J. Stone, R.W. Gerwien, P. Bento et al., Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles, *Proc. Natl. Acad. Sci. USA* **100**(16) (2003), 9608–9613.
- [40] B. Wu, T. Abbott, D. Fishman, W. McMurray et al., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, *Bioinformatics* **19**(13) (2003), 1636–1643.
- [41] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson et al., Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *J. Chem. Inf. Comput. Sci.* **43**(6) (2003), 1947–1958.
- [42] L. Breiman, *Out-of-bag estimation*, Technical Report, Department of Statistics, UC Berkeley, 1996.
- [43] M.C. Papadopoulos, P.M. Abel, D. Agranoff, A. Stich et al., A novel and accurate diagnostic test for human African trypanosomiasis, *Lancet* **363** (2004), 1358–1363.