

Supplement to

# 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

## MALDI Application Data

A MALDI spectrum of a lung cancer patient serum (Wang, Howard et al. 2003) was kindly provided by Dr. Edward Patz, Jr. (Duke University, NC).

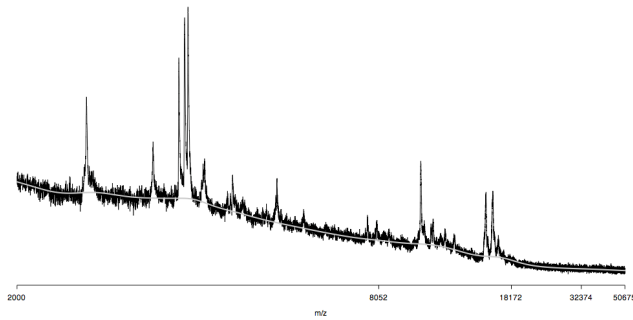


Fig.5. Baseline correction of MALDI data of a lung cancer serum. The baseline resulted from the HbBr algorithm is in grey.

## Parametric Smoothing

In addition to the nonparametric model implemented in HbBr and PROcess, the decaying shape of the baseline (Fig. 2) suggests that an exponentially decaying model of

$$B(t) = \gamma e^{-\lambda t} + \alpha,$$

or an inverse power model of

$$B(t) = \frac{1}{\lambda t + \gamma} + \alpha,$$

might fit the baseline. Thus, we compared five models (Table S1) in terms of their ability to accurately estimate the signal after removing the baseline, and their stability against changing levels of noise.

Table S1. Comparison of five models over blank data.

Baseline Estimation Models	Inaccuracy <sup>a</sup>	Variability <sup>a</sup>
(A) Nonparametric, HbBr	2889 ± 130 <sup>*</sup>	253 ± 110
(B) Nonparametric, PROcess, median	2770 ± 139 <sup>*</sup>	305 ± 20
(C) Nonparametric, PROcess, default	5383 ± 278	3870 ± 167
(D) Parametric, exponential	5829 ± 390	3732 ± 164
(E) Parametric, inverse power	7316 ± 757	3732 ± 164

(F) Nonparametric, PROcess, 960  $\pm$  76 2185  $\pm$  89  
median, window size=2, no  
smooth constraint

---

<sup>a</sup> Inaccuracy is defined as the difference between the estimated signal and the true signal. Variability 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.

<sup>\*</sup> Significantly lower than Model C, D, and E by t-test,  $p < 0.01$ .

Since a blank sample is a special case, one might be concerned that an overly complex baseline, such as those resulted from nonparametric models, will paradoxically result in a better estimation of the zero signal by deflating noise into the baseline. Indeed, model F, with a very small window size of 2 (the default size is 187) and no smoothing constraints, has the best performance in terms of accurately estimating the zero signal (Table S1, column two). To solve this paradox, we additionally evaluated how stable the baseline estimation is, when the noise level is reduced by a wavelet denoising algorithm. *Variability* is defined as the difference in baseline estimates before and after denoising (Table 1, column three). We found that the over-fitted baseline model F is not stable, comparing to model A and B. Note that the two parametric models of D and E have the same instability measurement because their simplicities make them similarly robust to noise. Also note that model F (using median) is more stable than model C, D, E (using the lower 5th percentile) because the median avoids the noise-leakage problem.

The noise-leakage also explains why model C, D and E under-perform in terms of signal accuracy: the noise elevates the signal estimate (c.f. Fig 1B). Replacing the lower 5th percentile with the median improves accuracy (model B and F), although model F is not stable because it over-fits the baseline. Parametric models D and E cannot provide an accurate estimation because their simplicity might not describe the complex mechanism of baselines. Thus, we will not further pursue parametric modeling of the baseline in the following sections. Model A, which uses our HbBr algorithm instead of mathematical morphology, works equally well with the quick fix of using the median (model B). In the section 3.2, we demonstrate that in the presence of signals, model B is not robust, and we also explain how HbBr works.

## Reference

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.