Harvard University
Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 29

Simultaneous and Exact Interval Estimates for the Contrast of Two Groups Based on an Extremely High Dimensional Response Variable: Application to Mass Spec Data Analysis

Yuhyun Park* Sean R. Downing† Cheng Li Dr.‡
William C. Hahn** Philip W. Kantoff†† L. J. Wei‡‡

*Harvard University, ypark@hsph.harvard.edu

†Lank Center for Genitourinary Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Sean_Downing@dfci.harvard.edu

‡Harvard University, cli@hsph.harvard.edu

**Lank Center for Genitourinary Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, William_Hahn@dfci.harvard.edu

††Lank Center for Genitourinary Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Philip_Kantoff@dfci.harvard.edu

‡‡Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper29

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Simultaneous and exact interval estimates for the contrast of two groups based on an extremely high dimensional response variable: Application to Mass Spec data analysis

Yuhyun Park [1,2]*†, Sean R. Downing [3]†, Cheng Li [1,2], William C. Hahn [3], Philip W. Kantoff [3], and L.J.Wei [2]

[1]Department of Biostatistics, Dana-Farber Cancer Institute, [2]Department of Biostatistics, Harvard School of Public Health, [3]Lank Center for Genitourinary Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, U.S.A.

ABSTRACT

Motivation: Analysis of high-throughput proteomic/genomic data, in particular, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) data and microarray data, has led to a multitude of techniques aimed at identifying potential biomarkers. For SELDI-TOF MS data, most of these techniques rely on arbitrary, user-defined rules for identifying peaks and do not control for global error, potentially leading to false positives.

Results: We have devised a simultaneous confidence bands method capable of detecting potential biomarkers, while controlling for overall Type I error, in high-dimensional genomic/proteomic datasets that discriminate two treatment groups using a permutation scheme. For example, for the SELDI-TOF MS data, we treat the entire spectrum as a stochastic process and construct $(1-\alpha)$ confidence bands for the mean differences between groups. Furthermore, peaks were identified based on the maximal differences between the groups as determined by the confidence bands. The analysis method herein described gives both qualitative (significance in terms of p-value) and quantitative data (magnitude of difference). The Clinical Proteomics Programs Databank's ovarian cancer dataset and data from in-house samples containing known spiked-in proteins were analyzed. We were able to identify potential biomarkers similar to those described in previous analysis of the ovarian cancer data, however, while these markers are highly significant between cancer and normal groups, our analysis indicated the absolute difference between the two groups was minimal. In addition, we found additional markers than those previously described with greater differences in average intensities. The proposed confidence bands method successfully detected the spiked-in peaks, as well as, secondary peaks generated by adducts and double-charged species. We also illustrate our method utilizing paired gene expression data from a prostate cancer microarray experiment by constructing confidence bands for the fold changes between cancer and normal samples.

Availability: The R and C codes are available from the authors.

Contact: park@jimmy.harvard.edu

*to whom correspondence should be addressed

†These authors contributed equally to this work

1 INTRODUCTION

Microarray and mass spectrometry technologies require effective statistical/computational methods that find important biomarkers, defined as features (genes/proteins), differentially expressed between two groups of samples within thousands of features. An inherent problem in analyzing such high-dimensional data is the detection of false positives if one uses separate statistical tests for each feature using traditional p-value cutoffs of 0.01 or 0.05, due to a large number of features which are potentially correlated with each other by unknown fashion. This multiplicity problem has been widely studied in the statistical literature (Benjamini & Hochberg 1995, Lehmann. 1986, Hochberg & Tahmhane 1987, Westfall & Young. 1993) and recent papers analyzing microarray experiments (Dudoit *et al.* 2003, Efron *et al.* 2002, Golub *et al.* 1999, Pollard & van der Laan. 2003, Tusher *et al.* 2001).

However, most of the ideas introduced for biomarker detection in high-dimensional data, so far, were originated from hypothesis testing, rather than interval estimates for the difference between two groups. While such two-sample tests only detect a qualitative significance of difference (p-value) that measures how likely markers are to distinguish between two sample groups, interval estimates are much more informative by giving quantitative importance of the markers (multitude of contrast between groups) on top of qualitative importance. For example, if the 95% confidence interval for the mean difference in expression between two groups for gene A is (20, 40), then gene A is a statistically significant difference between groups with a p-value less than 0.05 and the magnitude of difference in expression can range from 20 to 40 with 95% confidence. However if that for gene B is (-20, 10) indicating the mean expression difference for gene B between groups can be 0, then the p-value for testing gene B is greater than 0.05. Likewise, $(1-\alpha)$ interval estimates corresponds to the testing with a level of α. Usefulness of such interval estimation is well acknowledged in the statistical literature, however, simultaneous interval estimation for high-dimensional data has not been widely considered due to complications arising from the unknown dependence between structures within a large number of features.

We propose a new and simple algorithm based on permutation method to construct $(1-\alpha)$ simultaneous and exact confidence bands for any contrast assessment between groups with high-dimensional data sets. The proposed method allows for visualization of the possible range of difference in protein/gene abundance between groups with statistical significance while simultaneously controlling for overall type I error. We consider comparisons not only between two independent samples, but also dependent paired samples. One of the most intuitive contrast measurements would be mean difference between two groups and it is usually tested by *t*-statistics. Our method can be flexibly generalized to construct confidence bands for general two- or one-sample test statistics, such as Wilcoxon test statistics.

