Genome analysis

Simultaneous and exact interval estimates for the contrast of two groups based on an extremely high dimensional variable: application to mass spec data

Yuhyun Park1,2,*†, Sean R. Downing3,†, Dohyun Kim4, William C. Hahn3, Cheng Li1,2, Philip W. Kantoff3 and L.J. Wei1,2

1Department of Biostatistics, Dana-Farber Cancer Institute, 2Department of Biostatistics, Harvard School of Public Health, 3Lank Center for Genitourinary Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA and 4Department of Statistics, Seoul National University, Seoul, Korea.

Received on March 27, 2006; revised on March 12, 2007; accepted on March 28, 2007

Advance Access publication April 25, 2007

Associate Editor: Alfonso Valencia

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. Most of the statistical techniques for comparing two groups are based on qualitative measures such as \( P \)-value. A quantitative way such as interval estimation for the contrasts of two groups is more appealing.

Results: We have devised a simultaneous confidence bands method capable of detecting potential biomarkers, while controlling for overall confidence coverage level, in high-dimensional datasets that discriminate two treatment groups using a permutation scheme. For example, for the SELDI-TOF MS data, we deal with the entire spectrum simultaneously 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 (\( P \)-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.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Availability: R-package, ‘seie.zip’ (license: GNU GPL), is publicly available at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/

Contact: parkyuhyun@gmail.com

Supplementary information: For supplementary data, please refer to Bioinformatics online.

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 cut-offs 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 and Hochberg, 1995; Hochberg and Tamhane, 1987; Lehmann, 1986; Westfall and Young, 1993) and recent papers analyzing microarray experiments (Dudoit et al., 2003; Efron et al., 2002; Golub et al., 1999; Pollard and 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. An interval estimate provides a set of possible values for a parameter of interest, which is more informative than a single \( P \)-value to quantify its ‘importance’ from statistical and also biological interpretation. For example, in comparing the mean gene-expression levels between two experimental groups, suppose that for Gene X, 0.95 confidence interval for the fold change is \((1.1, 1.2)\) whereas 0.95 interval for the fold change for Gene Y is \((4, 8)\). And suppose the first one has the Z score of six and the second has that of five. In this case, the fold change for X may not be interesting biologically.
even though its $P$-value is more highly significant. 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 simple algorithm based on permutation method to construct $(1 - \alpha)$ simultaneous and exact confidence bands for any contrast assessment between groups with high-dimensional datasets. 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 confidence coverage level. 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 datasets including microarray, CGH/SNP experiments or proteomics data. In particular, we found this method to be very useful for exploring proteome-wide biomarkers using Surface-enhanced laser desorption/ionization time-to-flight mass spectrometry (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 peak detection method based on the mean spectrum and the results can vary depending on the method, we use all the intensity data along the spectrum individual spectra.

To bypass the difficulties associated with peak detection method, we need to use all the intensity data 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 determine the ‘peaks (biomarkers)’ that have potential maximal differences between groups based on 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 datasets from the latest experiment for ovarian cancer in the Clinical Proteomics Programs Databank (http://clinicalproteomics.steen.com). This dataset has been discussed by several papers (Baggerl et al., 2004; Diamandis, 2004; Petricoin et al., 2002; Sorace and Zhan 2003), 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 or some specific transformations thereof from each study subject, that is, the response random vector for $A$ is $\{X_i\} = \{X_i, t \in I\}$ and is $\{Y_j\} = \{Y_j, t \in I\}$ for $B$, where $I$ is the index set. We assume that there exists an unknown constant vector $\{\Delta_0^t, t \in I\}$ such that $\{Y_j - \Delta_0^t\}$ has the same distribution as that of $\{X_i\}$. Let $\{X_j\}, j = 1, \ldots, m, t \in I$, be $m$ independent copies of $\{X_i\}$ and $\{Y_j\}, j = 1, \ldots, n, t \in I$, be $n$ independent copies of $\{Y_j\}$.

We are interested in constructing a $(1 - \alpha)$ simultaneous confidence band $\mathcal{J}$ for $\{\Delta_0^t\}$, where $0 < \alpha < 1$ i.e.

