

# Heuristic-based Baseline Removal Algorithm for SELDI

## Proteomics Data

Simon M. Lin<sup>\*</sup>, Pan Du, and Warren A. Kibbe

Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, 60611, USA

---

### ABSTRACT

**Motivation:** Baseline removal, as the first preprocessing step of SELDI data, critically influences subsequent analysis steps. Current baseline removal algorithms of SELDI data, which are based on mathematical morphology, result in biased signal estimates. Due to the parameterization of current algorithms for baseline removal, noise and spectral signal distributions bias the removal results, which may lead to seemingly interesting but ultimately irreproducible results in downstream analysis.

**Results:** We proposed a Heuristic Based Baseline Removal (HbBr) algorithm to model the baseline. HbBr first identifies the potential peak regions by utilizing first derivatives and amplitudes information as a fast heuristic, then down-weights peak regions before modeling the baseline with a nonparametric smooth curve. It outperformed mathematical morphology-based algorithms implemented in the PROcess package of Bioconductor, as judged by a series of benchmark experimental data sets and simulated data sets. We also found that HbBr is computationally more efficient than PROcess. Furthermore, we demonstrated that the HbBr algorithm, although designed for SELDI, yields a good baseline correction of MALDI data without adjusting any parameters.

**Availability:** The algorithm is implemented in R and will be included as an open source module in the Bioconductor project. <http://basic.northwestern.edu/publications/baseline/>

**Contact:** Simon M. Lin ([s-lin2@northwestern.edu](mailto:s-lin2@northwestern.edu)), Pan Du ([dupan@northwestern.edu](mailto:dupan@northwestern.edu)), and Warren A. Kibbe ([wakibbe@northwestern.edu](mailto:wakibbe@northwestern.edu)).

### 1 INTRODUCTION

Proteomics spectra often contain undesirable components, such as baseline and noise, in addition to the desired signal itself. We are particularly interested in Surface Enhanced Laser Desorption Ionization - Time Of Flight (SELDI-TOF, or SELDI in short) mass spectroscopy, which is utilized in clinical and cancer proteomics (Petricoin, Fishman et al. 2004). Usually, SELDI is used to profile the quantitative difference of intact proteins in specimens and thus to identify biomarkers of interest. Coombes and colleagues propose an additive model to partition a SELDI measurement into three components: the baseline, the signal, and the noise (distributed around zero) (Coombes, Tsavachidis et al. 2005). In this framework, all further inferences to associate SELDI signals with

clinical outcomes are based on the assumption that the baseline is appropriately removed. Otherwise, biased and inconsistently estimated signals may lead to seemingly interesting but ultimately irreproducible results (Baggerly, Morris et al. 2004; Hilario, Kalousis et al. 2006).

The SELDI baseline can be seen in the spectrum of a blank (zero protein) sample as a smooth and downward drifting curve moving from low  $m/z$  to high  $m/z$ . Previous studies have established the following three characteristics of the SELDI and Matrix Assisted Laser Desorption Ionization (MALDI) baselines (Baggerly, Morris et al. 2003; Wang, Howard et al. 2003; Hilario, Kalousis et al. 2006). First, the amplitude of the baseline may be much larger than the signal so that the fold change estimate will be downward-biased if we ignore the baseline. Second, the baseline is not flat in a spectrum so that the bias will be heterogeneous across the spectrum. Third, the baseline varies from spectrum to spectrum, even between replicate sample runs; in such a way it creates an unwanted source of variation. In the context of quantification, the goal of baseline correction is to enable the estimation of the true fold change of the molecules present in different biological samples by removing the component due to baseline in each spectrum. Although normalization will also play a role in this quantification process, we do not address this issue in this paper.

Experimentally, the baseline in SELDI is caused by a cloud of matrix molecules hitting the detector shortly after ionization (Malyarenko, Cooke et al. 2005). Unlike Raman spectroscopy, the SELDI baseline cannot be removed chemically by adding quenchers or pretreating the samples (Smith and Dent 2005) since it is inherent in the method. Moreover, the inherent run-to-run variance in the intensity and shape of the baseline excludes the possibility of eliminating baselines by a simple blank spectrum subtraction. Hence, the removal of the baseline from a SELDI spectrum requires numerical processing, either manual or automatic.

Because the peak width of desired signal varies both within and across spectra, and overlapping peaks can merge into a bigger peak cluster, automatic baseline removal is non-trivial. As a result, Jirasek and colleagues argue that the manual method of baseline correction, as guided by visual inspection, is still useful and important (Jirasek, Schulze et al. 2004). It reflects our experience: in many cases, it is hard to fit a seemingly simple baseline into a mathematical model because of the presence of clustered peaks.