The confidence bands method can be applied to different data sets including microarray, CGH/SNP experiments, or proteomics data. In particular, we found this method to be very useful for exploring proteome-wide biomarkers using SELDI-TOF MS technology.

Conventional methods for analyzing SELDI-TOF MS data first detect the peaks in a spectrum generated for each sample after calibration and then align these peaks across the samples. Next, peaks related to a mass-to-charge ratio (m/z) that discriminate groups based on testing of peak intensities are determined. However, the peak detecting methods are controversial because they are ad-hoc and the results can vary due to user-defined parameters such as signal-to-noise ratio. To tackle this problem, researchers have used "fingerprints" (Fung *et al.* 2001 and Vlahou *et al.* 2001), however, this method does not allow for individual peak labeling and subsequent protein identification. Morris *et al.* 2005 introduced a new peak detection method based on the mean spectrum and demonstrated that the usage of the spectrum average leads to greater sensitivity and specificity while eliminating the difficult and intrinsically error-laden step of matching detected peaks on individual spectra.

In the confidence bands method, we define a new concept for peaks (biomarkers) based on the proposed confidence bands. We use all the intensity data and treat them as a stochastic process along the spectrum without a peak calling procedure. After constructing $(1-\alpha)$ confidence bands for the mean difference between groups along the entire spectrum, we automatically search for peaks, with mass accuracy adjustment, that have the maximal differences between groups as determined by the confidence bands. In this way, we can obtain the statistically meaningful peaks that discriminate two groups with qualitative and quantitative significance.

First we used well-known SELDI-TOF MS data sets from the latest experiment for ovarian cancer in the Clinical Proteomics Programs Databank (http://clinicalproteomics.steem.com). This dataset has been discussed by several papers (Petricoin *et al.* 2002, Sorace & Zhan 2003, Baggerl *et al.* 2004, Diamandis 2004), and has been controversial. We re-analyzed the data and compared our results with the results from these papers. We also performed a spike-in experiment with samples from 91 prostate cancer patients to assess the accuracy of our methods.

For demonstrating our method for matching paired samples and other high-dimensional data, we also constructed confidence bands for the fold change between cancer and normal samples for a prostate cancer microarray experiment with 46 matching pairs that was used in Singh *et al.* 2002.

2 STATISTICAL METHODS

Suppose that we are interested in making inferences about the difference between two groups, A and B, based on a large set of measurements from each study subject. That is, the response random vector for A is $\{X(t)\} = \{X(t), t \in I\}$ and is $\{Y(t)\} = \{Y(t), t \in I\}$ for B, where I is the index set. Furthermore, we assume that there exists an unknown constant vector $\{\Delta_0(t), t \in I\}$ such that $\{Y(t) - \Delta_0(t)\}$ having the same distribution as that of $\{X(t)\}$. Let $\{X_i(t)\}, i = 1, \cdots, m$, be m independent copies of $\{X(t)\}$ and $\{Y_j(t)\}, j = 1, \cdots, n$ be n independent copies of $\{Y(t)\}$. We are interested in constructing a $(1-\alpha)$ simultaneous confidence band $\mathcal{J}$ for $\{\Delta_0(t)\}$, where $0 < \alpha < 1$. That is,

$$\text{pr}(\Delta_0(t) \in \mathcal{J}, \text{for all } t \in I) \geq 1 - \alpha. \tag{2.1}$$

Now, let $\bar{Y}(t)$ and $\bar{X}(t)$ be the sample means of $X(t)$ and $Y(t)$, respectively. Then, for each fixed t, a pointwise, two-sided $(1-\alpha)$ confidence interval for $\Delta_0(t)$ is

$$\hat{\Delta}(t) \pm z_{\alpha/2} S(t), \tag{2.2}$$

where $\hat{\Delta}(t) = \bar{Y}(t) - \bar{X}(t)$, $S(t)$ is a type of sample standard error for $\hat{\Delta}(t)$, and z_α, the upper 100α percentage point of the standard normal. Obviously, to obtain a $(1-\alpha)$ confidence band $\mathcal{J}$ in (2.1), one needs to replace $z_{\alpha/2}$ in (2.2) with larger cutoff values, say, c_α and d_α such that the interval

$$\mathcal{J} = (\hat{\Delta}(t) - c_\alpha S(t), \hat{\Delta}(t) + d_\alpha S(t)) \tag{2.3}$$

contains the true $\Delta_0(t)$ for all $t \in I$ with probability $1-\alpha$. Unfortunately, since we do not know the dependence structure among the components of $\{X(t)\}$ nor of $\{Y(t)\}$, the cutoff points c_α and d_α are difficult, if not impossible, to obtain analytically even for the case with large sample sizes m and n.

Here, we utilize a simple permutation idea to obtain these cutoff point c_α and d_α. First, for a generic random quantity Q, let the low case q be its observed value. Let c be a positive real value and let $\Delta_c(t) = \hat{\delta}(t) - cs(t)$. If $\{\Delta_c(t), t \in I\}$ is the true value of $\{\Delta_0(t)\}$, then $\{y_j(t) - \Delta_c(t)\}, j = 1, \cdots, n$, and $\{x_i(t)\}, i = 1, \cdots, m$, were generated from the same distribution. Let $\{Y_j^*(t)\}$ be a random sample with size n drawn from the finite population composed of $\{y_j(t) - \Delta_c(t)\}, j = 1, \cdots, n$, and $\{x_i(t)\}, i = 1, \cdots, m$, and let $\{X_i^*(t)\}, i = 1, \cdots, m$ be the random sample which is the complement of $\{Y^*(t)\}$. Moreover, let $\bar{Y}^*(t)$ and $\bar{X}^*(t)$ be the sample means for $\{Y^*(t)\}$ and $\{X^*(t)\}$, respectively. Then, $c \leq c_\alpha$ if

