

## *A new method for biomarker detection:*

### **Quantitative assessment for contrasts between groups of subjects based on Mass Spec data**

Yuhyun Park<sup>1</sup>, Sean R. Downing<sup>2</sup>, Cheng Li<sup>1,2</sup>, Phillip W. Kantoff<sup>2</sup>, L.J. Wei<sup>1</sup>

**Keywords:** SELDI, Differentially expressed proteins, Confidence bands, Multiple comparison

## **1 Introduction.**

The recent surge in surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) mass spectrometry (MS) technology has made it possible to explore proteome-wide biomarkers. One of the most frequently asked questions for such proteome-wide studies is how to find a list of biomarkers, in this instance defined as proteins differentially expressed between two groups of samples. Conventional methods first detect the *peaks* in protein spectrum for each sample after calibrations and then align these peaks across the samples. Next they find peaks related to a mass-to-charge ratio ( $m/z$ ) that discriminate groups based on testing of peak intensities. There are two major concerns with this approach. First, for SELDI-TOF MS data, the peak detecting methods are controversial because they are ad-hoc and the results can vary due to user defined parameters. Second, since a large number of proteins, potentially correlated with each other by unknown fashion, are tested simultaneously with a relatively small number of samples, it is expected to have a lot of false positives in detecting *statistically significant* biomarkers if one uses separate statistical tests for each features using traditional  $p$ -value cutoffs of 0.01 or 0.05.

**To tackle these problems, we propose a new and simple algorithm based on permutation method to visualize the possible range of difference in protein abundance between groups with statistical significance while guarding against false positives simultaneously by constructing confidence bands of the contrast between groups. We also define a new concept for *peaks* (biomarkers) based on the proposed confidence band method.** These confidence bands allow investigators to understand both a *qualitative* statistical significance of the difference in expression ( $p$ -value) between two groups while controlling the overall type I error rate and the *quantitative* significance which is possible range of magnitude of differences. This kind of quantitative assessment is quite appealing since SELDI-TOF technology is not capable of detecting any serum component at concentrations of less than  $1 \mu\text{g} / \text{mL}$  and statistically very 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 [2].

We illustrate our new method with well-known SELDI-TOF MS data sets from the latest experiment for ovarian cancer in the Clinical Proteomics Programs Databank [1],[3],[4],[5]. We also performed a spike-in experiment with samples from 91 prostate cancer patients to assess the accuracy of our methods. Our confidence bands methods can be applied to any high-dimensional data for comparison between two groups, such as gene expression datasets.

## **2 Software and files.**

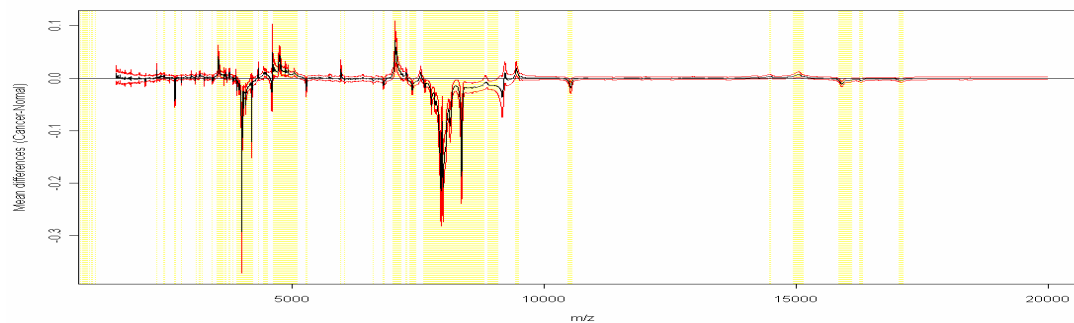
The software in a format of R and C programs is available from the authors. The raw spike-in data will be posted at the DFCI website to provide scientific community to develop and validate algorithms for SELDI MS experiments.

---

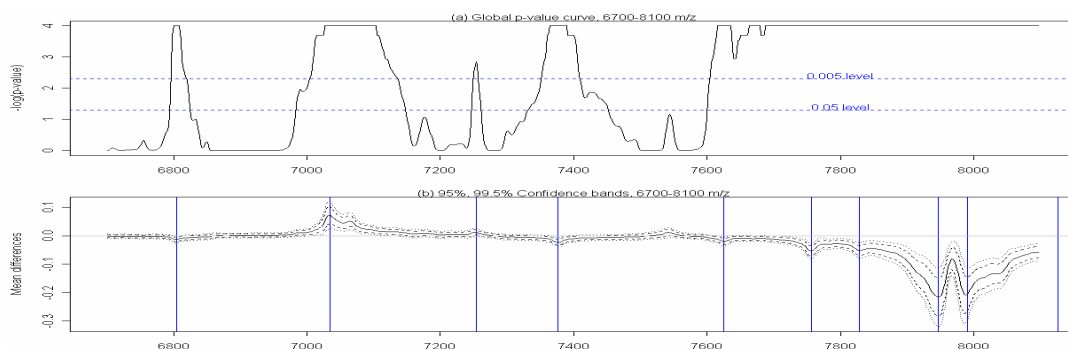
<sup>1</sup> Harvard University, 677 Huntington Ave. Boston, MA 02115, Email: ypark@hsph.harvard.edu

<sup>2</sup> Dana Farber Cancer Institute, 44 Binney Street Boston, MA 02115 Email: Sean\_Downing@dfci.harvard.edu

### 3 Figures and tables.



**Figure 1: 95% proteom-wide confidence bands for Ovarian cancer data (Difference between cancer and normal samples) Red curves are 95% confidence bands, black curve in the middle is the observed difference in protein intensities, and the yellow vertical lines are the significant regions from the testing.**



**Figure 2. Zoom-in figure (6800-8000 m/z) for global p-value curve (a) and confidence bands (b). The blue vertical lines in (b) are indicating the positions of final detected markers significant at the level 0.05 in this region. It is important to note that the confidence bands give more information than global p-value curve (a conventional method). Whereas the global p-value curve only test if two groups are different to each other with level of  $\alpha$ , the confidence bands actually show the magnitude of difference with the corresponding  $1 - \alpha$  confidence. For example, the peak at 6800 m/z was found to be very significant in p-value curves and as significant as the peaks around 7990 m/z ( $p\text{-value} \leq 0.0001$ ). However, the confidence bands indicate that the relative abundance of proteins in the cancer patients at 6800 m/z may not be large enough to be as significant as the relative abundance of proteins in the normal patients at 7990 m/z. Moreover p-value curves alone do not give information about the precise position of the actual biomarker.**

### 4 References

- [1] Baggerly KA, Morris JS & Coombes KR 2004 Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments. *Bioinformatics* 20 777–785
- [2] 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.
- [3] <http://clinicalproteomics.steem.com>
- [4] Petricoin EF, Ardekani AM, Hitt BA, Leviine PH, Fusaro VA, Steinbuerg SM, Mills GB, Simone C, Fishman DA, Kohn EC & Liotta LA 2002 *Lancet*, 359, 572-577.
- [5] Sorace J & Zhan M 2003 A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 4