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

Let $\hat{X}_t$ and $\hat{X}_t$ 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

$$\hat{\Delta}_t \pm z_{\alpha/2} S_t,$$  \hspace{1cm} (2)

where $\hat{\Delta}_t = \hat{Y}_t - \hat{X}_t$, $S_t$ is a sample standard error estimate for $\hat{\Delta}_t$, and $z_{\alpha}$ the upper $100\alpha$ percentage point of the standard normal. Obviously, to obtain a $(1 - \alpha)$ confidence band $\mathcal{J}$ in (1), one needs to replace $z_{\alpha/2}$ in (2) with larger cut-off values, say, $c_\alpha$ and $d_\alpha$ such that the interval

$$\mathcal{J} = (\hat{\Delta}_t - c_\alpha S_t, \hat{\Delta}_t + d_\alpha S_t) \quad (3)$$

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 cut-off points $c_\alpha$ and $d_\alpha$ 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 cut-off point $c_\alpha$ and $d_\alpha$. First, for a generic random quantity $Q$, let the lower case $q$ be its observed value.

Let $c$ be a positive real value and let $\Delta^t = \hat{\Delta}_t - c S_t$. If $\{\Delta^t, t \in I\}$ is the true value of $\{\Delta_0^t\}$, then $\{Y_j - \Delta^t, j = 1, \ldots, n\}$ and $\{X_i, i = 1, \ldots, m\}$, were generated from the same distribution. Let $\{Y^t\}$ be a random sample with size $n$ drawn from the finite population composed of $\{Y_j - \Delta^t, j = 1, \ldots, n\}$ and $\{X_i, i = 1, \ldots, m\}$ and let $\{X^t\}$, $i = 1, \ldots, m$ be the random sample which is the complement of $\{Y^t\}$. Moreover, let $\hat{Y}^t$ and $\hat{X}^t$ be the sample means for $\{Y^t\}$ and $\{X^t\}$, respectively.

Then $c \leq c^*$ if

$$\text{pr} \left( \sup_{i \in I} \frac{\hat{Y}^t - \hat{X}^t}{S_{\hat{Y}^t}} \leq \sup_{i \in I} \frac{\hat{\delta}_t - \Delta^t}{S_{\hat{\delta}_t}} \right) \geq 1 - \alpha/2, \quad (4)$$

where the probability is generated from the random samples $\{Y^t\}, j = 1, \ldots, n$, and $\{X^t\}, i = 1, \ldots, m$. If the left hand side of (4) is less than $(1 - \alpha/2)$, then $c$ is greater than $c^*$. Note that for the present case, $\sup_{i \in I} (\hat{\delta}_t - \Delta^t)/(S_{\hat{\delta}_t}) = c$. To obtain a good approximation to $c^*$, in practice, first we let $c = z_{\alpha/2}$ and if (4) is satisfied, we then increase $c$ by a small value, say, 0.01 and check (4) again. We continue this process until (4) is violated.
To obtain the upper bound \( d_u \), we let \( d \) be a positive number and \( \Delta^d_t = \hat{\delta} + d \). We let \((Y^*_j, j = 1, \ldots, n)\) be a random sample for a population consisting of \((y_j - \Delta^d_t), j = 1, \ldots, n\) and \((x_j), j = 1, \ldots, m\). Moreover, let \((X^*_i), i = 1, \ldots, m\) be the complement of the sample \((Y^*_j)\). Then, \( d \leq d_u \) if

\[
\Pr \left[ \inf_{i \in I} \left\{ \frac{1}{s_i} \left( \bar{y}_i - \bar{X}^*_i \right) \right\} < -\inf_{i \in I} \left\{ \frac{1}{s_i} (\hat{\delta} - \Delta^d_t) \right\} \right] \geq 1 - \alpha/2.
\]

Otherwise, \( d > d_u \). Note that for the present case, \( \inf_{i \in I} (\hat{\delta} - \Delta^d_t)/s_i = d \). Again, one may start with \( d = z_{\alpha/2} \), check (5) and continue this process by adding a small value to \( d \) if necessary.

These confidence bands (3) can be graphically displayed with two curves along \([t, t \in I]\) indicating the possible range of \( \Delta^d_t \) with \((1 - \alpha)\) confidence. Unlike the \( P \)-value curve across \( t \in I \), this confidence band is quite useful to identify features whose, for example, mean marker values between two groups are statistically and also biologically significant. Note that the confidence band for \( \Delta^d_t \) is from the same study subject, but under different experimental conditions, say, \( A \) and \( B \). Let \((X_t, Y_t), t = 1, \ldots, n\) be independent copies of \((X, Y, t \in I)\). The confidence bands for \( \Delta^d_t \) can be obtained via a similar argument.