$$\text{pr}(\sup_{t \in I}\{\frac{\bar{Y}^*(t) - \bar{X}^*(t)}{s(t)}\} < \sup_{t \in I}\{\frac{\hat{\delta}(t) - \Delta_c(t)}{s(t)}\}) \geq 1 - \alpha/2, \tag{2.4}$$

where the probability is generated from the random samples $\{Y_j^*(t)\}, j = 1, \cdots, n$, and $\{X_i^*(t)\}, i = 1, \cdots, m$. If the left hand side of (2.4) is less than $(1-\alpha/2)$, then c is greater than c_α. To obtain a good approximation to c_α, in practice, first, we let $c = z_{\alpha/2}$ and if (2.4) is satisfied, we then increase c by a small value, say, 0.01 and check (2.4) again. We continue this process until (2.4) is violated.

To obtain the upper bound d_α, we let d be a positive number and $\Delta_d(t) = \hat{\delta} + ds(t)$. We let $\{Y_j^*(t)\}, j = 1, \cdots, n$, be random sample for a population consists of $\{y_j(t) - \Delta_d(t)\}, j = 1, \cdots, n$ and $\{x_i(t)\}, i = 1, \cdots, m$. Moreover, let $\{X_i^*(t)\}, i = 1, \cdots, m$ be the complement of the sample $\{Y^*(t)\}$. Then, $d \leq d_\alpha$ if

$$\text{pr}(\inf_{t \in I}\{\frac{\bar{Y}^*(t) - \bar{X}^*(t)}{s(t)}\} > -\inf_{t \in I}\{\frac{\hat{\delta}(t) - \Delta_d(t)}{s(t)}\}) \geq 1 - \alpha/2. \tag{2.5}$$

Otherwise, $d > d_\alpha$. Again, one may start with $d = z_{\alpha/2}$, check (2.5) and continue this process by adding a small value to d if necessary.

These confidence bands (2.3) can be graphically displayed with two curves along $\{t, t \in I\}$ indicating the possible range of the difference between mean expression of two groups with $(1-\alpha)$ confidence. Therefore, if (2.3) for t does not include 0, such t is significantly differentially expressed with

a level of α (in other words, p-value $< \alpha$). Here, a collection of such significant t values is denoted by the significant region T.

Consider the case that $\{X(t)\}$ and $\{Y(t)\}$ are from the same study subject, but under different experimental conditions, say, A and B. Let $\{X_i(t), Y_i(t)\}, i = 1, ...n$, are independent copies of $\{X(t), Y(t), t \in I\}$. The confidence bands for $\{\Delta_0(t)\}$ can be obtained via a similar argument. However, the random vectors $\{X_i^*(t), Y_i^*(t)\}$ in (2.4), $i = 1, \cdots, n$, are generated by permuting $\{x_i(t), t \in I\}$ and $\{y_i(t) - \Delta_c(t), t \in I\}$ randomly within the ith subject, $i = 1, \cdots, n$. For (2.5), $\{X_i^*(t), Y_i^*(t)\}, i = 1, \cdots, n$ are generated by permuting $\{x_i(t)\}$ and $\{y_i(t) + \Delta_d(t)\}, i = 1, \cdots, n$ randomly.

Note that the confidence band for $\{\Delta_0(t)\}$ is constructed by inverting a test statistic based on $\bar{Y}(t) - \bar{X}(t) - \Delta_0(t)$, which can be replaced by any two sample or one sample (for paired observations) test statistic $W_t(\Delta_0(t)) = W(\underline{Y}(t) - \underline{\Delta}_0(t), \underline{X}(t))$, where $\underline{X}(t)$ and $\underline{Y}(t)$ are the vectors of the random samples $\{X_i(t)\}, i = 1, \cdots, m$, and $\{Y_j(t)\}, j = 1, \cdots, n$, respectively, and $\underline{\Delta}_0(t)$ is a $n \times 1$ vector whose components are $\Delta_0(t)$. For example, W may be the standard two or one-sample Wilcoxon test statistic. Let $\hat{\Delta}(t)$ be the consistent estimator by solving the equation $W_t(\Delta(t)) = 0$, and $S(t)$ is the standard error estimate of $W_t(\hat{\Delta}(t))$. Then, the cutoff points c_α and d_α for the confidence band $\mathcal{J}$ in (2.3) can be obtained via the above iterative procedure. Specifically, to check whether $c \leq c_\alpha$, we replace (2.4) by

$$\text{pr}(\sup_{t\in I}\{\frac{W(\underline{Y}^*(t), \underline{X}^*(t))}{s(t)}\} < \sup_{t\in I}\{\frac{w(\underline{y}(t) - \underline{\Delta}_c(t), \underline{x}(t))}{s(t)}\}) \geq 1-\alpha/2,$$

and to check whether $d \leq d_\alpha$, we replace (2.5) by

$$\text{pr}(\inf_{t\in I}\{\frac{W(\underline{Y}^*(t), \underline{X}^*(t))}{s(t)}\} > -\inf_{t\in I}\{\frac{w(\underline{y}(t) - \underline{\Delta}_d(t), \underline{x}(t))}{s(t)}\}) \geq 1-\alpha/2,$$

where $\underline{Y}^*(t)$ and $\underline{X}^*(t)$ are the vectors consisting of the random samples $\{Y_j^*(t)\}$ and$\{X_i^*(t)\}$, respectively, and $\underline{\Delta}_c(t)$ and $\underline{\Delta}_d(t)$ are $n \times 1$ vectors whose components are $\Delta_c(t)$ and $\Delta_d(t)$, respectively.