---

<sup>\*</sup>To whom correspondence should be addressed.

To remove the peaks, the mainstream SELDI baseline correction algorithms (Baggerly, Morris et al. 2003; Sauve and Speed 2004; Li 2006) are based on the theory of mathematical morphology (Dougherty 1993). We will omit a discussion of time-frequency based methods, such as Fourier filtering or wavelets, since they are not reported in the literature of baseline correction of SELDI. Mathematical morphology implementations include either static or moving window samplings of the data. A prerequisite for effectively applying mathematical morphology is that most of the data points in a window (also called a structuring element) of the spectrum are non-peaks. Thus, the baseline is taken to be the intensity of the spectrum in low percentile (for example, the 5th percentile) within this window. This estimate can be constrained further by smoothing with a nonparametric model such as a loess curve, as implemented in the PROcess package (Li 2006).

There are two major problems with this approach. First, the choice of the low percentile underestimates the true baseline and thus “leaks” part of the noise into the signal estimation (Fig. 1B). We refer to this effect as “noise-leakage”. Although one may choose the median instead of the 5th percentile to quickly correct this problem, we will show in section 3.2 that it is problematic when peaks exist in the window. Second, if we accept the underestimated baseline by using a low percentile, the bias introduced by this approach is noise-dependent and  $m/z$  range-dependent (Fig. 1A).

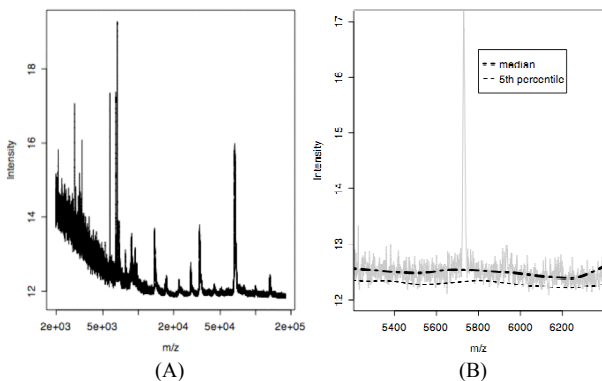


Fig. 1. Difficulties in estimating the SELDI baseline. (A) The raw spectrum of a human serum. Note that the noise at lower  $m/z$  end is higher than the noise at the higher end. (B) A zoomed-in peak region with baseline estimates using the median or the 5th percentile of a moving window.

We have approached the baseline correction problem with a different strategy. We believe the difficulty in automatically estimating baselines is in the discernment of the baseline (non-peak regions) from the peak regions. If we can subtract the peaks from the spectrum, we can use standard methods to model the baseline using a loess curve, a spline, or a faster kernel-based smoothing algorithm (Hastie, Tibshirani et al. 2001). To do this, we have implemented the Heuristic Based Baseline Removal (HbBr) method to exclude the peak regions that interfere with baseline estimation.

To test the effectiveness of this method, we evaluated different algorithms based on the following assumptions:

- After baseline removal, the estimated signal should be closer to the true signal.

- In the simulated data, estimated baseline should be closer to the true baseline.
- After baseline removal, the estimated fold change should be closer to the true fold change.
- The consistency of technical replicates should be improved after baseline removal.

These evaluations use the benchmark datasets collected recently by Pawitan and colleagues (Tan, Ploner et al. 2006), and the Virtual Spectrometer designed by Coombes (Coombes, Kooman et al. 2005). The Pawitan Blank Dataset enables us to study the characteristics of the baseline in the absence of signals, whereas the Pawitan Spike-in Dataset establishes the basis of differential detection. The Virtual Spectrometer allows us to simulate spectra and then to test how the algorithms recover the true baselines in simulated spectra.

In the existing literature, baseline removal algorithms have not been systematically compared, although it has been acknowledged as a critical step in the SELDI data analysis pipeline (Baggerly, Morris et al. 2004). We have established a set of evaluation criteria for baseline removal methods using benchmark and simulated data. Based on these evaluations, we demonstrate that HbBr outperforms mathematical morphology algorithms implemented in the open-source package of PROcess.

