BMC Bioinformatics

BioMed Central

Research article

A data review and re-assessment of ovarian cancer serum proteomic profiling

James M Sorace*[1,2,3] and Min Zhan[4]

Address: [1]Department of Pathology and Laboratory Services, Veterans Administration Maryland Health Care System, Baltimore, 21201, USA, [2]Department of Information Systems, University of Maryland Baltimore County, Baltimore County Maryland, 21250, USA, [3]Department of Pathology, University of Maryland School of Medicine, Baltimore, 21201, USA and [4]Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, Baltimore, 21201, USA

Email: James M Sorace* - jmsorace@ix.netcom.com; Min Zhan - MZHAN@epi.umaryland.edu

* Corresponding author

Published: 9 June 2003

BMC Bioinformatics 2003, **4**:24

Received: 28 March 2003
Accepted: 9 June 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/24

Abstract

Background: The early detection of ovarian cancer has the potential to dramatically reduce mortality. Recently, the use of mass spectrometry to develop profiles of patient serum proteins, combined with advanced data mining algorithms has been reported as a promising method to achieve this goal. In this report, we analyze the Ovarian Dataset 8-7-02 downloaded from the Clinical Proteomics Program Databank website, using nonparametric statistics and stepwise discriminant analysis to develop rules to diagnose patients, as well as to understand general patterns in the data that may guide future research.

Results: The mass spectrometry serum profiles derived from cancer and controls exhibited numerous statistical differences. For example, use of the Wilcoxon test in comparing the intensity at each of the 15,154 mass to charge (M/Z) values between the cancer and controls, resulted in the detection of 3,591 M/Z values whose intensities differed by a p-value of 10^{-6} or less. The region containing the M/Z values of greatest statistical difference between cancer and controls occurred at M/Z values less than 500. For example the M/Z values of 2.7921478 and 245.53704 could be used to significantly separate the cancer from control groups. Three other sets of M/Z values were developed using a training set that could distinguish between cancer and control subjects in a test set with 100% sensitivity and specificity.

Conclusion: The ability to discriminate between cancer and control subjects based on the M/Z values of 2.7921478 and 245.53704 reveals the existence of a significant non-biologic experimental bias between these two groups. This bias may invalidate attempts to use this dataset to find patterns of reproducible diagnostic value. To minimize false discovery, results using mass spectrometry and data mining algorithms should be carefully reviewed and benchmarked with routine statistical methods.

Background

The early diagnosis of ovarian cancer has the potential to dramatically reduce the mortality associated with this disease. Recently, the use of surface-enhanced laser desorption/ionization (SELDI) time-of-flight mass spectrometry profiling of patient serum proteins, combined with advanced data mining algorithms, to detect protein patterns associated with malignancy, has been reported as a

promising field of research to achieve the goal of early cancer detection [1–5]. Several reports have detailed the ability of this proteomic method to diagnose the difference between ovarian cancer [6–8], prostate cancer [9–13], and bladder cancer [13,14]. Much of the effort in these analyses has focused on the use of a variety of data mining tools such as the evaluation of prostate cancer using peaks in the mass to charge (M/Z) region between 2 K and 40 K combined with boosted decision tree analysis [10] to try to detect patterns that allow the diagnosis of cancer versus non-cancer. The use of similar technology to evaluate bladder cancer has also been reported [13,14]. Thus, this field represents an active area of current research. For example, a recent report by the Clinical Proteomics Program Databank has demonstrated that the use of genetic algorithms coupled with clustering analysis has resulted in rule sets that can predict ovarian cancers (including samples from patients with stage 1 disease) with 100% sensitivity and 96% specificity [6]. These results have been extended by the same group to include a larger series of ovarian cancer patients as well as prostate cancer patients [7,9]. The Clinical Proteomics Program Databank has provided three sets of ovarian cancer data to the scientific community without restriction. These data sets include Lancet Ovarian Data 2-16-02 used in the study noted above [6]. This study consisted of a total of 100 control, 100 cancer, and 16 benign disease samples run on a Ciphergen H4 ProteinChip array (since discontinued). The samples were manually processed. The data was posted after baseline subtraction. The second data set, Ovarian Dataset 4-3-02 consist of the same samples as the first but the samples were run on a Ciphergen WCX2 ProteinChip array. The samples were manually prepared and the data was posted with baseline subtraction. A model diagnostic rule based on this dataset is published on the website, but no data is given regarding the rules sensitivity or specificity. In this report, we analyze the third Ovarian Dataset 8-7-02 and corresponding sample information downloaded from the Clinical Proteomics Program Databank website [7]. This set of data consists of serum profiles of 162 subjects with ovarian cancer and 91 non-cancer control subjects. The cancer group may be further divided into 28 stage 1 patients, 20 stage 2 patients, 99 stage 3 patients, 12 stage 4 patients, and 3 no stage specified patients. For each subject a set of data consisting of intensities at 15,154 distinct M/Z values ranging from 0.0000786 to 19995.513 was available for analysis. This dataset was constructed using the Ciphergen WCX2 ProteinChip array. All the steps of preparing the chips for sample analysis were preformed robotically, and the raw data without baseline subtraction was posted for download. A model rule claiming 100% sensitivity and specificity is also given. Additional details of experimental data collection may be found at the Clinical Proteomics Data Bank [5]. In addition to the various methods of preparing and running the samples on the mass spectrometer, the optimal steps in processing the raw data from the mass spectrometer for further analysis have not been standardized and remain a fertile area for investigation [15]. We choose the deliberately simple strategy of using Wilcoxon test on the raw data to better understand the underlying properties of the data set. We consider this simple approach a "benchmark method" to which other methods can be compared. Further, we use Wilcoxon test and stepwise discriminant analysis on a training subset consisting of 80 cancer patients and 45 controls, randomly chosen from the original data set, to develop rules to classify a test set consisting of the remaining cancer and control subjects. Disease classifiers of great sensitivity and specificity could be readily constructed by visual inspection and manual binning of M/Z values based on the p-values of the Wilcoxon test combined with classical stepwise discriminant analysis. The ability of these rules to classify disease and normal samples were comparable to the model rule published on this dataset at the Clinical Proteomics Program Databank website which was developed using a proprietary genetic algorithm. Further, in examining all M/Z values, the M/Z values that discriminated best between ovarian cancer and control were all found to be less than 500, an area of the spectrum often discarded as noise [10]. These findings are useful for several reasons. First, the statistical methods used in this study are readily available, widely understood, and can be cheaply implemented. Secondly, a vast amount of mathematical research and practical experience underlies their interpretation. Finally, they can be used to discover unexpected patterns present in the data set. These patterns may be missed by machine learning methods that are narrowly focused on diagnostic classification, and do not present the researcher with a broad overview of the data. As a result of these traditional studies, a better understanding of the weaknesses and possible strengths of serum proteomic profiling becomes apparent.