3 APPLICATION TO MASS SPECTROMETRY DATA ANALYSIS

3.1 Finding the Potential Biomarkers (peaks) for SELDI-TOF MS data

The mass-axis of the SELDI-TOF MS output shifts from experiment to experiment by approximately η =0.2-0.5% of the m/z values. Based on the confidence bands, we searched for peaks that may be potential biomarkers with, say η =0.5%, mass accuracy. These peaks best discriminated two groups both qualitatively and quantitatively among the significant regions found in the test. First, we defined the minimum-potential-change (MPC) at t m/z as the minimal absolute change between the mean intensities at t m/z of two groups with confidence. Therefore if the two confidence bounds at t m/z include zero, then MPC should be 0 and if not, then MPC would be either the lower bound for positive mean difference at t m/z or the absolute value of the upper bounds for negative mean difference at t m/z. Let's say that the m/z value with the largest MPC is t_1. We then updated T by removing all satisfying $(1-\eta)*t_1 \leq t \leq (1+\eta)*t_1$. This process was repeated until there existed unique t per each $\pm\eta$ window in the significant region T. Furthermore, we excluded the m/z values that were not apexes among remaining T by examining the slopes of the curves of observed difference of mean intensities of two groups. We concluded that the final T was the list of important biomarkers.

3.2 Application to Petricoin's ovarian SELDI-TOF MS dataset

We analyzed the latest SELDI-TOF MS data from the ovarian cancer study available in the Clinical Proteomics Programs Databank with our method. This set of data consists of serum profiles of 162 subjects with ovarian cancer and 91 non-cancer control subjects. For each subject, a set of data consisting of intensities at 15,154 distinct m/z values ranging from 0.0000786 to 19,995.513 was available for analysis. This dataset was constructed using Ciphergen WCX2 ProteinChip Arrays. Preparation of chips for sample analysis was performed robotically and the raw data, without baseline subtraction, was posted for download. We used the normalization method outlined in the Clinical Proteomics Databank by scaling the intensities value between 0 and 1. Additional details of experimental data collection can be found at the Clinical Proteomics Programs Databank. We analyzed the ovarian cancer dataset with all 11,003 m/z data points within the m/z range of I = [1,500 m/z, 20,000 m/z] for 91 normal and 162 tumor samples. The intensity measures within the rage below 1500 m/z were discarded due to the effects of matrix. Table 1 shows the cut-off points, c_α and d_α, for the confidence bands, with levels of, 0.005, 0.01, and 0.05 with the total permutation number of M = 10,000.

Table 1. The cut-off values of 99.9%, 99.5%, 99% and 95% confidence bands for Petricoin's ovarian cancer dataset

$(1-\alpha)$	99.9%	99.5%	99%	95%
CB (c_α,d_α)	(4.10, -3.85)	(3.50, -3.50)	(3.30, -3.25)	(2.95, -2.90)
Bonferroni	5.40	5.11	4.97	4.65

The cut-off values were relatively smaller than the Bonferroni's adjusted cut-off values which are the cut-off values under the conservative assumption that intensities for individual m/z points are independent. Figure 1 shows the 95% confidence bands for the differences in mean intensity between cancer and normal patients. We obtained the significant region T (the shaded regions in Figure 1) by excluding t values including 0 within the 95% confidence bands for further analysis.

We next followed the procedure of the section 3.1 with a mass precision of 0.5% and found 48 biomarkers in the region of I as significant peaks with 95% confidence. Figure 2 illustrates (a) adjusted p-value curve by maxT method (Westfall & Young, 1993); A higher p-value curve indicates greater significance of difference between the two groups at that m/z value, and (b) 95%, 99.5% confidence bands in the region of [6700m/z, 8100m/z]. The outer dotted lines are 99.5% and the inner dashed lines are 95% confidence bands since larger ranges for the difference are produced for higher confidence level.

It is important to note that the confidence bands give more information than global p-value curve. Whereas the conventional p-value curve only tests if two groups are different to each other with level of α, the confidence bands actually show the magnitude of differences with the corresponding confidence of $1 - \alpha$. Moreover, p-value curves alone do not give information about the precise m/z position for