## 2 METHODS

### 2.1 The Mathematical Model of a Mass Spectrum

We modeled a spectrum  $y(t)$  as a function of the mass-to-charge ( $m/z$ ) ratio  $t$ ,

$$y(t) = B(t) + S(t) + \varepsilon \quad (1)$$

The observed spectrum  $y(t)$  is a sum of three components: the baseline  $B(t)$ , the signal  $S(t)$ , and the error term  $\varepsilon$ . Following Morris et al. (2005), we assumed that the errors are Gaussians of zero-mean with the variance a function of  $t$ , i.e.,  $\varepsilon \sim N(0, \sigma^2(t))$ . We found this error model adequately describes the heterogeneous noise structure in SELDI experiments, namely, noises at the lower  $m/z$  end is higher than the noises at the higher end.

### 2.2 The HbBr Algorithm

The key idea of the Heuristic Based Baseline Removal (HbBr) method is to locate peak regions first, and then these peak regions are down-weighted when modeling the baseline with a smooth curve. As shown in Equation (1), the peak regions by definition include the component  $S(t)$ , while the non-peak regions basically are only composed of baseline  $B(t)$  and noise  $\varepsilon$ . Supposing the noise  $\varepsilon$  has a zero mean, by fitting these non-peak regions with a smooth curve, the baseline  $B(t)$  can be estimated. As discussed in the Introduction, direct identification of peak regions by thresholding in the raw spectrum is difficult because of the existence of the baseline. Instead, the HbBr algorithm utilizes the first derivative as a heuristic to locate the peak regions. By taking the first derivative, the regions of the spectra that are due to baseline are close to zero while the fast changing peaks are non-zero. To correctly label the relatively flat region usually at the tail of a clumped peak, the HbBr algorithm also takes advantage of the raw spectrum amplitude to refine the calling of the non-peak regions. If the identified non-peak regions based on derivatives have higher intensity amplitudes than the surrounding non-peak regions, they are still called as peak regions. By integrating the first derivative and the amplitude information of the raw spectrum in the heuristic, the HbBr algorithm consists of the following three steps.

#### Step 1. Down sampling the spectrum

Since the baseline is very slow changing, down sampling the spectrum will not affect the estimation of the baseline and greatly speed up the subsequent calculations.

*Step 2. Heuristics to exclude  $y(t)$  at peak locations (cf. Figure 3)*

(a). Smooth  $y(t)$  with a smoother. We use Friedman's supersmoother (Friedman 1984). Briefly, it calculates a modified version of a weighted moving average (Hastie, Tibshirani et al. 2001).

(b). Calculate the absolute value of the first derivative with certain time lag  $d$ ,  $dy = |y(t) - y(t-d)|$ . The time lag  $d$  controls the sensitivity to the rate of amplitude change.

(c). Calculate the dilation (Dougherty 1993) of  $dy$  using a running 90th percentile moving average. This will be an indication of a peak region.

(d). Calculate the morphological opening (Dougherty 1993) of  $dy$ . This will be used as the threshold to identify peaks. The dilation above the morphological opening is initially regarded as a peak region; otherwise it is a non-peak region.

(e). Check the data points in each non-peak region  $nonPeak_i$  based on the amplitude of the raw spectrum (cf. Figure 4).

- Fit a linear model based on the raw spectrum in the extended non-peak regions surrounding  $nonPeak_i$  (within a certain distance);
- Calculate the standard deviation,  $\sigma_{residue}$ , of the residues of the linear fitting;
- Calculate the average distance,  $dist_i$ , from data points in  $nonPeak_i$  to the fitted line;
- If  $dist_i > 3\sigma_{residue}$ , then reset this non-peak region as a peak region.

(f). Set the weights in non-peak regions,  $w[i] = 1/(dy[i]/\max(dy) + 0.01)$ ,  $i, dy \in non\text{-}peak\ region$ ; whereas weights in the peak regions,  $w[i] = 0$ ,  $i \in peak\ region$ . Basically, it down-weights the peak regions, and over-weights the non-peak regions with low derivatives when fitting the baseline in step 3.

*Step 3. Model the baseline with a smoother based on  $y_1(t)$  with weight  $w(t)$  as calculated in step 2.*

(a). The current implementation includes two smoothers, smooth spline and supersmoother (Hastie, Tibshirani et al. 2001). The smooth spline provides better global smoothness, while the supersmoother can fit better to the local structures. All the results were based on the smooth spline implementation.

### 2.3 Evaluation with the SAD

To evaluate how much an estimated variable is deviated from its true value, we use the Sum of Absolute Differences (SAD). For example, the SAD of the baseline is

$$SAD = \sum_t |\hat{B}(t) - B(t)|, \quad (2)$$

where  $\hat{B}(t)$  is the estimated baseline and  $B(t)$  is the known baseline. This SAD measurement, which uses the absolute difference, is more resistant to outliers than the commonly used sum of square differences. Similar to the sum of square differences, to keep it simple, we do not rescale the SAD. Thus, the value of SAD will be comparable only within a data set. Note that the SAD is also known as the Manhattan distance in the context of pattern recognition.

