Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 11

Issues of Processing and Multiple Testing of SELDI-TOF MS Proteomic Data

Merrill D. Birkner[*] Alan E. Hubbard[†] Mark J. van der Laan[‡]
Christine F. Skibola[**] Christine M. Hegedus[††] Martyn T. Smith[‡‡]

[*]Division of Biostatistics, School of Public Health, University of California, Berkeley, mbirkner@gene.com

[†]Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

[‡]Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

[**]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, chrisfs@berkeley.edu

[††]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, chegedus@berkeley.edu

[‡‡]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, martynts@berkeley.edu

Issues of Processing and Multiple Testing of SELDI-TOF MS Proteomic Data*

Merrill D. Birkner, Alan E. Hubbard, Mark J. van der Laan, Christine F. Skibola, Christine M. Hegedus, and Martyn T. Smith

Abstract

A new data filtering method for SELDI-TOF MS proteomic spectra data is described. We examined technical repeats (2 per subject) of intensity versus m/z (mass/charge) of bone marrow cell lysate for two groups of childhood leukemia patients: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). As others have noted, the type of data processing as well as experimental variability can have a disproportionate impact on the list of "interesting" proteins (see Baggerly et al. (2004)). We propose a list of processing and multiple testing techniques to correct for 1) background drift; 2) filtering using smooth regression and cross-validated bandwidth selection; 3) peak finding; and 4) methods to correct for multiple testing (van der Laan et al. (2005)). The result is a list of proteins (indexed by m/z) where average expression is significantly different among disease (or treatment, etc.) groups. The procedures are intended to provide a sensible and statistically driven algorithm, which we argue provides a list of proteins that have a significant difference in expression. Given no sources of unmeasured bias (such as confounding of experimental conditions with disease status), proteins found to be statistically significant using this technique have a low probability of being false positives.

KEYWORDS: proteomics, mass-spectrometry, multiple testing, preprocessing, leukemia, tail probability

---

*Merrill D. Birkner is a trainee of the U.C. Berkeley Superfund Basic Research program. We thank Professor Patricia Buffler, Principal Investigator of the Northern California Childhood Leukemia Study, for access to the biological samples. This study was supported by NIH grant P42ES04705.

1 Introduction

This article presents the application of a preprocessing and multiple testing technique on a proteomic dataset, with the aim of determining differentially expressed proteins between two leukemia subtypes. This proteomic data is not a straightforward (exact) measurement of underlying protein abundances and is victim to sources of experimental variability (for instance, see Baggerly et al. (2004)), which cause problems, especially when one is interested in finding proteins that are related to the disease of interest. As vendors have done (e.g., Ciphergen Biosystems), we provide a series of processing steps that are meant to minimize the sources of nuisance variation (e.g. baseline drift) and therefore facilitate the process of finding related proteins. We rely on having technical replicate measures of the samples on a biologic unit (in our case, a child), which provide a convenient motivation for choosing optimal processing parameters using a statistical criteria. These preprocessing techniques are important, especially when one is interested in using this data for subsequent statistical analyses, in this case multiple testing procedures. Although not discussed in this paper, optimal designs should insure that the data is not confounded by experimental variation, which is most efficiently done by design (for instance, making sure that either experimental conditions are homogenous for all samples or that at least samples are evenly distributed across experimental conditions with respect to the factors of interest).

In this article, we discuss two classes of data processing/filtering problems that are typical of proteomic data: pre-processing and selection of proteins of interest by multiple testing. In Section 2 we first discuss our processing algorithm and give some arguments why it should be relatively robust (provide reproducible results) and also suggests augmentations that will make it more flexible. We then follow with a discussion of multiple testing in general and a newly introduced method that provides accurate and yet not overly conservative control for experimentwise (Type I) error rates. We conclude the paper with the analysis of the childhood leukemia data and a short discussion.

1.1 Data Application

The methods described in this article can be applied to a variety of protein based datasets. In this article, the study is based on the analysis of array-based proteomic data obtained by surface enhanced laser desorption ionization mass spectrometry (SELDI-TOF MS) of childhood leukemia sam-

ples. Two sets of samples of bone marrow cell lysate from children with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) were used, as originally described in Hegedus et al. (2005). Leukemia is a group of cancers characterized by the uncontrolled proliferation of blood precursor cells of the myeloid or lymphoid lineage. ALL and AML are the most common leukemias among children representing approximately 31 percent of cancers in children under 14 years of age (Smith et al., 2005). These leukemias are further classified by immunophenotypic and cytogenetic characteristics. Cases with high hyperdiploidy (greater than 50 chromosomes) and those harboring t(12;21) constitute a majority of childhood ALL (Greaves, 2002). With the exception of a few known risk factors such as benzene and radiation exposure, little is known about the causes of leukemia. Researchers are interested in determining differences in protein expression between leukemic subtypes in order to develop a biomarker that can distinguish subtypes and investigate possible mechanisms of leukemogenesis. Previous microarray and proteome studies have successfully identified such markers (Golub et al., 1999; Valk et al., 2004; Ohmine et al., 2001; Kohlmann et al., 2004; Yeoh et al., 2002; Ross et al., 2003, 2004; Cui et al., 2004, 2005; Hegedus et al., 2005; Issaq et al., 2002). However, few studies have used SELDI-TOF MS for proteomic analysis of bone marrow from childhood leukemia cases. Here we have used raw data from SELDI-TOF MS analysis of bone marrow described in Hegedus et al. (2005). The bone marrow cell lysate from ALL and AML cases was analyzed to generate data consisting of mass-to-charge ratios (m/z) representing individual proteins and their corresponding intensities, which represent the relative abundance.