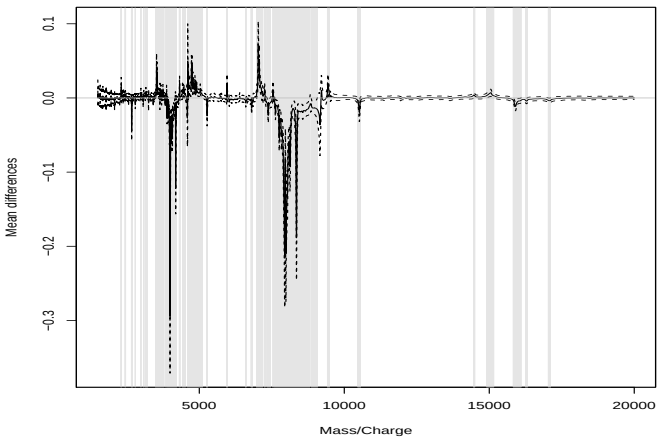

Fig. 1. 95% proteome-wide confidence bands for Petricoin's ovarian MS dataset

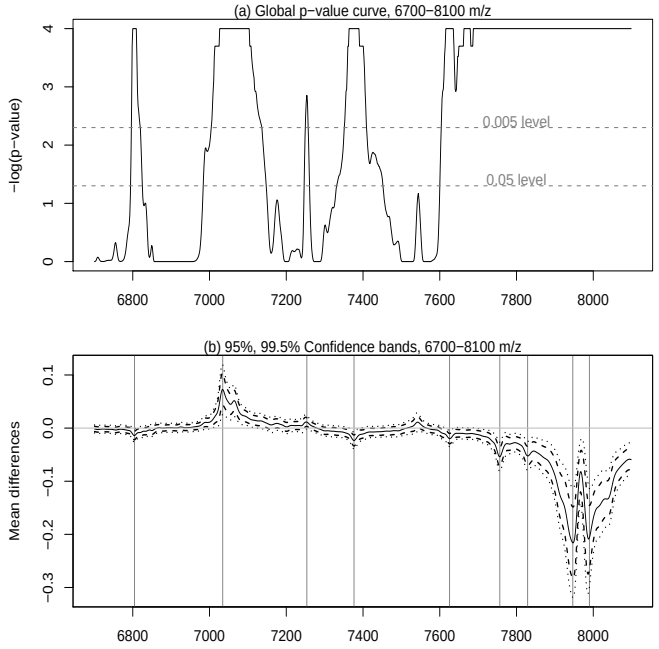

Fig. 2. Zoom-in figure within the region of [6,700 m/z, 8,100 m/z] for (a) global p-value curve and (b) 95%, 99.5% confidence bands for Petricoin's ovarian MS dataset

potential biomarkers. For example, the peak at 6800 m/z was found to be as significant in the p-value curve as the region 7700 - 8000 m/z (p-value $\leq$ 0.0001). However, the confidence bands indicate that the region 7700-8000 m/z contains four individual peaks, each with a greater magnitude of difference that the single peak at 6800 m/z. In this way, we were able to find potential biomarkers based on actual contrast between two groups, not arbitrary ad-hoc ways of defining peaks. We sorted 48 detected markers based on their minimum potential changes, (The data can be found in the supplementary documentation at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/), and we compared the top five significant biomarkers

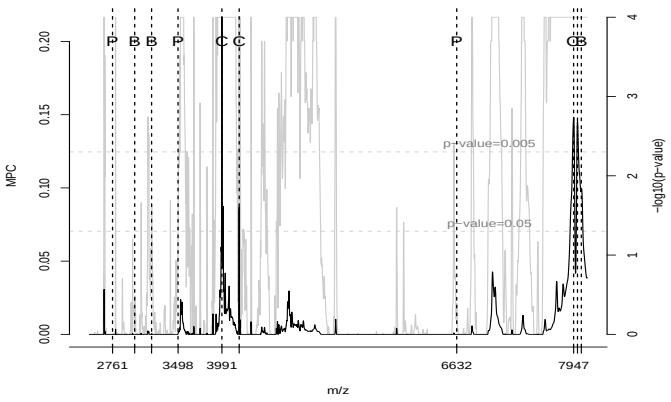

Fig. 3. Minimum potential changes (MPC) and p-values for detected peaks by CB (C), Petricoin's (P), and Baggerly's (B) methods

detected by our method with significant peaks reported in the Clinical Proteomics Databank and those from Baggerly *et al.* We report their observed mean difference (cancer-normal) and 95% confidence bands in Table 2.

Table 2. Top five significant markers at the level of 0.05 detected by Baggerly's, Petricoin's, and the CB methods

CB method		Petricoin's method		Baggerly's method	
m/z	95% CB	m/z	95% CB	m/z	95% CB
3900	(-0.37, -0.21)	2761	(-0.003, 0.005)	1531	(-0.006, 0.017)
7947	(-0.28, -0.15)	19643	(0.001, -0.003)	3010	(-0.001, 0.007)
7990	(-0.27, -0.15)	6632	(-0.009, 0.002)	3200	(0.001. 0.009)
8351	(-0.24, -0.14)	14052	(-0.002, 0.004)	8033	(-0.17, -0.10)
4185	(-0.15, -0.09)	3498	(-0.001, 0.010)		

Figure 3 shows the MPC curves (left y-axis and black solid curves) and p-value curves (right y-axis and grey curves) for m/z values which were determined to be significant in these two papers. Although there was no direct match between peaks identified by Petricoin *et al.* and the CB method, Petricoin's peaks at 2761, 3498 and 6632 were within 1% m/z of CB method detected peaks that have significant p-values but relatively small MPC (Figure 4). This may be due to incorrect peak-calling. Only two m/z at 3200 and 8033 detected by Baggerly *et al.* 2004 were found to be significant in our methods, however, the magnitude of difference between the groups at these two m/z values was again low when compared to the top markers identified using the CB method.

3.3 Application to the Spike-in Study

Plasma samples from 91 prostate cancer patients were divided into 7 age matched groups consisting of thirteen samples. The groups were labeled A-G. Groups B-F were spiked with five proteins at 1X, 2X, 5X, and 10X concentrations in a "Latin Square"

formation (Table 3). The minimal concentration of each of the spiked proteins allowing for a detectable peak in plasma was previously determined (data not shown). The minimal concentration for each protein was labeled as 1X and was found to be 1 fmol/μL for cytochrome c (from bovine heart, Sigma, St. Louis, MO), 10 fmol/μL for ubiquitin (from bovine red blood cells, Sigma), lysozyme (from chicken egg white, Sigma), and myoglobin (from horse heart, Sigma), and 100 fmol/μL for trypsinogen (from bovine pancreas, Sigma). The volume of spiked-in proteins was fixed at 10% of the plasma volume. Group A was not spiked with protein, however, an equal volume of diluent was added. Group G contained all five proteins at maximal (10X) concentrations.