Results and Discussion

Based on the initial training set, the intensity at each of the consecutive 15,154 M/Z values was first compared using a two-sided Wilcoxon test (see methods) Figure 1 shows the pattern of the resulting two-sided Wilcoxon test p-values generated on a training set consisting of 80 cancer patients and 45 controls randomly chosen from the larger data set, with the M/Z values on the x-axis and the negative logarithm (base 10) of the Wilcoxon test p-values on the y-axis. There are a total of 685 distinct M/Z values differing between the cancer and control populations with a p-value of less than 10^{-6}. Also of note in this distribution is that all M/Z values with a Wilcoxon p-value less than 10^{-12} are found at M/Z values of less than 500. The significance of this finding will be discussed further below.

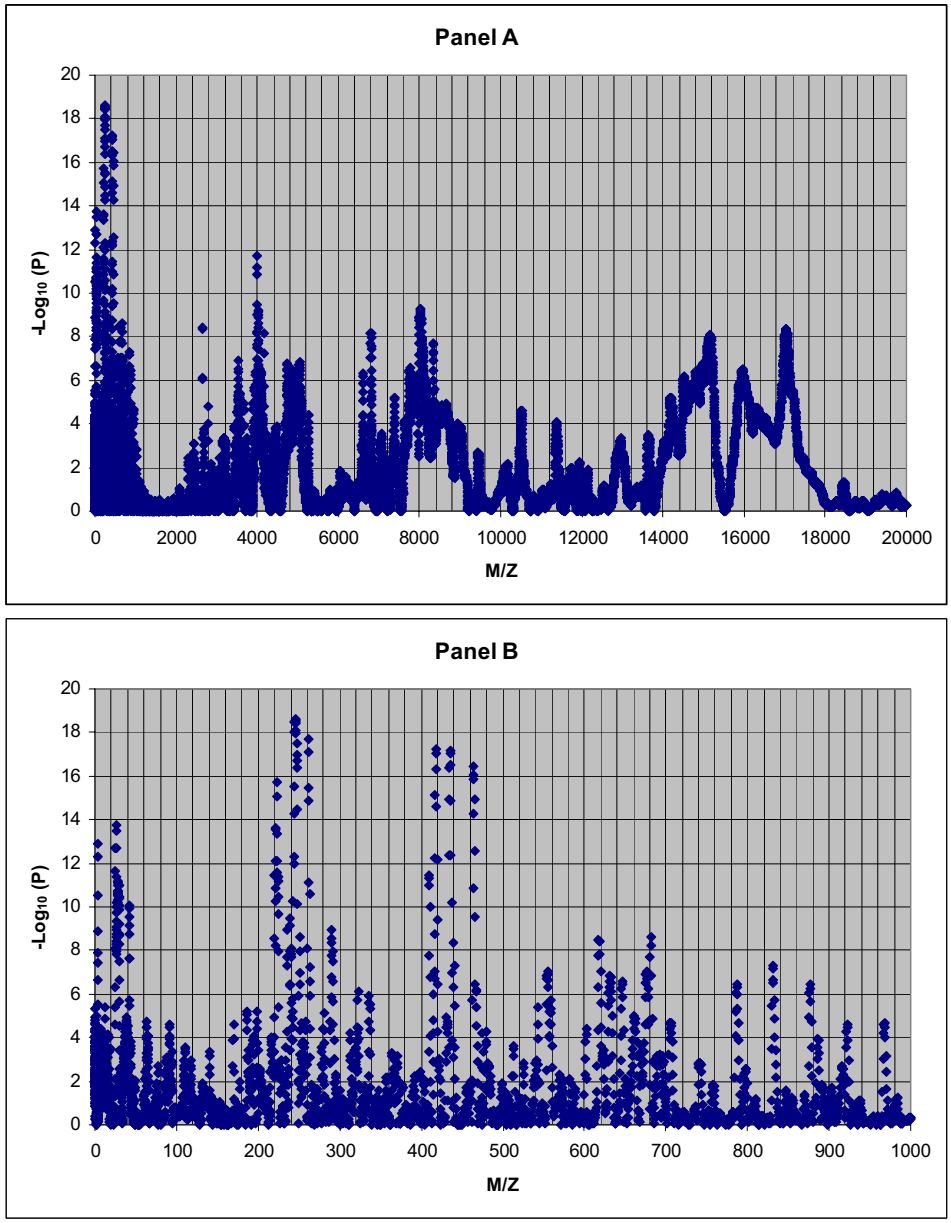

Figure 1

Training Set Wilcoxon p-values by M/Z value. Wilcoxon p-values between normal and cancer members of the training set were calculated for every M/Z value. The Y axis represents negative the Log (base 10) of the p-value. Panel A: The X axis represents M/Z values between 0 and 20,000. Panel B: The X axis represents M/Z values between 0 and 1000. The following control spectra were used for the initial training set: daf-0181 daf-0182 daf-0183 daf-0188 daf-0189 daf-0192 daf-0193 daf-0195 daf-0196 daf-0197 daf-0198 daf-0200 daf-0201 daf-0202 daf-0205 daf-0207 daf-0210 daf-0211 daf-0212 daf-0217 daf-0218 daf-0220 daf-0223 daf-0226 daf-0230 daf-0234 daf-0235 daf-0241 daf-0242 daf-0244 daf-0247 daf-0248 daf-0250 daf-0251 daf-0252 daf-0258 daf-0259 daf-0261 daf-0262 daf-0263 daf-0267 daf-0269 daf-0270 daf-0279 daf-0280 The following cancer spectra were used for the initial training set. daf-0601 daf-0602 daf-0606 daf-0608 daf-0609 daf-0612 daf-0617 daf-0618 daf-0619 daf-0620 daf-0621 daf-0625 daf-0627 daf-0632 daf-0633 daf-0634 daf-0635 daf-0636 daf-0643 daf-0644 daf-0651 daf-0654 daf-0655 daf-0656 daf-0657 daf-0661 daf-0662 daf-0663 daf-0664 daf-0666 daf-0667 daf-0669 daf-0673 daf-0675 daf-0682 daf-0683 daf-0687 daf-0688 daf-0691 daf-0692 daf-0697 daf-0698 daf-0701 daf-0702 daf-0703 daf-0705 daf-0706 daf-0707 daf-0708 daf-0709 daf-0716 daf-0718 daf-0719 daf-0726 daf-0727 daf-0729 daf-0731 daf-0733 daf-0735 daf-0737 daf-0740 daf-0744 daf-0751 daf-0752 daf-0753 daf-0754 daf-0755 daf-0756 daf-0757 daf-0758 daf-0760 daf-0761 daf-0762 daf-0764 daf-0768 daf-0770 daf-0773 daf-0776 daf-0778 daf-0780

Table 1: Development of Diagnostic Rule 1.