### 2.4 Simulation of Mass Spectra

We used Coombes's Virtual Mass Spectrometer to simulate signal  $S(t)$  (Coombes, Kooman et al. 2005). Briefly, the time-of-flight was simulated using a physical model of the charged protein particles. We added the baseline as a two-parameter ( $\gamma$  and  $\lambda$ ) exponentially decaying function

$$B(t) = \gamma e^{-\lambda t} \quad (3)$$

Based on a recent study of the HUP0 sera (Rai, Stemmer et al. 2005), we simulated 60 proteins in each spectrum, with their position randomly distributed in the  $m/z$  range of 3000 to 73000 Da. We simulated 100 spectra in total; each spectrum contains 103,496 data points.

### 2.5 Benchmark Data Sets

**Pawitan Blank Dataset.** Four blank spectra were obtained from CIPHERGEN SAX2 chips using buffer only and low laser intensity for ionization. Data were collected from 3K to 200 KDa. Details were reported by Tan et al. (2006).

**Pawitan Spike-in Dataset.** A single protein was spiked into normal human serum and then measured with low laser intensity using CIPHERGEN WCX2 chips. Data were collected from 2K to 180K Da. Bovine insulin (5733 Da) was chosen as the spike-in because the human serum is blank at  $m/z$  from 5700 to 5800. We selected experiments with spike-ins at 1, 2, 4, and 6  $\mu$ l. Each spike-in concentration was performed in independent duplicates. Details were reported by Tan et al. (2006).

## 3 RESULTS

### 3.1 Baseline Characteristics from the Blank Data

We started our investigation with a very simple case using the Pawitan Blank Dataset, where the true signal  $S(t)$  is zero, and thus the measurement  $y(t)$  is just baseline and noise (Fig. 2).

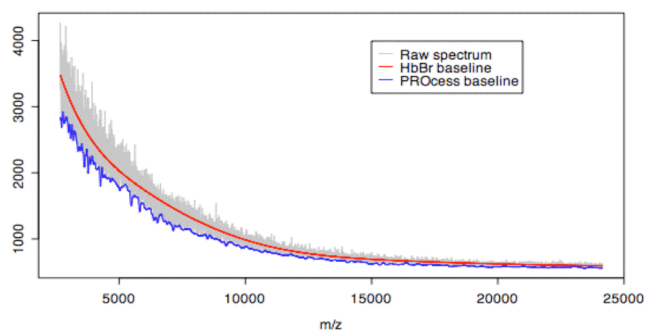


Fig. 2. Pawitan Blank Data.

Following equation 1, we estimated the signal by

$$\hat{S}(t) = y(t) - \hat{B}(t),$$

and then compared it with the true signal  $S(t)$ , of which  $S(t) \equiv 0$ . This evaluation is based on the premise that after baseline correction, the estimated signal should be closer to the true signal. We used a robust estimate of the Sum of Absolute Differences (SAD) to evaluate *inaccuracy*, which measures how the estimation is deviated from the truth. A small SAD indicates better performance of the algorithm. In addition, the baseline estimate should not be af-

ected when the level of noise changes. We define *instability* as the difference in baseline estimates before and after denoising the raw data. The instability of baseline estimate indicates an oversensitivity caused by the over fitted baseline.

We compared four models (Table 1) in terms of their ability to accurately estimate the signal after removing the baseline, and their stability against changing levels of noise. By definition, the mathematical morphology based model B traces the bottom of the measurements. Thus, it leaked noise into the signal estimate, which contributed to its poor accuracy of estimating the signal. For the same reason, when the noise level changed, the baseline estimate also dramatically differed, as measured by the variability in column two of Table 1.

Table 1. Comparison of four models over blank data.

Baseline Estimation Models	Inaccuracy <sup>a</sup>	Instability <sup>a</sup>
(A) HbBr	2889 $\pm$ 130*	253 $\pm$ 110
(B) PROcess, default	5383 $\pm$ 278	3870 $\pm$ 167
(C) PROcess, median	2770 $\pm$ 139*	305 $\pm$ 20
(D) PROcess, median, window size=2, no smooth constraint	960 $\pm$ 76	2185 $\pm$ 89

<sup>a</sup> Inaccuracy is defined as the difference between the estimated signal and the true signal. Instability is defined as the difference in baseline estimations before and after denoising. These differences are estimated by the SAD, an outlier-resistant measurement of the difference. See methods for the formula. Mean and standard deviation are calculated from four independent blank spectra.