Let \((X_t, Y_t), t = 1, \ldots, n\) be generated by permuting \((x_0, y_0, t \in I)\) randomly within the \( i \)th subject, \( i = 1, \ldots, n \). For \((5), (X^*_i, Y^*_i), i = 1, \ldots, n\) are generated by permuting \((x_0)\) and \((y_0 + \Delta^d_t), i = 1, \ldots, n\) randomly.

Note that the confidence band for \( \Delta^d_t \) is constructed by inverting a test statistic based on \( \bar{Y}_i - \bar{X}^*_i \), which can be replaced by any two sample or one sample (for paired observations) test statistic \( W \). Let \( \Delta^d_t \) be the consistent estimator by solving the equation \( W(\Delta_t) = 0 \) and \( S_t \) is the standard error estimate of \( W(\Delta_t) \). Then, the cut-off points \( c_{\alpha} \) and \( d_{\alpha} \) for the confidence band \( J \) in (3) can be obtained via the above iterative procedure. Specifically, to check whether \( \epsilon \leq \epsilon_0 \), we replace (4) by

\[
\Pr \left[ \sup_{i \in I} \left\{ \frac{W(Y^*_i, X^*_i)}{S_t} \right\} < \sup_{i \in I} \left\{ \frac{m(Y^*_i - \Delta^d_t, X^*_i)}{S_t} \right\} \right] \geq 1 - \alpha/2,
\]

and to check whether \( d \leq d_u \), we replace (5) by

\[
\Pr \left[ \inf_{i \in I} \left\{ \frac{W(Y^*_i, X^*_i)}{S_t} \right\} > -\inf_{i \in I} \left\{ \frac{m(Y^*_i - \Delta^d_t, X^*_i)}{S_t} \right\} \right] \geq 1 - \alpha/2,
\]

where \( Y^*_i \) and \( X^*_i \) are the vectors consisting of the random samples \((Y^*_j)\) and \((X^*_j)\), respectively and \( \Delta^d_t \) and \( \Delta^d_t \) are \( n \times 1 \) vectors whose components are \( \Delta^d_t \) 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 ~0.2-0.5% of the \( m/z \) values. Let \( \eta \) be a mass accuracy of a spectrum. We searched for peaks that may be potential biomarkers with \( \eta = 0.5\% \) mass accuracy based on the confidence bands. 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 \( t \) \( m/z \). Now, we denote \( T \) as a set of collection of the \( m/z \) values with the positive MPC values. Let us say that the \( m/z \) value with the largest MPC is \( t_t \). We then updated \( T \) by removing all satisfying \((1 - \eta) \times t_t \leq t \leq (1 + \eta) \times t_t \). 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 in mean intensities between 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 15154 distinct \( m/z \) values ranging from 0.0000786 to 19995.513 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 11003 \( m/z \) data points within the \( m/z \) range of \( I = [500 \text{ m/z}, 20000 \text{ m/z}] \) for 91 normal and 162 tumor samples. The intensity measures within the range below 1500 \( m/z \) were discarded due to the effects of matrix. Table 1 shows the cut-off points, \( c_{\alpha} \) and \( d_{\alpha} \), for the confidence bands, with levels of 0.005, 0.01 and 0.05 with the total permutation number of \( M = 10000 \). Using an Intel (R) Pentium (R) D 3.00 GHz CPU and 2.00 GB RAM, the algorithm took 8 min and 56 s for 10000 permutations, and 90 min and 12 s for 100000 permutations. However, it has been determined that there was no practical difference between 1000 and 10000 permutations. The estimated \( c_{\alpha} \) and \( d_{\alpha} \) obtained from 1000 permutations typically differ by less than 0.1 from those obtained from 10000 permutations.