Consecutive M/Z	M/Z Value	Bin Range Consecutive M/Z	Wilcoxon p-value Training Set	Rule 1	Wilcoxon p-value Entire Data Set
6782	4003.645	6781–6783	1.8685E-12	S	8.98721E-27
2311	464.3617	2308–2314	3.6867E-17	S	6.76511E-34
2237	435.0751	2234–2242	6.822E-18	S	3.895E-37
2193	418.1136	2190–2196	5.6991E-18	S	3.91174E-34
2171	409.7594	2170–2172	3.6168E-12		3.28383E-25
1736	261.8864	1734–1739	1.9206E-18	S	1.22566E-35
1681	245.53704	1673–1691	2.2891E-19	S	7.24111E-38
1600	222.4183	1598–1608	1.8911E-16		2.01896E-33
1594	220.7513	1593–1596	2.3886E-14		5.52587E-30
576	28.70048	562–582	6.82E-12		2.60148E-24
544	25.58989	541–547	1.9179E-14		8.67451E-30
181	2.7921478	181–183	1.2929E-13	S	1.21243E-27

Consecutive M/Z is the numerical order of the M/Z value between 1 and 15,154. The M/Z values were sorted by p-values and the lowest 100 were arbitrarily selected. The M/Z values were then binned as described in the text, and the most significant consecutive M/Z score from each of the 12 bins was selected. M/Z values that were selected by the stepwise discriminant analysis are designated a "S" in the Rule 1 column. The Wilcoxon p-values calculated from the training set (used to derive the rule) and calculated from the entire data set are shown in their respective columns.

In order to determine if these data could be used to separate normal from cancer, we used three strategies to develop rules for diagnostic classification. First, all data points regardless of M/Z value were sorted from most to least significant (according to the two-sided Wilcoxon test p-values) and the 100 M/Z values with the lowest p-values were chosen. These 100 M/Z values were then separated into distinct bins by sorting on consecutive M/Z and requiring a separation of at least 1 M/Z value to start the next bin (12 bins were detected in this process). The M/Z value with the smallest p-value in each bin was selected. The results are shown in Table 1. Next, stepwise discriminate analysis was performed, and 7 M/Z values were selected for Rule 1 (of note, all but one M/Z value was below 500). When this rule was applied to the entire data set, test and training inclusive, all 162 cancer and 91 controls were appropriately classified without error for 100% sensitivity and specificity. Given that the interpretation of low M/Z values maybe problematic, we next focused attention on a set of rules which met the following requirements. First, the M/Z value had to exceed 2000, and the Wilcoxon test P-value had to be less than 10^{-6}. A total of 462 M/Z values from the training set met these criteria. As shown in Table 2, a total of 30 bins were detected by sorting on consecutive M/Z values as above, and the most significant p-value from each bin was selected for stepwise discriminant analysis. Thirteen M/Z values were retained in Rule 2. In the training set, one subject in the cancer group was misclassified as normal and one in the control group was misclassified as cancer. In the test set, two subjects from the control group were misclassified as cancer. Therefore for the test set, the sensitivity was 82/82 or 100% and the specificity was 43/45 or 95.7%. For the test and training set combined, one subject was misclassified in the cancer group as normal and three subjects were misclassified in the control group as cancer. Thus for this rule the overall sensitivity was 161/162 or 99.4 % and its overall specificity was 88/91 or 96.7 %. Finally, Rule 3 was constructed using the 30 M/Z values in Rule 2 combined with four M/Z values 409.75936, 418.11364, 435.0751, and 464.3617 (all also used in Rule 1). This was done because prior studies have indicated the possible presence of low molecular weight biomarkers in ovarian cancer (see below). When this set of M/Z values was subjected to stepwise discriminant analysis, seven variables at M/Z values of 418.1136, 435.0751, 464.3617, 4003.645, 4906.962, 6599.8232, and 6801.495 were retained. When Rule 3 was applied to the entire data set, test and training inclusive, all 162 cancer and 91 controls were appropriately categorized without error for 100% sensitivity and specificity. The actual classification schema for all three rules is shown in Table 3. The results presented with these three rules were all achieved in the first attempt. No effort was made to further optimize these rules. We next interchanged the test and training sets and used the same three rule development strategies. This resulted in:

1) Rule 1 with M/Z values of 2.8234234, 222.41828, 410.13727, 417.73207, 435.07512, 4027.2999, and 8035.0581, achieved 100% sensitivity and specificity on both the test and training sets.

2) Rule 2 with M/Z values of 3676.3951, 3937.7816, 4003.6449, 4440.095, 5269.0367, 10511.699, 14182.82, and 17019.433. This rule achieved 100% sensitivity and specificity on the training set. However sensitivity and

Table 2: Development of Diagnostic Rule 2.

Consecutive M/Z	M/Z	Wilcoxon p-value	Rule 2
5534	2665.397	4.06E-09	S
6372	3534.072	1.26E-07	
6753	3969.469	4E-07	S
6772	3991.844	6.87E-09	S
6782	4003.645	1.87E-12	S
6802	4027.3	6.21E-10	S
6814	4041.526	1.86E-07	
6823	4052.213	8.33E-07	
6827	4056.967	9.38E-07	S
6836	4067.673	3.9E-07	
6852	4086.742	4.11E-07	
6934	4185.17	6.56E-09	
7383	4744.889	1.71E-07	S
7449	4830.124	2.89E-07	
7468	4854.802	8.22E-07	
7508	4906.962	5.45E-07	
7606	5035.93	1.41E-07	
8707	6599.823	4.96E-07	
8839	6801.495	6.46E-09	S
9439	7756.437	2.66E-07	
9457	7786.054	4.58E-07	S
9483	7828.934	6.23E-07	
9607	8035.058	4.94E-10	
9793	8349.266	2.04E-08	S
12910	14511.46	6.4E-07	
13036	14796.14	3.95E-07	S
13113	14971.48	1.76E-07	
13201	15173.13	7.88E-09	
13537	15955.47	3.13E-07	S
13987	17034.05	4.53E-09	S

The M/Z values were sorted by M/Z values greater than 2,000 and p-values less than 10^{-6}. Consecutive M/Z is the numerical order of the M/Z value between 1 and 15,154. The M/Z values were then binned as described in the text, and the most significant consecutive M/Z score from each of the 30 bins was selected. M/Z values that were selected by the stepwise discriminant analysis are designated "S" in the rightmost column.

specificity fell on the test set to 96.25% and 91.11% respectively.

3) Rule 3 with M/Z values of 417.73207, 435.07512, 2666.361, 2674.0769, 3937.7816, 3991.8435, 4821.0481, 4839.2088, 5269.0367, 7627.1183, 14182.82, and 17019.433. This rule achieved 100% sensitivity and specificity on the training set. On the test set it achieved a sensitivity of 100% and a specificity of 97.8%. We have used a strategy identical to that used in Rule 1 to further analyze this data. First, a randomly ordered list of cancer spectra and a randomly ordered list of control spectra were prepared. Next, we assigned the first 20% of each list to a test set and the remaining 80% to a training set. The process was repeated five times assigning the next consecutive 20% of each list for the test set on each occasion. The results were very similar to those above with all five rules achieving 100% sensitivity and specificity. This data is posted as additional data file Supplement1.xls.