2 Data Pre-Processing

In this section, we first give the specific structure of the leukemia, SELDI-TOF MS spectral proteomic data for our childhood leukemia subjects. We then discuss how this structure can be utilized for optimally smoothing the intensity vs. m/z data to derive summary intensity measurements for each child for a common set of m/z values.

DOI: 10.2202/1544-6115.1198

2.1 Data Structure

The dataset consists of two replicates each of AML ($n = 7$) and ALL ($n = 13$). We are interested in obtaining an intensity value for a specific number of unique m/z values, averaged over the replicates.

2.2 Background Drift Correction

For this type of proteomic data, there is often a drift in the apparent background values in raw m/z-intensity data (see top row of Figure 1 as an example). Optimally, we would like the value for all non-peak m/z values to be at 0. In addition, a procedure should take advantage of the smoothness (in our case, the background declines in a linear fashion). That is, a reasonable low-dimensional model can be fit to this minimum. Our solution is to use quantile regression, which models the trend in the p^{th} quantile of an outcome versus a predictor variable(s) (Koenker and Bassett, 1978): $F^{-1}_{Y|X}(p \mid X = x) = g(x \mid \beta)$, where X is the explanatory variable, p is the percentile $\in (0, 1)$, $F_{Y|X}(y \mid X = x) \equiv P(Y \leq y \mid X = x)$ and $g(x \mid \beta)$ is some function of x and coefficients, β, for instance, $g(x \mid \beta) = \beta_0 + \beta_1 x$, X is the explanatory variable (in our case, m/z) and Y the outcome (intensity). One can not model the minimum, so we have chosen a very small quantile ($p = 0.02$) and we have modelled the background as a linear decline, but in practice models of arbitrary complexity can be applied, e.g., a high order polynomial basis. The background corrected intensities are simply $Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$, where Y is the original intensity and X is the corresponding m/z ratio. The results are shown on the second line of Figure 1. Because this procedure can borrow information from adjacent m/z values when determining the baseline correction at a particular m/z, and because the baseline drift is typically quite smooth, this procedure should in theory provide a relatively robust method for baseline-correction.

2.3 Smoothed Intensity

Signal filtering procedures are used to reduce the noise from the signal in the observed profile of intensity versus m/z values. This process therefore produces a "true" profile of intensity values with a minimum of noise. Our method of filtering takes advantage of the replicate nature of our design. Therefore, each biologic replicate is analyzed twice resulting in two protein

spectra per child (two technical replicates). Our method emphasizes reproducible peaks as opposed to features that are unique to only one sample. We use an estimate of the underlying true (noiseless) m/z curve on one sample to predict intensities on the other sample, using a rectangular kernel smoother (Härdle, 1990). These smoothers which estimate the curve at a particular point, can be thought of as a simple, local weighted average of the intensities in a small neighborhood of m/z ratios defined by the width of the neighborhood, referred to as the bandwidth. The nature of the weight is referred to as a kernel. In the protein application the function is not smooth, but consist of a set of unpredictable peaks surrounded by flat areas with nearly no signal. Thus, a natural choice of a kernel is a simple, uniform weight over a small box or rectangular kernel, otherwise known as the uniform kernel. The width of the box is the bandwidth, and presumably the width chosen will represent the measurement error on the m/z axis. Note, we used the function **ksmooth()** in R (using a box kernel) to estimate the kernel smooth.

The next problem involves choosing the optimal bandwidth to use in the smoothing algorithm. We will utilize the fact that we have two replicates per sample. We invoke recently developed theory for the optimality of cross-validation for choosing the "best" estimator from a set of candidate estimators (van der Laan and Dudoit, 2003). The method of cross-validation is based on building an algorithm, or "training an algorithm" on one sample of data and subsequently testing the trained algorithm on an independent set of data (or in this case, the other sample of data). The kernel bandwidth is chosen by using a simple cross-validation technique on the replicates that attempts to minimize the mean-squared error of prediction (MSPE). Specifically, the smoothing algorithm is applied, using one of the candidate bandwidth values, on one of the replicates of a biological sample (subject). Once this smoothing algorithm is applied we predict the intensities of its matched replicate. We then reverse the roles of the two replicates and train, or apply, the smoothing algorithm on the second replicate and test the performance on the first replicate. Again, this is performed by predicting the intensities on the first replicate. Each time the smoothing algorithm is applied, or "trained", on the other replicate the mean squared error (MSPE) is recorded for each bandwidth. We therefore are interested in determining how well the algorithm predicts the values of the other replicate. This is then repeated over all samples/replicates. The average MSPE is calculated for each bandwidth and the bandwidth with the smallest average MSPE is chosen.

DOI: 10.2202/1544-6115.1198

2.3.1 Bandwidth

We developed a flexible procedure which considered a cross-validation based model selection routine that allows the bandwidth used from smoothing intensities to vary by m/z value, based on the fact that the error in m/z might not be constant, but itself have some drift. Although more complicated models can be used, we choose to examine bandwidths that changed linearly with m/z value (Note: $x = m/z$):

$$h = \beta_0^* + \beta_1^* x$$

where h is the bandwidth, and (β_0^* and β_1^*) define the model. Both β_0^* and β_1^* are now chosen by cross-validation over a grid of possible values that include:

$\beta_0^* = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$

$\beta_1^* = (0, 0.00001, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.3)$.