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 Fig. 1) by excluding \( t \) values including 0 within the 95% confidence bands for further analysis.
Table 1. The cut-off values of 99.9, 99.5, 99 and 95% confidence bands for Petricoin’s ovarian cancer dataset

<table>
<thead>
<tr>
<th>(1−α)</th>
<th>99.9%</th>
<th>99.5%</th>
<th>99%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB (c*,d*)</td>
<td>(4.10,−3.85)</td>
<td>(3.50,−3.50)</td>
<td>(3.30,−3.25)</td>
<td>(2.95,−2.90)</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>5.40</td>
<td>5.11</td>
<td>4.97</td>
<td>4.65</td>
</tr>
</tbody>
</table>

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

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

It is important to note that the confidence bands give more information than a global \( P \)-value curve. Whereas the conventional \( P \)-value curve only conveys qualitative evidence on how strongly two groups are different from each other, the confidence bands actually show the magnitude of differences with a confidence of \( 1−\alpha \). Moreover, \( P \)-value curves alone do not give information about the precise \( m/z \) position for potential biomarkers. For example, the peak at 6800 \( m/z \) was found to be as significant on 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.

We sorted 48 detected markers based on their minimum potential changes, (The data can be found in the Supplementary Material at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/), and we compared the top five significant biomarkers 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.

Figure 3 shows the MPC curves (left \( y \)-axis and black solid curves) and \( P \)-value curves (right \( y \)-axis and gray 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 (Fig. 3). 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 seven 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/\( \mu l \) for cytochrome c (from bovine heart, Sigma, St. Louis, MO), 10 fmol/\( \mu l \) for ubiquitin (from bovine red blood cells, Sigma), lysozyme (from chicken egg white, Sigma) and myoglobin.
Using a Biomek 2000 (Beckman Coulter, Fullerton, CA, USA), CM10 ProteinChip Arrays (Ciphergen Biosystems, Fremont, CA, USA) were washed two times with 150 μl CM Low Stringency buffer (Ciphergen) with shaking for 5 min at room temperature. Following the wash step, 90 μl of buffer was aliquotted onto each spot of the array and 10 μl of sample was then added. The arrays were shaken for 30 min 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 min 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 seven 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/dhci/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 ±0.5% precisions can be found in the Supplementary Material.)

To assess the accuracy of detecting known proteins using the confidence band method, we compared all 21 possible pairwise comparisons for groups A–G without knowledge of the five spiked-in proteins. Only the researcher conducting the SELDI-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.

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 Fig.). 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 calibra-

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

<table>
<thead>
<tr>
<th>CB method</th>
<th>Petricoin’s method</th>
<th>Baggerly’s method</th>
</tr>
</thead>
<tbody>
<tr>
<td>m/z</td>
<td>95% CB</td>
<td>m/z</td>
</tr>
<tr>
<td>3900</td>
<td>(−0.37, −0.21)</td>
<td>2761</td>
</tr>
<tr>
<td>7947</td>
<td>(−0.28, −0.15)</td>
<td>19643</td>
</tr>
<tr>
<td>7990</td>
<td>(−0.27, −0.15)</td>
<td>6632</td>
</tr>
<tr>
<td>8351</td>
<td>(−0.24, −0.14)</td>
<td>14052</td>
</tr>
<tr>
<td>4185</td>
<td>(−0.15, −0.09)</td>
<td>3498</td>
</tr>
</tbody>
</table>

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

Table 3. ‘Latin Square’ design and protein concentrations

<table>
<thead>
<tr>
<th>Groups</th>
<th>Cytochrome c</th>
<th>Ubiquitin</th>
<th>Lysozyme</th>
<th>Myoglobin</th>
<th>Trypsinogen</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0X</td>
<td>0X</td>
<td>0X</td>
<td>0X</td>
<td>0X</td>
</tr>
<tr>
<td>B</td>
<td>0X</td>
<td>1X</td>
<td>2X</td>
<td>5X</td>
<td>10X</td>
</tr>
<tr>
<td>C</td>
<td>1X</td>
<td>2X</td>
<td>5X</td>
<td>10X</td>
<td>0X</td>
</tr>
<tr>
<td>D</td>
<td>2X</td>
<td>5X</td>
<td>10X</td>
<td>0X</td>
<td>1X</td>
</tr>
<tr>
<td>E</td>
<td>5X</td>
<td>10X</td>
<td>0X</td>
<td>1X</td>
<td>2X</td>
</tr>
<tr>
<td>F</td>
<td>10X</td>
<td>0X</td>
<td>1X</td>
<td>2X</td>
<td>5X</td>
</tr>
<tr>
<td>G</td>
<td>10X</td>
<td>10X</td>
<td>10X</td>
<td>10X</td>
<td>10X</td>
</tr>
</tbody>
</table>

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

(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.