The presence of statistically significant signals at M/Z values less than 500 was unexpected as some investigators, in their systems, conservatively disregard data beneath M/Z values of 2000 as possible noise [12]. To further investigate this, we first repeated the calculation of 2-sided Wilcoxon test p-values at each of the 15,154 M/Z values using the entire data set (see Figure 2). The trends noted in the training set were present in the entire data set, although with increased statistical significance. For example 3,591 of the 15,154 M/Z values had mean intensities that varied between cancer and control with a p-value of 10^{-6} or less. In a sample of a panel consisting of 15,154 independent random sets of measurements split between cancer and control, using Wilcoxon test 15,154 times with an individual significance level of 10^{-6}, the number of false positives is expected to be 0.015. It is very small. Alternatively, in the above setting the chance that at least one of the 15,154 measurements would have a p-value less than 10^{-6} is approximately 1.5%. Thus it is extremely unlikely that a

Table 3: Classification rules

Classification Rule 1 based on the intensities at the 7 M/Z values:
Let
a_1 = (1303,5.66302,5.48787.19.60743,-8.88828,-30.47983,-0.34510)',
a_2 = (1413,6.44028,6.36701,20.84677,-11.04580,-32.04436,0.03553)',
c_1 = -2984,
c_2 = -3521.
Let X be the vector that represents the intensities from a subject at the 7 M/Z values:
2.7921478, 245.53704, 261.8864, 418.1136, 435.0751, 464.3617, 4003.645. Classify X into the cancer group if
$(a_1 - a_2)' X + (c_1 - c_2) \geq 0$;
Otherwise classify X into the control group.

Classification Rule 2 based on the intensities at the 13 M/Z values:
Let
a_1 = (24.58884,-15.09887,0.95772,-3.16411,-10.93854,31.7966,8.11259,6.59602,-53.15727, -7.49888,149.18784,399.67258,112.83481)',
a_2 = (26.26343,-19.06632,1.07975,-1.89482,-9.83188,29.47779,7.69470,8.54597,-49.99409, -7.32192,142.08982,389.53254,116.55493)',
c_1 = -1386,
c_2 = -1312.
Let X be the vector that represents the intensities from a subject at the following 13 M/Z values: 2665.397, 3969.469, 3991.844, 4003.645, 4027.3, 4056.967, 4744.889, 6801.495, 7756.437, 8349.266, 14796.14, 15955.47, 17034.05.
Classify X into cancer if
$(a_1 - a_2)' X + (c_1 - c_2) \geq 0$.
Otherwise classify X into control.

Classification rule #3 based on 7 M/Z values:
Let
a_1 = (6.04377,3.42186,-1.99804,0.23374,2.46593,-1.87559,17.37384)',
a_2 = (7.35920,1.86527,-1.32486,0.92386, 1.18336,4.76619,9.95349)',
c_1 = -218.19592,
c_2 = -254.33039.
Let X be the column vector that represents the intensities from a subject at the following 7 M/Z values:
418.1136, 435.0751, 464.3617,4003.645, 4906.962, 6599.823, 6801.495.
Classify X into cancer if
$(a_1 - a_2)' X + (c_1 - c_2) \geq 0$.
Otherwise classify X into control.

false positive would occur by chance alone in a 15,154 member test set. The finding of significant signals at M/Z values less than 500 is consistent with two of the seven M/Z values used in the model rule published at the Clinical Proteomics Program Databank website that was developed on the same data set (see Table 4), specifically M/Z values 435.46452 (p-value = 9.08×10^{-37} 2nd most significant of 9 values in its bin) and 465.56916 (p-value = 2.50×10^{-28} 6th most significant of the 7 M/Z values found in its bin) [7]. These values correspond to M/Z values of 435.0751 and 464.3617 used in Rule 1. As shown in Figure 3, each of these two pairs of M/Z values are surprisingly effective at separating the 162 cancer subjects from the 91 control subjects with an advantage noted with the first pair (compare panel A with B). Even more interesting is the finding that significant M/Z values found in the first Rule 1 included the M/Z values of 2.7921478 and 245.53704. As shown in Figure 3, panel C, these two values can also significantly separate the 162 cancer subjects from the 91 control subjects. The interpretation of these values is problematic, given the low M/Z values involved. In order to evaluate these findings, we first investigated whether data normalization as described at the Clinical Proteomics Data Bank [5] could influence the Wilcoxon test p-values found using the raw data (see methods). Several points were chosen, and no effect was noted on the p-values (see Table 5). We further analyzed several selected low M/Z values, less than 500. In this process, the cancer and control data were pooled. The pooled data were randomly partitioned between a set containing 91 members and a set containing 162 members. The Wilcoxon test was then run on the randomized set. The process was repeated 10,000 times, and the lowest p-values were chosen. As shown in Table 5, the lowest p-values generated by the permutation process were on the order of 0.0001, as expected given the number of permutations tested. Thus, it is highly unlikely that either data normalization or a chance distribution could have accounted for the highly significant p-values noted in the M/Z region less than 500. Finally, it is interesting to note that the remaining five values in the Clinical Proteomics Program Databank model rule all have M/Z values greater than 2000 and relatively high p-values. Specifically the remaining values are (note that the p-values are calculated from the entire data set):