For each combination of β_0^* and β_1^* the smoothing algorithm is trained on one replicate of a biological sample (subject) and used to predict the intensities of its matched replicate, just as done above (the constant bandwidth choice is simply fixing $\beta_1^* = 0$ and selecting over the β_0^*). As described in the previous cross-validation section, we reverse the roles of the two replicates and train the smoothing algorithm on the second replicate and test it on the first replicate. The MSPE is recorded each time the algorithm is trained on the second replicate for each combination of β_0^* and β_1^*. This is then repeated over all samples/replicates. The average MSPE is calculated for each combination of β_0^* and β_1^*. The (β_0^*, β_1^*) set which result in the smallest MSPE is chosen, in our case ($\beta_0^* = 2, \beta_1^* = 0.009$).

The original m/z values are then averaged within windows corresponding to the respective bandwidth. This is done in order to produce a smaller set of unique m/z values since many of the original values are in groups which are in close proximity to each other. Although our discussion focuses on a simple linear model (linear in m/z), competing models of greater complexity can be used (e.g., higher order polynomials) and the same technique can be used to choose the respective parameters of the model using a similar cross-validation technique.

A byproduct of this bandwidth selection method, combined with quantile regression (where a very small percentile, such as 0.02, is chosen) is that the post-smoothed intensities will robustly have a zero baseline. This is because 1) quantile regression for a relatively small quantile essentially models the minimum, 2) after correction (subtracting off the quantile regression trend) the non-peak m/z values will vary randomly around zero and these non-peak areas, which make up most of the function of intensity versus m/z values play an important role in bandwidth selection 3) because the quantile regression trends are fit independently for each technical replicate, the pre-smoothed function will vary (relatively) randomly around 0, and thus 4) a bandwidth that results in predicted values of 0 in non-peak areas will be favored. Thus, an attractive byproduct of this method of using quantile regression, rectangular kernel smoothing with cross-validated bandwidth selection (where each technical replicate serves as a validation sample) is a well-defined, zero, background level.

2.4 Defining Protein Expression Peaks

At this point the data have the structure of two intensity vs. m/z functions for each biological sample. However, to test the association of some characteristic of the patient (in our case, subtype of leukemia) with protein expression, we need to have a common set of m/z values and associated expressions for each biological sample. We propose deriving a common m/z set as follows: 1) for each technical replicate, the entire list of unique (above background) m/z values is ranked, 2) any m/z values within the chosen bandwidth (either constant or varies with m/z as discussed above) is lumped together and labeled now with a new m/z value, which is simply the average of all those m/z values within the new m/z group, 3) after re-labeling all the m/z values for each of the replicates, we derive a single expression vs. m/z profile for each biological sample by averaging the respective replicates at the new m/z values. This resulted in our case (for the leukemia data) in a set of unique values contains 204 and 100 m/z ratios, using the constant and variable bandwidth methods, respectively.

DOI: 10.2202/1544-6115.1198

3 Multiple Testing

The final step after creating a data matrix that consists of processed protein intensities for each independent biological sample (the columns) and each unique m/z value, is to select the m/z values that are significantly associated with some phenotypic trait. In our case, we are interested in associating the intensities of each m/z value with the type of leukemia (ALL vs. AML). We want to choose those proteins for which we have relatively high confidence that they are truly different (i.e., different in mean intensities) between the two groups, using a multiple testing procedure (MTP). In general, multiple testing procedures consist of several steps. The first step consists of choosing an appropriate parameter of interest (e.g., mean difference in intensities in the two groups). Secondly, one must specify the null hypothesis that relates this parameter to the question of interest (e.g., the mean difference is 0). The test statistic for which the null distribution is known, (at least asymptotically (e.g., the two-sample t-statistic) is then specified. In our specific case, the test is performed for each row of the data matrix (m/z value). Finally, one decides upon an appropriate experimentwise error rate to control (e.g., the number of false positives or Type I errors) and in turn chooses a method to control this rate. The parameters of interest, resulting null hypotheses, test statistic and Type I error rate are choices for the investigator. Once these are chosen, one can debate the merits of various MTPs, specifically, which provide accurate Type I error control under assumptions the investigator is willing to make and among these, which have the greatest power.

There are several Type-I error rates which may be of interest to the investigator: 1) The family wise error rate (FWER), which controls the probability of rejecting more than one false positive; 2) generalized family wise error rate (gFWER), which controls the probability of rejecting more than a user defined number, k, false positives; 3) tail probability of the proportion of false positives (TPPFP), which controls the proportion of false positives to total rejections at a user defined value q, $q \in (0, 1)$; 4) False Discovery Rate (FDR), or controlling the mean of the proportion of false positives to total rejections. FWER is a conservative error rate, and often too conservative for most biological applications; thus less stringent methods which will allow some false positives, but at a given number or proportion, may be more conducive to scientific application. A method controlling the TPPFP is attractive especially since it deals with the proportion of false positives to total rejections, instead of an absolute number of false rejections. It will allow

some false positives as long as the probability of the proportion of false positives to total rejections is less than or equal to q, with probability 0.05. Also, as compared to the FDR methods, TPPFP controls the actual proportion of false positives to total rejections, whereas the FDR controls that proportion on average, therefore making a method controlling the TPPFP favorable in some settings, particularly since the expected number of false positives can be highly variable (e.g. when the test statistics are highly dependent).