Following the addition of proteins, 20 μl of each plasma sample was diluted with 30 μl 9M urea and incubated at 4°C for 30 min in order to denature proteins. The samples were further diluted with 150 μl 1M urea and subsequently stored at −8°C until analyzed by SELDI-TOF MS.
4 APPLICATION TO MICROARRAY DATA
(PAired SAMPLES)

We applied our confidence bands 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 article 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.

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 normalization with 95% confidence bands (The details of the analysis can be found at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data) Preprocessing of the spectra made no difference regarding the detection of the spiked-in proteins. The raw data from our spike-in study is posted at http://research2.dfci.harvard.edu/dfci/MS_spike-in_data/

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

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

*also found in top 10 by Ciphergen Express analysis.

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 measure both qualitative and quantitative importance of potential markers by exploiting simultaneous confidence bands that reflect the true random fluctuation in the difference between groups of samples. Our algorithm bypasses the difficulties associated with estimating high-dimensional parameters to obtain the interval estimates of the difference between groups by reducing the problem into estimating two parameters, (c_a, d_a) in (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. Sorace and Zhan, (2003) found markers using the latest dataset from Clinical Proteomics (the same set used for our analysis), however, the majority of their markers were <500 m/z and were likely to be artifacts or experimental bias as these are within the noise signal of the energy absorbing matrix.

Thus, our confidence band method employed an alternative approach to detect potential biomarkers throughout the m/z region. Most of the previous papers (Morris et al., 2005; Yasui et al., 2003) have developed various peak detection methods and classification rules. Conventional approaches for MS data analysis usually take these two steps.

However, our method processes the data without an ad-hoc-based peak detection method and directly identifies statistically meaningful markers that discriminate between two groups based on (1−α) simultaneous confidence bands. Our method is theoretically well justified in preventing overall false positives. Our method is simple and efficient without involving a complex classification rule which often yields non-biologically relevant markers and it provides a powerful visualization tool for potential bio-markers with both qualitative and quantitative importance. Our method was
successful in detecting robust, biologically relevant, as well as statistically relevant, potential biomarkers via a spike-in study, regardless of background subtraction and normalization. Moreover, our method can also be informative to eliminate experimental artifacts or biological variations from individual patient effects. According to Diamandis (2004), SELDI-TOF technology is not capable of detecting any serum component at concentrations of <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. Our confidence bands can easily reveal the quantitatively small differences even though these artifacts may be found to be qualitatively significant markers. Therefore, researchers can remove potential artifacts among the list of significantly detected markers by sorting them by values of corresponding MPCs or upper confidence bounds and by putting high priority in markers with larger values. Moreover, usage of mean spectra in confidence bands effectively normalizes possible biological variations from individual sample effects. Benefits of using average spectra over individual spectra were also well studied in Morris et al. 2005. Based on our findings, we claim that our method is useful for detecting biologically relevant biomarkers while being less prone to false positives.

Our confidence bands were originally designed for two-class problems. In the case of multi-class (>2 classes) problems, our method can be used via various combinations of pair-wise comparisons depending on the purpose of the analyses. For example, suppose that there are three classes of samples: ovarian, breast cancer and normal samples, and suppose that the interests of a researcher are in identifying proteins that distinguish ovarian cancer samples or breast cancer samples from normal samples. One can use the normal samples as a baseline to compare with each of the others with a multiple comparison adjustment.
While the spike-in experiment did not generate a clean dataset 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 unfractonated 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 ~66 500 Da 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 Fig. 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 1X and 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 worthwhile to consider constructing confidence bands for fold-change by taking the log-transformation of the spectra intensities. 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.

ACKNOWLEDGEMENTS

We thank reviewers for their through reviews and helpful comments. The work is supported by Claudia Adams Barr Program in Cancer Research (Y.P., C.L.) and partially supported by the US NIH grants for I2B2 and AIDS (L.J.W.).

Conflict of Interest: none declared.

REFERENCES

Morris,J.S. et al. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics, 21, 1764–1775.