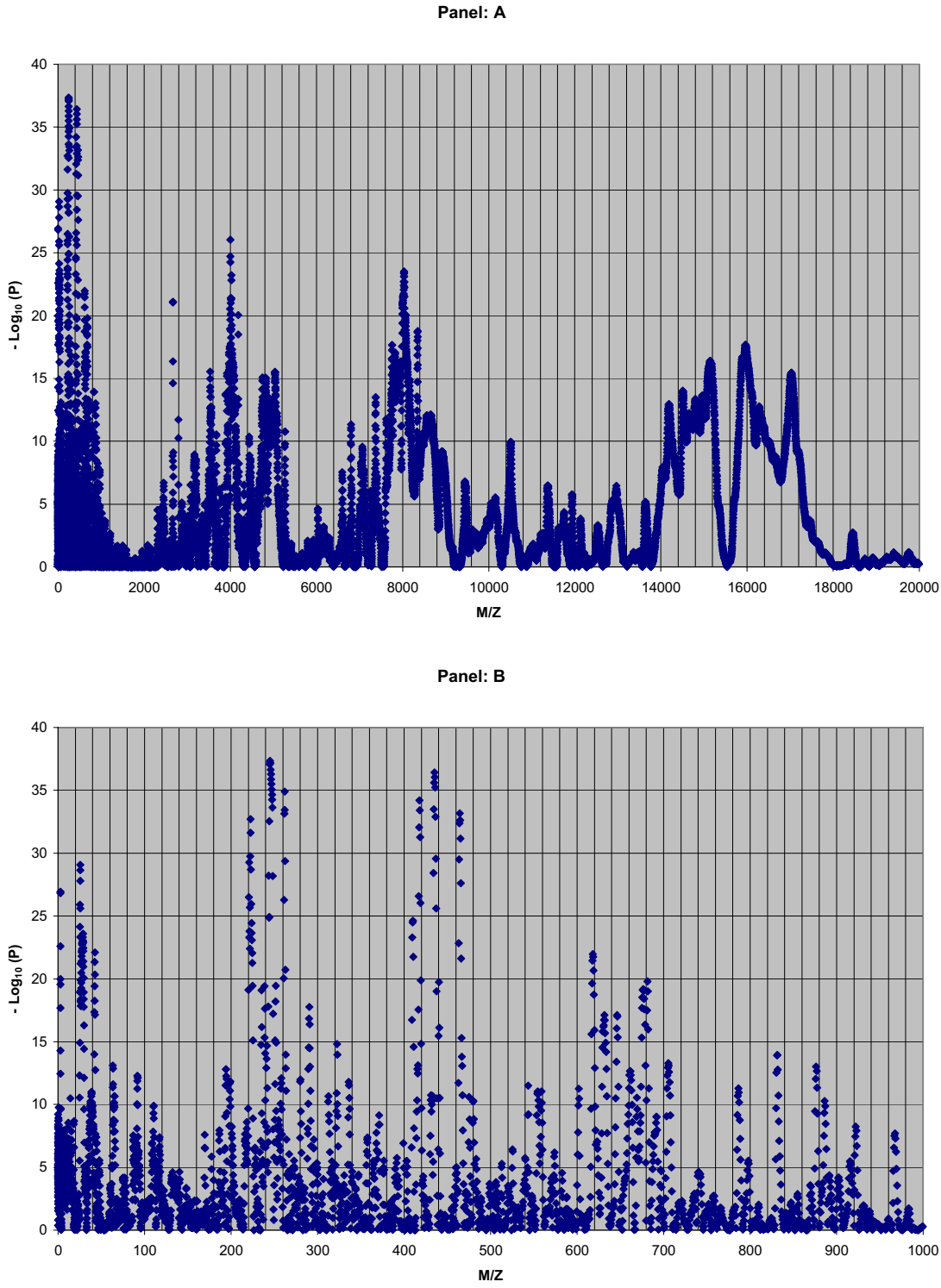

Figure 2
Wilcoxon P-Values by M/Z Value for Entire Dataset .Wilcoxon p-values between normal and cancer members of the entire dataset set were calculated for every M/Z value. The Y axis in negative the Log base 10 of the p-value. Panel A: the x-axis represents M/Z from 0 to 20,000. Panel B: the x-axis represents M/Z from 0 to 1,000.

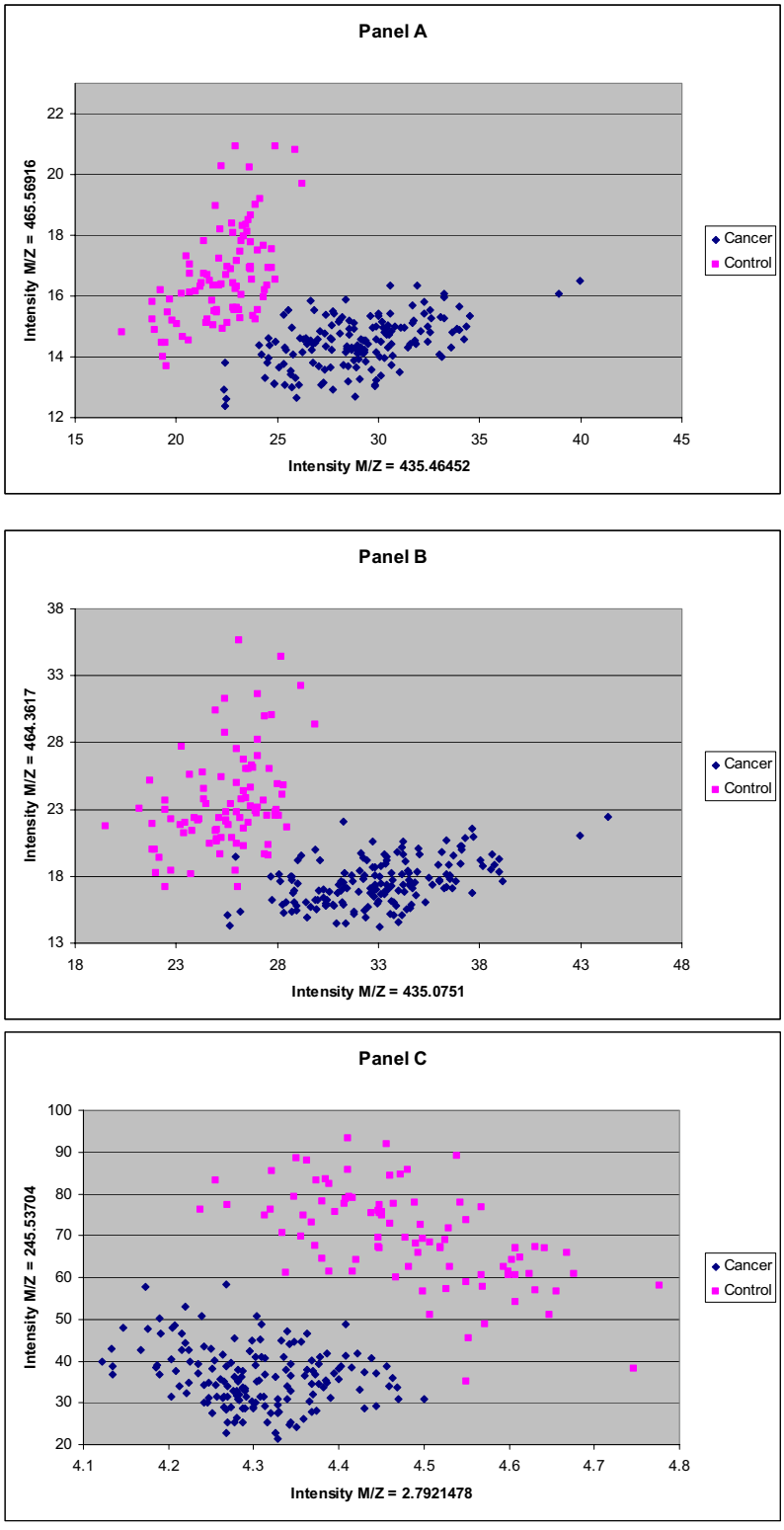

Figure 3
Diagnostic value of Low M/Z values. Scatter plots of the 162 cancer subject versus 91 normal subjects. Panel A represents 2 M/Z values from the Clinical Proteomics Program Database while Panel B and Panel C are both derived from Rule 1. See text for details.