This article presents a data application of the E-Bayes/Bootstrap TPPFP approach, outlined in detail in van der Laan et al. (2005). This approach controls the TPPFP at a user defined level q, with probability $1 - \alpha$. van der Laan et al. (2005) outlines this procedure and provides finite and asymptotic rationale of the proposed procedure, as well as simulations showing the method is more powerful and less conservative in the finite setting, relative to competing TPPFP procedures. Since this method is less conservative, we are apt to properly reject more null hypotheses at a nominal α level as compared to other more conservative methods. We refer the reader to van der Laan et al. (2005) for a detailed description of simulation results which illustrate the non-conservative nature of this method as compared to a variety of existing methods. In this article, this technique will be applied to two separate datasets, which are described in detail in Section 4.

3.1 TPPFP

The E-Bayes/Bootstrap TPPFP method aims to control the proportion of false positives to total rejections at a user defined level q, with probability $1 - \alpha$. As discussed in van der Laan et al. (2005), the recently developed, resampling based E-Bayes/Bootstrap TPPFP approach has proven to be less conservative and thus more powerful, as compared to other methods, such as the augmentation approach outlined in van der Laan et al. (2004b) and the Lehmann and Romano (2003) TPPFP techniques. The procedure involves 1) specifying a conditional distribution for a guessed set of true nulls, given the data, which asymptotically is degenerate at the true set of nulls; and 2) specifying a generally valid null distribution for the vector of test-statistics proposed in Pollard and van der Laan (2003), and generalized in subsequent articles Dudoit et al. (2004), van der Laan et al. (2004a), and van der Laan et al. (2004b). The finite and asymptotic results are outlined in the van der Laan et al. (2005) as well as relevant simulations, which illustrate comparisons of the power and error rate of this procedure

DOI: 10.2202/1544-6115.1198

in various situations. The statistical details of this technique are described below.

Let $X_1, ...X_n$ be i.i.d. observations and $X \sim P$. We will define $H_{0j}, j = 1, ..., m$ as the m null hypotheses about P, $H_{0j} : P \in M_j$. We will define $T_n = (T_n(1), \ldots, T_n(m))$ as the test-statistics corresponding to null hypotheses $H_1, ..., H_m$ for each m/z value, in the respective datasets, with m corresponding to the number of tests performed. This vector of test statistics has an unknown distribution Q_n. Given a user supplied q and $\alpha \in (0, 1)$, the procedure selects a common cut-off c_n such that,

$$Pr\left(\frac{\sum_{j=1}^{m} I(T_n(j) > c_n, j \in \mathcal{S}_0)}{\sum_{j=1}^{m} I(T_n(j) > c_n)} > q\right) \leq \alpha,$$

where $j \in \mathcal{S}_0$ indicates a null hypothesis, and $T_n(j) > c_n$ indicates a rejection of H_{0j}.

3.1.1 E-Bayes/Bootstrap TPPFP Approach

Our method for choosing c involves controlling the tail probability of a random variable $\tilde{r}_n(c)$ defined as:

$$\tilde{r}_n(c) = \frac{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n})}{\sum_j I(\tilde{T}_n(j) > c, j \in \mathcal{S}_{0n}) + \sum_j I(T_n(j) > c, j \notin \mathcal{S}_{0n})}.$$

$\tilde{r}_n(c)$ represents a guessed proportion of false positives among rejections, defined by drawing a random set $\mathcal{S}_{0n}$ (a guessed set of true null hypotheses $\mathcal{S}_0$) and a draw $\tilde{T}_n$ from a null distribution for the test-statistic vector. We want $\tilde{r}_n(c)$ to dominate in distribution the true proportion of false positives: $\frac{\sum I(T_n(j)>c, j \in S_0)}{\sum I(T_n(j)>c)}$. Clearly the random variable $\tilde{r}_n(c)$ is defined by the proposed definition of $\tilde{T}_n(j)$ and $\mathcal{S}_{0n}$.

Derivation of $\tilde{T}_n(j)$:

In order to estimate $\tilde{T}_n$, we bootstrap the data $(X_1^{\#}, \ldots, X_n^{\#})$ B^* times (e.g. $B^* = 10,000$). For each iteration, we recalculate the m test-statistics. This $m \times B^*$ matrix, $\tilde{T}_n^*$, represents a draw from the test-statistic vector under the empirical distribution P_n. We then calculate the row-specific means and center the $\tilde{T}_n^*$ matrix at its null value. Each column of this matrix specifies

a draw of $\tilde{T}_n = (\tilde{T}_n(j) : j = 1, \ldots, m)$.

Derivation of $B_n(j) = I(j \in S_{0n})$:

We will define the distribution of our guessed set of nulls $\mathcal{S}_{0n}$, and describe how this random set is drawn. This random set is defined by drawing a null or alternative status for each of the test statistics. The working model for defining the distribution of the guessed set $\tilde{\mathcal{S}}_{0n}$ will assume $T_n(j) \sim p_0 f_0 + (1 - p_0) f_1$, a mixture of a null density f_0 and alternative density f_1. Let $B(j)$ represent the underlying Bernoulli random variable, such that $f_0 \sim (T_n(j)|B(j) = 0)$, is the density of $T_n(j)$ if $H_0(j)$ is true, and $f_1 \sim (T_n(j)|B(j) = 1)$ is the density of $T_n(j)$ if $H_0(j)$ is false.