Table 3. "Latin Square" design and protein concentrations

Groups	Cytochrome c	Ubiquitin	Lysozyme	Myoglobin	Trypsinogen
A	0X	0X	0X	0X	0X
B	0X	1X	2X	5X	10X
C	1X	2X	5X	10X	0X
D	2X	5X	10X	0X	1X
E	5X	10X	0X	1X	2X
F	10X	0X	1X	2X	5X
G	10X	10X	10X	10X	10X

Proteins 1X= Ubiquitin (1 fmol/μL), Cytochrome/Lysozyme/Myoglobin (10 fmol/μL), Trypsinogen(100 fmol/μL)

Following the addition of proteins, 20 μL of each plasma sample was diluted with 30 μL 9 M urea and incubated at 4°C for 30 minutes in order to denature proteins. The samples were further diluted with 150 μL 1 M urea and subsequently stored at -80°C until analyzed by SELDI-TOF MS.

Using a Biomek 2000 (Beckman Coulter, Fullerton, CA), CM10 ProteinChip Arrays (Ciphergen Biosystems, Freemont, CA) were washed two times with 150 μL CM Low Stringency buffer (Ciphergen) with shaking for 5 minutes at room temperature. Following the wash step, 90 μL of buffer was aliquoted onto each spot of the array and 10 μL of sample was then added. The arrays were shaken for 30 minutes at room temperature to allow for protein binding to the surface chemistry. Subsequently, the diluted samples were removed and the arrays were washed three times with 150 μL buffer, with shaking for 5 minutes per wash at room temperature, and rinsed twice with 200 μL water. The arrays were air dried and 1 μL of sinapinic acid (Ciphergen) was added twice to the arrays. The samples were analyzed on a PBSIIc SELDI-TOF mass spectrometer (Ciphergen) per manufacturer's instructions at a laser setting of 190, detector setting of 7, and a digitizer rate of 1000.

The data was baseline subtracted and normalized by total-ion current using Ciphergen Express (Ciphergen). From the five purity spectra that contain a single spiked-in protein (http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/), we found that SELDI-TOF MS experiments often produce secondary peaks, in addition to expected peaks, generated by multiple-charged species or matrix adducts for each of the five proteins. Moreover, we observed several peaks generated from contaminants within the pure spiked-in proteins. We also observed the intensities for group G, which contained all five proteins in maximal concentrations, were generally lower than those for other groups with maximal concentration most likely due to ion suppression. (The mean intensity curves of groups A-G in the m/z regions of each spiked-in protein with $\pm$0.5% precisions can be found in the supplementary documentation.) To assess the accuracy of detecting known proteins using the confidence band method, we compared all 21 possible pair-wise comparisons for groups A to G without knowledge of the five spiked-in proteins. Only the researcher conducting the SEDI-TOF MS knew the protein concentrations in each group and he did not take part in the analysis. We found 133 peaks as significant with 95% confidence. Table 4 reports the top ten detected peaks sorted primarily by the number of comparisons in which the peaks were detected as significant and secondarily by the largest MPC among the results from all comparisons between groups.

Table 4. Top 10 detected peaks from the CB methods with 95% confidence for the spiked-in experiment

m/z	95% CB	Comparison groups	Identity
8473	(9.17, 12.46)	ca	Myoglobin+2H+
12234*	(6.02, 8.73)	ga	Cytochrome c
16952*	(39, 4, 9.17)	ca	Myoglobin
14301*	(36.34, 45.92)	da	Lysozyme
15203*	(-11.69, -9.17)	cb	Trypsinogen Impurity
23979*	(8.48, 12.61)	ba	Trypsinogen
7146*	(5.97, 8.39)	da	Lysozyme+2H+
14521*	(4.72, 7.03)	da	Lysozyme+SPA
15380	(2.30, 3.85)	fc	Trypsinogen Impurity
14691	(1.11, 1.74)	ga	Lysozyme+2SPA

* also found in top 10 by Ciphergen Express analysis

Ciphergen's biomarker detection algorithm, Ciphergen Express, found 124 significant markers with a level of 0.05 [data are not shown]. Their top 10 detected peaks also contained four of the five spiked-in proteins. Curiously, ubiquitin was not found to be one of the top ten most significant peaks. This may have been due to the decrease in intensity observed at higher concentrations for ubiquitin (supplementary figure). The resultant lower MPC resulted in ubiquitin being ranked as the 12th most significant peak, as determined by our analysis.