Table 4: Clinical Proteomics Program Databank Example Ovarian Rule.

Consecutive M/Z Bin	M/Z-Value	P2_Wil
5632	2760.6685	0.239533474
15020	19643.409	0.521014657
2314	465.56916	2.49791E-28
8728	6631.7043	9.00537E-4
12704	14051.976	1.79156E-08
2238	435.46452	9.07922E-37
6339	3497.5508	1.40316E-06

Consecutive M/Z values and Wilcoxon p-values based on the entire dataset for the rule present on the Clinical Proteomics Program Databank website.

Table 5: Normalization and Permutation Analysis of Low M/Z Values

M/Z	Permutated P-Wilcoxon	Normalized P-Wilcoxon	Actual P-Wilcoxon
2.792148	8.30646E-4	1.2124E-27	1.21243E-27
25.58989	0.210484E-4	NT	8.67451E-30
245.537	0.571414E-4	NT	7.24111E-38
418.1136	1.99231E-4	NT	3.91174E-34
435.0751	3.34994E-4	3.895E-37	3.895E-37
464.3617	0.263008E-4	6.7651E-34	6.76511E-34
4003.645	0.292578E-4	NT	8.98721E-27
15526.93	2.079E-4	NT	0.741858713

Normalization and Permutation Analysis of Low M/Z Values Normalization and permutation analysis (smallest p-value of the 10,000 iterations per M/Z point tested) were carried out on selected M/Z points. See text for details. NT = Not tested.

1) M/Z = 2760.6685, p = 0.24

2) M/Z = 19643.409, p = 0.52

3) M/Z = 6631.7043, p = 9.0×10^{-4}

4) M/Z = 14051.976, p = 1.8×10^{-8}

5) M/Z = 3497.5508, p = 1.4×10^{-6}

By contrast all the M/Z values used in Rule 1 have p-values less than 10^{-26} (compare Table 1 with Table 4).

There are several non-exclusive explanations for the presence of significant P-values in M/Z region less than 500. First, these may actually represent biomarkers that correlate with ovarian cancer. The disease process may influence the serum concentration of lipids, or other small molecules that either bind to the chip directly or through a complex formation with other macromolecules (e.g., binding to a receptor). For example, the lysophospholipids represent a class of compounds that have an important role in extracellular signaling. Lysophophatidic Acid (LPA) is a member of this class of compounds, and its plasma levels have been proposed as a potential biomarker for ovarian cancer [16,17]. LPA is a family of related molecules with molecular weights in the vicinity of 400 to 600 Daltons, and a variety of LPA species has been reported to be increased in malignant ascites from patients with ovarian cancer as detected by electrospray ionization mass spectrometry (ESI-MS) [18]. LPA related species have also been reported to be increased in plasma samples from patients with ovarian cancer using a combination of thin layer chromatography (to isolate an "LPA band" from patient plasma) followed by ESI-MS. This study reported significant LPA increases in cancer samples with increased intensities noted at M/Z values of 409, 433–437, 457, 481–482, 571, 599, and 619. This report also reviews the evidence that these M/Z values are consistent with LPA family members [19]. Figure 4 shows the average intensities and p-values for both the cancer and control groups in the region between M/Z values of 410 to 470. Among other features, an increase in the mean intensity for cancers at a peak centered at an M/Z of 459 is noted. However, also of note in this region are:

1) M/Z = 464.3617 with a p-value less than 6.8×10^{-35}, that correlates with a shoulder in a secondary peak at

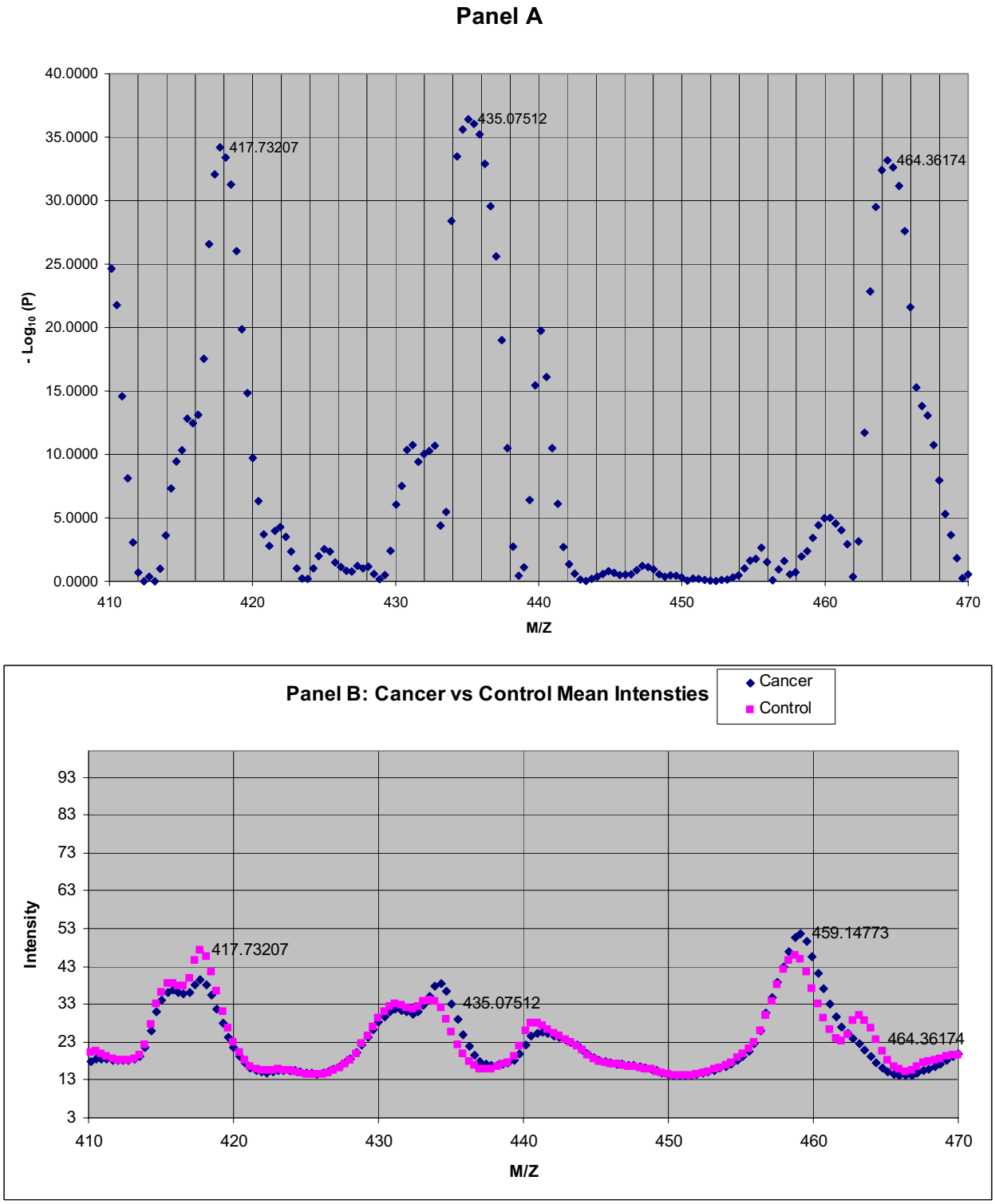

Figure 4

P-values and Intensities for M/Z values between 410 and 470. The p-values and mean intensities of cancer and control groups (entire set) for M/Z values between 410 and 470 are shown in panels A and B respectively. Selected data points are labelled with their M/Z values directly to the right of the points, see text for details.

about 463, that is decreased in cancer patients (average intensity 17.5 for cancer versus 23.6 for controls).