Under this working model, the posterior probability defined as the probability that $T_n(j)$ came from a true H_{0j}, given its observed value $T_n(j)$, can now be calculated:

$$P(B(j) = 0|T_n(j)) = p_0 \frac{f_0(T_n(j))}{f(T_n(j))}.$$

We will use this posterior probability as the Bernoulli probability on H_{0j} being true, given the test statistic, where we have to specify or estimate p_0, f_0 and f. Since f_0 plays the role of the density of test-statistics under the null hypothesis, in some situations f_0 is simply known, e.g., $f_0 \sim N(0, 1)$. However, in cases where the marginal distribution of $T_n(j)$ is not known if H_{0j} is true, one can use a kernel density (**density()** in R with a given kernel and bandwidth) on the mean centered elements in the matrix representing B draws of $\tilde{T}_n$. The elements from this matrix are pooled into a vector of length $m * B^*$ in the kernel density function. In order to estimate the density f, we can again apply a kernel smoother on the bootstrapped test statistics, before they are mean centered. Again, the elements of the matrix are pooled into a vector of length $m * B^*$ in the kernel density function.

Finally, p_0 represents the proportion of nulls $\mid \mathcal{S}_0 \mid /m$ and typically the user might employ a conservative p_0^* for this true proportion of nulls. The most conservative prior, $p_0^* = 1$, will be used throughout this paper. Now, given T_n, we can define the random set

$$\mathcal{S}_{0n} = \{j : C(j) = 1\}, C(j) \sim Bernoulli\left(\min\left(1, p_0^* \frac{f_0(T_n(j))}{f(T_n(j))}\right)\right).$$

Given the data $X_1, \ldots, X_n$ (i.e., P_n), $\mathcal{S}_{0n}$ and $\tilde{T}_n$ are drawn independently.

DOI: 10.2202/1544-6115.1198

We will now draw $(\mathcal{S}_{0n}, (\tilde{T}_n(j))$ B^* times, and each time calculate the corresponding realization of $\tilde{r}_n(c)$, where T_n is fixed at the true original test statistics. This provides us with a sample of B^* realizations of $(\tilde{r}_n^b(c) : c \geq 0)$, $b = 1, \ldots, B^*$, conditional on the data P_n (and thus, conditional on T_n as well).

The cut-off c is set so that the tail probability, at a user supplied level q, of the random variable, $\tilde{r}_n(c)$, equals α. To do so, we will then choose c such that average over B^* draws of both $\tilde{T}_n(j)$ and $\mathcal{S}_{0n}(j)$ equals α.

Specifically, we set

$$c_n = \inf \left\{ c : \frac{1}{B^*} \sum_{b=1}^{B^*} I(\tilde{r}_n^b(c) > q) \leq \alpha \right\}.$$

3.1.2 Augmentation Technique

An augmentation TPPFP procedure was also applied to the protein dataset (van der Laan et al., 2004b). This augmentation corresponds to merely adding the $[\frac{q}{1-q} r_0]$ most significant rejections to the rejection set of the FWER method, where r_0 is the set of initial rejections from the FWER procedure. The FWER procedure is based on the resampling-based null distribution $\tilde{T}_n$ described above. The maximum values over the columns of this matrix are used to compare the test-statistics. The resulting adjusted p-values corresponding to these t-statistics control FWER and the previously described augmentation controls TPPFP. Further detail of this method can be found in Pollard and van der Laan (2003).

3.2 Adjusted p-values

A convenient way to display the results of a MTP is by reporting the adjusted p-values in a ordered list corresponding to their relative significance. Both the E-Bayes/Bootstrap TPPFP and Augmentation techniques provide adjusted p-values as a summary measure for each test. Adjusted p-values provide a measure of the probability of making a Type-I error taking into account that one made multiple tests. The j^{th} adjusted p-value can be interpreted as the nominal alpha level one would use to just reject the j^{th} specific test-statistic. Displaying these adjusted p-values provide a summary measure of the tests and therefore makes them easier to compare.

4 Existing Preprocessing and Testing Methods

4.1 Ciphergen Biomarker Wizard®

One of the existing methods which is used to preprocess and test for differentially expressed intensity peaks with mass spectrometry data is the Ciphergen's proprietary Biomarker Wizard® software (Ciphergen Biosystems, Fremont, CA, USA). This software analyzes and compares mass spectra to determine potential biomarkers. Though the results of these methods are not directly compared in this paper, we briefly outline the competing method in this section.

Ciphergen's baseline adjustment procedure is based on a segmented convex hull algorithm, which eliminates the baseline error resulting in a spectrum without noise. The user also has options in Ciphergen's software to edit the fitting width and therefore change the shape of the baseline. The user can also choose to use a smoothing function before fitting the baseline. The filtering procedure provided by the Ciphergen software allows the user to choose the width of the peak over which to average to eliminate noise. This is user driven and statistically chosen, as with our proposed method.

Ciphergen's spectral alignment aligns sections of the spectra together. It initially selects a reference spectrum and then filters the peaks to calibrate to this reference spectrum. It uses either manual or automatic peak detection. Manual peak detection corresponds to allowing the user manually choose which peaks visually look important or of interest. If the automatic option is chosen, the user can also manually add additional peaks to the selected set of peaks. Clustering is then performed by clustering areas which are 0.03 percent of the m/z ratio (this percentage is user defined and the default is 0.03 percent). The peak width is also taken into consideration when averaging these regions. Finally, the Ciphergen software records Mann-Whitney or Kruskal-Wallis tests for each cluster. These are the only statistical tests which are reported with a respective adjusted p-value.

The Biomarker Wizard® software may not always be an optimal technique when applying it to spectral samples. As noticed in the Hegedus et al. (2005) paper, the packaged technique results in too many peaks which are claimed to be significant. We are interested in narrowing down the group of significant peaks to determine the groups that are significant when adjusting

DOI: 10.2202/1544-6115.1198

for multiple comparisons, therefore removing possibilities of false positives with a given probability. In addition, biologists are interested in a smaller set of peaks which encompass the group that would be beneficial to further test and identify. As later stated, the two procedures both detect similar peaks but the new method allows the user to obtain a targeted group which is significant when taking into account the multiplicity of tests performed. We used the Biomarker Wizard® software to confirm the significant peaks. In addition, the results were compared to peaks found in the Hegedus et al. (2005) paper, which used a similar dataset but preprocessed and analyzed the data with the Biomarker Wizard® software.