Baggerl *et al.* (2004) discussed the problems behind calibration, background subtraction, and normalization of data. In order to address these potential problems, we further analyzed our spiked-in data set to examine the effects of background subtraction and normalization. The total-ion current method of normalization assumes that the total amount of proteins in each sample may vary due to sample handling or instrument sensitivity. In general, with the exception of known disease states, the protein concentration of blood samples falls within a narrow range, but since the amount of proteins in the spike-in study may vary across the samples in different groups overall intensities may have been over-normalized with the total ion current method. We analyzed the raw data without background subtraction and normalization, and obtained 137 significant markers which contained 94 out of 133 markers detected from the analysis with the background subtraction and

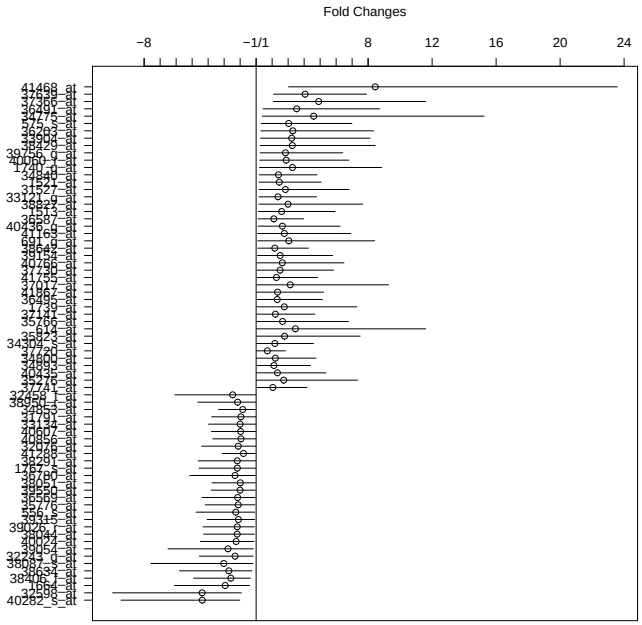

Fig. 4. 99% Confidence bands of the fold-change between paired tumor-normal samples of Singh *et al.*'s prostate cancer microarray experiment

normalization with 95% confidence bands (The detail of analysis can be found at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data). Spiked-in proteins were detected as with spectrum processing. The raw data from our spike-in study is posted at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/.

4 APPLICATION TO MICROARRAY DATA (PAIRED SAMPLES)

We applied our confidence band method to a previously published prostate cancer microarray experiment (Affymetrix H95Av2 containing 12,600 probesets) (Singh *et al.* 2002). Of the 52 prostate tumor and 50 normal samples, 46 were matching pairs. The paper ignored this dependency within the same patients and obtained 456 differently expressed genes at the level of 0.001 using the signal-to-noise method of Golub *et al.* 1999. We re-analyzed this data with our confidence bands method with the paired data and found 71 genes to be significantly different between cancerous and normal tissue at the level of 0.001. Our smaller number of significant genes indicates that previous analysis likely yielded many false-positives due to incorrect multiple-comparisons and neglect of the dependency of matching samples. Figure 4 shows 99% confidence bands of 71 significant genes sorted by their minimum-potential-change.

5 DISCUSSION

In this article, the approach we took for finding biomarkers in high-dimensional genomic/proteomic data is quite different from methods previously reported. Rather than determining qualitatively significant markers (with small p-value), we attempt to measure both qualitative and quantitative importance of markers by exploiting simultaneous confidence bands that reflect the true random fluctuation in the difference between groups of samples. Our algorithm cleverly bypasses the curse of dimension for estimating the high-dimensional parameters to obtain the interval estimates of the difference between groups, by reducing the problem into estimating two parameters, (c_α, d_α) in (2.3). The proposed method is flexible and can be extended to general two/one-sample test statistics. For example, if one wants to use a rank-based test, one can construct confidence bands for the median difference between groups. Our method also allows for the analysis of paired samples.

Although our method is applicable to any high-dimensional data, we extensively studied SELDI-TOF MS data as our major application. While SELDI-TOF MS technology has been acclaimed as one of the most powerful new frontiers among disease diagnosis technologies utilizing blood-borne proteins, inconsistent results from previous publications indicated the necessity of new, refined approaches to derive optimal biomarkers that can be both interpreted and accepted by the scientific community. Petricoin *et al.* (2002)'s claim that they can predict the presence of ovarian cancer using SELDI-TOF MS data with 100% accuracy has brought great attention, as well as, controversy. Zhu *et al.* (2003) also studied the early dataset used by Petricoin *et al.* (2002) and published a non-overlapping set of markers. Baggerly *et al.* (2004) also pointed out that the markers detected by Petricoin *et al.* (2002) were not significant in terms of t-statistics, leading to suspicion of their importance as discriminating biomarkers between cancer and normal samples. Solace & Zhan (2003) found markers using the latest dataset from Clinical Proteomics (the same set used for our results), however, the majority of their markers were less than 500 m/z and were likely to be artifacts or experimental bias as these are within the noise signal of the energy absorbing matrix. Our confidence bands method provides a new, alternative way to detect the potential biomarkers throughout the whole m/z region. Our method provides a powerful visualization tool for detecting potential markers with both qualitative and quantitative importance of markers without arbitrary peak-calling. Furthermore, our method detects a precise position for each peak that discriminates between two groups based on MPC with mass-accuracy adjustment. According to Diamandis (2004), SELDI-TOF technology is not capable of detecting any serum component at concentrations of less than 1 μg/mL and statistically significant markers with such small differences are often detected due to artifacts related to the nature of the clinical samples used or the MS instruments. Therefore, it is quite informative to sort the list of potential biomarkers by corresponding MPCs indicating the relative magnitude of protein abundance compared to control samples.