\* Significantly lower than Model B by t-test,  $p < 0.01$ .

Replacing the lower 5th percentile (in model B) with the median (in model C) improved accuracy and stability, since the median is more robust against noise. A further tuning of the parameters (in model D) to remove the smoothing constraints resulted in a paradoxically better estimation of the zero signal. This is because the blank sample is a special case, where over-fitted baseline can deflate noise. To solve this paradox, we additionally evaluated how stable the baseline estimation is, when the noise level is reduced by a wavelet denoising algorithm using a data analytical threshold of  $\alpha=0.05$  (Gencay, Selcuk et al. 2002). We found that the over-fitted baseline from model D is not stable, compared with models A and C.

In summary, Model A, which uses the HbBr algorithm instead of mathematical morphology, works equally well with the quick fix of using the median (model C). In the following section, we will demonstrate that model C is not robust when there are peaks, and we will also explain how HbBr works.

### 3.2 Baseline Estimates from Simulated Data

The blank data is a special case lacking any biological significance. To increase the complexity, we would like to test the baseline algorithms in the presence of protein signals  $S(t)$ . As discussed in the introduction, it is impossible to get an experimentally measured signal without an accompanying baseline. Moreover, because the intensity and shape of the baseline varies from spectrum to spectrum, we cannot dissect  $S(t)$  from the experimentally measured  $y(t)$  by subtracting a blank spectrum.

To evaluate the performance of our algorithm and PROcess in handling this problem, we devised a simulation set where we know the components of  $S(t)$  and  $B(t)$  for each spectrum. This will help us compare the estimated baseline with the known components. The simulation set consists of 100 simulated spectra with exponentially decaying baselines with decay characteristics representing typical experimental variance and differing levels of signal  $S(t)$ . As shown in Table 2, the inaccuracy of HbBr is significantly lower than PROcess-median and PROcess-default ( $p < 0.01$ ). We also noticed that our HbBr implementation runs more than four times faster than the PROcess algorithms.

Table 2. Comparison of three algorithms over simulated data.

Baseline Estimation Models	Inaccuracy <sup>a</sup>	Run time <sup>b</sup>
HbBr	85167 $\pm$ 16744*	7 min
PROcess, median	549807 $\pm$ 74275	36 min
PROcess, default	2856889 $\pm$ 39896	45 min

<sup>a</sup> Mean and standard deviation of SAD from 100 simulated spectra.

<sup>b</sup> Total clock running time of 100 spectra on a Mac G4 computer.

\* Significantly lower than PROcess-median and PROcess-default by t-test,  $p < 0.01$ .

HbBr performs better than PROcess because it uses a different model to exclude potential peaks from the baseline. As illustrated in Fig. 3, the peak regions were identified by taking the first derivative, and then down-weighted before being fitted to a smooth line. In contrast to the conventionally used mathematical morphology methods, typified by PROcess, which trace the lower percentile in a moving window, HbBr will more closely model the middle of the baseline, relatively insensitive to the level of noise.

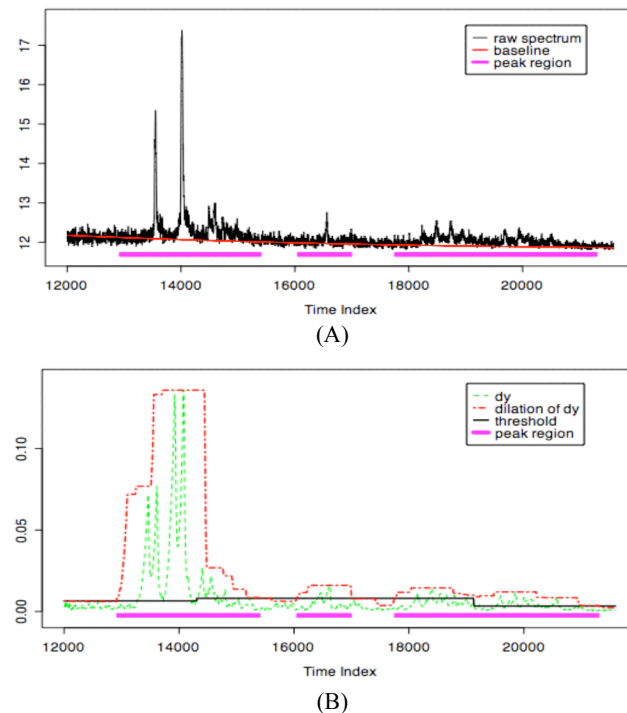


Figure 3. The heuristic of the HbBr algorithm. (A) The raw spectrum, identified peak regions, and estimated baseline. (B) The mathematical morphology method was applied to derivatives instead of raw data, in order to identify peak regions.