4.2 Additional Existing Methods

In addition to Ciphergen's method, several other existing methods are used to analyze SELDI-TOF proteomic data. We will briefly mention some of the statistical methods that are currently being used to analyze SELDI-TOF spectra. The first method focuses on the alignment of spectra. Jeffries (2005) developed two algorithms that are used to align SELDI-TOF data among samples. One algorithm is built to be used with Ciphergen SELDI data, whereas the other algorithm can be generalized to other types of data. The algorithms standardize various measurements taken from an instrument over time points and if the spectra are not consistent. This process determines the appropriate spectral alignment window to use to determine if the peaks reflect the same proteins. This realignment was done before the data preprocessing step. This method is compared to using the 0.3 percent alignment window of Biomarker Wizard, and therefore is an alternative method of alignment. Other methods of spectral preprocessing are popular throughout the literature. One such method is described in Liu and Li (2005). In this article, the authors have devised a process to find cancer biomarkers by using decision lists, which are forms of decision trees. The authors create decision lists to discover biomarkers, which are based on classification by mass spectrometry spectrum. The advantage of this procedure is that it can be used directly in clinical screening. In addition, this method had proven to perform better as compared to current methods such as support vector machines or decision tree methods. Other preprocessing techniques include a method proposed by Ressom et al. (2005). This technique processes mass spectral data and subsequently use a machine learning method (support vector machines and particle swarm optimization) to select optimal m/z values and subsequently

align the peaks. The combination of these methods is used for biomarker selection. Finally, Coombes et al. (2005) developed an algorithm to process mass spectrometry profiles. The spectra are denoised with the undecimated discrete wavelet transform and then evaluated to assess their consistency and reproducibility properties. This method is used to isolate the noise of the spectrum. A baseline correction and normalization are then applied to the spectrum. The peaks that are within 0.3 percent in relative mass are then combined. The authors claim that this method finds more peaks (and more reproducible peaks) as compared to the Ciphergen software.

5 Data Applications

In the following section, we will present the application of the E-Bayes/Bootstrap TPPFP approach, as well as the van der Laan et al. (2004b) Augmentation technique. Firstly, we will describe the results of the multiple testing application to the dataset preprocessed with the constant bandwidth method. This will be followed by the variable bandwidth results.

5.1 Application to AML/ALL data: Constant Bandwidth Preprocessing

The following results and discussion correspond to a case where we set $\beta_1^* = 0$ and therefore applied the preprocessing cross-validation technique to the β_0^* values. This was performed as a naive manner of choosing a constant bandwidth, though biologically we should be using the variable bandwidth method (the results are described in the following sections). The difference in the mean intensities of the AML versus the ALL samples at each of the 204 m/z ratios is tested. The test-statistics will be defined as: $T_n(j) = \sqrt{n}\frac{(\hat{\mu}^{AML}(j)-\hat{\mu}^{ALL}(j))}{\hat{\sigma}_{AML/ALL}(j)}, j = 1, ..., 204$, where $\hat{\sigma}^2_{AML/ALL}$ is the pooled variance of the two samples. The null hypothesis is that $(\hat{\mu}_{AML} - \hat{\mu}_{ALL}) = 0$ and the alternative hypothesis is that $(\hat{\mu}_{AML} - \hat{\mu}_{ALL}) \neq 0$. We would like to note that a non-parametric test might have been more appropriate in this situation on account of the non-normal data and small sample size. The E-Bayes/Bootstrap TPPFP procedure is used to determine those m/z ratios that have significantly different mean intensities between AML and ALL, while controlling the proportion of false positives to total rejections at a level $q = 0.1$, with probability 0.95 ($\alpha = 0.05$).

DOI: 10.2202/1544-6115.1198

Table 1: Constant Bandwidth: Adjusted p-values; Top 10 m/z Ratios:

m/z	E-Bayes/Bootstrap TPPFP ($q = 0.1$)	Augmentation ($q = 0.1$)
4968.104	0.039	0.051
3333.169	0.043	0.0595
4941.165	0.0491	0.1515
3201.327	0.215	0.352
8457.161	0.3197	0.437
3281.276	0.3404	0.4535
3908.681	0.3586	0.460
2908.314	0.3605	0.4615
10527.394	0.3897	0.467
10509.961	0.3999	0.467

There are 20 m/z values out of the 204 with an unadjusted p-value less than $\alpha = 0.05$. With the TPPFP augmentation method no m/z are rejected at an $\alpha = 0.05$ and two are rejected at an $\alpha = 0.1$ level. The E-Bayes/Bootstrap TPPFP rejects 3 m/z ratios at an $\alpha = 0.05$ and also three are rejected at an $\alpha = 0.1$ level (reference Table 1). The proprietary Biomarker Wizard® software (Ciphergen Biosystems, Fremont, CA, USA) also found these masses to be significant, based on another algorithm, not accounting for multiple testing. These were found through the software's autodetection method. The autodetection method marked a peak as significant if it had a signal to noise ratio greater than 2, and was present in at least 25 percent of the samples. These results show that in this situation the E-Bayes/Bootstrap TPPFP method is less conservative as compared to the Augmentation technique.