2) M/Z = 435.0751 with a p-value of less than 3.9×10^{-37}, that corresponds to a peak with increased intensity in cancer (average intensity 33 for cancer versus 25.5 for controls).

3) M/Z = 417.73207 with a p-value less than 6.2×10^{-35}, that corresponds to a peak that is decreased in cancer (average intensity 39.5 for cancer versus 47.4 for controls)

The identity of the molecules responsible for these differences cannot be determined from this data. However, it is possible that in some cases they may relate to the LPA family of molecules, or to alterations in proteins that bind LPA family members.

Other explanations for the presence of statistically significant bands of low M/Z include degradation products of higher molecular weight macromolecules or a matrix effect. For example, if a set of proteins exist that are expressed at different levels between cancer and control subjects but have a common domain, then a common product ion of lower M/Z may be generated that would represent a summation of all the changes in expression of the group of proteins, and might thus have greater statistical significance than the changes associated with any single high M/Z value. Similarly, a set of low M/Z molecules (e.g., energy-absorbing molecule or matrix) that interacts differently in a protein environment that differs markedly between cancer and control could hypothetically generate a similar phenomenon. However, it is difficult to apply any of the above explanations to the very low M/Z values such as 2.7921478 and 245.53704, although in the last case an extremely small organic molecule is possible.

Alternatively, there maybe some unexpected experimental bias or systematic error that accounts for low M/Z discrimination. This could occur at any experimental step, and might include medication or lifestyle change that occurs in patients who learn they have a cancer diagnosis, variation in sample collection, processing and preservation, as well as bias introduced at the time of analysis. In the case of LPA, increased plasma levels may be associated with platelet activation. Another group trying to repeat the observations of increased levels of LPA associated with ovarian cancer concluded that there was no diagnostic value in the assay, and attributed the discrepant findings as possibly related to different sample centrifugation protocols used by the two groups to remove platelets from the samples prior to analysis [20]. However, LPA continues to be actively evaluated for its clinical utility [21].

Conclusions

Serum proteomic profiling is a new approach to cancer diagnosis. However it confronts a challenging environment, as it combines measurement technologies that are new in the clinical setting with novel approaches to processing and interpreting high dimensional data. Further, controlling large clinical studies can be challenging even in more established settings. Nevertheless, it represents an advance in the ability to diagnose and understand illness. The results presented in this study are useful for several reasons. First, in regard to disease classification, advanced data mining techniques should be benchmarked against traditional methods when possible. Further identical training sets should be defined for such a comparison as results may very depending on the samples chosen for inclusion in the training set. The development of disease classifiers using routine analysis proved to be straightforward, and resulted in excellent performance in both the test and training sets (e.g. 100% sensitivity and specificity for Rules 1 and 3 in the first training set). In particular these preliminary data suggest that these two rules may be specific enough to scale to larger population trials without generating an unacceptably high false positive rate. This study also confirms that a classifier could be developed with M/Z values greater than 2000. This indicates that information regarding the difference between cancer and control is present throughout the entire M/Z region studied, a result entirely consistent with the observed Wilcoxon test p-values. Secondly, routine analysis allows investigators to rapidly review the data for their general trends, and correlate the findings with other information. The findings of significant discrimination between cancer and control groups at low M/Z values indicates that attention should be focused in this region. In particular, if experimental bias and noise effects can be excluded, this region may prove to offer the optimum for ovarian cancer diagnostic test development. On the other hand, if bias cannot be excluded, the possibility must be entertained that higher M/Z values may also have been similarly affected. In order to address these issues, consideration may be given to using mass spectrometry methods with increased sensitivity in the low M/Z region. The experimental conditions used to physically bind the serum samples to the chip prior to analysis may also prove critical, and should be consistent with those used in collecting the current data set. Also, the possibility that the changes in the low M/Z region may represent an additive effect caused by differing protein environments between cancer and normal may be approached by intentionally spiking samples with panels of known proteins, and determining if there is an effect on the spectra in the low M/Z region. The use of internal standards to normalize this type of experimental system in general may also be considered. As with all clinical test development, confirmation of results in independent laboratories running

blinded samples will remain the gold standard in ruling out the possible effects of bias, unless the sample set itself contains the bias. Particular attention should be paid to pre-analytic causes of bias that may influence the serum proteome. In particular the coagulation and complement systems should be considered as potential sources of noise in this context, as both are activated during serum sample collection and generate low molecular weight products. These products are undesirable for two reasons. First, if a putative tumor biomarker (e.g. LPA) is a member of a pathway altered during serum sample collection, changes between plasma levels of cancer and control subjects may be obscured. Secondly, the generation of activation products may simply complicate the spectrum. Also, sample collection practices should be rigorously defined, and include submitting matched control and cancer samples from all centers participating in the study. Matching for age and menopausal status should be considered. For example, in the data set used in this study, the mean age of the control group was 47 years and the cancer group 60 years. It is noteworthy that the average age of menopause is approximately 51 years [22]. This may introduce a bias in the results reported in this study as well as all others derived from this dataset. Finally, the steps associated with sample collection, processing, and binding to the chip may represent a particularly fertile area for research. Any combination of such steps may significantly alter the molecular subset of the sample that can be successfully analyzed.

However, the ability to discriminate between cancer and control based on the M/Z values of 2.79 and 245.5 reveals the presence of a significant experimental bias not related to disease pathology, that likely involves machine noise and matrix effects. This is particularly true of the M/Z value at 2.79 which represents a bias of the mass spectrometer instrument itself. If this is the case the higher M/Z regions may also be affected. These findings indicate that any rule derived from this data set, including the ones presented in this paper, may be detecting differences in experimental bias and not disease pathology. Investigators in this field may minimize their chances of false discovery by careful experimental design and by using routine statistical methods to both overview the data (in an intentional search for bias) as well as a benchmark for comparison with other data mining algorithms.