A common difficulty for many baseline removal algorithms is the heavy tail of a peak cluster. The tail signal typically elevates (overestimates) the baseline prediction and therefore underestimates the signal present in these regions (Figure 4A). We solve this problem by using amplitude information of the raw data in addition to the derivatives. Figure 4 compares the baseline estimates with (Figure 4B) and without (Figure 4A) step 2(e) in the HbBr algorithm.

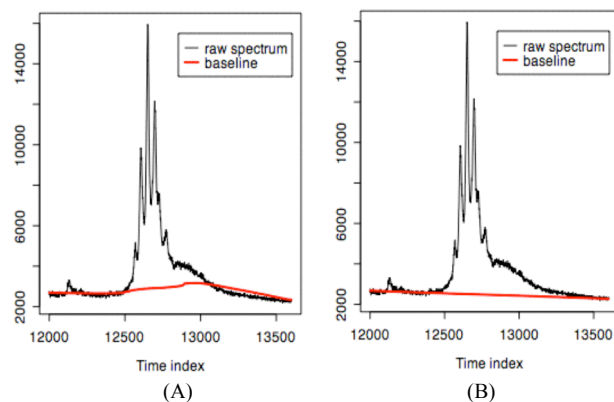


Figure 4. Comparing baseline estimations before (A) and after (B) integrating amplitude information

### 3.3 Effects of Baseline Correction on Fold-change Estimation

The goal of SELDI is to measure the biological differences from sample to sample. From an analytical chemistry point of view, an ideal estimator will result in estimates of fold changes that are exactly the same as the true fold changes. We used the Pawitan Spike-in Dataset (Tan, Ploner et al. 2006) to assess the impact of baseline correction on estimating true fold changes of signals (Figure 5).

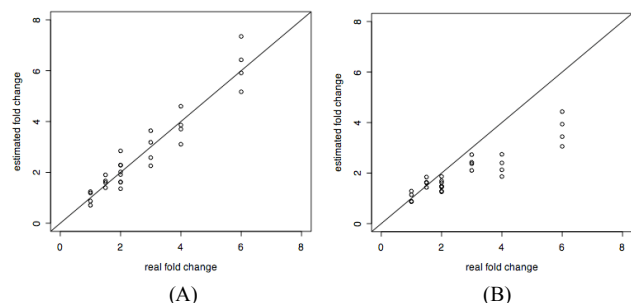


Fig. 5. Estimated v.s. real fold change. (A) HbBr. (B) PROcess-default.

To compare different algorithms, we calculated the inaccuracy of each signal estimator. Briefly, the baselines are corrected by different algorithms; and then subjected to the same normalization by q-spline (Workman, Jensen et al. 2002) and the same quantification method using the area under the curve from  $m/z$  5720 to 5745. The HbBr algorithm outperforms PROcess because it does not intro-

duce as much baseline correction error when quantifying the signal in each spectrum (Fig. 5 and Table 3).

Table 3. Error in estimating the fold change.

Baseline Estimation Models used in each signal estimator	Inaccuracy <sup>a</sup>
HbBr	11.1
PROcess, median	18.4
PROcess, default	23.6
No baseline correction	49.0

<sup>a</sup> Measured by SAD.

### 3.4 Human Serum Application Datasets

After assessing the HbBr algorithm with simulated and benchmark experimental data sets, we applied it to human serum data. Since the same human serum was used as a background for the 14 spectra in the Pawitan Spike-in Dataset, appropriate baseline correction will improve the reproducibility of the signals outside of the spike-in region. As shown in Table 4, HbBr results in higher correlation coefficients among the replicates, providing another measure of performance that is consistent with the previous evaluations.

Table 4. Correlations among technical replicates of human serum increased after baseline correction.

Baseline Estimation Models	25 <sup>th</sup> percentile	Median	75 <sup>th</sup> percentile
HbBr	0.94	0.96	0.97
PROcess, default	0.91	0.93	0.94

Pairwise correlations among 14 spectra in the 6 KDa to 180 KDa region (outside of the spike-in) of human serum (from the Pawitan Spike-in dataset).