The mass to charge ratios have yet to be identified as unique proteins. However, researchers plan to follow this analysis and identify the most significant mass to charge ratios by purification and MS/MS.

Table 2: Variable Bandwidth: Adjusted p-values; Top 6 m/z Ratios:

m/z	E-Bayes/Bootstrap TPPFP ($q = 0.1$)	Augmentation ($q = 0.1$)
4967.375	<0.0001	0.001
3336.293	0.051	0.089
2908.006	0.092	0.122
3201.008	0.156	0.291
5174.152	0.171	0.312
9956.193	0.238	0.340

5.2 Application to AML/ALL data: Variable Bandwidth Preprocessing

The variable bandwidth preprocessing and multiple testing procedures were also applied to the same AML/ALL dataset used with the constant bandwidth method. The test statistics are created in the same manner, with the only difference being the preprocessing steps. In total, there are 109 m/z ratios which are tested between the AML and ALL samples.

There are 9 m/z values out of the 109 with an unadjusted p-value less than $\alpha = 0.05$. With the TPPFP augmentation method one m/z is rejected at an $\alpha = 0.05$ and two are rejected at an $\alpha = 0.1$ level. The E-Bayes/Bootstrap TPPFP rejects one m/z ratios at an $\alpha = 0.05$ and also three are rejected at an $\alpha = 0.1$ level. The results are displayed in Table 2. Similarly with the previous example, the m/z ratios which are found to be significant with this procedure marked as "differently expressed" using the Biomarker Wizard® software, though this software does not adjust for the multiplicity of the tests performed. (Note that the variable bandwidth method as compared to the constant bandwidth method is more consistent in finding peaks similar to those peaks found using Biomarker Wizard® software (Ciphergen Biosystems, Fremont, CA, USA)).

6 Discussion

Unless one knows the true underlying data-generating mechanism for their particular technology and experimental design, it is hard to argue that one

DOI: 10.2202/1544-6115.1198

set of processing steps yields universally superior results relative to a competitor. However, we have proposed a series of processing steps and multiple testing procedures that are flexible, take advantage of technical replicates and have some optimal properties (e.g., cross-validation for bandwidth selection and the empirical Bayes approach for controlling TPPFP). In addition, TPPFP is an appropriate Type-I error rate to control in many biological applications. This error rate is less conservative than the family-wise error rate and allows the user to define the upper limit on the proportion of false positives to total rejections. The application of the E-Bayes/Bootstrap TPPFP approach seemed to be an appropriate multiple testing method since it was less conservative as compared to the augmentation technique, by rejecting more m/z values. We suggest that the applied example as well as the simulations presented in van der Laan et al. (2005) demonstrate that the E-Bayes/Bootstrap TPPFP approach is a more powerful technique to control the proportion of false positives to total rejections at a given level q, as compared to various other methods controlling the TPPFP. Finally, the significant m/z values found with this analysis were also seen as significant peaks using the Biomarker Wizard® software, though the latter procedure does not take into account the multiplicity of the tests being performed. This technique therefore is a flexible method to preprocess mass spectrometry data as well as provide an appropriate testing method which will take into account the multiplicity of m/z levels being tested.

References

K.A. Baggerly, J.S. Morris, and K.R. Coombes. Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Datasets from Different Experiments. *Bioinformatics*, 22;20(5):777–785, 2004.

K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M. C. Hung, and H. M. Kuerer. Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform. *Proteomics*, 5(16):4107–4117, 2005.

J.W. Cui, J. Wang, K. He, B.F. Jin, H.X. Wang, W. Li, L.H. Kang, M.R. Hu, H.Y. Li, M. Yu, B.F. Shen, G.J. Wang, , and X.M. Zhang. Proteomic Analysis of Human Acute Leukemia Cells: Insight into their Classification. *Clinical Cancer Research*, 10(20):6887–6896, 2004.

J.W. Cui, J. Wang, K. He, B.F. Jin, H.X. Wang, W. Li, L.H. Kang, M.R. Hu, H.Y. Li, M. Yu, B.F. Shen, G.J. Wang, and X.M. Zhang. Two-dimensional Electrophoresis Protein Profiling as an Analytical Tool for Human Acute Leukemia Classification. *Electrophoresis*, 26(1):268–279, 2005.

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 386(5439): 531–537, 1999.

M. Greaves. Childhood Leukaemia. *British Medical Journal*, 324(7332):283–287, 2002.

W. Härdle. *Applied Nonparametric Regression*. Economic Series Monographs, No. 19. Cambridge University Press, Cambridge, U.K., 1990.

DOI: 10.2202/1544-6115.1198

C.M. Hegedus, C.F. Skibola, L. Zhang, R. Shiao, S. Fu, E.A. Dalmasso, C. Metayer, G.V. Dahl, P.A. Buffler, and M.T. Smith. Proteomic Analysis of Childhood Leukemia. *Leukemia*, 19:1713–1718, 2005.

H.J. Issaq, T.D. Veenstra, T.P. Conrads, and D. Felschow. The SELDI-TOF MS Approach to Proteomics: Protein Profiling and Biomarker Identification. *Biochemical and Biophysical Research Communications*, 292(3): 587–592, 2002.

N. Jeffries. Algorithms for Alignment of Mass Spectrometry Proteomic Data. *Bioinformatics.*, 15;21(14):3066–3073, 2005.

R. Koenker and G.S. Bassett. Regression Quantiles. *Econometrica*, 46:33–50, 1978.

A. Kohlmann, C. Schoch, S. Schnittger, M. Dugas, W. Hiddemann, W. Kern, and T. Haferlach. Pediatric Acute Lymphoblastic Leukemia (ALL) Gene Expression Signatures Classify an Independent Cohort of Adult ALL Patients. *Leukemia*, 18(1):63–71, 2004.