Methods

A training set was formed by randomly sampling 45 spectra out of the 91 controls and 80 spectra out of the 162 cancer cases (see Figure 1). Those spectra that were in the original data set but not in the training set were considered in a 'test' set. Two-sided Wilcoxon test was used to compare the intensity between the controls and cancers in the training set at each of the 15,154 M/Z values. We then selected a subset of the M/Z values with the lowest Wilcoxon test p-values (see the Results section for details). We sorted on consecutive M/Z values to get bins. A separation of at least one M/Z value was required to start the next bin. The lowest p-value in each bin was selected and the corresponding M/Z value was used in stepwise discriminant analysis to determine the subset of M/Z values that best discriminated cancer from control in the training set. The criteria were applied to the test data set, and sensitivity and specificity were computed. All the analyses were performed in SAS Version 8.2 (A statistical package from SAS Institute Inc., Cary, NC, USA) on a personal computer. Wilcoxon test was performed using NPAR1WAY procedure in SAS, stepwise discriminant analysis was performed using STEPDISC procedure in SAS, and discriminant analysis was performed using DISCRIM procedure in SAS [23].

To normalize the data, the procedure outlined by the Clinical Proteomics Program Databank was used [24]. The cancer and control values for each M/Z were given respective labels, and the data were then pooled and normalized using the formula NV = (V-Min)/(Max - Min). In this expression, Min is the minimum intensity of the pooled samples, Max represents the maximum intensity found in the pooled samples, and NV represents the normalized value. Using this procedure, the data intensities will all fall between 0 and 1. The data points were sorted into cancer and controls, and the p-values were calculated.

Authors' contributions

JMS conceived of the studies and developed the initial process for selecting diagnostic rules. MZ performed the statistical analysis and further refined the rules with stepwise discrminant analysis. All authors read and agreed with the final manuscript.

Additional material

Additional File 1

This Excel file contains 5 work sheets. Each sheet contains the most significant 100 MZ values from a training set consisting of a different 80% subset of the data. The MZ values have been sorted into bins and the most significant MZ value from each bin (marked with a "1" in column E) was used in stepwise dicriminant analysis. MZ values retained in the final rule are indicated with an "s" in column F.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-4-24-S1.xls]

Acknowledgements

We thank Christopher Gocke MD, Jules J. Berman PhD MD, G. William Moore MD PhD, and Robert Rohwer PhD for their critical reading of the manuscript.

References

1. Michener CM, Ardekani AM, Petricoin EF 3rd, Liotta LA and Kohn EC: **Genomics and proteomics: application of novel technology to early detection and prevention of cancer** *Cancer Detect Prev* 2002, **26:**249-255.
2. Petricoin EF, Zoon KC, Kohn EC, Barrett JC and Liotta LA: **Clinical proteomics: translating benchside promise into bedside reality** *Nat Rev Drug Discov* 2002, **1:**683-695.
3. Srinivas PR, Verma M, Zhao Y and Srivastava S: **Proteomics for cancer biomarker discovery** *Clin Chem* 2002, **48:**1160-1169.
4. Herrmann PC, Liotta LA and Petricoin EF 3rd: **Cancer proteomics: the state of the art** *Dis Markers* 2001, **17:**49-57.
5. **Clinical Proteomics Data Bank** [http://clinicalproteomics.steem.com/ppatterns.php]
6. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC and Liotta LA: **Use of proteomic patterns in serum to identify ovarian cancer** *Lancet* 2002, **359:**572-577.
7. **Clinical Proteomics Data Bank** [http://clinicalproteomics.steem.com/download-ovar.php]
8. Rai AJ, Zhang Z, Rosenzweig J, Shih IeM, Pham T, Fung ET, Sokoll LJ and Chan DW: **Proteomic Approaches to tumor marker discovery** *Arch Pathol Lab Med* 2002, **126:**1518-26.
9. Petricoin EF 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC and Liotta LA: **Serum proteomic patterns for detection of prostate cancer** *J Natl Cancer Inst* 2002, **94:**1576-1578.
10. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, Feng Z, Semmes OJ and Wright GL Jr: **Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients** *Clin Chem* 2002, **48:**1835-1843.
11. Cazares LH, Adam BL, Ward MD, Nasim S, Schellhammer PF, Semmes OJ and Wright GL Jr: **Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry** *Clin Cancer Res* 2002, **8:**2541-2552.
12. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z and Wright GL Jr: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men** *Cancer Res* 2002, **62:**3609-3614.
13. Adam BL, Vlahou A, Semmes OJ and Wright GL Jr: **Proteomic approaches to biomarker discovery in prostate and bladder cancers** *Proteomics* 2001, **1:**1264-1270.
14. Vlahou A, Schellhammer PF, Mendrinos S, Patel K, Kondylis FI, Gong L, Nasim S and Wright GL Jr: **Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine** *Am J Pathol* 2001, **58:**1491-1502.
15. Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC and Coombes KR: **A comprehensive Approach to the analysis of MALDI-TOF proteomics spectra from serum samples** *Proteomics* .
16. Xu Y, Shen Z, Wiper DW, Wu M, Morton RE, Elson P, Kennedy AW, Belinson J, Markman M and Casey G: **Lysophosphatidic acid as a potential biomarker for ovarian and other gynecologic cancers** *JAMA* 1998, **280:**719-723.
17. Xu Y, Xiao YJ, Baudhuin LM and Schwartz BM: **The role and clinical applications of bioactive lysolipids in ovarian cancer** *J Soc Gynecol Investig* 2001, **8:**1-13.
18. Xiao YJ, Schwartz B, Washington M, Kennedy A, Webster K, Belinson J and Xu Y: **Electrospray ionization mass spectrometry analysis of lysophospholipids in human ascitic fluids: comparison of the lysophospholipid contents in malignant vs nonmalignant ascitic fluids** *Anal Biochem* 2001, **290:**302-313.
19. Xiao Y, Chen Y, Kennedy AW, Belinson J and Xu Y: **Evaluation of Plasma Lysophospholipids for Diagnostic Significance Using Electrospray Ionization Mass Spectrometry (ESI-MS) Analysis** *Ann N Y Acad Sci* 2000, **905:**242-259.
20. Baker DL, Morrison P, Miller B, Riely CA, Tolley B, Westermann AM, Bonfrer JM, Bais E, Moolenaar WH and Tigyi G: **Plasma lysophosphatidic acid concentration and ovarian cancer** *JAMA* 2002, **287:**3081-3082.
21. **Atairgin Technologies Inc** [http://www.atairgin.com/atairgin.htm]
22. **What is Perimenopause?** [http://www.menopause-online.com/pmsormenopause.html]
23. *SAS Institute Inc., SAS/STAT User's Guide, Version 8, DISCRIM procedure, Pages 1011-1119, NPAR1WAY procedure, Pages 2505-2552, and STEPDISC procedure, Pages 3153-3179, Cary, NC: SAS Institute Inc.* 1999.
24. **Clinical Proteomics Program Database Detailed explanation of Proteome Quest for data analysis** [http://clinicalproteomics.steem.com/proteome-detail.php]