In addition, we found HbBr can be applied to MALDI profiling data, although it was originally designed for SELDI (see the supplement).

## 4 DISCUSSION

The difficulty in modeling the baseline for SELDI spectra is analogous to the difficulty in finding an accurate regression line in a noisy dataset that contains outliers. In both cases, the peaks or outliers can be easily recognized by a domain expert, but can be hard for a machine to identify automatically. That is why Jirasek and colleagues indicate that manual baseline removal is still useful and important (Jirasek, Schulze et al. 2004).

However, high throughput proteomics demand accurate, automated analyses with manual review only in exceptional cases. For example, a recent clinical project design included samples from 50 patients with 50 case-controls, where six fractions of serum were measured in duplicate on three different SELDI surfaces and scanned at two different laser powers, and resulted in 7200 spectra. Each step in the analysis pipeline has to be automated to handle this level of data generation. Moreover, the pipeline has to be computationally efficient. For 7200 spectra using our development machine it would require two days just to process these spectra using the PROcess algorithm alone. The increased computational efficiency of HbBr provides an additional advantage in a scenario such as this.

The improved performance of HbBr is due to the initial processing of the raw spectral data into its first derivative; it uses mathemati-

cal morphology only indirectly. Mathematical morphology, which is widely used in image analysis to enhance visual recognition (Dougherty 1993), is an effective method to separate peaks from the baseline when noise is not a significant component, or when there are not significant changes in the baseline across an image. However, as we have shown in the evaluations above, it is not appropriate for signal estimation in the three-component model of Equation 1. The major problem is that the direct application of mathematical morphology methods underestimates the baseline by tracing the bottom of an 'edge'. This in turn adds some of the noise energy present in the observed spectrum to the signal estimate. Instead of using mathematical morphology to identify the baseline (an estimate of the y-axis) directly, we use it only in the heuristic step to identify the potential peaks (an estimate of the x-axis). Thus, the noise-leakage problem associated with mathematical morphology does not affect the HbBr algorithm. As a result, HbBr gives a better prediction of the baseline in the presence of noise.

Another advantage of HbBr is that it can be applied directly to the raw data. In contrast, many mathematical morphology based algorithms recommend smoothing prior to baseline removal to improve the performance. However, smoothing itself will bring extra variability among spectra, since it is hard to decide how much smoothing is the good choice. A problematic smoothing of the raw spectra in this way removes information from the spectrum, such as small peaks, side peaks, and other features that could be instructive during later analysis or identification steps.

The HbBr heuristic relies on the first derivative. First derivatives, which have been used for a long time in peak detection due to their simplicity, are not robust with noisy data (Breton 1990). However, we only use them as a heuristic to exclude the *potential* regions of peaks. As shown in Figure 3, this method may overestimate some jittering regions as peaks and underestimate some peaks with small amplitudes. However, these misestimates have a minimal overall affect on the performance of our HbBr algorithm for the following two reasons. First, compared to true signal or noise, the baseline is a slowly changing curve, so overestimating the width of peak regions will not significantly effect the baseline estimation. Second, since the baseline will be an exponentially decaying function, a spline or other smoothing function can be very effectively applied to the baseline estimate to remove the contribution of a few poor baseline calls.

A frequent argument in the microarray literature against background correction is that the background is usually small and the corrected data might be negative. These two reasons are not applicable to SELDI baseline. First, the SELDI baseline is substantial; ignoring them will result in drastically inaccurate fold changes (Table 3 and Figure 5). Second, the baseline corrected spectrum (potentially with some points in the non-peak regions being negative) is not directly used to make inferences. Instead, extracted peak features (such as peak height or area under the curve) are used for biomarker discovery. The feature extraction step can ensure there are no negative estimators. As we have shown in Figure 5, baseline correction is required to truly reflect the biological changes.

We believe the four performance metrics we present will enable the benchmarking of future improvements in baseline removal, and this strategy can be extended to other analysis steps in the analysis pipeline for SELDI spectra. The four benchmarks we devised evaluate the major performance criteria for baseline correction algorithms: signal accuracy and baseline stability using the Pawi-

tan Blank Dataset, baseline accuracy using the simulated dataset, fold-change accuracy using the Pawitan Spike-in Dataset, and reproducibility using the technical replicates of human serum. We feel these benchmarks are more useful than, for instance, trying to measure the performance by the number of classification errors in a clinical data set. We feel the benchmarks presented are better, because classification error is a composite result of multiple processing steps, including transformation, normalization, peak identification, and peak quantification, all of which are potential confounders in the determination of the performance of baseline removal algorithms. In addition, with a clinical data set there are potentially unknown confounding variables (Coombes, Morris et al. 2005). Thus, we believe each preprocessing step should be evaluated independently using suitable "gold standard" data sets or simulations. Accordingly, we have investigated the peak identification problem in another study (Du, Kibbe, and Lin, 2006). In order to achieve better classification rates, signal normalization and peak quantification methods should undergo a similar performance assessment.