While the spike-in experiment did not generate a clean data set that contained only the spiked-in proteins as significant peaks in the spectrum, it was a valuable experiment to understand the nature of SELDI-TOF MS and to evaluate the performance of our CB method. The five proteins were initially chosen because unfractionated human plasma does not have peaks at the corresponding molecular weights on the Ciphergen cationic chip surface and all have isoelectric points at least 2 units above 4, the pH at which the low stringency wash was performed. While only the five spiked-in proteins were expected, the CB method detected 133 significant peaks. Most of the peaks were artifacts typical of the SELDI-TOF MS method, such as, EAM adducts, multiple charged species, and ion suppression (see Table 4 for examples of the former two). The albumin peak at approximately 66,500 daltons was found to be significantly different between the groups (data not shown). This

was due to ion suppression caused by an increase in cytochrome c levels as evidenced by groups E-G having the lowest levels of albumin and the highest levels of cytochrome c. Furthermore, some of the additional, unexpected peaks were found to be contaminants in the stocks of "pure" proteins. The stock of trypsinogen contained several peaks that could not be attributed to any of the typical artifacts and SDS-PAGE demonstrated that these excess peaks were in fact contaminating proteins. One of these impurities, at an m/z of 15,203, was found to be significantly different, in fact, it was one of the top ten peaks detected by both the CB and Ciphergen analysis methods. A second peak at 15,380 m/z, also identified as a top ten discriminating marker, may have been the EAM adduct of the 15,203 m/z peak as it was within 0.2% of the expected mass-to-charge ratio. Curiously, these peaks did not have a maximal MPC in comparison to group A as expected, but rather with group C. Group C was the experimental group that did not contain trypsinogen, but did contain all of the other proteins. The contaminants were detected at the highest levels in groups B and F containing 10X and 5X trypsinogen, respectively. It should be noted that two of the proteins did not exhibit a linear relationship between intensity and concentration when placed into plasma (Supplementary figure f). Ubiquitin demonstrated a biphasic response with increasing intensity to a maximal at 2X concentration and then a reduction in intensity at higher concentrations. Trypsinogen showed no difference in intensity between 1-5X concentration with a marked increase in intensity at 10X. Considering the fact that different spiked-in proteins yielded different levels of increment in intensities, it would be worth-while to consider constructing confidence bands for log-fold changes by talking the log-transformation of the spectra intensities. Our method was successful in detecting the spiked-in proteins robustly, regardless of background subtraction and normalization. The described experiment also demonstrates some of the potential hazards of conducting spike-in studies.

Simultaneous confidence band method is a well-established inference scheme in statistical literatures, however, it has not been exploited in bioinformatics thus far. This type of interval estimates for contrast between groups can be very attractive for genomic/proteomic datasets since this method allows investigators to visualize the potential differences of mean intensities between groups while guarding against false-positives due to multiple comparisons. It also yields meaningful biologically peaks rather than arbitrary, user-defined peaks for SELDI-TOF MS data.

REFERENCES

Baggerly, Keith A., Morris, Jeffrey S., and Coombes, Kevin R. (2004a). Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Data Sets from Different Experiments. *Bioinformatics*, **20**, 777-785.

Baggerly, Keith A., Edmonson, Sarah R., Morris, Jeffrey S., and Coombes, Kevin R. (2004b). High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-Related Cancer*, **11**(4):583-584.

Benjamini, Y., Hochberg, Y. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B, Methodological*, **57**, 289-300

Diamandis EP 2004 Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Molecular and Cellular Proteomics*, **3**, 367-378.

Dudoit, S., Shaffer, JP, Boldrick, JC. (2003) Multiplie hypothesis testing in microarray experiments, *Statistical Science*. **18**. 71-103.

Efron, B., Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 70-86

Fung, E.T., Thulasiraman, V., Weinberger, S.R., and Dalmasso, E.A. (2001) Protein biochips for differential profiling. *Current Opinion in Biotechnology* **12**(1):65-69.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Hochberg, Y., Tamhane, AC. (1987) Multiple comparison procedures, *John Wiley & Sons, Inc.*, New York.

Lehmann, E. L. (1986), Testing statistical hypotheses, *John Wiley & Sons, Inc*, New York.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, **21**: 1764 - 1775.

Petricoin EF, Ardekani AM, Hitt BA, Leviine PH, Fusaro VA, Steinbuerg SM, Mills GB, Simone C, Fishman DA, Kohn EC & Liotta LA (2002) Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. *Lancet*, **359**, 572-577.

Pollard KS, Van der Laan MJ. (2003) Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data. *Division of Biostatistics Working Paper 121. Berkeley, CA:University of California Berkeley.* Available: http://www.bepress.com/ucbiostat/paper121

Singh, D., Gebbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P.,Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W., (2002) Molecular determinants of prostate cancer behavior. *Cancer Cell*, **1**, 203-209.

Sorace, J., Zhan, M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, **4**:24. http://www.biomedcentral.com/1471-2105/4/24.

Tusher, VG, Tibshirani, R., Gilbert C. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc.Natl.Acad.Sci. U.S.A*, **98** 5116-5121.

Vlahou, A., Schellhammer, P.F., Mendrinos, S., Patel, K., Kondylis, F.I., Gong, L., Nasim,S., and Wright, G.L., Jr. (2001) Development of a novel proteomic approach for the detection of transitional carcinoma of the bladder in urine. *American Journal of Pathology* **158**(4):1491-1502.

Westfall, PH., Young, SS. (1993) Resampling-based multiple testing: examples and methods for P-value adjustment, *John Wiley & Sons, Inc.*, New York.

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 a proteomics dataset. *Bioinformatics*, **19**, 1636-1643.

Zhu, Wei, Wang, Xuena, Ma, Yeming, Rao, Manlong, Glimm, James, Kovach, John S. (2003) Detection of Cancer-Specific Markers amid Massive Mass Spectral Data. *PNAS*, **100**, 14666-14671.