E.L. Lehmann and J.P Romano. Generalizations of the Family-wise Error Rate. Technical report, Department of Statistics, Stanford University, 2003.

J. Liu and M. Li. Finding Cancer Biomarkers from Mass Spectrometry Data by Decision Lists. *Journal of Computational Biology*, 12(7):971–979, 2005.

K. Ohmine, J. Ota, M. Ueda, S. Ueno, K. Yoshida, Y. Yamashita, K. Kirito, S. Imagawa, Y. Nakamura, K. Saito, M. Akutsu, K. Mitani, Y. Kano, N. Komatsu, K. Ozawa, and H. Mano. Characterization of Stage Progression in Chronic Myeloid Leukemia by DNA Microarray with Purified Hematopoietic Stem Cells. *Oncogene*, 20(57):8249–8257, 2001.

K. S. Pollard and M. J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. A. Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo,

and R. Goldman. Analysis of Mass Spectral Serum Profiles for Biomarker Selection. *Bioinformatics*, 21(21):4039–4045, 2005.

M.E. Ross, X. Zhou, G. Song, S.A. Shurtleff, K. Girtman, W.K. Williams, H.C. Liu, R. Mahfouz, S.C. Raimondi, N. Lenny, A. Patel, and J.R. Downing. Classification of Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. *Blood*, 102(8):2951–2959, 2003.

M.E. Ross, R. Mahfouz, M. Onciu, H.C. Liu, X. Zhou, G. Song, S.A. Shurtleff, S. Pounds, C. Cheng, J. Ma, R.C. Ribeiro, J.E. Rubnitz, K. Girtman, W.K. Williams, S.C. Raimondi, D.C. Liang, L.Y. Shih, C.H. Pui, and J.R. Downing. Gene Expression Profiling of Pediatric Acute Myelogenous Leukemia. *Blood*, 104(12):3679–3687, 2004.

M.T. Smith, C.M. McHale, J.L. Wiemels, L. Zhang, J.K. Wiencke, S. Zheng, L. Gunn, C.F. Skibola, X. Ma, and P.A. Buffler. Molecular Biomarkers for the Study of Childhood Leukemia. *Toxicology and Applied Pharmacology*, 206(2):237–245, 2005.

P.J. Valk, R.G. Verhaak, M.A. Beijen, C.A. Erpelinck, S. Barjesteh van Waalwijk van Doorn-Khosrovani, J.M. Boer, H.B. Beverloo, M.J. Moorhouse, P.J. van der Spek, B. Lowenberg, and R. Delwel. Prognostically Useful Gene-expression Profiles in Acute Myeloid Leukemia. *New England Journal of Medicine*, 350(16):1617–1628, 2004.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive ϵ-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003. URL `www.bepress.com/ucbbiostat/paper130`.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004a. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

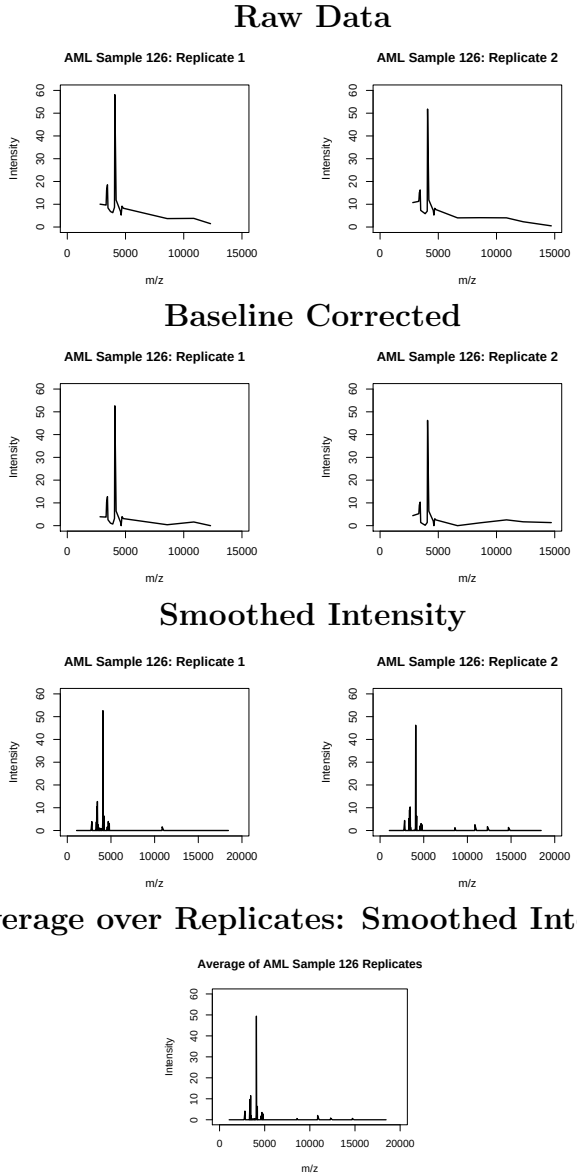
M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. Technical Report 1, 2004b. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

DOI: 10.2202/1544-6115.1198

M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. Resampling Based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL `http://www.bepress.com/sagmb/vol4/iss1/art29`. Article 29.

E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve L. Wong, and J.R. Downing. Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling. *Cancer Cell*, 1(2):133–143, 2002.

Figure 1: **Preprocessing Steps**



DOI: 10.2202/1544-6115.1198