## 5 CONCLUSIONS

Baseline removal is a first step in the SELDI data analysis pipeline. We devised a new heuristics-based algorithm, HbBr, to solve the noise-dependent bias introduced by mathematical morphology algorithms. In addition, the current HbBr is six times faster to run. We tested the accuracy and stability of this algorithm in reference and simulated data sets and determined its better performance. We also found the algorithm can be applied to MALDI data without any modifications.

## ACKNOWLEDGEMENTS

We thank Professor Jeannie Herrick for copy editing the manuscript.

## REFERENCES

- Baggerly, K. A., J. S. Morris, et al. (2004). "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments." *Bioinformatics* **20**(5): 777-85.
- Baggerly, K. A., J. S. Morris, et al. (2003). "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples." *Proteomics* **3**(9): 1667-72.
- Breton, R. G. (1990). *Chemometrics: applications of mathematics and statistics to laboratory systems*. New York, E. Horwood.
- Coombes, K. R., J. M. Kooman, et al. (2005). "Understanding the characteristics of mass spectrometry data through the use of simulation." *Cancer Informatics* **1**(1): 41-52.
- Coombes, K. R., J. S. Morris, et al. (2005). "Serum proteomics profiling--a young technology begins to mature." *Nat Biotechnol* **23**(3): 291-2.
- Coombes, K. R., S. Tsavachidis, et al. (2005). "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): 4107-17.
- Dougherty, E. R. (1993). *Mathematical morphology in image processing*. New York, M. Dekker.

- Du, P., W. A. Kibbe, et al. (2006). "Improved Peak Detection in Mass Spectrometry Spectrum by Incorporating Continuous Wavelet Transform-based Pattern Matching." Bioinformatics (in print).
- Friedman, J. H. (1984). "A variable span scatterplot smoother." Laboratory for Computational Statistics, Stanford University Technical Report No. 5.
- Gencay, R., F. Selcuk, et al. (2002). An introduction to wavelets and other filtering methods in finance and economics. San Diego, Academic Press.
- Hastie, T., R. Tibshirani, et al. (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. New York, Springer.
- Hilario, M., A. Kalousis, et al. (2006). "Processing and classification of protein mass spectra." Mass Spectrom Rev 25(3): 409-49.
- Jirasek, A., G. Schulze, et al. (2004). "Accuracy and precision of manual baseline determination." Appl Spectrosc 58(12): 1488-99.
- Li, Y. (2006). "BioConductor Vignette 'PROcess'." [www.bioconductor.org](http://www.bioconductor.org).
- Malyarenko, D. I., W. E. Cooke, et al. (2005). "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques." Clin Chem 51(1): 65-74.
- Petricoin, E. F., D. A. Fishman, et al. (2004). "Lessons from Kitty Hawk: from feasibility to routine clinical use for the field of proteomic pattern diagnostics." Proteomics 4(8): 2357-60.
- Rai, A. J., P. M. Stemmer, et al. (2005). "Analysis of Human Proteome Organization Plasma Proteome Project (HUPO PPP) reference specimens using surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry: multi-institution correlation of spectra and identification of biomarkers." Proteomics 5(13): 3467-74.
- Sauve, A. C. and T. P. Speed (2004). Normalization, baseline correction and alignment of high-throughput mass spectrometry data. Proceedings of the Genomic Signal Processing and Statistics workshop, Baltimore, MO, USA, <http://stat-www.berkeley.edu/users/terry/Group/publications/Final2Gensips2004Sauve.pdf>.
- Smith, E. and G. Dent (2005). Modern Raman spectroscopy : a practical approach. Hoboken, NJ, J. Wiley.
- Tan, C. S., A. Ploner, et al. (2006). "Finding regions of significance in SELDI measurements for identifying protein biomarkers." Bioinformatics.
- Wang, M. Z., B. Howard, et al. (2003). "Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry." Proteomics 3(9): 1661-6.
- Workman, C., L. J. Jensen, et al. (2002). "A new non-linear normalization method for reducing variability in DNA microarray experiments." Genome Biol 3(9): research0048.